CCIT Report #602
October 2006

On Successive Refinement With Causal Side Information at the Decoders

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 process, $\{X_i, Y_i, Z_i\}_{i=1}^{\infty}$, producing independent copies of a triplet of jointly distributed random variables (RVs). The $\{X_i\}$ part of the process – the source, is observed at the encoder, and is supposed to be reproduced at two decoders, decoder 1 and decoder 2, where the $\{Z_i\}$ and the $\{Y_i\}$ parts of the process are observed, respectively, in a causal manner. The communication between the encoder and the decoders is carried out across two memoryless channels in two successive communication stages. In the first stage, the compressed transmission is available to both decoders, but only decoder 1 reconstructs the source (according to the received data-stream and its causal side information $\{Z_i\}$). In the second stage, the second decoder reconstructs the source according to $\{Y_i\}$ and the transmissions of the encoder at both stages. It is desired to find necessary and sufficient conditions such that the distortions incurred (in each stage) will not exceed given thresholds. First, a single-letter characterization of achievable rates is derived for a pure source-coding problem with successive refinement and causal side information at the decoders. Then, for a joint source-channel coding setting, a separation theorem is proved, asserting that in the limit of long blocks, no optimality is lost by first applying lossy successive-refinement source coding, regardless of the channels, and then applying good channel codes to each one of the resulting bitstreams, regardless of the source. Next, conditions for a source to be successively refinable in two different senses are established, and finally, it is shown that the binary symmetric source is successively refinable.

Index terms - causal rate distortion function, channel capacity, joint source-channel coding, side information, source-channel separation, source coding, successive refinement.

1 Introduction

In the last two decades, the problem of multiple description has been attracting considerable attention in the Information Theory community, the Image Processing community and

*This work is part of A. Maor's Ph.D. dissertation.

other scientific communities. One instance of this problem is successive refinement of information [1]-[3]. Codes for successive refinement are codes designed for systems where source reconstruction is done in a number of stages, as is demonstrated in Fig. 1. Specifically,

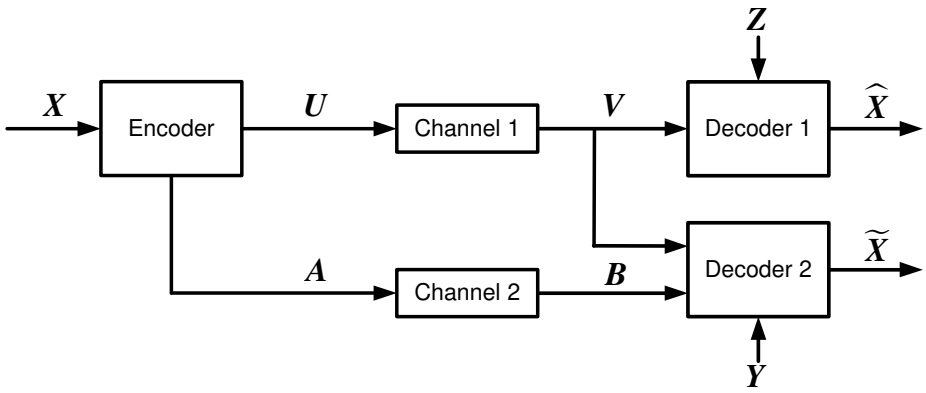

Figure 1: Two-stage successive refinement with side information.

there is a source which is encoded by a single encoder. In each stage, the encoder sends some amount of information to a decoder of that stage, which has access to all the previous transmissions of the encoder. The decoder bases its reconstruction of the source on all transmissions available to it and on additional side information (SI). The quality of reconstruction is measured with respect to some distortion measure. In the case of a pure source coding, the information transmitted by the encoder in each stage arrives at the decoder noiselessly, while in the case of noisy channels connecting the encoder and decoders, the transmission arrives to the decoder corrupted and thus, joint source-channel coding should be applied.

In [4], the problem of successive source coding was studied for the Wyner-Ziv setting: The encoder transmits a source sequence, $\boldsymbol{X}$, to two decoders in two successive stages. In the first step, a coarse description of the source is transmitted to the first-stage decoder at a relatively low rate, and the reconstruction at the decoder should satisfy a certain distortion constraint. The reconstruction is based on the received bitstream and on SI, correlated to the source, which is available at the decoder non-causally. At the second decoder, which has access to additional bits of description of $\boldsymbol{X}$, as well as to the first step bitstream, a reconstruction of higher quality is required. In other words, in the second stage, the encoder transmits refinement bits to the second-stage decoder and the decoder reconstructs $\boldsymbol{X}$ based on the bitstreams of both stages and on non-causal SI. In general, the second-

stage SI differs from that available at the first-stage decoder. For the case of degraded[1] SI available at successive decoders, necessary and sufficient conditions were provided in [4], in terms of single-letter formulas, for the achievability of information rates corresponding to given distortion levels of each communication step. Special attention was devoted in [4] to the case where the SI streams at the decoders are identical. For the case of identical SI available at all the decoders, the two-stage coding scheme was extended to include any finite number of stages. A source was referred to as successively refinable in [4] if no rate loss was incurred with multistage coding (relative to the Wyner-Ziv rate-distortion function). For a two-stage scheme, necessary and sufficient conditions were given in [4] for successive refinability. In the sequel, we will occasionally refer to the definition of successive refinability in the sense of [4] as strict successive refinability.

In [5], the noise-free setting of [4] was extended into a joint source-channel coding scenario, considering communication across independent memoryless channels. Similarly as in [4], the output of the channel corresponding to the coarse (first) description of $\boldsymbol{X}$ was also available to the refinement (second) decoder and each decoder had access to different SI correlated with the source. The main result of [5] was a separation theorem asserting that asymptotically, no optimality is lost by first applying lossy successive source coding, regardless of the channels and then applying good channel codes to each of the resulting bitstreams, regardless of the source and the SI.

The problem of multistage coding with degraded non-causal SI at the decoders has been further studied in [9] by Tian and Diggavi: Instead of considering per-stage rates (as was done in [4]), the focus in [9] was confined to cumulative (sum-) rates after each communication stage. Examining rate sums, Tian and Diggavi have been able to characterize the achievable rate distortion region for successive coding with degraded SI at the decoders for any number of stages as well as to provide conditions for a source to be successively refinable. The definition of successive refinability that was adopted in [9] was somewhat more relaxed than in [4], and hence was referred to as generalized successive refinability (to be defined precisely in the sequel). It was based on the work by Heegard and Berger [10], who showed that if a single encoder has to be designed to two or more decoders with different statistics of SI's, then there might, in general, be some rate loss at least in one of them. In

[1] A notion of degraded SI refers to availability of a more informative SI at the later stage of decoding and is defined for a two-stage scheme by the Markov chain $X \div Y \div Z$, with Z and Y being SI of the first and second stages, respectively.

this paper, we term the successive refinability in the sense on [9] by wide-sense successive refinability.

In [6], the (one-stage) problem of source coding with limited lookahead SI at the decoder has been studied. Among other results, for the case of zero lookahead, i.e., causal SI at the decoder, a single-letter characterization of the smallest rate needed to achieve a certain level of expected per-symbol distortion was provided. This scenario fits well the problem of sequential denoising or filtering: The causal SI stands for a noisy version of the source, which is processed "on-line" at the decoder. A noise-free description of this source is available at the encoder and it is compressed in a lossy manner and transmitted to the decoder in order to make the noise reduction efficient. An example of a practical application of a sequential target tracker can be found in [6].

In this paper, we extend the setups of [4], [5] and [6] in a combined manner: We consider the two-stage coding schemes proposed in [4] and [5], but, in order to reduce decoding complexity, we assume that the SI is available at each decoder causally. To simplify the exposition, we begin with the pure source coding part. We provide a single-letter description of the achievable rate region of a two-stage communication scheme. Next, we extend the noise-free setting into a joint source-channel coding setting and provide a single-letter characterization of the achievable region of distortions achieved at each stage. The main feature of this characterization is that it admits a separation principle. In this sense, our results are related to these of the non-causal setting [5]. We then establish conditions for a source to be successively refinable in two different senses and provide an example of a successively refinable source.

It should be pointed out that a number of interesting differences between causal and non-causal setups arise already in the noise-free setting: it turns out that in order to characterize completely the rate-distortion region achievable with causal SI at the decoders no special structure is required between SI available at each stage. This is unlike in the non-causal setting, where the single-letter characterization of the two-stage scheme was possible only for the case of degraded SI, and remains open for the case of general SI. Moreover, the causal two-stage setting is easily extendable to multi-stage situations, while in case of non-causal SI, such extension of the results is possible only for the case of identical SI available in all the refinement stages [4], or, only when considering per-stage rate sums and degraded SI [9]. The technique used in the direct proof of the causal setting differs substantially from

this of the non-causal case. Unlike in the non-causal case, [4], [9], where the Wyner-Ziv binning technique [8] was extended to fit the multi-stage scheme, when SI is available at the decoders causally, no binning is used in order to prove the achievability scheme. To demonstrate further a substantial difference between causal and non-causal settings, we consider an example of the binary symmetric source (BSS). It turns out that for certain distortions, the BSS is successively refinable when different causal SI is available at the decoders. This is unlike in the non-causal scenario studied in [9], where it was shown that when no SI is available at the first stage decoder, the BSS is wide-sense successively refinable, but not strictly successively refinable. These and additional differences between the causal and non-causal setups are detailed in the sequel.

The outline of the paper is as follows: In Section 2, we define notation conventions. A formal definition of the problem is provided in Section 3. In Section 4, we give the characterizations of the achievable rate regions and formulate the coding theorems for the successive-refinement two-stage source coding and the joint source-channel coding. The conditions for successive and wide-sense successive refinability are provided in Section 5 along with the example of a successively refinable source (BSS). Finally, the converse proofs are given in Section 8.

2 Notation Conventions and Preliminaries

We begin by setting up the notation. Throughout this paper, scalar random variables (RVs) will be denoted by capital letters, specific values they may take will be denoted by the corresponding lower case letters, and their alphabets, as well as most of the other sets, 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 subscript will be omitted. For example, the random vector $\boldsymbol{X} = X^N = X_1^N = (X_1, ..., X_N)$, ($N$-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 $\mathcal{X}$ will be denoted by $|\mathcal{X}|$. For $i > j$, x_i^j (or X_i^j) will be understood as the null string.

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 notation E will denote the expectation operator.

A distortion measure (or distortion function) is a mapping from the set $\mathcal{X} \times \hat{\mathcal{X}}$ into the set of nonnegative reals: $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathcal{R}^+$. The additive distortion, $d(\boldsymbol{x}, \hat{\boldsymbol{x}})$, between two vectors $\boldsymbol{x} \in \mathcal{X}^N$ and $\hat{\boldsymbol{x}} \in \hat{\boldsymbol{X}}^N$ is defined in a usual manner as

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

The information-theoretic quantities, used throughout this paper are denoted using the conventional notations [7]: 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 will be 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 $I(X; Y)$, where logarithms are defined to the base 2. The binary entropy function will be denoted by $h(\alpha) \stackrel{\triangle}{=} -\alpha \log \alpha - (1 - \alpha) \log(1 - \alpha)$ for $0 \leq \alpha \leq 1$.

3 System Description and Problem Definition

We refer to the communication system depicted in Fig. 2. Consider a source, $\{(X_i, Y_i, Z_i)\}_{i=1}^{\infty}$,

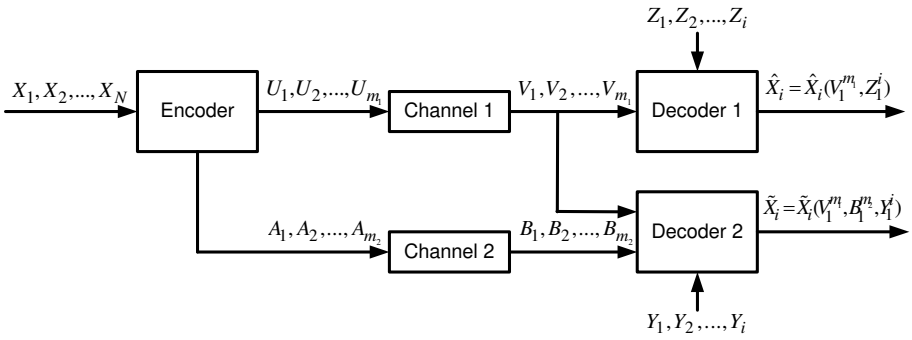

Figure 2: Two-stage communication scheme.

producing independent copies of a triple of RVs, (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 $\{X_i\}$ part of the $\{X_i, Y_i, Z_i\}$ process is observed at the encoder and is supposed to be reproduced at the other side, where the $\{Z_i\}$ and $\{Y_i\}$ parts of the process are observed at decoders 1 and 2, respectively, and are treated as "causal" SI streams, as will be formally defined in the sequel. The reproductions at decoders 1 and 2 take values in the finite sets, $\hat{\mathcal{X}}$ and $\tilde{\mathcal{X}}$, respectively. The

communication between the encoder and two decoders is carried out (in a fashion specified shortly) over two independent memoryless channels, channel 1, $P_{V|U}$ and channel 2, $P_{B|A}$. We denote by C_1 and C_2 the capacities of channel 1 and channel 2, respectively. Channel no. i (i=1,2) operates at the relative rate of ρ_i channel uses per source symbol.

The coding scheme operates as follows: Encoder 1 sends some amount of information to both decoders over Channel 1 and some additional information over Channel 2 to decoder 2 only. We consider block coding, i.e., an N-vector $\boldsymbol{x}$ is encoded into the Channel 1 input sequence $\boldsymbol{u} = (u_1, ..., u_{m_1})$ of length $m_1 \stackrel{\triangle}{=} \rho_1 N$ and into the Channel 2 input sequence $\boldsymbol{a} = (a_1, ..., a_{m_2})$ of length $m_2 \stackrel{\triangle}{=} \rho_2 N$. Decoder 1 receives a noisy version of $\boldsymbol{u}$, denoted by $\boldsymbol{v} = (v_1, ..., v_{m_1})$, and reconstructs each of the components of $\hat{\boldsymbol{x}} = (\hat{x}_1, ..., \hat{x}_N) \in \hat{\mathcal{X}}^N$, $\hat{x}_i$, according to $\boldsymbol{v}$ and $\boldsymbol{z}_1^i$, i.e., decoder 1 uses the SI available to it in a "causal" manner. Decoder 2 has access to $\boldsymbol{v}$ and also to the noisy version of information $\boldsymbol{a}$ conveyed to it by the encoder over channel 2, i.e., $\boldsymbol{b} = (b_1, ..., b_{m_2})$. Decoder 2 processes, thus, $\boldsymbol{b}$, $\boldsymbol{v}$ and the "causal" SI available to it to reproduce each of the components of $\tilde{\boldsymbol{x}} = (\tilde{x}_1, ..., \tilde{x}_N) \in \tilde{\mathcal{X}}^N$, $\tilde{x}_i = \tilde{x}_i(\boldsymbol{v}, \boldsymbol{b}, y_1^i)$. The quality of reconstruction in each of the decoders is judged in terms of the expectation of an additive distortion measure $d_1(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{N}\sum_{i=1}^N d_1(x_i, \hat{x}_i)$ and $d_2(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \frac{1}{N}\sum_{i=1}^N d_2(x_i, \tilde{x}_i)$, where $d_1(x, \hat{x})$ and $d_2(x, \tilde{x})$, $x \in \mathcal{X}$, $\hat{x} \in \hat{\mathcal{X}}$, $\tilde{x} \in \tilde{\mathcal{X}}$, are non-negative, bounded distortion measures.

Definition 1. *For a given memoryless source P_{XYZ} and two memoryless channels $P_{V|U}$ and $P_{B|A}$, an $(N, m_1, m_2, \Delta_1, \Delta_2)$ joint source-channel code for successive refinement with causal side information at the decoders consists of a first-stage encoder:*

$$f_1 : \mathcal{X}^N \to \mathcal{U}^{m_1}, \tag{2}$$

a sequence of N first-stage decoding functions

$$g_{1,i} : \mathcal{Z}^i \times \mathcal{V}^{m_1} \to \hat{\mathcal{X}}, \quad i = 1, ..., N, \tag{3}$$

a second-stage encoder

$$f_2 : \mathcal{X}^N \to \mathcal{A}^{m_2}, \tag{4}$$

and a sequence of N second-stage decoding functions:

$$g_{2,i} : \mathcal{Y}^i \times \mathcal{V}^{m_1} \times \mathcal{B}^{m_2} \to \tilde{\mathcal{X}}, \quad i = 1, ..., N, \tag{5}$$

such that

$$\frac{1}{N}\sum_{i=1}^{N} Ed_1(X_i, g_{1,i}(\mathbf{V}, Z^i)) \leq \Delta_1, \tag{6}$$

$$\frac{1}{N}\sum_{i=1}^{N} Ed_2(X_i, g_{2,i}(\mathbf{V}, \mathbf{B}, Y^i)) \leq \Delta_2, \tag{7}$$

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

Definition 2. *Given ρ_1 and ρ_2, a distortion pair (Δ_1, Δ_2) is said to be achievable if for every $\epsilon > 0$, and sufficiently large block-length N, there exists an $(N, N\rho_1, N\rho_2, \Delta_1+\epsilon, \Delta_2+\epsilon)$ joint source-channel code of successive refinement with causal side informations at the decoders for the source P_{XYZ} and the channels $P_{V|U}$, $P_{B|A}$. The distortion region, denoted $\mathcal{D}$, is the closure of the set of all achievable pairs (Δ_1, Δ_2).*

The characterization of the achievable region of $\mathcal{D}$ is related to the definition of the successive source coding with causal SI at the decoders. In terms of Definition 1, the channels 1 and 2 are binary and noise-free. The compression rate of the first stage (in bits per symbol) is $R_1 = \rho_1$, and the compression rate of the second stage is $R_2 - R_1 = \rho_2$ (R_2 being the total rate). Here we adopt the incremental definitions of information rates used in [4] and [5], which are different from those used in [1], [3] and [9] where the achievable rate-region is given via cumulative communication rates, i.e., (R_1, R_2). The differences between these two approaches are detailed in [4].

Many results of this paper relate to the noise-free case. Hence, in parallel to Definitions 1 and 2, we provide explicit definitions for the corresponding pure source coding problem:

Definition 3. *For a given source P_{XYZ}, an $(N, M_1, M_2, \Delta_1, \Delta_2)$, source code for successive refinement with causal SI at the decoders consists of a first-stage encoder:*

$$f_1 : \mathcal{X}^N \to \{1, 2, ..., M_1\}, \tag{8}$$

a sequence of N first-stage decoding functions:

$$g_{1,i} : \mathcal{Z}_1^i \times \{1, 2, ..., M_1\} \to \hat{\mathcal{X}}, \tag{9}$$

a second-order encoder:

$$f_2 : \mathcal{X}^N \to \{1, 2, ..., M_2\}, \tag{10}$$

and a sequence of N second-stage decoding functions:

$$g_{2,i} : \mathcal{Y}_1^i \times \{1, 2, ..., M_1\} \times \{1, 2, ..., M_2\} \to \tilde{\mathcal{X}}, \tag{11}$$

such that

$$\sum_{i=1}^{N} Ed_1 \left(X_i, g_{1,i} \left(Z^i, f_1 \left(\mathbf{X} \right) \right) \right) \leq N\Delta_1, \tag{12}$$

$$\sum_{i=1}^{N} Ed_2 \left(X_i, g_{2,i} \left(Y^i, f_1 \left(\mathbf{X} \right), f_2 \left(\mathbf{X} \right) \right) \right) \leq N\Delta_2, \tag{13}$$

$$\tag{14}$$

Definition 4. *Given a distortion pair $\mathbf{D} = (\Delta_1, \Delta_2)$, a rate pair (R_1, R_2) is said to be achievable with causal SI (Y, Z) at the decoders if for every $\delta > 0$, $\epsilon > 0$, and a sufficiently large block-length N, there exists an $\left(N, 2^{N(R_1+\delta)}, 2^{N(R_2-R_1+\delta)}, \Delta_1 + \epsilon, \Delta_2 + \epsilon\right)$ source code for successive refinement for the source P_{XYZ} with causal SI at the decoders.*

The collection of all $\boldsymbol{D}$-achievable rate pairs is the achievable rate region for successive refinement coding with causal SI and is denoted by $\mathcal{R}(\boldsymbol{D})$. The collection of all $(R_1, R_2, \Delta_1, \Delta_2)$-achievable rate-distortion quadruples is the achievable rate-distortion region, and is denoted by $\mathcal{RD}$.

Our first objective is to provide a single-letter characterization of $\mathcal{RD}$ and to propose strategies for (asymptotically) achieving any given point in $\mathcal{RD}$. The other objective of this work is to provide a single-letter characterization of $\mathcal{D}$, and in particular, to show that any given point in $\mathcal{D}$ can be achieved by separate source coding of the source part P_X, achieving rates is $\mathcal{RD}$, used in tandem with an optimal channel code [7] (independently of the source).

4 Main Result

In this section, we consider the problem of two-stage joint source-channel coding with a fidelity criterion. For the clarity of exposition, we begin with the pure source coding problem and then extend the setup to include communication over two memoryless capacity-limited channels, and formulate coding theorems for each of the cases.

4.1 Pure Source Coding

In this subsection, we give a single-letter characterization of $\mathcal{RD}$ for a given source P_{XYZ}. Let a distortion pair $\boldsymbol{D} \stackrel{\triangle}{=} (\Delta_1, \Delta_2)$ be given. Define $\mathcal{R}^*(\boldsymbol{D})$ to be the set of all rate pairs

(R_1, R_2) for which there exists a pair of random variables (W_1, W_2), taking values in finite alphabets, $\mathcal{W}_1$ and $\mathcal{W}_2$, respectively, such that the following conditions are simultaneously satisfied:

1. The following Markov chain holds:

$$(W_1, W_2) \div X \div (Y, Z). \tag{15}$$

2. There exist deterministic decoding functions $G_1 : \mathcal{Z} \times \mathcal{W}_1 \to \hat{\mathcal{X}}$ and $G_2 : \mathcal{Y} \times \mathcal{W}_1 \times \mathcal{W}_2 \to \tilde{\mathcal{X}}$, such that

$$Ed_1\left(X, G_1(W_1, Z)\right) \leq \Delta_1 \tag{16}$$

and

$$Ed_2\left(X, G_2(W_1, W_2, Y)\right) \leq \Delta_2. \tag{17}$$

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

$$|\mathcal{W}_1| \leq |\mathcal{X}| + 3 \tag{18}$$

and

$$|\mathcal{W}_2| \leq |\mathcal{X}| \cdot (|\mathcal{X}| + 3) + 1. \tag{19}$$

4. The rates R_1 and R_2 satisfy

$$R_1 \geq I(X; W_1) \tag{20}$$

and

$$R_2 - R_1 \geq I(X; W_2|W_1). \tag{21}$$

The main result of this subsection is the following:

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

$$\mathcal{R}(\mathbf{D}) = \mathcal{R}^*(\mathbf{D}). \tag{22}$$

The proofs of the direct and the converse parts are provided in Sections 7 and 8, respectively. The proof of the direct part follows the lines of the proof of Theorem 1 in [6], and,

interestingly, it comes from a rather standard random coding argument and appropriate application of the Markov Lemma [7, pp. 436, Lemma 14.8.1].

Discussion: First, let us notice the conceptual difference between the internal relations assumed to exist between the (partial) source P_X and SI data-streams in this work and these of [4]. Here, no constraints are imposed on relations between X, Y and Z and yet it is possible to obtain a single-letter representation of $\mathcal{R}(\boldsymbol{D})$. This is in contrast to the case of non-causal SI available at the decoders [4] and [9], where, similarly as in [10], it is not known how to fully characterize the achievable rate-distortion region for a general distribution of a source P_{XYZ} and only the case of the degraded SI was fully analyzed. A possible explanation to this rather unexpected difference between the causal and non-causal cases can be found in the achievability schemes used in each case. Specifically, while in [4], the reconstruction of the source block $\boldsymbol{X}$ was based on the entire SI sequence and binning played a key role in the proof, here, similarly as in [6], no binning is used, and the i-th reconstruction uses only the i-th SI symbol. Now, as soon as the complete index of an auxiliary codeword is transmitted to the decoders, only the relation between the auxiliary and the source components is relevant, while in case of binning, this structure is not sufficient to ensure the "correct" guessing of index of the auxiliary codeword. In fact, the extension of the result of Theorem 1 to any number of communication stages is straightforward and this is also in contrast to [4], where a full characterization of the multi-stage coding scheme was provided only for the case of identical SI and where it was shown that even the extension of the two-stage successively refinable degraded scheme is a non-trivial task. This is also unlike in [9], where, even by considering sum rates, a characterization of the achievable rate-distortion region was possible only under the assumption of degraded SI.

Another rather interesting difference between the causal and non-causal cases lies in the fact that in the causal case, two auxiliary random variables suffice to provide a complete characterization of $\mathcal{R}(\boldsymbol{D})$, while in [4] three auxiliary variables were required even for the case of degraded SI. As it turned out in [4], for a degraded source, the information transmitted in the first stage contains a part which cannot be used by the first decoder and can be utilized by the second decoder only at the second (refinement) stage. In the causal case, this payoff for utilizing the second stage does not exist - every communication stage contains the information which can benefit entirely to the decoder of that stage. The third auxiliary random variable of [4] helps, therefore, to define the rate-payoff of the first stage.

This payoff, in turn, exists due to the fact that the source is assumed degraded, because of the binning used in the achievable part of the proof. As stated in [6], "if the future us not allowed to be looked into, the past is useless", and in our case, there is a certain advantage in this uselessness, as it allows a complete characterization of the rate-distortion achievable region.

4.2 Joint Source-Channel Coding

The necessary and sufficient conditions for (Δ_1, Δ_2) to be the achievable distortion levels of the scheme depicted in Fig. 2 are described in the next theorem:

Theorem 2. *Given a DMS P_{XYZ}, the distortion levels (Δ_1, Δ_2) are achievable for successively refinable communication with causal SI at the decoders over stationary memoryless channels $P_{V|U}$ and $P_{B|A}$, if and only if there exist auxiliary RVs W_1 and W_2, satisfying (15), whose finite alphabets $\mathcal{W}_1$ and $\mathcal{W}_2$ are of cardinalities bounded by (18) and (19), respectively, and there exist deterministic decoding functions G_1 and G_2, satisfying (16) and (17), respectively, such that*

$$I(X;W_1) \leq \rho_1 C_1, \tag{23}$$

$$I(X;W_2|W_1) \leq \rho_2 C_2. \tag{24}$$

The similarity between the characterization of the region of achievable distortion levels of Theorem 2 and the characterization of $\mathcal{D}$ is self-evident. In fact, the only difference is that in each communication stage, the coding rates R_1 and $R_2 - R_1$ of the former are replaced by $\rho_1 C_1$ and $\rho_2 C_2$, respectively. The immediate conclusion from this observation is that the separation principle applies to our model.

The proof of the converse part appears in Section 8. The proof of the direct part comes directly from considering an asymptotically optimum two-stage source code (independent of the channel) followed by a reliable transmission code for each one of the channels (independent of the source), i.e., separate source and channel coding. If the distortion level of the two-stage source code, presented in Section 4.1, is chosen such that $R_1 < \rho_1 C_1$ and $R_2 - R_1 < \rho_2 C_2$, one may select constants $R_{s,1}$, $R_{c,1}$, $R_{s,2}$ and $R_{c,2}$ such that

$$NR_1 < NR_{s,1} = m_1 R_{c,1} < m_1 C_1 \tag{25}$$

$$\begin{aligned} N(R_2 - R_1) &< N(R_{s,2} - R_{s,1}) \\ &= m_2(R_{c,2} - R_{c,1}) \\ &< m_2 C_2. \end{aligned} \tag{26}$$

In the first (second) communication step, the encoder may then compress the information about X into $R_{s,1}$ ($R_{s,2} - R_{s,1}$) bits per symbol within distortion Δ_1 (Δ_2), and then map the resulting $NR_{s,1}$-bit $\big(N(R_{s,2} - R_{s,1})$-bit$\big)$ codewords into channel codewords of the same number of bits $m_1 R_{c,1} < m_1 C_1$ $\big(m_2(R_{c,2} - R_{c,1}) < m_2 C_2\big)$. Since $R_{c,1} < C_1$ and $(R_{c,2} - R_{c,1}) < C_2$, there exist reliable channel codes which cause asymptotically negligible additional distortion. Since a pair (Δ_1, Δ_2) can be chosen in such a way that R_1 is arbitrarily close to $\rho_1 C_1$ and $(R_2 - R_1)$ is arbitrarily close to $\rho_2 C_2$, all distortion levels for which $R_1 < \rho_1 C_1$ and $(R_2 - R_1) < \rho_2 C_2$ are achievable.

In the case of causal SI, the separation theorem holds similarly as in the case of non-causal SI at the decoders [5] for the degraded SI structure, but again, the results of Theorem 2 can be extended straightforwardly to any finite number of stages.

5 Successively Refinable Sources

In this section, we refer to the notion of a successively refinable source. In [4], a source was called successively refinable if in each communication stage, the achievable cumulative rate was equal to the Wyner-Ziv rate distortion function [8] of that stage, i.e., no rate loss was incurred with successive encoding. In parallel to [4], one way to define a successively refinable source is via achievability (in each communication stage) of the rate distortion function with causal SI, $R_{X|SI}(\cdot)$, defined in [6]:

Theorem 3. *[6] The rate distortion function for the case of causal side information Y and $\Delta \geq \Delta_{min}$ is given by*

$$R_{X|Y}(\Delta) = \min I(X; W), \tag{27}$$

where the minimum is over all functions $f : \mathcal{W} \times \mathcal{Y} \to \hat{\mathcal{X}}$, $|\mathcal{W}| \leq |\mathcal{X}| + 1$, and $P(w|x)$ such that

$$E\left(d\left(X, f\left(W, Y\right)\right)\right) \leq \Delta. \tag{28}$$

Using Theorem 3, the definition of a successively refinable source is the following:

Definition 5. *A source X is said to be successively refinable from Δ_1 to Δ_2 with causal SI if*

$$\big(R_{X|Z}(\Delta_1), R_{X|Y}(\Delta_2)\big) \in \mathcal{R}(\Delta_1, \Delta_2). \tag{29}$$

From Definition 5, it is then immediate to prove the following conditions for a source to be successively refinable with causal SI at the decoders.

Theorem 4. *A source X with causal SI (Z,Y) is successively refinable from Δ_1 to Δ_2*[2] *if and only if there exists a pair of random variables (W_1, W_2) and a pair of deterministic maps $G_1 : \mathcal{Z} \times \mathcal{W}_1 \to \hat{\mathcal{X}}$ and $G_2 : \mathcal{Y} \times \mathcal{W}_1 \times \mathcal{W}_2 \to \tilde{\mathcal{X}}$, such that the following conditions simultaneously hold:*

1. *$R_{X|Z}(\Delta_1) = I(X, W_1)$ and $Ed_1(X, G_1(W_1, Z)) \leq \Delta_1$,*
2. *$R_{X|Y}(\Delta_2) = I(X; W_1W_2)$ and $Ed_2(X, G_2(W_1, W_2, Z)) \leq \Delta_2$,*
3. *$(W_1, W_2) \div X \div (Y, Z)$ form a Markov chain.*

The proof of Theorem 4 follows straightforwardly from the definition of a successively refinable source and the rate-distortion function given by Theorem 3 and thus is omitted.

In [4], the achievable rate distortion region was determined only for the case of degraded SI. To complete the characterization of that region, a number of conditions were imposed on the Markov relations between the auxiliary variables, the source and the SI. Unlike in the non-causal setting, when causal SI is available at the decoders, no such conditions are needed and the characterization of the achievable region is not restricted to degraded SI. Also, it is clear what relation is needed between the auxiliary random variables so that the source will be successively refinable: the auxiliary variable that is used to achieve the causal rate-distortion function in the first stage must be a (stochastic) function of the auxiliary variable used at the second stage. This dependence is demonstrated explicitly well by the achievability scheme used for proving Theorem 1 - an index transmitted in the first stage is concatenated with the index sent in the refinement stage, and thus, a codeword used in the first stage of coding may be seen as a part of a "super"-codeword used in the second stage.

[2] Note that Δ_1 is not necessarily larger that Δ_2.

It is noted in [9] that the demand of achieving the Wyner-Ziv rate distortion function at each communication stage may be restrictive in certain cases. To demonstrate that, recall that in [10], a special instance of one-to-many coding was investigated - a single transmission was received by a group of (ordered) decoders, each having access to another SI and each basing a reconstruction of the source on the encoder transmission and its SI. A complete characterization of the achievable rate distortion region was provided in [10] only for a special case of degraded SI available at successive decoders, and it was shown that the minimum cumulative rate may exceed the corresponding Wyner-Ziv rate-distortion function. In [9], an alternative definition of successively refinable source is introduced for systems with degraded SI, where at each stage of communication, it is desired to achieve the Heegard-Berger rate-distortion function rather than that of Wyner and Ziv. This definition allows in [9] a complete characterization of successively refinable sources for coding with a degraded SI at the decoders.

It is of operative interest to derive a causal analog to the achievable rate-distortion region for the Heegard-Berger problem setting, i.e., when causal SI may be present or absent at the decoders. Formally, we shall adapt the following definition:

Definition 6. *For a given source P_{XYZ}, an $(N, M, \Delta_1, \Delta_2)$, source code with causal SI at the decoders consists of an encoder:*

$$f : \mathcal{X}^N \to \{1, 2, ..., M\}, \tag{30}$$

a sequence of N decoding functions of the first decoder:

$$g_{1,i} : \mathcal{Z}^i \times \{1, 2, ..., M\} \to \hat{\mathcal{X}}, \tag{31}$$

and a sequence of N decoding functions of the second decoder:

$$g_{2,i} : \mathcal{Y}^i \times \{1, 2, ..., M\} \to \tilde{\mathcal{X}}, \tag{32}$$

such that

$$\sum_{i=1}^{N} Ed_1\left(X_i, g_{1,i}\left(Z^i, f\left(\mathbf{X}\right)\right)\right) \leq N\Delta_1, \tag{33}$$

$$\sum_{i=1}^{N} Ed_2\left(X_i, g_{2,i}\left(Y^i, f\left(\mathbf{X}\right)\right)\right) \leq N\Delta_2, \tag{34}$$

$$\tag{35}$$

Definition 7. *Given a distortion pair* $\mathbf{D} = (\Delta_1, \Delta_2)$*, a rate* R *is said to be achievable with causal SI* (Y, Z) *at the decoders if for every* $\delta > 0$*,* $\epsilon > 0$*, and a sufficiently large block-length* N*, there exists an* $(N, 2^{N(R+\delta)}, \Delta_1 + \epsilon, \Delta_2 + \epsilon)$ *source code for the source* P_{XYZ} *with causal SI at the decoders.*

Having this definition, in parallel to [9], a successively refinable source is defined in terms of achieving 'causal' Heegard-Berger rates with successive coding. It is easy to prove that the achievable rate-distortion region for the (two-decoder) causal instance of the Heegard-Berger problem is given by the following Theorem.

Theorem 5. *The rate distortion function for the case of causal side information* Z *and* Y *available at the first and the second decoder, respectively, and distortion constraints* Δ_1 *and* Δ_2*, is given by*

$$R^{HB}_{X|ZY}(\Delta_1, \Delta_2) = \min I(X; W_1 W_2), \tag{36}$$

where the minimum is over all functions $G_1 : \mathcal{W}_1 \times \mathcal{Y} \to \hat{\mathcal{X}}$*,* $G_2 : \mathcal{W}_1 \times \mathcal{W}_2 \times \mathcal{Y} \to \tilde{\mathcal{X}}$*,* $|\mathcal{W}_1| \leq |\mathcal{X}| + 2$*,* $|\mathcal{W}_2| \leq |\mathcal{X}| \cdot (|\mathcal{X}| + 2)$*,* $(W_1, W_2) \div X \div (Y, Z)$*, and* $P(w_1, w_2|x)$ *such that*

$$Ed(X, G_1(W_1, Z)) \leq \Delta_1 \quad \textit{and} \quad Ed(X, G_2(W_1, W_2, Y)) \leq \Delta_2. \tag{37}$$

The achievability scheme used for the proof of Theorem 5 is very similar to the proof of Theorem 1, but instead of creating two successive messages, the encoder combines a single message out of the transmissions to both encoders (as are detailed in Section 7)) and each decoder uses the part of the message that is relevant to it. The converse part of the proof of Theorem 5 is also very similar to that of Theorem 1 (though, considers the sum-rate of the scheme) and modifications to the proof of Theorem 5 are outlined in Section 8.

Analogously to [9], we establish the following definition:

Definition 8. *A source* X *is said to be successively refinable in the wide sense with respect to causal SI,* (Y, Z)*, if*

$$\left(R^{HB}_{X|Z}(\Delta_1), R^{HB}_{X|ZY}(\Delta_1, \Delta_2)\right) \in \mathcal{R}(\Delta_1, \Delta_2). \tag{38}$$

The conditions for wide-sense successive refinablility are given by Theorem 4, with $R^{HB}_{X|Z}$ and $R^{HB}_{X|Y}$ replacing $R_{X|Z}$ and $R_{X|Y}$ in items (1) and (2), respectively. Here, similarly as in the case of non-causal SI available at the decoders, the source is successively refinable

if and only if it is successively refinable in the wide sense and the Heegard-Berger (causal) rate distortion function equals the El Gamal-Weissman rate distortion function for each communication stage. The proof of this result follows [9] and is thus omitted.

6 Example

We conclude the discussion on successive refinability with an example of a successively refinable source - the binary symmetric source (BSS). In this section, we confine ourselves to two stages.

The random variables in the following example are binary, taking values in $\{0, 1\}$. We begin with some notation: Let us denote by $\bigoplus$ the XOR operation, or the binary modulo-two addition, and by $\bigotimes$ the AND, or the binary multiplication. Also, denote by $h(\cdot)$ the binary entropy function and by $*$ the binary convolution, i.e., $\alpha * \beta = \alpha(1-\beta) + \beta(1-\alpha)$. Finally, denote by $h'(\cdot)$ the derivative of $h(\cdot)$ and by d_c the solution to the equation $(1 - h(d_c))/(d_c - \delta) = -h'(d_c)$.

6.1 Identical Side Information at the Decoders

We begin with the case of identical SI at the decoders. Specifically, let X be the BSS and let the SI Y be $Y = X \bigoplus N$, where N is an independent Bernoulli(δ) RV. Without loss of generality, assume that $\delta \leq 0.5$ and that $\Delta_2 \leq \Delta_1 \leq \delta$.

For a source to be successively refinable, the cumulative sum of rates at each stage must equal the El Gamal-Weissman rate-distortion function, matching the distortion constraint of that stage. In [6], it was shown that the rate-distortion function of the BSS is the following:

Example 1. *[6] Consider the case where X is the unbiased input to a BSC(δ), $0 \leq \delta \leq 1/2$, and Y is the corresponding output, and the distortion measure is the Hamming loss. The rate-distortion function in this case is given by*

$$R_{X|Y}(\Delta) = \begin{cases} 1 - h(\Delta) & 0 \leq \Delta \leq d_c \\ -h'(d_c)\Delta + h'(d_c)\delta & d_c < \Delta \leq \delta. \end{cases} \tag{39}$$

Consider now the following choice of auxiliary RVs, as is given in Fig. 3. This configuration was proposed in [4] for a similar problem of two-stage coding with successive refinement with a non-causal SI in both stages. In the configuration, $S \sim \text{Bernoulli}\left(\min\{\Delta_2, d_c\}\right)$ and is independent of (X, N) and $B_1 \sim \text{Bernoulli}\left(\frac{\delta - \max\{\Delta_2, d_c\}}{\delta - d_c}\right)$, independent of (X, N, S).

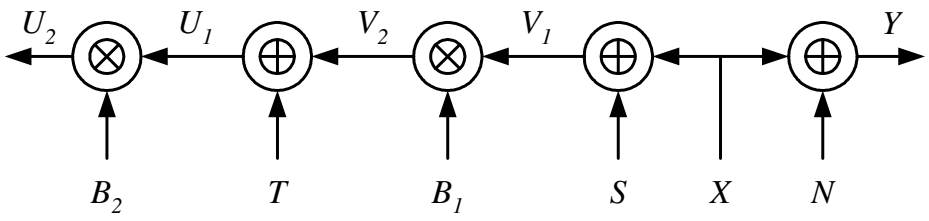

Figure 3: Two-stage scheme for communication with identical non-causal SI [4].

The combination $W_2 = (B_1, V_2)$ serves as the auxiliary RV used in the second (refinement) stage of communication. For the first (coarse) stage of communication, the following RVs are defined: $T \sim \text{Bernoulli}\,(\Pr\{T=1\})$, independent of (X, N, S, B_1) such that $\Pr\{T=1\} * \Pr\{S=1\} = \min\{\Delta_1, d_c\}$, and $B_2 \sim \text{Bernoulli}\left(\frac{\delta - \max\{\Delta_1, d_c\}}{p_0 - \max\{\Delta_2, d_c\}}\right)$, independent of (X, N, S, B_1, T). The auxiliary RV used at the first stage is constructed as follows: $W_1 = (B_1 \cdot B_2, U_2)$. Note that the following Markov chain holds: $W_1 \div W_2 \div X \div Y$. The reconstructions of the first and the second stages are generated according to

$$G_1(W_1, Y) = B_1 \cdot B_2 \cdot U_2 + (1 - B_1 \cdot B_2) \cdot Y, \tag{40}$$

and

$$G_2(W_2, Y) = B_1 \cdot V_2 + (1 - B_1) \cdot Y. \tag{41}$$

It turns out that the same construction of the auxiliary RVs (W_1, W_2) and reconstruction functions (G_1, G_2) attains optimum performance also in the causal setup (here $G_2(W_1, W_2, Y) = G_2(W_2, Y)$ due to the Markov chain $Y \div W_2 \div W_1$). It is then straightforward to show that for the Hamming distortion measure and the rate-distortion function given by (39),

$$Ed_1(X, G_1(W_1, Y)) \leq \Delta_1 \quad \text{and} \quad I(X; W_1) = R_{X|Y}(\Delta_1), \tag{42}$$

$$Ed_2(X, G_2(W_1, W_2, Y)) \leq \Delta_2 \quad \text{and} \quad I(X; W_1, W_2) = I(X; W_2) = R_{X|Y}(\Delta_2). \tag{43}$$

6.2 No SI at the First Decoder

We next consider the special case of different SI available at the decoders - assume that no SI is available at the first decoder and the above-defined Y is available causally at the second decoder. For the similar setting of no SI at the first decoder and non-causal SI at the second decoder, the BBS was shown in [9] to be successively refinable in the wide sense

but not strictly successively refinable. Unlike in the non-causal setting, when SI is available at the decoders causally, it turns out that, for a certain range of distortions, the BSS is successively refinable when there is no SI at the first stage. In this setting, the desired rate at the first stage is given by [7]

$$R(\Delta) = \begin{cases} 1 - h(\Delta) & 0 \leq \Delta \leq \frac{1}{2} \\ 0 & \Delta > \frac{1}{2}. \end{cases} \tag{44}$$

The desired sum-rate in the second stage is still given by (39). It is immediate to show that as long as $\Delta_2 \leq d_c$, $\Delta_2 \leq \Delta_1$, the desired system performance is attainable by the configuration depicted in Fig. 4: Here the RVs (N, S, T, B_1) are defined as above and $B_3 \sim$

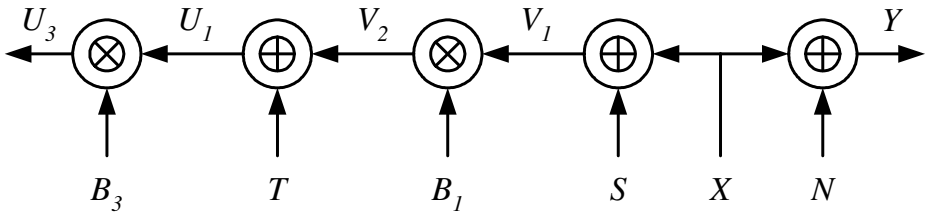

Figure 4: Two-stage scheme for communication without SI at the first stage.

Bernoulli $\left(1 - \frac{\max\{\frac{1}{2}, \Delta_1\} - \frac{1}{2}}{|\Delta_1 - \frac{1}{2}|}\right)$. The refinement stage remains as in the previous example, i.e., $W_2 = (B_1, V_2)$ and $G_2(W_1, W_2, Y) = B_1 \cdot V_2 + (1 - B_1) \cdot Y$. The first stage auxiliary RV is now $W_1 = (B_1 \cdot B_3, B_1 \cdot B_3 \cdot U_3)$. The reconstruction function of the first stage is then $G_1(W_1, Y) = B_1 \cdot B_3 \cdot U_3$.

7 Direct Proof

In this section, we use the definitions from [7] for all the calculations related to the method of types. Fix the DMS P_{XYZ}, δ, $\Delta_1 \geq Ed(X, \hat{X})$ and $\Delta_2 \geq Ed(X, \tilde{X})$ and the decoding functions $G_1(X, W_1)$ and $G_2(X, W_1, W_2)$.

We next describe the mechanisms of random code selection and the encoding and decoding operations.

Code Generation:

We first construct the codebook used at the first stage. For the first stage, 2^{NR_1}, $R_1 \geq I(X; W_1) + \epsilon_1 + \delta$, sequences $\{\boldsymbol{W}_1(k)\}$, $k \in [1, ..., 2^{NR_1}]$, are drawn independently from $T^{\delta}_{P_{W_1}}$. Let us denote the set of these sequences by $\mathcal{C}_1$. For each codeword $\boldsymbol{w}_1$, a set of 2^{NR_2}, $R_2 \geq I(X; W_2|W_1) + \epsilon_2 + \delta$, second-stage codewords $\{\boldsymbol{W}_2(j)\}$, $j \in [1, ..., 2^{NR_2}]$, are

independently drawn from $T^{\delta}_{P_{W_2|W_1}}(\boldsymbol{w}_1)$. We denote this set by $\mathcal{C}_2(\boldsymbol{w}_1)$ and its elements by $\{\boldsymbol{W}_2(\boldsymbol{w}_1, j)\}$. 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 $\boldsymbol{W}_1(k) = \boldsymbol{w}_1$ such that (s.t.) the pair $(\boldsymbol{x}, \boldsymbol{w}_1) \in T^{2\delta}_{P_{XW_1}}$, the index k is chosen for transmission at the first stage. Next, if the codebook $\mathcal{C}_2(\boldsymbol{w}_1)$ contains a sequence $\boldsymbol{W}_2(\boldsymbol{w}_1, j) = \boldsymbol{w}_2$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_1, \boldsymbol{w}_2) \in T^{3\delta}_{P_{XW_1W_2}}$, the index j is chosen for transmission at the second stage. If there exist a number of sequences $\boldsymbol{W}_1 \in \mathcal{C}_1$ and $\boldsymbol{W}_2 \in \mathcal{C}_2(\boldsymbol{W}_1)$ that are jointly typical with $\boldsymbol{x}$, the above described process is applied to the the first matching $\boldsymbol{W}_1(k) = \boldsymbol{w}_1$ found in $\mathcal{C}_1$ and $\boldsymbol{W}_2(\boldsymbol{w}_1, j)$ found in $\mathcal{C}_2(\boldsymbol{w}_1)$, respectively.

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 $\nexists\, \boldsymbol{W}_2(\boldsymbol{w}_1, j) = \boldsymbol{w}_2$ s.t. $(\boldsymbol{x}, \boldsymbol{w}_1, \boldsymbol{w}_2) \in T^{3\delta}_{P_{XW_1W_2}}$, an arbitrary error message is transmitted at both stages.

Decoding:
The decoder of the first stage retrieves the first-stage codeword according to its index and generates the reproduction by $\hat{X}_i = G_1\left(W_{1,i}(k), Z_i\right)$, $i \in [1, 2..., N]$. Similarly, the decoder of the second stage retrieves both the first-stage and the second-stage codewords and creates the reconstruction of the source according to $\tilde{X}_i = G_2\left(W_{1,i}\left(k\right), W_{2,i}\left(W_{1,i}\left(k\right), j\right), Y_i\right)$, $i \in [1, 2..., N]$.

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. $(\boldsymbol{x}, \boldsymbol{w}_1, \boldsymbol{w}_2) \in T^{3\delta}_{P_{XW_1W_2}}$. 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 applied twice [7, pp. 436, Lemma 14.8.1], the following is satisfied: $(\boldsymbol{X}, \boldsymbol{Z}, \hat{\mathbf{X}}) \in$

$T^{5\delta|\mathcal{W}_1\times\mathcal{W}_2|}_{P_{XZ\hat{X}}}$ and $(\boldsymbol{X},\, \boldsymbol{Y}, \tilde{\mathbf{X}}) \in T^{5\delta|\mathcal{W}_1\times\mathcal{W}_2|}_{P_{XY\tilde{X}}}$. The joint typicality of $(\boldsymbol{X}, \hat{\mathbf{X}})$ and $(\boldsymbol{X}, \tilde{\mathbf{X}})$ imposes that the distortion constraints (16) and (17) 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

$$P_e \leq P_{e_1} + P_{e_2} + P_{e_3}. \tag{45}$$

The fact that $P_{e_1} \to 0$ follows from the properties of typical sequences [7]. As for P_{e_2}, we have:

$$P_{e_2} \stackrel{\triangle}{=} \prod_{k=1}^{2^{NR_1}} \Pr\{(\boldsymbol{x},\, \boldsymbol{W}_1) \notin T^{2\delta}_{P_{XW_1}}\}. \tag{46}$$

Now, for every k:

$$\begin{aligned} \Pr\{(\boldsymbol{x},\, \boldsymbol{W}_1) \notin T^{2\delta}_{P_{XW_1}}\} &= 1 - \Pr\{(\boldsymbol{x},\, \boldsymbol{W}_1) \in T^{2\delta}_{P_{XW_1}}\} \\ &= 1 - \frac{|T^{2\delta}_{P_{XW_1}}|}{|T^{\delta}_{P_{W_1}}|} \\ &= 1 - 2^{-N[I(X;W_1)+\epsilon_1]}, \end{aligned} \tag{47}$$

where the last equation follows from the size of typical sequences as are given in [7]. Substitution of (47) into (46) and application of the well-known inequality $(1-v)^N \leq \exp(-vN)$, provides us with the following upper-bound for $N \to \infty$:

$$P_{e_2} \leq \left[1 - 2^{-N[I(X;W_1)+\epsilon_1]}\right]^m \leq \exp\left\{-2^{NR_1} \cdot 2^{-N[I(X;U)+\epsilon_1]}\right\} \to 0, \tag{48}$$

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

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

$$P_{e_3} \stackrel{\triangle}{=} \prod_{j=1}^{2^{NR_2}} \Pr\{(\boldsymbol{x}, \boldsymbol{w}_1,\, \boldsymbol{W}_2(w_1, j)) \notin T^{3\delta}_{P_{XW_1W_2}}\}. \tag{49}$$

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

$$\Pr\{(\boldsymbol{x}, \boldsymbol{w}_1,\, \boldsymbol{W}_2) \notin T^{3\delta}_{P_{XW_1W_2}}\} \leq 1 - 2^{-N[I(X;W_2|W_1)+\epsilon_2]}, \tag{50}$$

and therefore, substitution of (50) into (49) gives

$$P_{e_3} \leq \left[1 - 2^{-N[I(X;W_2|W_1)+\epsilon_2]}\right]^{2^{NR_2}} \leq \exp\left\{-2^{NR_2} \cdot 2^{-N[I(X;W_2|W_1)+\epsilon_2]}\right\} \to 0, \tag{51}$$

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 $\{\mathcal{C}_2\}$ that give rise to the reliable source reconstruction at both stages with communication rates R_1 and R_2.

8 Converse Proofs

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. We also adjust the proof of Theorem 2 to show that the necessity part of Theorem 5 holds.

Let (f_1, g_1, f_2, g_2) be given encoder and decoder functions for which $Ed_1(\boldsymbol{X}, \hat{\boldsymbol{X}}) \leq N\Delta_1$ and $Ed_2(\boldsymbol{X}, \tilde{\boldsymbol{X}}) \leq N\Delta_2$. In the proof, for the first and the second steps of the communication protocol, we examine the mutual information $I(\boldsymbol{X}; \boldsymbol{V})$ and $I(\boldsymbol{X}; \boldsymbol{B})$, respectively.

Due to the physical structure of the communication scheme, $\boldsymbol{X} \div \boldsymbol{U} \div \boldsymbol{V}$ is a Markov chain, and therefore, by the data processing inequality, we obtain

$$I(\boldsymbol{X}; \boldsymbol{V}) \leq I(\boldsymbol{U}; \boldsymbol{V}) \overset{(a)}{\leq} m_1 C_1, \tag{52}$$

where (a) follows from the capacity formula for a stationary memoryless channel [7]. On

the other hand,

$$
\begin{aligned}
I(\boldsymbol{X};\, \boldsymbol{V}) &= H(\boldsymbol{X}) - H(\boldsymbol{X}|\, \boldsymbol{V}) && (53)\\
&= \sum_{i=1}^{N} [H(X_i|X_1^{i-1}) - H(X_i|X_1^{i-1},\, \boldsymbol{V})] && (54)\\
&\overset{(a)}{=} \sum_{i=1}^{N} [H(X_i) - && (55)\\
&- H(X_i|X_1^{i-1}, Y_1^{i-1}, Z_1^{i-1},\, \boldsymbol{V})]\\
&= \sum_{i=1}^{N} I(X_i; X_1^{i-1}, Y_1^{i-1}, Z_1^{i-1},\, \boldsymbol{V}) && (56)\\
&\overset{(b)}{=} \sum_{i=1}^{N} I(X_i; \hat{W}_{1,i}) && (57)\\
&\overset{(c)}{=} NI(X_T; \hat{W}_{1,T}|T) && (58)\\
&\overset{(d)}{=} NI(X; \hat{W}_1|T) && (59)\\
&= N[I(X; \hat{W}_1, T) - I(X; T)] && (60)\\
&\overset{(e)}{=} NI(X; \hat{W}_1, T) && (61)\\
&\overset{(f)}{=} NI(X; W_1), && (62)
\end{aligned}
$$

where (a) follows from the fact that the source is memoryless and from the Markov chain $X_i \div (X_1^{i-1},\, \boldsymbol{V}) \div (Y_1^{i-1}, Z_1^{i-1})$; (b) by denoting $\hat{W}_{1,i} \overset{\triangle}{=} (X_1^{i-1}, Y_1^{i-1}, Z_1^{i-1},\, \boldsymbol{V})$; (c) by defining a time-sharing auxiliary random variable T, distributed uniformly over $\{1, ..., N\}$ independently of all other random variables in the system; (d) by further denoting $X \overset{\triangle}{=} X_T$ and $\hat{W}_1 \overset{\triangle}{=} \hat{W}_{1,T}$; (e) is due to the fact that the source is stationary and thus $I(X; T) = 0$; and finally, (f) follows by denoting random variable $W_1 \overset{\triangle}{=} (\hat{W}_1, T)$.

As for the second stage, from the capacity formula for a stationary memoryless channel [7], we obtain

$$I(\boldsymbol{X}; \boldsymbol{B}) \leq I(\boldsymbol{A}; \boldsymbol{B}) \leq m_2 C_2, \tag{63}$$

Also,

$$
\begin{aligned}
I(\boldsymbol{X};\boldsymbol{B}) &= H(\boldsymbol{B}) - H(\boldsymbol{B}|\boldsymbol{X}) && (64)\\
&\overset{(a)}{\geq} H(\boldsymbol{B}|\,\boldsymbol{V}) - H(\boldsymbol{B}|\boldsymbol{X}) + I(\,\boldsymbol{V};\boldsymbol{B}|\boldsymbol{X}) && (65)\\
&= H(\boldsymbol{B}|\,\boldsymbol{V}) - H(\boldsymbol{B}|\boldsymbol{X},\,\boldsymbol{V}) && (66)\\
&= I(\boldsymbol{X};\boldsymbol{B}|\,\boldsymbol{V}) && (67)\\
&= \sum_{i=1}^{N} I(X_i;\boldsymbol{B}|X_1^{i-1},\,\boldsymbol{V}) && (68)\\
&= \sum_{i=1}^{N}[H(X_i|X_1^{i-1},\,\boldsymbol{V}) && (69)\\
&- H(X_i|X_1^{i-1},\,\boldsymbol{V},\boldsymbol{B})] && (70)\\
&\overset{(b)}{=} \sum_{i=1}^{N}[H(X_i|X_1^{i-1},\,\boldsymbol{V},Y_1^{i-1},Z_1^{i-1}) && (71)\\
&- H(X_i|X_1^{i-1},\,\boldsymbol{V},\boldsymbol{B},Y_1^{i-1},Z_1^{i-1})] && (72)\\
&= \sum_{i=1}^{N} I(X_i;\boldsymbol{B}|X_1^{i-1},\,\boldsymbol{V},Y_1^{i-1},Z_1^{i-1}) && (73)\\
&\overset{(c)}{=} \sum_{i=1}^{N} I(X_i;W_{2,i}|\hat{W}_{1,i}) && (74)\\
&= NI(X;W_{2,T}|\hat{W}_{1,T},T) && (75)\\
&= NI(X;W_2|W_1) && (76)
\end{aligned}
$$

where (a) follows from the fact that conditioning reduces entropy and the Markov chain $\boldsymbol{V} \div \boldsymbol{X} \div \boldsymbol{B}$; (b) from the Markov chains $X_i \div (X_1^{i-1},\,\boldsymbol{V}) \div Y_1^{i-1}Z_1^{i-1}$ and $X_i \div (X_1^{i-1},\,\boldsymbol{V},\boldsymbol{B}) \div Y_1^{i-1}Z_1^{i-1}$ and the lines that follow (74) come from using the above-defined auxiliary random variables T, $\{W_{1,i}\}_{i=1}^{N}$ and W_1, denoting $W_{2,i} \overset{\triangle}{=} \boldsymbol{B}$ and finally, letting $W_2 \overset{\triangle}{=} W_{2,T}$.

Obviously, 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 (15) is satisfied.

From (52)-(62) and from (63)-(76) we obtain that

$$m_1 C_1 \geq NI(X;W_1) \tag{77}$$

and

$$m_2 C_2 \geq NI(X;W_2|W_1) \tag{78}$$

i.e., the conditions (23) and (24) of Theorem 2 hold.

We pause to adjust the above proof to the Heegard-Berger setting. Consider a noise-free scenario and denote by f the encoder function. Then the transmission rate is lower-bounded as follows:

$$\begin{aligned}
NR &\geq H(f) && (79)\\
&\geq I(\boldsymbol{X}; f) && (80)\\
&= H(\boldsymbol{X}) - H(\boldsymbol{X}|f) && (81)\\
&= \sum_{i=1}^{N}[H(X_i) - H(X_i|f, X_1^{i-1})] && (82)\\
&\overset{(a)}{=} \sum_{i=1}^{N}[H(X_i) - H(X_i|f, X_1^{i-1}, Y_1^{i-1}, Z_1^{i-1})] && (83)\\
&= \sum_{i=1}^{N} I(X_i; f, X_1^{i-1}, Y_1^{i-1}, Z_1^{i-1}) && (84)\\
&\overset{(b)}{=} \sum_{i=1}^{N} I(X_i; \hat{W}_{1,i}, \hat{W}_{2,i}) && (85)\\
&\overset{(c)}{=} NI(X_T; \hat{W}_{1,T}, \hat{W}_{2,T}|T) && (86)\\
&\overset{(d)}{=} NI(X; \hat{W}_1, \hat{W}_2|T) && (87)\\
&= N[I(X; \hat{W}_1, \hat{W}_2, T) - I(X; T)] && (88)\\
&\overset{(e)}{=} NI(X; \hat{W}_1, \hat{W}_2, T) && (89)\\
&\overset{(f)}{=} NI(X; W_1, W_2), && (90)
\end{aligned}$$

where in (b) we define the auxiliary RVs $\hat{W}_{1,i} \overset{\triangle}{=} (X_1^{i-1}, Z_1^{i-1}, f)$ and $\hat{W}_{2,i} \overset{\triangle}{=} (Y_1^{i-1})$, and (a)-(f) follow from the same reasons as (a)-(f) in the proof of (53)-(62), with some straightforward adjustments of RVs.

It is left to show that there exist functions G_1 and G_2 that satisfy (16)-(17). Denote by $g_{1,i}$ and $g_{2,i}$ the output of the decoders 1 and 2, respectively, at time $i = 1, ..., N$. The random variable W_1 contains $(X_1^{i-1}, Y_1^{i-1}, Z_1^{i-1}, \boldsymbol{V})$ and W_2 contains $\boldsymbol{B}$. By choosing the functions G_1 and G_2 as follows:

$$G_{1,T}(W_1, Y) = g_{1,T}(Y_1^T, \boldsymbol{V}) \tag{91}$$

and

$$G_{2,T}(W_1, W_2, Z) = g_{2,T}(Z_1^T, \boldsymbol{V}, \boldsymbol{B}) \tag{92}$$

we have for the average distortions

$$Ed(X, G_1(W_1, Y)) = \frac{1}{N}\sum_{i=1}^{N} Ed(X, g_{1,i}(Y_1^i, \boldsymbol{V})) \leq \Delta_1 \tag{93}$$

and

$$\begin{aligned} Ed(X, G_2(W_1, W_2, Z)) &= \frac{1}{N}\sum_{i=1}^{N} Ed(X, g_{2,i}(Z_1^i, \boldsymbol{V}, \boldsymbol{B})) \\ &\leq \Delta_2, \end{aligned} \tag{94}$$

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. To this end, we will use the support lemma [11], 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 $j = 1, ..., J$

$$\int_{\mathcal{P}(\mathcal{X})} q_j(Q)\mu(dQ) = \sum_{i=1}^{J} \alpha_i q_j(Q_i). \tag{95}$$

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 the lower bound to R_1:

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

and in the same manner, the lower bound to $R_2 - R_1$ becomes

$$I(X; W_2|W_1) = H(X|W_1) - H(X|W_1, W_2). \tag{97}$$

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 lower bounds to R_1 and $R_2 - R_1$, 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. 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 $\{1, 2, ..., m\}$, $m \stackrel{\triangle}{=} |\mathcal{X}|$:

$$q_i(Q) = \sum_{w_2} Q(x, w_2), \ i \stackrel{\triangle}{=} x = 1, 2, ..., m-1, \tag{98}$$

$$q_m(Q) \ = \ -\sum_{x,w_2} Q(x, w_2) \log \sum_{w_2} Q(x, w_2), \tag{99}$$

and

$$q_{m+1}(Q) \ = \ \sum_{x,w_2} Q(x, w_2) \log Q(x|w_2). \tag{100}$$

Also, we define

$$q_{m+2}(Q) = \sum_{z} \min_{\hat{x}} \sum_{x,w_2} Q(x, w_2) P(z|x) d_1(x, \hat{x}), \tag{101}$$

and

$$q_{m+3}(Q) = \sum_{y} \min_{\tilde{x}} \sum_{x,w_2} Q(x, w_2) P(y|x) d_2(x, \tilde{x}). \tag{102}$$

which along with (99) and (100) 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 $|W_1| = m + 3 = |\mathcal{X}| + 3$ and it satisfies simultaneously:

$$\sum_{w_1} \Pr\{W_1 = w_1\} q_i(P(\cdot|w_1)) = P_X(x), \ i = 1, 2, ..., m-1, \tag{103}$$

$$\sum_{w_1} \Pr\{W_1 = w_1\} q_m(P(\cdot|w_1)) = H(X|W_1), \tag{104}$$

$$\sum_{w_1} \Pr\{W_1 = w_1\} q_{m+1}(P(\cdot|w_1)) = H(X|W_1, W_2), \tag{105}$$

$$\sum_{w_1} \Pr\{W_1 = w_1\} q_{m+2}(P(\cdot|w_1)) = \min_{G_1} Ed(X, G_1(Z, W_1)), \tag{106}$$

$$\sum_{w_1} \Pr\{W_1 = w_1\} q_{m+3}(P(\cdot|w_1)) = \min_{G_2} Ed(X, G_2(Y, W_1, W_2)). \tag{107}$$

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 $m = |\mathcal{X}| \cdot |\mathcal{W}_1| - 1$ constraints to preserve the joint distribution of (X, W_1) just defined and 2 more constraints to preserve the second-stage rate and distortion. 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

$$|\mathcal{W}_2| \leq |\mathcal{X}| \cdot |\mathcal{W}_1| + 1. \tag{108}$$

This completes the proof of the converse part.

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] A. El Gamal and T. Weissman, "Source Coding with Limited Side Information Lookahead at the Decoder", to appear in *IEEE Trans. on Inform. Theory*. Available at **http://www.stanford.edu/ tsachy/research.html**.

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

[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. 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**.

[10] 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.

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