ISIT'2013 1569718397

The Zero-Delay Wyner-Ziv Problem

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

***Abstract*—We consider the zero-delay version of the Wyner-Ziv problem. The zero-delay constraint translates into causal (sequential) encoder and decoder pairs as well as the use of instantaneous codes. 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 be 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.**

I. Introduction

We consider the following source coding problem. Symbols produced by a discrete memoryless source are to be encoded, transmitted noiselessly and reproduced by a decoder which has access to side information (SI), correlated to the source. Operation is in real-time, that is, the encoding of each symbol and its reproduction by the decoder must be performed without any delay. The average distortion between the source and the reproduced symbols is constrained to be smaller than some predefined constant. Such a model is motivated by many practical source-coding systems where delay is not tolerable, such as live multimedia streaming.

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 encoder 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.

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 showed 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]. 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 mposed 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 subsets in each partition were constrained to 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 subsets which are not convex.

Zero-delay codes form a subclass of the class of causal codes, as defined by Neuhoff and Gilbert [7]. In [7], entropy coding is used on the whole sequence of reproduction symbols, introducing arbitrarily long delays. In the zero-delay case, entropy coding has to be instantaneous, symbol-by-symbol (possibly taking into account past transmitted symbols). It was shown in [7] that for a discrete memoryless source (DMS), the optimal causal encoder consists of time-sharing between no more than two scalar encoders. Weissman and Merhav [8] extended [7] to the case where SI is also available at the decoder, or encoder or both. The discussion in [8] 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 [9].

The results of [7] 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 [8] can be adapted to zero-

†This research was supported by the Israeli Science Foundation (ISF) grant no. 208/08.

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 [8].

This paper has several contributions. The first is the extension of [7] to zero-delay with decoder SI and average distortion constraint, where unlike [8], we do not restrict the usage of the SI. We show that results in the spirit of [7] 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 [7] 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 [9], where SI lookahead was shown to improve the performance. Not surprisingly, our results place the optimal performance of zero-delay systems far below the classical source coding results. We ask which of the zero-delay constraints (causal encoder/decoder, instantaneous codes) are causing this performance degradation. 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, 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 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.

The rest of this paper is organized as follows. In Section II we give the formal setting and notation used throughout the paper. In Section III, we state and discuss the main contributions of this paper. In Section IV we explain by example which of the constraints that stem from the zero-delay model is causing the performance degradation, compared to classical arbitrary delay results. Finally, the proof of Theorem 2 is given in Section V.

II. 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 integer i, x^i will denote the vector $(x_1, \ldots, x_i)$. The source alphabet, $\mathcal{X}$, as well as all other alphabets in the sequel, is finite. The probability distribution over $\mathcal{X}$, will be denoted by $P_X(\cdot)$. When there is no room for ambiguity, we use $P(x)$ instead of $P_X(x)$.

We investigate the following zero-delay problem, depicted in Fig. 1. An encoder observes X_t and transmits a compressed version, W_t, to a decoder which 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 with zero-delay. 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 $R \stackrel{\triangle}{=} \limsup_{n\to\infty} \frac{1}{n} \boldsymbol{E} L_n$. Our goal is to find the tradeoffs between R and D.

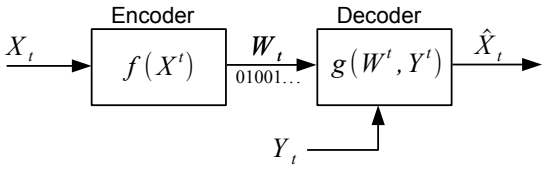

Fig. 1: System model

Since no delay is allowed, W_t must be encoded by an instantaneous code. Note that in general $P(w_t, y_t) \neq P(w_t)P(y_t)$ and therefore, we will need to consider instantaneous coding of W_t in the presence of correlated SI at the decoder. We restrict the coding of W_t to be error-free. The remainder of this section will be devoted to definitions that are needed for zero error transmission in the presence of SI.

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 will 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

$$L_Y(X) = \min\left\{\overline{l_y}(\phi) :\ \phi \text{ is an RI protocol}\right\}. \qquad (1)$$

Upper and lower bounds on $L_Y(X)$ in terms of the entropy of the optimal coloring are given in [2]. Finding a single-letter expression for $L_Y(X)$ is an open problem. We will use $L_Y(X)$ as a figure of merit and our results will be single-letter expressions, in terms $L_Y(X)$. In Fig. 2, 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 $P_X(\cdot)$.

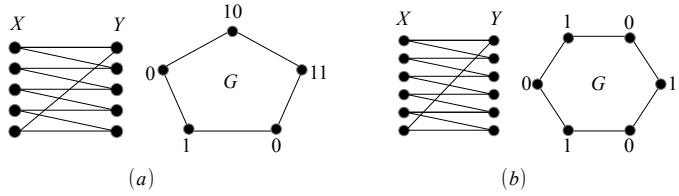

Fig. 2: Example of bipartite graphs of $P(x,y)$ along with their associated characteristic graphs G and a RI protocol for 5 (a) and 6 (b) letter alphabets with "typewriter" SI.

In Fig 2a, 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. 2b, however, $\chi(G) = 2$ and indeed the optimal RI protocol uses a 2-coloring scheme.

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.

In the proof of 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 (x,y) such that $p(x,y,z) > 0$. For any $\phi \in \Phi(z)$, let $\bar{l}_Y(\phi|z) \stackrel{\triangle}{=} \sum_{x\in\mathcal{X}} P(x|z)|\phi(x)|$ be the average length when $Z = z$. Similarly, let

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

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 $L_Y(X|Z) \le L_Y(X)$ 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 $y \in \mathcal{Y}$.

III. Main Results

In this section, we state and discuss the main results of this work. The 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

$$R_{ZD}(D) = \min_{h,f} L_Y(f(X)) \qquad (3)$$

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 $\boldsymbol{E}d(X,h(Y,Z)) \le D$ (obviously, $|\mathcal{Z}| \le |\mathcal{X}|$). Finally, denote the lower convex envelope of $R_{ZD}(D)$ by $\underline{R}_{ZD}(D)$. In (3), each possible f in the search domain, along with its optimal h, will incur some average distortion. Since there is only a finite number of such functions f, the $R-D$ plain contains a finite number points. The lower convex envelope of these points will give us $\underline{R}_{ZD}(\cdot)$, which is therefore piecewise-linear.

The first result of this paper is the following theorem:

Theorem 1. $\mathcal{R}_{ZD}(D) = \underline{R}_{ZD}(D)$.

The proof of the direct of this and the other theorems in this work is easily shown by time-sharing scalar encoder/decoder pairs. The converse of this theorem is given in Section V. 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. As in the example in Fig. 2, there are cases where increasing the encoder output alphabet beyond the minimal number of subsets that achieves distortion D improves performance. Specific examples can be found in [10].

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

The theorem states that allowing the decoder to observe the future SI symbols will not result in a 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 [9], where it was shown that SI look-ahead can improve performance. The proof of the converse of this theorem, which also covers the converse of Theorem 1, is given in Section V.

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-versa. 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., $\hat{X}_t = g_t(W_t, Y^n)$ (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. $\mathcal{R}_{s-d}(D) = \mathcal{R}_{s-e}(D) = \underline{R}_{ZD}(D)$.

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 [11] to include variable rate, look-ahead and SI. The proof of Theorem 3, is omitted due to the space limitations. The full proofs as well as examples of actual SI-aware quantizers can be found in [10].

IV. What causes the performance degradation in the zero-delay regime?

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. For example, without SI, for a uniform binary source and Hamming distortion measure, we compare the the optimal curve of $R = 1 - h_2(D)$, to the straight line $R = 1 - 2D$ which is the optimal performance of a zero-delay system. While it is not surprising that the zero-delay constraint causes performance degradation, in this section we ask 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 this degradation. Had we alleviated one of these constraints, can classical 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. For simplicity, we first assume that the reconstruction alphabet is binary and then describe the extension to general finite alphabets. We look at the following system: An encoder observes the whole source sequence X^n, but can send no more than one bit per transmission. The decoder is 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 zero-delay settings of Theorems 1,2, we only alleviated the constraint on the encoder. We show that, at least in some cases, such a system can achieve the classical, arbitrary delay, rate distortion performance. This in turn shows that the "blame" for the performance degradation falls solely on the causal encoders in these cases. 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 the classical block-coding schemes. Such a system can represent a streaming scenario, where the encoder knows the whole stream in advance and we want zero decoding delay, meaning that every received symbol can be immediately utilized. Our scheme works as follows: A classical rate-distortion random code [12] is used. The encoder finds the first codeword $(\hat{X}^n)$ in the codebook which is distortion-typical (as defined in [12]) and starts to transmit the *reproduction symbols*, $\hat{X}_t$, sequentially, using one bit in each transmission (in contrast to the classical encoder which will send the index of the codeword in a block). The decoder sequentially outputs $\hat{X}_t$ as it receives it. If we send the whole n bits of the reconstruction sequence, the rate of this scheme will, of course, be one bit per sample. However, the idea here, is that after receiving enough $\hat{X}_t$'s, 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. In fact, for any $\epsilon > 0$, it is enough to send only $n \cdot \alpha$ bits with $\alpha = R(D)/H(\hat{X}) + \epsilon$. This means that whenever $H(\hat{X}) = 1$ (for example when X is binary and uniform and the Hamming distortion measure is used), we can achieve the classical rate-distortion performance with this scheme. The following lemma, which is proved in [10], makes the described scheme work:

Lemma 1. *Given a codeword $\hat{X}^n$, the probability to draw another codeword in a codebook of 2^{nR} codewords with the same $n \cdot \alpha$ first symbols as $\hat{X}^n$ vanishes double exponentially fast if $\alpha > \frac{R}{H(\hat{X})}$.*

This lemma guarantees that when using a random code, the added error event (compared to the classical scheme) in which after sending $n\left(R(D)/H(\hat{X}) + \epsilon\right)$ bits, the decoder cannot detect the correct codeword, has negligible effect on

the average distortion. We therefore know that the average distortion is less than $D+\epsilon$ since essentially, we used the classical rate-distortion codebook and only transmitted the codeword differently. While for binary reconstruction symbols we achieve the optimal performance only when $H(\hat{X})=1$, for larger alphabets, if variable rate coding of $\hat{X}_t$ is allowed, it can be shown that this scheme either achieves the classical rate-distortion performance or is bounded away from it by less than a bit. The extra bit is due to the use of instantaneous codes and not due to the sequential decoding. We believe that the extension of this scheme to the Wyner-Ziv setting is possible as well. However, the analysis is much more involved and is yet to be done.

V. Proof of the converse of Theorem 2

At every stage, the encoder sends a message W_t which 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 $nR \geq \sum_{t=1}^n L_{Y^t}(W_t|W^{t-1})$. To avoid the complex SI structure, which depends on the time instant 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:

$$\begin{aligned}
nR &\geq \sum_{t=1}^n L_{Y_t}(W_t|X^{t-1},Y^{t-1}) \\
&\geq \sum_{t=1}^n L_{Y_t}(W_t|X^{t-1},Y^{t-1},Y_{t+1}^n) \qquad (4)\\
&= \sum_{t=1}^n \int L_{Y_t}(W_t|x^{t-1},y^{t-1},y_{t+1}^n)d\mu(x^{t-1},y^{t-1},y_{t+1}^n) \\
&= \sum_{t=1}^n \int L_{Y_t}(f_t(X_t,x^{t-1},y^{t-1}))d\mu(x^{t-1},y^{t-1},y_{t+1}^n) \qquad (5)
\end{aligned}$$

where in (4) 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 $R_{ZD}(D)$, (3). 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 definition of $R_{ZD}(D)$. We therefore have

$$\begin{aligned}
nR &\geq \sum_{t=1}^n \int L_{Y_t}(f_t(X_t,x^{t-1},y^{t-1}))d\mu(x^{t-1},y^{t-1},y_{t+1}^n) \\
&\geq \sum_{t=1}^n \int R_{ZD}(\boldsymbol{E}[d(X_t,g_t(f_t(X_t,x^{t-1},y^{t-1}),x^{t-1},y^{t-1}, \\
&\qquad Y_t),y_{t+1}^n)|x^{t-1},y^{t-1},y_{t+1}^n])d\mu(x^{t-1},y^{t-1},y_{t+1}^n) \qquad (6)\\
&\geq \sum_{t=1}^n \int \underline{R}_{ZD}(\boldsymbol{E}[d(X_t,g_t(f_t(X_t,x^{t-1},y^{t-1}),x^{t-1},y^{t-1}, \\
&\qquad Y_t,y_{t+1}^n))|x^{t-1},y^{t-1},y_{t+1}^n])d\mu(x^{t-1},y^{t-1},y_{t+1}^n) \qquad (7)\\
&= \sum_{t=1}^n \underline{R}_{ZD}\left(\boldsymbol{E}\left[d(X_t,g_t(f_t(X^t,Y^{t-1}),X^{t-1},Y^n))\right]\right) \qquad (8)\\
&= \sum_{t=1}^n \underline{R}_{ZD}\left(\boldsymbol{E}\left[d(X_t,\hat{X}_t)\right]\right) \\
&\geq n\underline{R}_{ZD}\left(\frac{1}{n}\sum_{t=1}^n \boldsymbol{E}\left[d(X_t,\hat{X}_t)\right]\right) \qquad (9)\\
&\geq n\underline{R}_{ZD}(D), \qquad (10)
\end{aligned}$$

where (6) follows from the definition of $R_{ZD}(D)$ and the discussion following (5), (7) follows from the definition of $\underline{R}_{ZD}(D)$, (8) and (9) follow from the convexity of $\underline{R}_{ZD}(D)$. Finally, (10) 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.

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

[8] 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.

[9] 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.

[10] Y. Kaspi and N. Merhav, "Zero-delay and causal single-user and multi-user lossy source coding with decoder side information," *Submitted to IEEE Transactions on Information Theory 2013, CoRR*, vol. abs/1301.0079, 2013.

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

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