Zero-Delay and Causal Single-User and Multi-User Lossy Source Coding with Decoder Side Information*

Yonatan Kaspi and Neri Merhav

Department of Electrical Engineering

Technion - Israel Institute of Technology

Technion City, Haifa 32000, Israel

Email: {kaspi@tx, merhav@ee}.technion.ac.il

January 1, 2013

Abstract

We consider zero-delay single-user and multi-user source coding with average distortion constraint and decoder side information. The zero-delay constraint translates into causal (sequential) encoder and decoder pairs as well as the use of instantaneous codes. For the single-user setting, we show that optimal performance is attained by time sharing at most two scalar encoder-decoder pairs, that use zero-error side information codes. Side information lookahead is shown to useless in this setting. We show that the restriction to causal encoding functions is the one that causes the performance degradation, compared to unrestricted systems, and not the sequential decoders or instantaneous codes. Furthermore, we show that even without delay constraints, if either the encoder or decoder are restricted a-priori to be scalar, the performance loss cannot be compensated by the other component, which can be scalar as well without further loss. Finally, we show that the multi-terminal source coding problem can be solved in the zero-delay regime and the rate-distortion region is given.

1 Introduction

The classical source coding theorems and their converse counterparts provide fundamental limits which are usually asymptotic in the sense that they can

be achieved by systems that introduce delay (imposed by operating on blocks) and/or require complexity that grows exponentially. In many practical scenarios, however, delay is not tolerable. Such systems include many real-time source coding applications such as streaming live multimedia. In such applications, not only should the data be encoded and decoded in real-time, but also there is no way to improve the reconstruction of previously decoded data as new data arrive. While zero-delay is clearly motivated by fast real-time systems, it can also be motivated by extremely slow systems that observe data once in a long while and must act on this data before new data will arrive. This work focuses on the fundamental limits of such zero-delay systems with the addition of side information (SI) which is available to the decoder.

Zero-delay operation means that each time a new source symbol is observed, a message must be sent to the decoder. The decoder must decode the message and reconstruct the source symbol before the next message will arrive. This, in turn, translates into three constraints: Primarily, the encoder functions must be causal functions of the source symbols. Secondly, the code with which the encoders sends the messages to the decoder must be instantaneous, meaning the decoder can detect the end of each codeword before the whole codeword of the next symbol will arrive. Alternatively, the decoder must be able to parse the bit-stream which is composed of the received codewords in a causal manner. Finally, the decoder must be a causal function of the encoder messages and its SI.

In this work, we consider both single- and multi-user scenarios. We start by describing the single-user setting and related work and then continue to the multiterminal scenarios. In the single-user setting, we consider the following source coding problem: Symbols produced by a discrete memoryless source (DMS) are to be encoded, transmitted noiselessly and reproduced by a decoder without delay. The decoder has access to SI correlated to the source. The average distortion between the source and the reproduced symbols is constrained to be smaller than some predefined constant.

When no distortion is allowed, this problem falls within the scope of zero-error source coding with SI, which was initially introduced by Witsenhausen in [1]. Witsenhausen considered fixed-length coding and characterized the side-information structure as a confusability graph defined on the source alphabet. With this characterization, fixed-length SI codes were equivalent to colorings of the associated graph. Alon and Orlitsky [2] considered variable-rate codes for the zero-error problem. Two classes of codes were considered and lower and upper bounds were derived for both the scalar and infinite block length regimes. The work of Alon and Orlitsky was further extended by Koulgi *et. al* [3], who showed that the asymptotic zero-error rate of transmission is the

complementary graph entropy of an associated graph. It was also shown in [3] that the design of optimal code is NP-hard and a sub-optimal, polynomial time algorithm was proposed. The combination of zero-error codes and maximum per-letter distortion was considered in [4]. As we will see in the sequel, the zero-error source coding with SI is relevant to our setting as well, since the encoder can rely on the SI when sending its messages.

When the source alphabet is finite and distortion is allowed, scalar quantizer design boils down to finding the best partition of the source alphabet into disjoint subsets. The number of such subsets will be governed by the constraints imposed on the system (distortion, rate, encoder's output entropy etc.). In [5], Muresan and Effros proposed an algorithm for finding good partitions in various settings which include the variable-rate scalar Wyner-Ziv [6] setting. However, the optimality of the partitions relied on the convexity of the subsets. Namely, the subsets in each partition must be intervals in the source alphabet. It was noted by the authors that this requirement is too strong in the scalar Wyner-Ziv setting and there are many cases where the optimal partition contains non-convex subsets. We demonstrate such a scenario in the examples section of this work. Bounds on the performance of scalar, fixed-rate source codes with decoder SI were recently given in [7].

Real–time codes form a subclass of the class of causal codes, as defined by Neuhoff and Gilbert [8]. In [8], entropy coding is used on the whole sequence of reproduction symbols, introducing arbitrarily long delays. In the real time case, entropy coding has to be instantaneous, symbol–by–symbol (possibly taking into account past transmitted symbols). It was shown in [8] that for a DMS, the optimal causal encoder consists of time-sharing between no more than two scalar encoders. In [9], Weissman and Merhav extended [8] to include SI at the decoder, encoder or both. The discussion in [9] was restricted, however, only to settings where the encoder and decoder could agree on the reconstruction symbol (i.e., the SI was used for compression, but not in the reproduction at the decoder). Non-causal coding of a source when the decoder has causal access to SI (with possibly a finite look-ahead) was considered by Weissman and El-Gamal [10]. Gaarder and Slepian [11] gave structure theorems for fixed-rate encoders and decoders that are time-invariant finite-state devices.

The results of [8] for causal coding can be adapted to zero-delay coding by replacing the arbitrary long delay entropy coding with zero-delay Huffman coding, thus showing that time-sharing at most two scalar quantizers, followed by Huffman coding, is optimal. When the SI is available to both the encoder and decoder, the results of [9] can be adapted to zero-delay in a similar manner, where at most two scalar quantizers followed by Huffman coding are used for every possible SI symbol. The setting where the decoder can use the SI both

to decode the compressed message and to reproduce the source was left open in [9]. As will be seen in the sequel, our results on zero-delay can be easily adapted back to the causal setting and answer some of the questions left open by [9].

This paper has several contributions. In the single-user setting, the first contribution is the extension of [8] to zero-delay with decoder SI and average distortion constraint, where unlike [9], we do not restrict the usage of the SI. We show that results in the spirit of [8] continue to hold here in the sense that it is optimal to time-share at most two scalar encoders and decoders. However, unlike the encoders in [8] that use Huffman codes in the zero-delay setting, here, the encoders transmit their messages using zero-error SI instantaneous codes, as defined in [2] (and will be properly defined in the sequel). Secondly, we show that there is no performance gain if the decoder has non-causal access to the SI (lookahead) and in fact, only the current SI symbol is useful. This is in contrast to the arbitrary delay and causal SI setting of [10], where SI lookahead was shown to improve the performance. These results place the optimal performance of zero-delay systems far below the classical source coding results where arbitrary delay is allowed. We further ask which of the zero-delay constraints (causal encoder, instantaneous code and causal decoder) are causing this degradation in performance. It is shown that if we remove the constraint on the encoder and allow it to observe the whole sequence in advance but force instantaneous codes (restricting the number of bits in each transmission) and a causal decoder, at least in some cases, the classical rate-distortion performance can be obtained. This suggests that the "blame" for the relatively poor performance of the zero-delay systems falls, at least in these cases, on the restriction to causal encoding functions. The scheme we use to show the last point is surprisingly simple, but to the best of our knowledge it is novel nonetheless. Finally, we show that in the zero-delay setting, if we a-priori restrict attention to scalar decoders (that use only the current encoder message and SI symbol), scalar encoders will do as well as encoders that observe the whole source sequence in advance. Similarly, if we restrict the encoders to be scalar, scalar decoders will do as well as decoders that introduce delay and generally use all the encoder messages to reproduce the symbols. This means that the simplicity of one of the components (encoder / decoder) cannot be compensated by the complexity of the other component.

Moving on to multiterminal scenarios, the zero-delay constraint should be carefully defined. Specifically, suppose we have several non-cooperating encoders and one decoder. How does the decoder receive the bit-streams from each of the encoders and can it decode the message from one of the users before it starts decoding the other message? If the decoder must decode the messages simultaneously, then each encoder should use an independent instantaneous code which can depend only on the past. However, if for example, one of the messages can

be decoded first (say, the messages are interleaved in one bit-stream), the first decoded message, as well as all past messages, can serve as SI when decoding the other message. This idea can be generalized to breaking the encoders messages into small pieces and sending them according to a predefined protocol. We revisit the well known "multiterminal source coding problem" (two user Wyner-Ziv problem), where a source emits correlated pairs and each element of the pairs is observed by a different encoder. Each encoder sends a codeword to a joint decoder which reconstructs the current pair. The distortion between the reconstructions and the original variables should not exceed a given threshold. The challenge here is to find the set of achievable rates and distortions. This problem remained an open problem for over three decades (relevant literature review for this problem will be given in Section 4). We consider a zero-delay version of this problem, where both users and the decoder must operate with zero-delay, assuming that one of the users messages is decoded first and serves as SI to the other user (simultaneous decoding of both messages is a simpler problem and its solution follow immediately from the solution to our problem). We show that, unlike the arbitrary delay setting, the zero-delay problem can be readily solved and the rate distortion region can be characterized.

The remainder of this paper is organized as follows. In Section 2, we give the formal setting and notation used throughout the paper. Section 3 deals with single user problems. Multiterminal zero-delay source coding is handled in Section 4. We conclude this work in Section 5.

2   Preliminaries

We begin with notation conventions. Capital letters represent scalar random variables (RV's), specific realizations of them are denoted by the corresponding lower case letters, and their alphabet – by calligraphic letters. For a positive integers i, j, x_j^i will denote the vector $(x_j, \ldots, x_i)$. If $j = 1$, the subscript will be omitted. The probability distribution over a finite alphabet $\mathcal{X}$ will be denoted by $P_X(\cdot)$. When there is no room for ambiguity, we will use $P(x)$ instead of $P_X(x)$.

For a joint distribution $P(x, y)$, we say that $x, x' \in \mathcal{X}$ are *confusable* if there is a $y \in \mathcal{Y}$ such that $P(x, y) > 0$ and $P(x', y) > 0$. A characteristic graph G is defined on the vertex set of $\mathcal{X}$ and $x, x' \in \mathcal{X}$ are connected by an edge if they are confusable. The pair (G, P), denotes a probabilistic graph consisting of G together with the distribution P over its vertices (here P denotes the marginal on $\mathcal{X}$). We say that two vertices (x, x') are adjacent if there is an edge that connects them in G. The chromatic number of G, $\chi(G)$, is defined to be the smallest number of colors needed to color the vertices of G so that no two

adjacent vertices share the same color.

We focus only on (x, y) pairs with $P(x, y) > 0$ and thus restrict attention only to *restricted inputs* (RI) protocols, as defined in [2]. A protocol for transmitting X when the decoder knows Y, henceforth referred to as an RI protocol, is defined to be a mapping $\phi : \mathcal{X} \to \{0, 1\}^*$ such that if x and x' are confusable then $\phi(x)$ is neither equal to, nor a prefix of $\phi(x')$. An encoder that uses an RI protocol will be referred to as a SI-aware encoder. The length in bits of $\phi(x)$ will be denoted by $|\phi(x)|$. Note that for restricted inputs, the prefix condition should be kept only over edges of G. Namely, for every $y \in \mathcal{Y}$, the prefix condition should be kept over the subset $\{x : P(x, y) > 0\}$. The fact that the same $x \in \mathcal{X}$ can be contained in multiple such subsets, but can have only a single bit representation, complicates the search for the optimal RI protocol. Let $\bar{l}_Y(\phi) \stackrel{\triangle}{=} \sum_{x \in \mathcal{X}} p(x)|\phi(x)|$, where the subscript emphasizes that Y is known to the decoder. Let

$$L_Y(X) = \min \left\{ \bar{l}_y(\phi) : \ \phi \text{ is an RI protocol} \right\}. \tag{1}$$

In degenerate cases, where given a SI symbol there is only one possible source symbol with positive probability, no bits are needed to be sent and we set $|\phi(x)| = 0$. In this case, the decoder knows that the next message in the bit-stream will be that of the next source symbol, so synchronization will be maintained. Upper and lower bounds on $L_Y(X)$, in terms of the entropy of the optimal coloring, are given in [2]. There is no known closed form expression for $L_Y(X)$. We will use $L_Y(X)$ as a figure of merit and our results will be single-letter expressions in terms $L_Y(\cdot)$. In Fig.1, we give an example of bipartite graphs, formed by two joint distributions $P(x, y)$ where an edge connects (x, y) if $P(x, y) > 0$ along with the characteristic graphs and the optimal RI protocols for a uniform $P_X(\cdot)$.
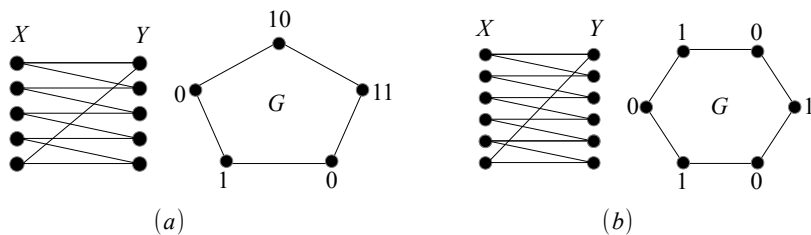

Figure 1: Example of bipartite graphs of $P(x, y)$ along with their associated characteristic graphs G and a RI protocol for 5 (a) and 6 (b) letter alphabets with "typewriter" SI.

In Fig.1a, we used 4 different bit representations for the source symbols.

These bit representations are not prefix-free, but are easily seen to be uniquely decodable with the SI. The optimal bit representations imply a 4-coloring scheme for G, although $\chi(G) = 3$. In Fig.1b, however, $\chi(G) = 2$ and indeed the optimal RI protocol uses a 2-coloring scheme. In Section 3.4, we will return to this example, as well as another example where increasing the quantizer output alphabet *reduces* the rate.

When the graph G is complete, the prefix condition should be kept for all $x \in \mathcal{X}$, thus reducing the RI protocol to regular prefix coding. In this case, $L_Y(X)$ is equal to the average Huffman codeword length of X. This is, of course, also true when SI is not available at the decoder and we can think of the of a bipartite graph where all elements of $\mathcal{X}$ are connected to the same constant y. The average Huffman length of the source X will be denoted by $L(X)$.

In the proof the converses of our theorems, we use a "genie" that reveals common information to both encoder and decoder, thus we define a conditional RI protocol. Let the triplet (X, Y, Z) be distributed with some joint distribution $P(x, y, z)$. The information that is known to both parties will be denoted by Z, while X,Y continue to play the roles of source output and the SI respectively. For any $z \in \mathcal{Z}$, let $\Phi(z)$ denote the set of conditional RI protocols for z. Namely, the set of all RI protocols for source X and SI Y such that $p(x, y, z) > 0$. For any $\phi \in \Phi(z)$, let $L_Y(\phi|z) \overset{\triangle}{=} \sum_{x \in \mathcal{X}} P(x|z)|\phi(x)|$ be the average length when $Z = z$. Similarly, let

$$L_Y(X|Z = z) = \min \left\{ \bar{l}_Y(\phi|z) : \phi \in \Phi(z) \right\}. \tag{2}$$

Finally, let $L_Y(X|Z) = \boldsymbol{E} L_Y(X|Z = z)$ where the expectation is with respect to $P_Z(\cdot)$ and we used the same abuse of notation which is commonly used with the notation of conditional entropy. It follows that $L_Y(X|Z) \leq L_Y(X)$ since the set of RI protocols which are valid without the common knowledge of Z is contained in the set of conditional RI protocols which are valid when Z is known at both ends. In the special case where $Z = Y$, i.e., the SI is known to both parties, the RI protocol for each y reduces to designing a Huffman code according to $P_{X|Y}(\cdot|y)$ for every $y \in \mathcal{Y}$. We will denote conditional Huffman length by $L(X|Z)$ and it is given by

$$L(X|Z) = \sum_z P(z) \min_{l(\cdot)} \sum_x P(x|z)l(x) \tag{3}$$

where in the minimization we consider all functions $l : \mathcal{X} \to \mathbb{R}^+$ that satisfy Kraft's inequality.

3 Single-User Zero-Delay Problems

We investigate the following zero-delay problem. A DMS is emitting pairs of random variables (X_t, Y_t) according to $P(x, y)$, The alphabets of X_t and Y_t, $\mathcal{X}$ and $\mathcal{Y}$, as well as all alphabets in the sequel are finite. At time t, an encoder observes X_t and transmits a compressed codeword, W_t, to a decoder which also observes Y_t. The decoder produces $\hat{X}_t \in \hat{\mathcal{X}}$, a reproduction of X_t, where $\hat{\mathcal{X}}$ is the reproduction alphabet. Given a constant D and a distortion measure $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$, it is required that $\limsup_{n \to \infty} \frac{1}{n} \boldsymbol{E} \sum_{t=1}^{n} d(X_t, \hat{X}_t) \leq D$. Operation is in real-time. This means that the transmitted data, W_t, can be a function only of the encoder's observations no later than time t, namely, X^t. Similarly, the decoder's estimate, $\hat{X}_t$ is a function of (W^t, Y^t). Let L_n denote the total number of bits sent after observing n source symbols. The rate of the encoder is defined by $R \overset{\triangle}{=} \limsup_{n \to \infty} \frac{1}{n} \boldsymbol{E} L_n$. Our goal is to find the tradeoffs between R and D. Since no delay is allowed, W_t must be encoded by an instantaneous code. The model is depicted in Fig.2.
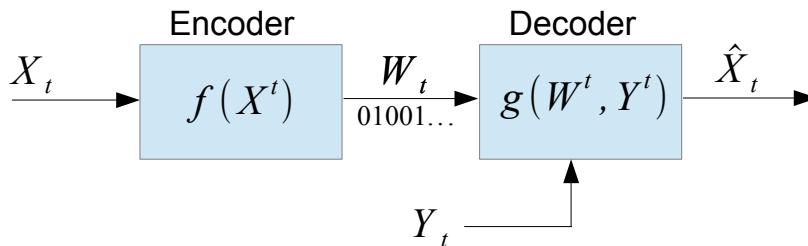

Figure 2: Basic model

In most block coding schemes, there is a probability of failure of the coding system (for example, the probability that the input will not be typical in schemes based on typicality). In these schemes, this probability becomes small as the block length increases and thus effectively does not affect the average distortion. In the zero-delay regime, however, there is no equivalent to failure since the encoding is sequential. An error at time t will mean the decoder failed to parse the bit-stream sent by the encoder correctly. However, such an error event is catastrophic and will render the encoder future messages either meaningless or ambiguous. We therefore restrict our model to one where synchronization is maintained between the encoder and decoder at all times. This means that the encoder messages, W_t, are sent without error. Note that in general $P(w_t, y_t) \neq P(w_t)P(y_t)$ (since (X_t, Y_t) are not independent) and therefore, we will need to consider instantaneous coding of W_t in the presence of correlated SI at the decoder.

We start with a fully zero-delay system and give the rate-distortion function for this system. Later, we relax the zero-delay constraint on the SI and allow lookahead, showing that there is no performance gain. Furthermore we give the optimal performance when one of the components (encoder or decoder) is forced to be scalar, while no constraints are imposed on the other components (decoder or encoder). Subsection 3.1 summarizes the main results and discusses the results of this section. Proofs of Theorem 2 and 3 are given in Subsections 3.2 and 3.3, respectively. Examples of SI aware quantizers that result from these theorems are given in Subsection 3.4. Finally, in Subsection 3.5, we revisit the causal setting as treated in [8] and [9] and derive the equivalent of Theorem 2 for the causal setting, which was left open in [9].

3.1   Main Results

A pair (R, D) is said to be achievable if there exists a rate-R encoder with causal encoding functions $W_t = f_t(X^t), t = 1, 2, \ldots$, and a decoder with causal reproduction functions, $\hat{X}_t = g_t(W^t, Y^t)$, such that the average distortion is smaller than D. Let $\mathcal{R}_{ZD}(D)$ denote the infimum over all rates that are achievable with a given D, where the subscript stands for zero-delay. Let

$$R_{ZD}(D) = \min_{h,f} L_Y(f(X)) \tag{4}$$

where the minimization is over all deterministic functions $h : \mathcal{Z} \times \mathcal{Y} \to \hat{\mathcal{X}}$ and $f : \mathcal{X} \to \mathcal{Z}$ such that $\boldsymbol{E}d(X, h(Y, Z)) \leq D$ (obviously, $|\mathcal{Z}| \leq |\mathcal{X}|$). Finally, denote the lower convex envelope of $R_{ZD}(D)$ by $\underline{R}_{ZD}(D)$.

The first result of this paper is the following theorem:

Theorem 1.

$$\mathcal{R}_{ZD}(D) = \underline{R}_{ZD}(D). \tag{5}$$

This theorem is proved along with Theorem 2 in Subsection 3.2.

Remarks:

1. Theorem 1 implies that optimal performance is attained by time-sharing at most two scalar SI-aware quantizers along with scalar decoders. The role of the function f is to partition the source alphabet into subsets. Note that there is no sense in creating overlapping subsets since it will only increase the uncertainty at the decoder (increase the distortion) while adding edges to the characteristic graph of $f(X)$ with Y (thus increasing the rate). For discrete X, there is a finite number of such partitions. For every possible partition, every choice of a possible h will define a point in the $R - D$ plane. This point will be calculated

by finding the best RI-protocol for the partition defined by f and calculating the average distortion with (f, h). Since there is a finite number of such f, h, there are a finite number of points on the $R - D$ plane. The lower convex envalope of these points gives us $\underline{R}_{ZD}(D)$, which will be piecewise linear.

2. For a given D, the partition of the source alphabet with the minimal number of subsets which achieves distortion D is not necessarily optimal and increasing the number of subsets (thus enlarging the quantizer's alphabet size) can *reduce* the average rate. To see this, assume that we have found f, h where f partitions the alphabet into the minimal number of subsets for this specific D. If at least one of these subsets contains more than one element, we can always find f' and h' such that f' will divide this subset into two disjoint subsets, while h' will be defined the same as h for the original subset and therefore f', h' will have the same average distortion as the original f, h. While it seems intuitive that the rate needed for f' will be higher that the rate needed for f, this is not necessarily true. The reason is that the characteristic graph of $f(X)$ is fully connected, while this is not always the case for the characteristic graph of $f'(X)$. To see why the characteristic graph of $f(X)$ is fully connected, note that if two subsets are not connected in the characteristic graph of $f(X)$, then they are not confusable. Therefore, we can combine these subsets into one subset without loss of performance since h can act differently for the $Y's$ associated with the original subsets. But this contradicts the assumption that f was the partition with the minimal number of subsets for the given D. If the characteristic graph of $f'(X)$ is not fully connected, the SI can be used to alleviate the prefix conditions on the code for $f'(X)$ and shorten its length. Therefore, increasing the output size of the quantizer might reduce the average length. We demonstrate such a phenomenon in Section 3.4.3.

3. There is no loss of generality in the restriction to deterministic encoders since $L_Y(Z)$ is a concave functional of $\{P(z|x)\}$ while the distortion is linear $\{P(z|x)\}$. Therefore, optimizing over the whole convex set of stochastic encoders (represented by distributions $\{P(z|x)\}$) is equivalent to optimizing only over the extreme points of this set, which are the deterministic encoders.

4. The result of this theorem relies heavily on the fact that no delay/ encoder lookahead is allowed. If, for example, the encoder will be allowed a lookahead of one symbol, i.e., $W_t = f(X^{t+1})$, then one can think of the problem as a zero-delay coding of a Markov source, with symbols $\tilde{X}_t = (X_t, X_{t+1})$. The results presented in [12] suggest that the optimal encoder in this case should use $\tilde{X}_t$ *and* its previous messages[1]. While [12] does not rule out the existence of simpler

---

[1]Although the result of [12] holds for a cost function which is a linear combination of the distortion and the average length, it can be adopted to the constrained setting as well when assuming that there is common randomness between encoder and decoder, see also [13] for an example.

optimal encoders, it seems intuitive that for Markov sources the encoder should not be scalar.

Let $\mathcal{R}_{ZD}^{\boldsymbol{y}}(D)$ denote the infimum over all rates that are achievable with a given D, with the same encoders as before and decoders that can use the whole SI sequence, i.e., $\hat{X}_t = g_t(W^t, Y^n)$. We have the following theorem:

Theorem 2.

$$\mathcal{R}_{ZD}^{\boldsymbol{y}}(D) = \underline{R}_{ZD}(D). \tag{6}$$

We prove Theorem 2 in Section 3.2. Theorem 1 is a special case of Theorem 2 and the direct part of the proof, as apparent from the theorems, is the same for both theorems.

The theorem states that allowing the decoder to observe the future SI symbols will not result in any performance gain. This is in contrast to the setting of non-causal access to the source and causal access to the SI at the decoder, treated in [10], where it was shown that SI look-ahead can improve performance.

Now assume that the encoder, in addition to the source X_t, observes causally another process, $S_t \in \mathcal{S}$, which is jointly distributed with the source and SI according to some joint distribution $P(S^n, X^n, Y^n) = \prod_{t=1}^{n} P(S_t, X_t, Y_t)$, i.e., the encoder has access to SI as well as the source. The goal remains to reconstruct only the source X_t with distortion not exceeding D. Let us define $\mathcal{R}_{ZD,EncSI}(D)$ to be the infimum over all rates that are achievable with a given D, where the encoders are causal functions $W_t = f_t(S^t, X^t)$ and the decoders are as in Theorem 2, i.e., causal in $\{W_t\}$ and access the whole SI. Let us define

$$R_{ZD,EncSI}(D) = \min_{h,f} L_Y(f(X, S)) \tag{7}$$

where the minimization is over all deterministic functions $h : \mathcal{Z} \times \mathcal{Y} \to \hat{\mathcal{X}}$ and $f : \mathcal{X} \times \mathcal{S} \to \mathcal{Z}$ such that $\boldsymbol{E}d(X, h(Y, Z)) \leq D$. We have the following corollary which follows from Theorem 2, and is proved exactly in the same way

Corrolary 1.

$$\mathcal{R}_{ZD,EncSI}(D) = R_{ZD,EncSI}(D). \tag{8}$$

Unlike the case without the SI, where causal encoder SI cannot help (since it can be simulated by the encoder even if it is not available), when the decoder has SI, encoder SI indeed helps and can be viewed as extra encoder insight on the decoder SI. In the extreme case when $S_t = Y_t$, we obtain back the results of [9], adopted to the zero-delay setting. Note that with S at the encoder, (x, x') will be confusable of there exist (y, s) such that $P(s, x, y) > 0$ and $P(s, x', y) > 0$

and thus, even if no distortion is allowed, S can be used to reduce the number of edges in the characteristic graph G and thus reduce the constraints on the RI protocol, achieving a lower average rate.

From Theorems 1 and 2, it is apparent that as long as the encoding function is causal in the source, and the decoder is causal in the encoder messages, scalar pairs of encoders and decoders (codecs) are optimal. This suggests, as stated in the Introduction, that the optimal performance achievable by zero-delay systems is by far inferior to the performance of systems that allow arbitrary delay. The inquisitive reader might ask at this point which of the three constraints imposed by our zero-delay model (causal encoding function / causal decoding functions / sequential messages with instantaneous code) is to be "blamed" for the rather poor performance given by Theorems 1 and 2. Had we alleviated one of these constraints, can optimal rate-distortion performance be achieved?

We now show by example that at least without SI at the decoder, there are sources and distortion measures for which the answer to the last question is affirmative. To see this, think of a system with an encoder that observes the whole sequence X^n. The encoder can send no more than $\log|\hat{\mathcal{X}}|$ bits per transmission and the decoders are causal (sequential) in the encoder's messages, meaning that $\hat{X}_t$ is calculated using the data received in the first t encoder transmissions. Compared to the systems described above, we only alleviated the constraint on the encoder. Note that we have to restrict the number of bits the encoder can send in each transmission, otherwise, there is no meaning to the decoder being sequential since the encoder can send the description for $\hat{X}^n$ in a single transmission, as done in block-coding schemes. Such a scheme can be extremely useful for streaming purposes, where the whole stream is known in advance to the encoder and it should be streamed to a decoder without decoding delay. Our scheme works as follows: A classical rate-distortion random code [14] is used. The encoder finds the first codeword ($\hat{X}^n$) in the codebook which is distortion-typical (as defined in [14]) and starts to transmit the *reproduction symbols*, $\hat{X}_t$, sequentially, using $\log|\hat{\mathcal{X}}|$ bits each transmission (in contrast to the classical encoder which will send the index of the codeword in a block). The decoder outputs $\hat{X}_t$ as it receives it. The idea here, is that after receiving enough symbols, the decoder can detect the specific codeword in the codebook since, with high probability, no other codeword will have the same prefix and the rest of the reproduction symbols can be reproduced without further transmissions from the encoder. We show in Appendix A that such a scheme can can achieve optimal rate distortion performance ($R(D)$ bits per source sample) in some cases and is less than one bit away from optimality in the remainder cases. Tree codes are another example of sequential schemes (see [15] and references therein), but there, the fact that the rate is constrained to be a the natural logarithm of an

integer usually forces the schemes to work in small blocks, unlike the scheme presented above. The point we make here is that at least for some combinations of sources and distortion functions, causality constraints only on the decoders and the use of instantaneous codes does not affect the performance.

Next, we ask what is the best attainable performance if we are constrained to use a scalar decoder/encoder, when the other component (encoder/decoder) is unconstrained. For example, we have a scalar encoder, but the decoder can wait until it received (W^n, Y^n) and only then output $\hat{X}^n$ or vice-verse. Such a limitation is motivated by practical constraints imposed on the encoding/decoding devices, for example, the encoder is a simple sensor but the decoder which receives data from another sensor as SI can be as complex as needed for optimal performance. Note that for a scalar decoder, although we allow a non-causal encoder, we still require that the encoder will send a message every time instance and the decoder will reconstruct $\hat{X}_t$ according to this message and SI. Can the simplicity of one of the elements be compensated by the complexity of the other? The next theorem will assert that the answer to this question is negative.

Let $\mathcal{R}_{s-d}(D)$ denote the infimum over all rates that are achievable with a given D, when non-causal encoders are allowed, i.e., $W_t = f_t(X^n)$, but the decoders are restricted to be scalar in W_t, i.e., $\hat{X}_t = g_t(W_t, Y^n)$ (the s-d subscript stands for scalar decoder). Similarly, let $\mathcal{R}_{s-e}(D)$ denote the infimum over all rates that are achievable with a given D, when non-causal decoders are allowed, i.e., $\hat{X}_t = g_t(W^n, Y^n)$, but the encoders are restricted to be scalar in X_t, i.e., $W_t = f_t(X_t)$. We have the following theorem:

Theorem 3.

$$\mathcal{R}_{s-d}(D) = \mathcal{R}_{s-e}(D) = \underline{R}_{ZD}(D). \tag{9}$$

Theorem 3 states that when either the encoder or decoder are constrained to be scalar, the other side can be scalar as well without any performance loss. Note that for the scalar decoder case, this is true even if the decoder is scalar only in the encoder's messages but has full SI look-ahead. This theorem extends Theorem 5 of [16] to include variable rate, look-ahead and SI.

The next two subsections will be devoted to the proofs of Theorem 2 and the right hand side of Theorem 3. The proof that $\mathcal{R}_{s-d}(D) = \underline{R}_{ZD}(D)$, which follows the same lines, is deferred to Appendix B. Examples are given in Subsection 3.4. Finally, the causal setting will be treated in Section 3.5.

3.2 Proof of Theorem 2

3.2.1 Converse part

At every stage, the encoder sends a message W_t. This message is in general a function of X^t. The decoder at time t has already received W^{t-1} and has access to Y^n. Only the current and past SI, Y^t, serves as SI when sending W_t since Y_{t+1}^n is independent of X^t. Therefore, we have

$$nR \geq \sum_{t=1}^{n} L_{Y^t}(W_t | W^{t-1}). \tag{10}$$

To avoid the complex SI structure, which depends on the time instance t, we use a genie aided scheme. At each time instant, a genie reveals all past SI symbols to the encoder and all past source symbols to the decoder. With this "genie aided" feedback and feed–forward, at each time instant, only Y_t serves as SI which is not known to both parties. Therefore, the minimal average number of transmitted bits at each stage is lower bounded by $L_{Y_t}(W_t | X^{t-1}, Y^{t-1})$. For any sequence of encoders which are functions of (X^t, Y^{t-1}) and any sequence of reproduction decoders which are functions of (W_t, X^{t-1}, Y^n), satisfying the distortion constraint, we have:

$$nR \geq \sum_{t=1}^{n} L_{Y_t}(W_t | X^{t-1}, Y^{t-1})$$

$$\geq \sum_{t=1}^{n} L_{Y_t}(W_t | X^{t-1}, Y^{t-1}, Y_{t+1}^n) \tag{11}$$

$$= \sum_{t=1}^{n} \int L_{Y_t}(W_t | x^{t-1}, y^{t-1}, y_{t+1}^n) d\mu(x^{t-1}, y^{t-1}, y_{t+1}^n)$$

$$= \sum_{t=1}^{n} \int L_{Y_t}(f_t(X_t, x^{t-1}, y^{t-1}) | x^{t-1}, y^{t-1}, y_{t+1}^n) d\mu(x^{t-1}, y^{t-1}, y_{t+1}^n)$$

$$= \sum_{t=1}^{n} \int L_{Y_t}(f_t(X_t, x^{t-1}, y^{t-1})) d\mu(x^{t-1}, y^{t-1}, y_{t+1}^n) \tag{12}$$

where in (11) we used the fact that conditioning reduces the average length. In the next line, $\mu(\cdot)$ denotes the joint probability mass function of its arguments and the last equation is true since X_t is independent of $(X^{t-1}, Y^{t-1}, Y_{t+1}^n)$. Now, $f_t(X_t, x^{t-1}, y^{t-1})$ can be seen as a specific choice of $f(X_t)$ in the definition of $R_{ZD}(D)$, (4). This, along with the fact that we know that $Y^{t-1} = y^{t-1}, Y_{t+1}^n = y_{t+1}^n$ and $X^{t-1} = x^{t-1}$, makes the decoding function
$\hat{X}_t = g_t(f_t(X_t, x^{t-1}), x^{t-1}, y^{t-1}, Y_t, y_{t+1}^n)$ a specific choice of $h(\cdot, \cdot)$ in the defi-

nition of $R_{ZD}(D)$. We therefore have

$$nR \geq \sum_{t=1}^{n} \int L_{Y_t}(f_t(X_t, x^{t-1}, y^{t-1}))d\mu(x^{t-1}, y^{t-1}, y_{t+1}^n)$$

$$\geq \sum_{t=1}^{n} \int R_{ZD}(\boldsymbol{E}[d(X_t, g_t(f_t(X_t, x^{t-1}, y^{t-1}), x^{t-1}, y^{t-1}, Y_t), y_{t+1}^n)|x^{t-1}, y^{t-1}, y_{t+1}^n]) \times$$

$$d\mu(x^{t-1}, y^{t-1}, y_{t+1}^n) \tag{13}$$

$$\geq \sum_{t=1}^{n} \int \underline{R}_{ZD}(\boldsymbol{E}[d(X_t, g_t(f_t(X_t, x^{t-1}, y^{t-1}), x^{t-1}, y^{t-1}, Y_t, y_{t+1}^n))|x^{t-1}, y^{t-1},$$

$$y_{t+1}^n])d\mu(x^{t-1}, y^{t-1}, y_{t+1}^n) \tag{14}$$

$$\geq \sum_{t=1}^{n} \underline{R}_{ZD}\left(\int \boldsymbol{E}[d(X_t, g_t(f_t(X_t, x^{t-1}, y^{t-1}), x^{t-1}, y^{t-1}, Y_t, y_{t+1}^n))|x^{t-1}, y^{t-1},\right.$$

$$\left. y_{t+1}^n]d\mu(x^{t-1}, y^{t-1}, y_{t+1}^n)\right) \tag{15}$$

$$= \sum_{t=1}^{n} \underline{R}_{ZD}\left(\boldsymbol{E}\left[d(X_t, g_t(f_t(X^t, Y^{t-1}), X^{t-1}, Y^n))\right]\right)$$

$$= \sum_{t=1}^{n} \underline{R}_{ZD}\left(\boldsymbol{E}\left[d(X_t, \hat{X}_t)\right]\right)$$

$$\geq n\underline{R}_{ZD}\left(\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{E}\left[d(X_t, \hat{X}_t)\right]\right) \tag{16}$$

$$\geq n\underline{R}_{ZD}(D), \tag{17}$$

where (13) follows from the definition of $R_{ZD}(D)$ and the discussion following (12), (14) follows from the definition of $\underline{R}_{ZD}(D)$, (15) and (16) follow from the convexity of $\underline{R}_{ZD}(D)$. Finally, (17) follows from the monotonicity of $\underline{R}_{ZD}(D)$. Combining the above with the direct part given in the next subsection, we also proved that feedback of the SI and feed-forward of the source cannot improve performance here.

3.2.2 Direct part

The direct part of the theorem is obtained by time–sharing two scalar SI-aware quantizers. By definition of $\underline{R}_{ZD}(D)$, we have that there exist (D_1, D_2, λ) such that $D = \lambda D_1 + (1 - \lambda)D_2$ and $(f_1, h_1), (f_2, h_2)$ that are the achievers of $R_{ZD}(D_1)$ and $R_{ZD}(D_2)$, respectively, such that $\lambda R_{ZD}(D_1) + (1-\lambda)R_{ZD}(D_2) = \underline{R}_{ZD}(D)$. Let ϕ_1, ϕ_1 be the optimal protocols for $Z_{1,t} = f_1(X_t), Z_{2,t} = f_2(X_t)$ respectively. Also, let $k_n \leq n$ be a non-decreasing sequence of integers such that $\lim_{n\to\infty} \frac{k_n}{n} = \lambda$. For every n, we use (f_1, h_1) for the first k_n stages and

(f_2, h_2) for the rest of the n–block. The resulting $Z_{i,t}, i = 1, 2$, are coded with the optimal protocols ϕ_1 or ϕ_2. The average distortion of this scheme is given by

$$
\frac{1}{n} \sum_{t=1}^{n} \boldsymbol{E} d(X_t, g(Y^t, Z^t))
$$
$$
= \frac{k_n}{n} \boldsymbol{E} d(X_t, h_1(Y_t, Z_{1,t})) + \frac{n - k_n}{n} \boldsymbol{E} d(X_t, h_2(Y_t, Z_{2,t}))
$$
$$
\leq \frac{k_n}{n} D_1 + \frac{n - k_n}{n} D_2 \tag{18}
$$

and therefore, $\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{E} d(X_t, g(Y^t, Z^t)) \leq D$. The rate of the code is given by

$$
\frac{1}{n} \boldsymbol{E} L_n = \frac{1}{n} \sum_{t=1}^{k_n} \boldsymbol{E} |\phi_1(Z_{1,t})| + \frac{1}{n} \sum_{t=k_n+1}^{n} \boldsymbol{E} |\phi_2(Z_{2,t})|
$$
$$
= \frac{k_n}{n} L(f_1(X)) + \frac{n - k_n}{n} L(f_2(X)) \tag{19}
$$

Therefore,

$$
R = \lim_{n \to \infty} \frac{1}{n} \boldsymbol{E} L_n
$$
$$
= \lambda L_Y(f_1(X)) + (1 - \lambda) L_Y(f_2(X))
$$
$$
= \underline{R}_{ZD}(D). \tag{20}
$$

3.3 Proof of Theorem 3

In this subsection, we prove the first part of Theorem 3, namely, we show that $\mathcal{R}_{s-d}(D) = \underline{R}_{ZD}(D)$. The proof that $\mathcal{R}_{s-e}(D) = \underline{R}_{ZD}(D)$ follows the same lines and is deferred to Appendix B. Since the direct part is the same as the direct part of Theorem 2, we need to prove the converse only.

We use the same genie-aided scheme as in the proof of Theorem 2 and prove a stronger converse: For any sequence of encoders, with possible access to the whole source sequence and past SI, $W_t = f_t(X^n, Y^{t-1})$ and any sequence of decoders with access to the whole SI sequence, past source symbols and the current encoder output, $\hat{X}_t = g_t(W_t, X^{t-1}, Y^n)$, that achieve average distortion

D (i.e., $\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{E}d(X_t,\hat{X}_t) \leq D$), we have

$$nR \geq \sum_{t=1}^{n}\int L_{Y_t}(W_t|X^{t-1},Y^{t-1})$$

$$\geq \sum_{t=1}^{n}\int L_{Y_t}(W_t|X^{t-1},X_{t+1}^n,Y^{t-1},Y_{t+1}^n)$$

$$= \sum_{t=1}^{n}\int L_{Y_t}(W_t|x^{t-1},x_{t+1}^n,y^{t-1},y_{t+1}^n)d\mu(x^{t-1},x_{t+1}^n,y^{t-1},y_{t+1}^n)$$

$$= \sum_{t=1}^{n}\int L_{Y_t}(f_t(x^{t-1},X_t,x_{t+1}^n,y^{t-1})|x^{t-1},x_{t+1}^n,y^{t-1},y_{t+1}^n)\times$$

$$d\mu(x^{t-1},x_{t+1}^n,y^{t-1},y_{t+1}^n)$$

$$\geq \sum_{t=1}^{n}\int L_{Y_t}(f_t(x^{t-1},X_t,x_{t+1}^n,y^{t-1}))d\mu(x^{t-1},x_{t+1}^n,y^{t-1},y_{t+1}^n) \qquad (21)$$

where the last step, as before, relied on the memorylessness of the source. Now, the rate in the inner expression cannot be smaller than the rate of the optimal scalar system (with all conditioned elements serving as index of functions), which achieves the distortion achieved by the given decoding function, $h_t(W_t,y^{t-1},Y_t,y_{t+1}^n)$. Therefore,

$$nR \geq \sum_{t=1}^{n}\int L_{Y_t}(f_t(x^{t-1},X_t,x_{t+1}^n,y^{t-1}))d\mu(x^{t-1},x_{t+1}^n,y^{t-1},y_{t+1}^n)$$

$$\geq \sum_{t=1}^{n}\int \underline{R}_{ZD}\left(\boldsymbol{E}\left[d(X_t,g_t(f_t(x^{t-1},X_t,x_{t+1}^n,y^{t-1}),x^{t-1},y^{t-1},Y_t,y_{t+1}^n))\right.\right.$$

$$\left.\left.|x^{t-1},x_{t+1}^n,y^{t-1},y_{t+1}^n\right]\right)d\mu(x^{t-1},x_{t+1}^n,y^{t-1},y_{t+1}^n)$$

$$\geq \sum_{t=1}^{n}\underline{R}_{ZD}\left(\int \boldsymbol{E}\left[d(X_t,g_t(f_t(x^{t-1},X_t,x_{t+1}^n,y^{t-1}),x^{t-1},y^{t-1},Y_t,y_{t+1}^n))\right.\right.$$

$$\left.\left.|x^{t-1},x_{t+1}^n,y^{t-1},y_{t+1}^n\right]d\mu(x^{t-1},x_{t+1}^n,y^{t-1},y_{t+1}^n)\right) \qquad (22)$$

$$\geq \sum_{t=1}^{n}\underline{R}_{ZD}\left(\boldsymbol{E}\left[d(X_t,g_t(f_t(X^n),X^{t-1},Y^n))\right]\right)$$

$$\geq \underline{R}_{ZD}\left(\sum_{t=1}^{n}\boldsymbol{E}\left[d(X_t,\hat{X}_t)\right]\right) \qquad (23)$$

$$\geq \underline{R}_{ZD}(D) \qquad (24)$$

where in (22) and (23) we used the convexity of $\underline{R}_{ZD}(\cdot)$ and its monotonicity in (24).

3.4 Examples

3.4.1 Lossless transmission

It is interesting to relate the above results to the lossless case which was considered in [2]. Since $D = 0$ cannot be achieved by time-sharing positive distortions, we get that

$$L_Y(X) = \min_{h, f : h(y, f(x)) = x} L_Y(f(X)). \tag{25}$$

Let $Z = f(X)$. Any f which is a coloring of the characteristic graph G and h, which is the mapping from color and y back to x, are valid candidates in the optimization problem of (25). If f is not a valid coloring, meaning that two connected x_1, x_2 will result in the same z, then there is no h which can result in zero error. In essence, we are looking for the coloring for which the restricted inputs protocol will produce the smallest rate. Note that when searching for the best coloring, our performance will be affected only by the characteristic graph G_z which will be built with the "source" (Z, Y). If f is a minimal coloring, i.e, $z \in \{1, 2, \ldots, \chi(G_z)\}$, then G_z is complete. To see this, note that for a minimal coloring, if z_1 and z_2 are not connected, then these colors can be combined and this reduces the number of colors, contradicting the fact that $f(x)$ is a minimal coloring. Remember that a complete graph reduces the RI protocol to Huffman coding. This means that the SI is not helping us to code the colors. Therefore, looking for colorings which will induce an incomplete G_z (i.e., non-optimal coloring) will allow us to use the SI not only to reduce the alphabet of the encoder output, but also for the coding of its output (namely, relax the prefix condition on the codewords when the graph is complete). In the example of Fig.1a, we used a 4-coloring scheme (we had 4 different bit representations for the vertices of G) and not the optimal 3-coloring. Indeed, G_z for the 4-coloring is not complete. For a uniform source, we get an average rate of 1.4 with the 4 coloring and if we had used a 3-coloring we would get an average rate of 1.6.

3.4.2 Uniform source, fully connected SI model

Let the reconstruction alphabet be the same as the source alphabet. We use the Hamming distortion measure ($d(x, \hat{x}) = 0$ if $x = \hat{x}$ and $d(x, \hat{x}) = 1$ otherwise). The encoder partitions the source alphabet into disjoint subsets $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k$, $k \leq |\mathcal{X}|$. When the encoder observes a new source symbol, x, it sends the index of the subset containing x, using an RI protocol, as defined in Section 2. With the Hamming distortion measure, the average distortion is equal to the probability of error. Therefore, the optimal decoder is the maximum likelihood

decoder, namely:

$$\hat{x} = \underset{x}{\operatorname{argmax}} P(y, z|x) = \underset{x \in \mathcal{A}_z}{\operatorname{argmax}} P(y|x).$$

where z is the subset index, sent by the encoder.

Let $|\mathcal{X}| = |\mathcal{Y}| = M$ and let for a small constant p, $P(X = a) = \frac{1}{M}$, $P(y = \alpha|x = a) = 1 - p$ if $\alpha = a$ and $P(y = \alpha|x = a) = \frac{p}{M-1}$ for any $\alpha \neq a$. With this choice of the joint distribution, since the bipartite graph of $\{P(x,y)\}$ is fully connected, then so is the bipartite graph of $\{P(y,z)\}$, regardless of the choice of Z. Therefore, the RI protocol used to describe the index of the subsets is reduced to a Huffman code for Z.

It is shown in [7], that for this distortion measure and source, only the number of partitions, and not their content (i.e., the actual alphabet letters in each subset), affects the average distortion. The distortion as a function of the number of partitions, K, is given by $\frac{p}{M-1}(|\mathcal{X}| - K)$. It turns out that in this case, $\underline{R}_{ZD}(D) = L(X) - \frac{L(X)}{p}D$, which is obtained by time-sharing the two trivial quantizers: the one that does not send information ($R = 0, D = p$) and the lossless quantizer ($R = L(X), D = 0$).

3.4.3   Uniform source, given SI model

We continue with the Hamming distortion measure and a uniform source with $|\mathcal{X}| = 5$. The channel from X to Y is given in Fig.3, along with the characteristic graph of X. In this example we set for $\alpha \in \mathcal{X}$: $P(y = \alpha|x = \alpha) = 1 - p$ and
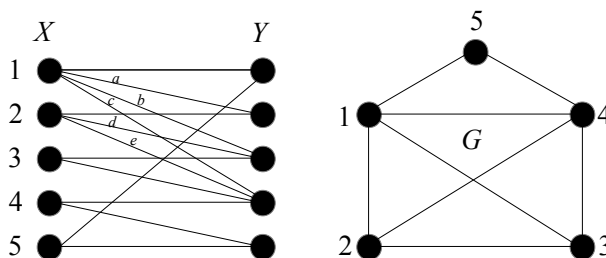

Figure 3: $P(y|x)$ and the resulting characteristic graph.

$a = \frac{1}{2}p, b = \frac{5}{12}p, c = \frac{1}{12}p, d = \frac{3}{4}p, e = \frac{1}{4}p$. Note that the chromatic number of G is 4. This means that any partitioning of the alphabet of X into less than 4 subsets will incur a lossy reconstruction. Unlike the previous example where the SI could be used only in the reconstruction, but not to reduce the length of the transmission (since G was fully connected), here it will be used for both. Note that as in the example of Fig.1, the optimal rate for lossless transmission is

actually obtained by using more subsets than the chromatic number of X. The optimal two subset partition ($|\mathcal{Z}| = 2$) is $\{1,4\}, \{2,3,5\}$, yielding an average distortion of $\frac{13}{60}p$. The rate for this (and any binary) partition is 1. The optimal 3-subset partition is $\{1,4\}, \{2,5\}, \{3\}$ yielding an average distortion of $\frac{1}{60}p$. The average rate for this partitioning is $8/5$. However, in this case it is beneficial to split $\{2,5\}$ and obtain a lower rate of $7/5$ (using the SI to alleviate the prefix requirement). Although we use more subsets, the rate is reduced, as discussed in the remarks that followed the statement of Theorem 1. In Fig.4, we compare $\underline{R}_{ZD}(D)$ to the performance of a system that uses the minimal number of subsets for each D and uses Huffman codes, i.e., a system that uses the SI only for the reproduction but not for the compression.
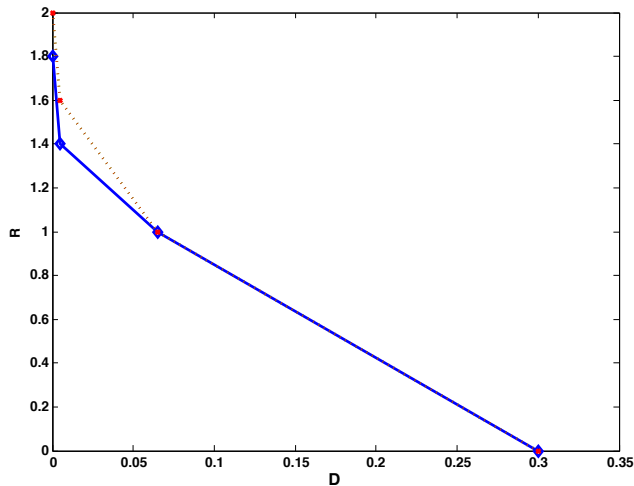

Figure 4: $\underline{R}_{ZD}(D)$ (solid) compared to a system that uses the SI only for reproduction (dotted) with $p = 0.3$.

3.5 The Causal Setting

In this section, we revisit the causal setting considered in [8] and [9]. In this setting, we still restrict the encoding and decoding functions to be causal functions of the their respective inputs, however, we do not restrict the delay which is imposed by the system. Effectively, this means that we drop the instantaneous code constraint and allow more sophisticated coding techniques which introduce delays. In [8], [9], the SI was either not present or was not used in the reconstruction of $\hat{X}_t$ (unless it was available at both the encoder and decoder) and therefore, the encoder could calculate the reproduction symbols. This in turn, enabled to recast the system into a reproduction coder (which creates $\left\{\hat{X}_t\right\}$) followed by a lossless entropy coder, which coded $\hat{X}^n$ into a message that was

sent to the decoder (this is where the delay was introduced). However, when the SI is available only at the decoder and used in the reconstruction, the encoder cannot reproduce $\left\{\hat{X}_t\right\}$ and therefore the encoder and the reconstruction decoder must be decoupled. In [9], the model we treat here was left unsolved. The combination of encoding that introduces delay with causal decoding functions is less motivated in practical terms than the zero-delay models that were introduced in the previous subsections. However, it can represent situations where the lossless coding of the encoder messages is done separately and then streamed to a simple causal decoder. The results which show that scalar codecs are optimal support such a model in retrospect.

In contrast to the zero-delay model, where synchronization must be kept between encoder and decoder over the bit-stream and therefore no decoding errors are allowed, here, since block coding is allowed, we have no such constraint. A block error with a vanishing probability will translate into a vanishing additional distortion as with the classical rate-distortion solutions.

Let $\mathcal{R}_c(D)$ denote the infimum over all rates such that there exist causal encoders $W_t = f_t(X^t)$, decoders which are causal in the encoder's messages $\hat{X}_t = g_t(W^t, Y^n)$ and define

$$R_c(D) = \inf_{f,h} H(f(X)|Y) \tag{26}$$

where the minimum is over all functions $f : \mathcal{X} \to \mathcal{Z}$ and $h : \mathcal{Z} \times \mathcal{Y} \to \hat{\mathcal{X}}$ such that $\boldsymbol{E}d(X, h(f(X), Y)) \leq D$. Finally, define $\underline{R}_c(D)$ to be the lower convex envelope of $R_c(D)$. We have the following theorem:

Theorem 4.

$$\mathcal{R}_c(D) = \underline{R}_c(D). \tag{27}$$

We will only outline the proof since after a few steps it is similar to the proofs of the previous theorems when replacing the average length functional with the entropy functional. In the direct part, we again timeshare at most two scalar encoders and then code the resulting blocks using a Slepian-Wolf [17] code. For the converse, we start with the converse to the lossless source coding with SI, stating that $nR \geq H(W^n|Y^n)$. While this step is valid when the SI in known at both sides when losslessly coding W^n, and might seem a loose lower bound here, we know from the Slepian-Wolf theorem that there is essentially no loss when the SI is available only at the decoder and a vanishing probability of

block error is allowed. We have:

$$
\begin{aligned}
nR &\geq H(W^n|Y^n) \\
&= \sum_{t=1}^{n} H(W_t|W^{t-1}, Y^n) \\
&\geq \sum_{t=1}^{n} H(W_t|W^{t-1}, X^{t-1}, Y^n) \\
&= \sum_{t=1}^{n} H(W_t|X^{t-1}, Y^n) \\
&= \sum_{t=1}^{n} \int H(f_t(X_t, x^{t-1})|x^{t-1}, y^{t-1}, Y_t, y_{t+1}^n) d\mu(x^{t-1}, y^{t-1}, y_{t+1}^n) \\
&= \sum_{t=1}^{n} \int H(f_t(X_t, x^{t-1})|Y_t) d\mu(x^{t-1}, y^{t-1}, y_{t+1}^n) \qquad (28)
\end{aligned}
$$

From here, one can identify that $R_c(D)$ can further lower bound the conditional entropy in the last equation. The rest of of the proof essentially follows the steps of the proof of Theorem 2 after equation (13).

4   Multiterminal Problem

In this section we revisit the multiterminal source coding problem (two-user Wyner-Ziv problem), where two correlated random variables are observed separately by two non-cooperative encoders who communicate with a decoder. The decoder needs to reconstruct both sources and the distortion between the reconstructions and the original variables should not exceed a given threshold. As mentioned in the Introduction, this problem has been open for the general setting for over three decades. There are specific cases in which the arbitrary delay problem can be solved for a general source and distortion measure. When no distortion is allowed, this is the Slepian-Wolf problem [17] and when one of the variables is known to the decoder, this is the original Wyner-Ziv problem [6]. Other examples include the source coding with side information of Ahlswede-Korner and Wyner [18], [19] where arbitrary distortion is allowed for one of the sources and the other source should be reconstructed losslessly. Finally, Beger and Yeung [20] considered a setting where one of the sources is to be perfectly reconstructed and the other source should be reconstructed with a distortion constraint (their setting subsumes all previous examples). Recent results for specific sources or distortion measures include the achievable rate distortion region for the quadratic Gaussian multiterminal source coding problem, given by Wagner, Tavildar, and Viswanath [21] and the characterization of the region

under logarithmic loss given by Courtade and Weissman [22].

We consider a zero-delay version of the multiterminal source coding problem. We show that, unlike the arbitrary delay setting, when the zero delay constraint is imposed, this problem can be readily solved and the rate distortion region can be characterized.

4.1 Formal Definition and Main Result

We start by formally stating the problem. Let (X_t, Y_t) be emitted by a memoryless source. Each of the symbols are encoded by f_t^x, f_t^y respectively to produce $W_t = f_{x,t}(X^t), Z_t = f_{y,t}(Y^t)$. A decoder, $g_t : \mathcal{W}^t \times \mathcal{Z}^t \to \hat{\mathcal{X}} \times \hat{\mathcal{Y}}$, uses (W^t, Z^t) to reproduce the pair $(\hat{X}_t, \hat{Y}_t)$. Let $(g_{x,t}, g_{y,t})$ denote the reproduction functions of X, Y respectively. Given two distortion measures, $d_x : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+, d_y : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}^+$, it is required that $\frac{1}{n} \sum_{t=1}^n \boldsymbol{E} d_x(X_t, \hat{X}_t) \leq D_x$ and $\frac{1}{n} \sum_{t=1}^n \boldsymbol{E} d_y(Y_t, \hat{Y}_t) \leq D_y$ simultaneously. Each encoder encodes its message with an instantaneous code. As noted in the Introduction, there is more than one way to define the zero-delay decoding in a multiterminal setting. We allow either Z_t or W_t to be decoded first by the decoder and therefore the code of the later symbol (and generally all past decoded symbols) can serve as SI. Let l_t^x, l_t^y denote the number of bits transmitted until the time that X_t and Y_t are encoded (inclusive), respectively. The system model is depicted in Fig.5.
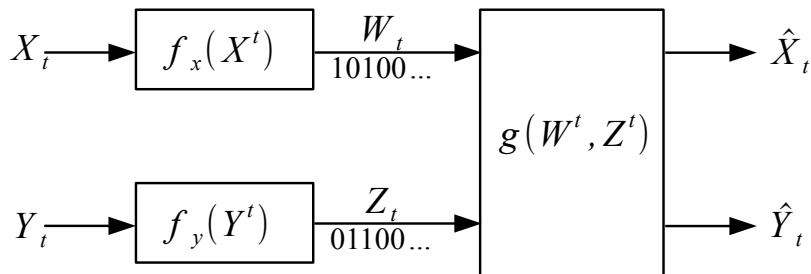

Figure 5: Two users system model

Let Ξ denote the subset of $\mathbb{R}^4$ consisting of all quadruples (R_x, R_y, D_x, D_y) such that for any given $\epsilon > 0$ there exists a sequence functions $\{f_{x,t}\}, \{f_{y,t}\}, \{g_t\}$ and an integer n such that

   i. $\frac{1}{n} \boldsymbol{E} l_n^x \leq R_x + \epsilon$

   ii. $\frac{1}{n} \boldsymbol{E} l_n^y \leq R_y + \epsilon$

   iii. $\frac{1}{n} \sum_{t=1}^n \boldsymbol{E} d_x(X_t, \hat{X}_t) \leq D_x + \epsilon$

   iv. $\frac{1}{n} \sum_{t=1}^n \boldsymbol{E} d_y(Y_t, \hat{Y}_t) \leq D_y + \epsilon$

Our goal is to characterize the above operationally defined region by single-letter information-theoretic functionals. Towards this goal, let us define Δ_{YX}^* to be the set containing quadruples (R_x, R_y, D_x, D_y) such that there exists random variables (U, V, Q) that belong to finite alphabets $\mathcal{U}, \mathcal{V}, \mathcal{Q}$, $|\mathcal{Q}| \leq 5$, $|\mathcal{U}| \leq |\mathcal{X}| \times |\mathcal{Q}|$, $|\mathcal{V}| \leq |\mathcal{Y}| \times |\mathcal{Q}|$ which are jointly distributed with (X, Y) and satisfy

1. $U = f_x(X, Q), V = f_y(Y, Q)$, for some functions $f_x : \mathcal{X} \times \mathcal{Q} \to \mathcal{U}, f_y : \mathcal{Y} \times \mathcal{Q} \to \mathcal{V}$, Q is independent of (X, Y).

2. $\exists g_x, g_y$ such that $g_x : \mathcal{U} \times \mathcal{V} \times \mathcal{Q} \to \hat{\mathcal{X}}, g_y : \mathcal{U} \times \mathcal{V} \times \mathcal{Q} \to \hat{\mathcal{Y}}$ satisfying $\boldsymbol{E} d_x(X, g_x(U, V, Q)) \leq D_x$ and $\boldsymbol{E} d_y(Y, g_y(U, V, Q)) \leq D_y$.

3. $R_x \geq L_V(U|Q)$

4. $R_y \geq L(V|Q)$

The subscript in Δ_{YX}^* stands for the order of transmission, where the message of the Y-encoder, V, is decoded first and serves as SI to the message from the X-encoder. We similarly define Δ_{XY}^* in the same way but with reversed transmission order, i.e., items 3,?? is defined as $R_x \geq L(U|Q)$ and $R_y \geq L_U(V|Q)$. Finally let $\Delta^* = \Delta_{XY}^* \cup \Delta_{YX}^*$ and define $\overline{\Delta^*}$ to be the closure of Δ^*.

The main results of this section is the following theorem:

Theorem 5. $\Xi = \overline{\Delta^*}$.

Theorem 5 basically states that using scalar encoders and decoders is optimal. The role of the encoder that transmits first is to send a message that will not only allow the joint reconstruction, but also to serve as SI for the decoding of the other message, thus reducing the average length which is needed for the other message. In the proof of the converse, we will show that there is no performance gain if each of the encoders receive the observations of the other encoder with delay (without delay, this becomes a single user problem for the source (X, Y) with two distortion measures). There are several obvious extensions to this problem. For example, We can consider breaking each user's message into small pieces and consider all protocols that order the transmission of the pieces between the users, where all previously transmitted pieces serve as SI. Then, we can optimize over all possible such protocols (since the alphabets of X and Y are finite, there is a finite number of such protocols). Additionally, SI can be available to the decoder. Also, the encoder that transmits first might share some information with the second encoder at a given rate. The extension to such scenarios follows the same lines of proof as those of Theorems 1-5 and is therefore omitted. Another variant of this problem is a two-way problem where the X-encoder and Y-encoder communicate, where each user reproduces

the other source to within a given distortion. It can be shown that in such a problem the past is, again, irrelevant and scalar encoders are optimal. In such a scenario, the interaction can be used both to assist the reconstruction and to reduce the average length which is needed to convey each message, as shown by Orlitsky in [23].

We prove the converse and direct parts in the next two subsections, respectively.

4.2   Converse:

We prove the converse assuming Z_t is sent first. The proof when W_t is sent first is the same. Once again, we use a "genie aided" scheme, where each encoder has access to (X^{t-1}, Y^{t-1}) when encoding X_t or Y_t. We will show that even when each encoder knows the past symbols of the other source, still scalar encoders are optimal. With this scheme the only element that is unknown to each of the encoders is the current message of the other encoder and the decoder's data is shared by both encoders.

For any sequence of encoders $\{f_{x,t}(X^t, Y^{t-1})\}$, $\{f_{y,t}(X^{t-1}, Y^t)\}$ and decoders $\{g_{x,t}(W^t, Z^t)\}$, $\{g_{y,t}(W^t, Z^t)\}$ achieving $(R_x, R_y, D_x, D_y) \in \Xi$ the following holds:

$$
\begin{aligned}
nR_y &\geq \sum_{t=1}^{n} L(Z_t | X^{t-1}, Y^{t-1}) \\
&= \sum_{t=1}^{n} L(f_{y,t}(X^{t-1}, Y^{t-1}, Y_t) | X^{t-1}, Y^{t-1}) \\
&= \sum_{t=1}^{n} L(f_{y,t}(\tilde{Q}_t, Y_t) | \tilde{Q}_t)
\end{aligned}
\tag{29}
$$

where we set $\tilde{Q}_t = (X^{t-1}, Y^{t-1})$, which is independent of (X_t, Y_t). We continue by letting M be a random variable distributed uniformly over $\{1, 2, \ldots, n\}$, independent of all other variables. Continuing (29) we have

$$
\begin{aligned}
R_y &\geq \sum_{t=1}^{n} L(f_{y,t}(\tilde{Q}_t, Y_t) | \tilde{Q}_t) \\
&= \frac{1}{n} \sum_{t=1}^{n} L(f_{y,t}(\tilde{Q}_t, Y_t) | \tilde{Q}_t, M = t) \\
&= L(\tilde{f}_y(M, \tilde{Q}_M, Y_M) | \tilde{Q}_M, M) \\
&= L(\tilde{f}_y(Q, Y) | Q) \\
&= L(V | Q)
\end{aligned}
\tag{30}
$$

where $Q \triangleq (\tilde{Q}_M, M)$, and $V \triangleq \tilde{f}_y(Q, Y)$ and we used the fact that Y's distribution does not depend on the time.

Using the same definitions of $M, \tilde{Q}_t, Q, V$ and following the same steps, we have for the rate of the X-encoder:

$$
\begin{aligned}
R_x &\geq \frac{1}{n} \sum_{t=1}^{n} L_{Z_t}(W_t | X^{t-1}, Y^{t-1}) \\
&= \frac{1}{n} \sum_{t=1}^{n} L_{f_{y,t}(X^{t-1}, Y^{t-1}, Y_t)}(f_{x,t}(X^{t-1}, Y^{t-1}, X_t) | X^{t-1}, Y^{t-1}) \\
&= \frac{1}{n} \sum_{t=1}^{n} L_{f_{y,t}(\tilde{Q}_t, Y_t)}(f_{x,t}(\tilde{Q}_t, X_t) | \tilde{Q}_t) \\
&= \frac{1}{n} \sum_{t=1}^{n} L_{f_{y,t}(\tilde{Q}_t, Y_t)}(f_{x,t}(\tilde{Q}_t, X_t) | \tilde{Q}_t, M = t) \\
&= L_{\tilde{f}_y(M, \tilde{Q}_M, Y_M)}(\tilde{f}_x(M, \tilde{Q}_M, X_M) | \tilde{Q}, M) \\
&= L_{\tilde{f}_y(Q, Y)}(\tilde{f}_x(Q, X) | Q) \\
&= L_V(U | Q)
\end{aligned}
\tag{31}
$$

where $U \triangleq \tilde{f}(Q, X)$.

For the distortion we have that:

$$
\begin{aligned}
D_x &\geq \frac{1}{n} \sum_{t=1}^{n} d_x(X_t, \hat{X}_t) \\
&= \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{E} d_x(X_t, g_{x,t}(W^t, Z^t)) \\
&\geq \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{E} \left[d_x(X_t, \tilde{g}_{x,t}(X^{t-1}, Y^{t-1}, f_{x,t}(X^{t-1}, Y^{t-1}, X_t), f_{y,t}(X^{t-1}, Y^{t-1}, Y_t))) \right]
\end{aligned}
\tag{32}
$$

$$
\begin{aligned}
&= \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{E} \left[d_x(X_t, \tilde{g}_{x,t}(\tilde{Q}_t, f_{x,t}(\tilde{Q}_t, X_t), f_{y,t}(\tilde{Q}_t, Y_t))) \right] \\
&= \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{E} \left[d_x(X_t, \tilde{g}_{x,t}(\tilde{Q}_t, f_{x,t}(\tilde{Q}_t, X_t), f_{y,t}(\tilde{Q}_t, Y_t))) | M = t \right] \\
&= \boldsymbol{E} \left[d_x(X_M, \tilde{g}_x(M, \tilde{Q}_M, \tilde{f}_x(M, \tilde{Q}_M, X_M), \tilde{f}_y(M, \tilde{Q}_M, Y_M))) \right] \\
&= \boldsymbol{E} \left[d_x(X, \tilde{g}_x(Q, U, V)) \right]
\end{aligned}
\tag{33}
$$

where in (32) we used the fact that giving the decoder access to (X^{t-1}, Y^{t-1}) can only decrease the average distortion. In exactly the same manner, by replacing

the roles of X and Y, we get

$$D_y \geq \boldsymbol{E}\left[d_y(Y, \tilde{g}_y(Q, U, V))\right]. \tag{34}$$

The bound on the cardinality of Q is a consequence of Caratheodory's theorem [24]: As can be seen in (35)-(38), the point

$$(L_V(U|Q), L(V|Q), \boldsymbol{E}d_x(X, g_x(U, V, Q)), \boldsymbol{E}d_y(X, g_y(U, V, Q)))$$

lies in the convex hull of the set $\{L_V(U|Q = q), L(V|Q = q), D_x(q), D_y(q)\}$ which is a subset of $\mathbb{R}^4$. Therefore, by Caratheodory's theorem, this region will not change if we restrict the alphabet size of Q to 5.

4.3 Direct:

In order to show that $\overline{\Delta^*} \subseteq \Xi$ we timeshare scalar quantizers. For every quadruple $(R_x, R_y, D_x, D_y) \in \overline{\Delta^*}$ We have

$$L_V(U|Q) = \sum_{q=1}^{|\mathcal{Q}|} P(Q = q) L_V(U|Q = q)$$

$$\stackrel{\triangle}{=} \sum_{q=1}^{|\mathcal{Q}|} P(Q = q) R_x(q), \tag{35}$$

$$L(V|Q) = \sum_{q=1}^{|\mathcal{Q}|} P(Q = q) L(V|Q = q)$$

$$\stackrel{\triangle}{=} \sum_{q=1}^{|\mathcal{Q}|} P(Q = q) R_y(q), \tag{36}$$

$$\boldsymbol{E}d_x(X, g(U, V, Q)) = \sum_{q=1}^{|\mathcal{Q}|} P(Q = q) \boldsymbol{E}\left[d_x(X, g_x(U, V, Q))|Q = q\right]$$

$$\stackrel{\triangle}{=} \sum_{q=1}^{|\mathcal{Q}|} P(Q = q) D_x(q), \tag{37}$$

$$\boldsymbol{E}d_x(X, g(U, V, Q)) = \sum_{q=1}^{|\mathcal{Q}|} P(Q = q) \boldsymbol{E}\left[d_y(Y, g_y(U, V, Q))|Q = q\right]$$

$$\stackrel{\triangle}{=} \sum_{q=1}^{|\mathcal{Q}|} P(Q = q) D_y(q). \tag{38}$$

Now, for every $q \in \mathcal{Q}$, we will use the scalar encoders $f_x(X, q), f_y(Y, q)$ which define U and V (using rates $R_x(q), R_y(q)$ respectfully) and the scalar decoders $g_x(U, V, q)$ and $g_y(U, V, q)$ to reconstruct $\hat{X}, \hat{Y}$ with average distortions $D_x(q)$

27

and $D_y(q)$ respectfully. Now for n large enough, we can find $n_1, n_2, \ldots, n_{|\mathcal{Q}|}$ such that $\sum_{q=1}^{|\mathcal{Q}|} n_q = n$ and n_q/n is arbitrarily close to $P(Q = q)$. For each $q \in \mathcal{Q}$, we use the encoder and decoders pertaining to $Q = q$ for n_q coding stages. With this timesharing we can approach $(R_x + \epsilon, R_y + \epsilon, D_x + \epsilon, D_y + \epsilon)$ for any $\epsilon > 0$ if n is large enough.

5 Conclusions

In this work we derived the fundamental limits of zero-delay lossy source coding with SI. It was shown that for the single-user setting, results in the spirit of [8] continue to hold in the sense that time-sharing scalar quantizers, followed by SI aware encoders is optimal. These codes have the feature that increasing the encoder's alphabet size can reduce the average length and not the more intuitive opposite, which is true when SI is not available and Huffman codes are used. Furthermore, we showed that in the multiterminal setting, the zero-delay variant of the multiterminal source coding problem can be readily solved, unlike the arbitrary-delay problem, which is open for three decades. We also discussed possible extensions which can be proved following the same methods we used in this paper. Although we discussed only finite alphabets, we believe that extending the results of this paper to continuous alphabets in the multiterminal setting or continuous alphabets with discrete SI (for example if the decoder has another quantized version of the source as SI) in the single-user setting should be straightforward.

It was shown that without SI, there are cases where there is no loss in forcing causal/sequential decoding, compared to standard rate-distortion decoding which waits for the whole block to arrive before producing the reconstruction. In other all other cases, we showed that the redundancy, compared to optimal rate-distortion performance is less than one bit when variable-rate coding is allowed. We used a simple scheme that relied on the classical random rate-distortion codebook construction, and only changed the way the encoder and decoder operate. However, trying to do the same when SI is available results in a substantial performance degradation. It is an interesting question whether there is an unavoidable loss when restricting to sequential decoding in these cases or whether our scheme, which works without SI, is too naive when SI is available.

Appendix

A   Details of the Sequential Scheme

In this appendix, we show the analysis of the sequential scheme which was briefly described in Section 3.1. This scheme uses a classical rate-distortion random code [14]. The encoder finds the first codeword, $\hat{X}^n$, in the codebook which is distortion-typical (as defined in [14]) and starts to transmit the *reproduction symbols*, $\hat{X}_t$, sequentially, using $\log |\hat{\mathcal{X}}|$ bits each transmission (in contrast to the classical encoder which will send the index of the codeword in a block). The decoder outputs $\hat{X}_t$ as it receives it. The idea here, is that after receiving $n \cdot \alpha$ symbols, the decoder can detect the specific codeword in the codebook since, with high probability, no other codeword will have the same prefix and the rest of the reproduction symbols ($\hat{X}^n_{n\alpha+2}$) can be reproduced without further transmissions from the encoder.

We start by analyzing the probability that given a sequence, if we now randomly draw a codebook, we will draw another sequence with the same prefix. Formally, Let $\hat{X}$ be drawn with an i.i.d source $P(\hat{x})$. We first draw a single sequence of length n according to $P(\hat{\boldsymbol{x}}) = \prod_{t=1}^{n} P(\hat{x}_t)$, then we independently draw 2^{nR} sequences using the same source. We ask what is the probability that none of the 2^{nR} sequences start with the same $n\alpha$ symbols prefix of the first sequence? We average this over the probability to draw a specific sequence in the first drawing. If we found an α for which this probability vanishes as n grows, this will give us a bound on the number of symbols we need to send in our sequential scheme in order to identify the correct codeword. Let this probability be denoted by P_c. We will treat only the first $n\alpha$ symbols since we do not care about the contents of the rest. We use the method of types. $P_{\boldsymbol{x}}$ will denote the empirical distribution of $\boldsymbol{x}$, i.e., $P_{\boldsymbol{x}}(x) = N_{\boldsymbol{x}}(x)/n\alpha$ where $N_{\boldsymbol{x}}(x)$ counts the number of occurrences of the symbol x in $\boldsymbol{x}$. $H(P_{\boldsymbol{x}})$ will denote the empirical entropy of the empirical distribution $P_{\boldsymbol{x}}(\cdot)$. We use standard results of the method of types which can be found in [24].

After we have the first sequence, the probability that it will not be the

outcome of the next drawing is $1 - 2^{-n\alpha(D(P_{\boldsymbol{x}}||P)+H(P_{\boldsymbol{x}}))}$. Therefore,

$$
\begin{aligned}
P_c &\doteq \sum_{\boldsymbol{x}} P(\boldsymbol{x}) \left(1 - 2^{-n\alpha(D(P_{\boldsymbol{x}}||P)+H(P_{\boldsymbol{x}}))}\right)^{2^{nR}} \\
&\doteq \sum_{P_{\boldsymbol{x}}} 2^{-nD(P_{\boldsymbol{x}}||P)} \left(1 - 2^{-n\alpha(D(P_{\boldsymbol{x}}||P)+H(P_{\boldsymbol{x}}))}\right)^{2^{nR}} \\
&\geq \left(1 - 2^{-n\alpha H(P)}\right)^{2^{nR}}
\end{aligned}
\tag{A.1}
$$

where the second equality is true since the the expression in the first line depends only on the type-class of $\boldsymbol{x}$ and the last inequality is true since from the sum over all type-classes, we took only $P_{\boldsymbol{x}} = P$ (we assume that n is large enough so we can get as close as we want to P, even if it is not rational). The last expression converges to unity double exponentially fast if $\alpha H(P) > R$ or equivalently, $\alpha > \frac{R}{H(P)}$.

We now use this result to analyse a coding system with sequential decoders. We draw a codebook consisting of $2^{n(R(D)+\epsilon)}$ codewords according to the prior on $\hat{X}$ that is calculated from the channel that achieves the (standard) rate distortion function. Given a source sequence, X^n, to be encoded, the encoder looks for a sequence in the codebook, $\hat{X}^n$, that is distortion typical (see [14]) with the source sequence and starts sending the reproduction symbols $\hat{X}_t$, $t = 1, 2, \ldots, n\alpha$, sequentially, $\log|\hat{\mathcal{X}}|$ bits at a time. The decoder outputs the reproduction symbols sequentially as it receives their description. With the above result, if $\alpha = \frac{R(D)}{H(\hat{X})} + 2\epsilon$, there is no other codeword in the codebook with the same prefix with probability that converge to unity double exponentially fast. Therefore, after receiving $n\left(\frac{R(D)}{H(P)} + 2\epsilon\right)\log|\hat{\mathcal{X}}|$ bits, with very high probability, the decoder knows the whole sequence $\hat{X}^n$ and can output the rest. For the whole n-block, we send no more than $\frac{1}{n}\left[n\left(\frac{R(D)}{H(\hat{X})} + 2\epsilon\right)\right]\log|\hat{\mathcal{X}}|$ bits per source symbol. This means that whenever the rate-distortion achieving prior is uniform, we achieve the optimal rate-distortion performance with this simple sequential scheme (since $H(\hat{X}) = \log|\hat{\mathcal{X}}|$). Otherwise, we are away from optimality by a factor of $\frac{\log|\hat{\mathcal{X}}|}{H(\hat{X})}$. This factor can be reduced if we allow variable rate coding, i.e., relax the constant number of bits per transmission constraint to a constraint on the average number of bits per transmission. If variable rate coding of $\hat{X}_t$ is allowed, we can upper bound the average number of bits we send each transmission by $H(\hat{X}) + 1$. In this case the the number of bits per source symbol we send is upper bounded by $R(D) + \frac{R(D)}{H(\hat{X})} \leq R(D) + 1$, meaning we are less than one bit away from optimal performance with this simple sequential scheme for any source and distortion measure. Note that, essentially, we used the same random code used in the classical analysis of the achievability of

the rate-distortion results. The error event is the union of the events that the random code will not cover a specific source sequence with distortion D and the event that the codeword that describes this source sequence with distortion D has another codeword with the same prefix. The complete analysis that shows that there exists a codebook for which our scheme will work is the same as the one found in [14, Chapter 10.5] and is therefore omitted.

B Proof of Theorem 3

In this appendix, we prove the first part of Theorem 3, namely, we show that $\mathcal{R}_{s-e}(D) = \underline{R}_{ZD}(D)$. Once again, the direct part is the same as the direct part of Theorem 2, we need to prove only the converse.

When the encoder is scalar (memoryless), it cannot use the past transmitted messages as SI. Therefore, for each message, at least $L_{Y_t}(W_t)$ are sent. We have

$$
\begin{aligned}
nR &\geq \sum_{t=1}^{n} \int L_{Y_t}(f_t(X_t)) \\
&\geq \sum_{t=1}^{n} \int L_{Y_t}(f_t(X_t)|W^{t-1}, W_{t+1}^n, Y^{t-1}, Y_{t+1}^n) \\
&\geq \sum_{t=1}^{n} \int L_{Y_t}(f_t(X_t)|w^{t-1}, w_{t+1}^n, y^{t-1}, y_{t+1}^n) d\mu(w^{t-1}, w_{t+1}^n, y^{t-1}, y_{t+1}^n) \\
&\geq \sum_{t=1}^{n} \int L_{Y_t}(f_t(X_t)) d\mu(w^{t-1}, w_{t+1}^n, y^{t-1}, y_{t+1}^n)
\end{aligned}
\tag{B.1}
$$

Now, the rate in the inner expression cannot be smaller than the rate of the optimal scalar system (with all conditioned elements serving as index of functions), that achieves the same distortion as the given system (with decoder functions $h_t(w^{t-1}, W_t, y^{t-1}, Y_t, y_{t+1}^n)$

$$
\begin{aligned}
nR &\geq \sum_{t=1}^{n} \int L_{Y_t}(f_t(X_t)) d\mu(w^{t-1}, w_{t+1}^n, y^{t-1}, y_{t+1}^n) \\
&\geq \sum_{t=1}^{n} \int \underline{R}_{ZD}\left(\boldsymbol{E}\left[d(X_t, g_t(f_t(X_t), w^{t-1}, w_{t+1}^n, y^{t-1}, Y_t, y_{t+1}^n))\right]\right) \times \\
&\qquad d\mu(w^{t-1}, w_{t+1}^n, y^{t-1}, y_{t+1}^n)
\end{aligned}
\tag{B.2}
$$

From here, using the concavity and monotonicity of $\underline{R}_{ZD}(\cdot)$ and following the steps of the proof of Theorem 2 the proof will be concluded.

References

[1] H. Witsenhausen, "The zero-error side information problem and chromatic numbers (corresp.)," *IEEE Transactions on Information Theory*, vol. 22, no. 5, pp. 592 – 593, sep 1976.

[2] N. Alon and A. Orlitsky, "Source coding and graph entropies," *IEEE Transactions on Information Theory*, vol. 42, no. 5, pp. 1329–1339, September 1996.

[3] P. Koulgi, E. Tuncel, S. Regunathan, and K. Rose, "On zero-error source coding with decoder side information," *IEEE Transactions on Information Theory*, vol. 49, no. 1, pp. 99 – 111, Jan 2003.

[4] E. Tuncel, P. Koulgi, S. Regunathan, and K. Rose, "Zero-error source coding with maximum distortion criterion," in *Data Compression Conference, 2002. Proceedings. DCC 2002*, 2002, pp. 92 – 101.

[5] D. Muresan and M. Effros, "Quantization as histogram segmentation: Optimal scalar quantizer design in network systems," *IEEE Transactions on Information Theory*, vol. 54, no. 1, pp. 344 –366, jan. 2008.

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

[7] A. Reani and N. Merhav, "Data processing bounds for scalar lossy source codes with side information at the decoder," *Submitted to the IEEE Transaction on Information Theory 2012, CoRR*, vol. abs/1209.2066, 2012.

[8] D. Neuhoff and R. K. Gilbert, "Causal source codes," *IEEE Transactions on Information Theory*, vol. 28, no. 5, pp. 701–713, September 1982.

[9] T. Weissman and N. Merhav, "On causal source codes with side information," *IEEE Transactions on Information Theory*, vol. 51, no. 11, pp. 4003–4013, November 2005.

[10] T. Weissman and A. El Gamal, "Source coding with limited side information lookahead at the decoder," *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5218–5239, December 2006.

[11] N. T. Gaarder and D. Slepian, "On optimal finite-state digital transmission systems," *IEEE Transactions on Information Theory*, vol. 28, no. 3, pp. 167–186, March 1982.

[12] Y. Kaspi and N. Merhav, "Structure theorems for real–time variable–rate coding with and without side information," *IEEE Transactions on Information Theory*, vol. 58, no. 12, pp. 7135 – 7153, December 2011.

[13] S. Yuksel and T. Basar, "Causal coding of Markov sources with continuous alphabets," in *Analysis and Design of Nonlinear Control Systems*, A. Astolfi and L. Marconi, Eds. Springer Berlin Heidelberg, 2008, pp. 91–108.

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, 2006.

[15] N. Merhav, "On the statistical physics of directed polymers in a random medium and their relation to tree codes," *Information Theory, IEEE Transactions on*, vol. 56, no. 3, pp. 1345 –1350, March 2010.

[16] N. T. Gaarder and D. Slepian, "On optimal finite-state digital transmission systems," in *International Symposium on Information Theory, Grignano, Italy*, June 1979.

[17] D. Slepian and J. Wolf, "Noiseless coding for correlated information sources," *IEEE Transactions on Information Theory*, vol. 19, pp. 471–480, 1973.

[18] R. Ahlswede and J. Korner, "Source coding with side information and a converse for degraded broadcast channels," *Information Theory, IEEE Transactions on*, vol. 21, no. 6, pp. 629 – 637, nov 1975.

[19] A. Wyner, "On source coding with side information at the decoder," *Information Theory, IEEE Transactions on*, vol. 21, no. 3, pp. 294 – 300, may 1975.

[20] T. Berger and R. Yeung, "Multiterminal source encoding with one distortion criterion," *Information Theory, IEEE Transactions on*, vol. 35, no. 2, pp. 228 –236, mar 1989.

[21] A. Wagner, S. Tavildar, and P. Viswanath, "Rate region of the quadratic gaussian two-encoder source-coding problem," *Information Theory, IEEE Transactions on*, vol. 54, no. 5, pp. 1938 –1961, may 2008.

[22] T. A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *CoRR*, vol. abs/1110.3069, 2011.

[23] A. Orlitsky, "Average-case interactive communication," *Information Theory, IEEE Transactions on*, vol. 38, no. 5, pp. 1534 –1547, sep 1992.

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