A Universal Scheme for Wyner-Ziv Coding of Discrete Sources

Shirin Jalali, Student Member, IEEE, Sergio Verdú, Fellow, IEEE, and Tsachy Weissman, Senior Member, IEEE

Abstract—We consider the Wyner-Ziv (WZ) problem of lossy compression where the decoder observes a noisy version of the source, whose statistics are unknown. A new family of WZ coding algorithms is proposed and their universal optimality is proven. Compression consists of sliding-window processing followed by Lempel-Ziv (LZ) compression, while the decompressor is based on a modification of the discrete universal denoiser (DUDE) algorithm to take advantage of side information. The new algorithms not only universally attain the fundamental limits, but also suggest a paradigm for practical WZ coding. The effectiveness of our approach is illustrated with experiments on binary images, and English text using a low complexity algorithm motivated by our class of universally optimal WZ codes.

Index Terms—Discrete denoising, rate-distortion function, sliding-window coding, universal algorithm, Wyner-Ziv coding.

I. INTRODUCTION

Consider the basic setup shown in Fig. 1 consisting of a source with unknown statistics driving a known discrete memoryless channel (DMC), and a decoder that receives a compressed version of the source in addition to the noisy channel output. The goal is to minimize the distortion between the source and the reconstructed signal by optimally designing the encoder and decoder. This is the problem of rate-distortion coding with decoder side information, commonly known as Wyner-Ziv compression after the seminal paper [1]. Even without side information, the problem of finding universal practical schemes that get arbitrarily close to a given point on the rate-distortion curve is notoriously challenging (see [2],[3] and [4] for recently proposed practical schemes). Even when the discrete source distribution is known, no practical scheme is currently known to approach the rate-distortion function when the source has memory. Other than the region of low distortion, the rate-distortion function is not known even for a binary Markov source (see [5],[6],[7]).

As an example of practical motivation for the setup shown in Fig. 1, consider the problem of audio/video broadcasting where in addition to the analog signal, the decoder has access to some additional information transmitted in a digital channel for instance. In such setup, a legacy receiver only observes the output of the channel, while a more sophisticated receiver in addition has access to coded information which helps boosting reproduction fidelity. Thus, we can view the setup as one of universal systematic channel coding where the added “redundancy” is received error-free.

An alternative view of this problem is as a denoising problem where the denoiser, in addition to the noise-corrupted data, has access to a fidelity-boosting (FB) sequence conveyed to it via a channel of capacity $R$. Both viewpoints are equivalent because the source/channel separation theorem [8] guarantees that there is no loss in separating the source coding and channel coding operations at least under certain sufficient conditions on source and channel [9]. Therefore, the encoder is able to send any information with entropy less than the channel capacity almost losslessly to the decoder. Consequently, although in practice we would often have a channel of capacity $R$, we simply consider the encoder-channel-decoder chain as a noiseless bit pipe of rate $R$.

Note that in these two viewpoints the role of the main signal and fidelity-boosting signal is interchanged. In this paper, we adopt the latter, and suggest a new algorithm for WZ coding of a source with unknown statistics. We show that, for stationary ergodic sources, the algorithm is asymptotically optimal in the sense that its average expected loss per symbol converges to the minimum attainable expected loss.

The encoder of the proposed algorithm consists of a sliding-block (SB) coder followed by Lempel-Ziv (LZ) compression [10]. SB lossy compression is shown in [11] to be able to perform as well as conventional lossy block compression. We extend this result to the WZ coding setup, and show that the same result holds in this case as well. The reason we use SB codes instead of block codes in our algorithm is the special type of decoder we employ. The decoder is based on a modification of the discrete universal denoiser (DUDE) algorithm [12]. DUDE with FB (fbDUDE), to take advantage of the FB sequence. We prove that the optimality results of the original DUDE carry over to fbDUDE as well.

As mentioned before, in our setting we always assume that the channel transition matrix is known both to the encoder and the decoder. As argued in [12], the assumption of a known channel and an unknown signal is realistic in many practical scenarios. Furthermore, unlike the DUDE setting [12], in the setup of this paper the decoder can easily learn the channel, e.g., by having the encoder dedicate a negligibly small portion of its rate to describing the first few components of the source sequence which then act as a training sequence. Further still, if a modicum of feedback from decoder to encoder is allowed, then the encoder too can be informed of the channel arbitrarily precisely. Therefore, unlike in the DUDE setting where knowledge of the channel plays a key role [13],[14],
in our setting, at least decoder knowledge of the channel is not crucial, and our schemes can be easily modified to accommodate channel uncertainty.

Some progress towards practical WZ coding schemes has been made in recent years, as seen, e.g., in [15], [16], [17], [18], [19], [20]. The proposed schemes, however, operate under specific assumptions of a known (usually memoryless) source and side information channel. Practical schemes for more general source and/or channel characteristics have yet to be developed and, a fortiori, no practical universal schemes for this problem are known.

The problem of WZ coding of a source with unknown statistics was recently considered in [21], where existence of universal schemes in a setting similar to ours is established. In contrast, our schemes suggest a paradigm for WZ coding of discrete sources which is not only practical but is justified through universal optimality results.

The organization of the remainder of this paper is as follows. In Section II, the notation used throughout the paper is introduced. Section III presents the DUDE, the extension of the DUDE denoising algorithm [12] to take advantage of a FB sequence, and shows how the asymptotic optimality of the original DUDE carries over to this case as well. Section IV proposes SB WZ codes and proves a result on their relationship to WZ block codes. In Section V, our new WZ coding algorithm is presented and its optimality is established. In Section VI we present some experimental results, concluding in Section VII with a brief discussion of possible extensions of this work. Outlines of the proofs are given in the main body, with the full proofs relegated to the Appendix.

II. NOTATION

Let $\mathcal{X}$, $\hat{\mathcal{X}}$, and $\mathcal{Z}$ denote the source, reconstructed signal, and channel output alphabets respectively. In this paper, for simplicity, we restrict attention to

$$\mathcal{X} = \hat{\mathcal{X}} = \mathcal{Z} = \{\alpha_1, \ldots, \alpha_N\},$$

though our derivations and results carry over directly to nonidentical finite sets $\mathcal{X}$, $\hat{\mathcal{X}}$, and $\mathcal{Z}$. Bold low case symbols, e.g., $x$, $y$, $z$, denote individual sequences. The discrete memoryless channel is described by its transition matrix $\Pi$, where $\Pi(i, j)$ denotes the probability of getting $\alpha_j$ at the output of the channel when the input is $\alpha_i$. Recall that we assume that the matrix $\Pi$ is known both by the encoder and the decoder.

Let $\lambda: \mathcal{X} \times \hat{\mathcal{X}} \rightarrow \mathbb{R}^+$ be the loss function (fidelity criterion) which measures the loss incurred in denoising (decoding) a symbol $\alpha_i$ to another symbol $\alpha_j$, which will be represented by a $N \times N$ matrix, $\Lambda = \{\lambda(\alpha_i, \alpha_j)\}$. Moreover, let

$$\lambda_{\max} = \max_{i,j} \lambda(\alpha_i, \alpha_j),$$

and note that $\lambda_{\max} < \infty$, since the alphabets are finite. The normalized cumulative loss between a source sequence $x^n$ and reconstructed sequence $\hat{x}^n$, is denoted by

$$\rho_n(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} \lambda(x_i, \hat{x}_i).$$

Let $\pi_i$ and $\lambda_j$ denote the $i$-th column and the $j$-th column of $\Pi$ and $\Lambda$ matrices respectively, i.e.,

$$\Pi = [\pi_1 \ldots \pi_N], \quad \Lambda = [\lambda_1 \ldots \lambda_N].$$

For $N$-dimensional vectors $u$ and $v$, $u \odot v$ denotes the $N$-dimensional vector that results from componentwise multiplication of $u$ and $v$, i.e.,

$$u \odot v[i] = u_i v_i.$$  \hfill (2)

As in (2), we denote the $i$-th component of a vector by either a subindex or, when that could lead to some confusion, in square brackets.

III. DUDE WITH FIDELITY BOOSTING INFORMATION

The DUDE algorithm was proposed in [12] for universal noncausal denoising of a discrete signal corrupted by a known DMC. The DUDE is described by (3) and (4).

Remark: Although in [13] it is implicitly assumed that the transition matrix $\Pi$ is a square matrix, this assumption is not necessary. As noted in [12], as long as the rows of $\Pi$ are linearly independent, the results can be generalized to non-square matrices by replacing $\Pi^{-1}$ with $\Pi^T (\Pi \Pi^T)^{-1}$ in (3).

Linear independence of the rows of $\Pi$ requires the channel inputs to be identifiable, i.e., it is not possible to fake the output distribution corresponding to any of them using some input distribution over the other input symbols. This property, which holds for most channels of interest, will be assumed throughout the paper.

In short, DUDE works as follows. In its first pass through the noisy data, it estimates the conditional marginal distributions of the clean data given their noisy observation of $P_{X_n|Z^n}(\cdot|\cdot)$ by first estimating the bidirectional conditional probabilities $P_{Z_n|X^n}(\cdot|\cdot)$ by counting, and then using the invertibility of the DMC. Then in the second pass, it finds $\hat{x}_i$ based on these estimations. The DUDE denoising algorithm is non-causal and therefore each output depends on the whole noisy sequence. The following optimality results have been shown for DUDE [12].

1) Stochastic setting: the source is assumed to be stationary, and no further constraints are imposed on its distribution. Asymptotically DUDE performs as well as the best denoiser that knows the source distribution provided that the context length $\ell$ grows adequately with the data size.

2) Semi-stochastic setting: the source is assumed to be an individual sequence, and the only randomness is assumed to originate from the channel. Asymptotically,
DUDE performs as well as the optimal denoiser in the class of sliding-window denoisers for that particular sequence.

Better experimental performance than the DUDE has been achieved by alternative methods to estimate the bidirectional conditional probabilities [22], [23] and [24]. As we discussed in Section I, decoding for the WZ problem can also be considered as a denoising problem where the denoiser, in addition to the noisy signal, has access to a FB sequence designed by the source encoder to be as helpful as possible to the decoder. From this perspective, we are motivated to generalize DUDE so as to handle not only the output of the DMC, but the fidelity boosting information. A desirable feature of such a generalization is the optimality in senses analogous to those described above. A natural way to accomplish this, which we refer to as ΠDUDE is described in (5) and (6).

Note that the counting process is done simultaneously in both the noisy and FB sequences. Although seemingly more involved, the denoising algorithm described in (5) and (6), is simply the DUDE algorithm working on an enlarged context. For the ΠDUDE, the context of each symbol in the noisy signal in addition to the conventional DUDE context of noisy neighboring symbols, consists of the same context window of the FB sequence. Note that in contrast to the conventional context of noisy neighboring symbols, in the context window of the fidelity boosting sequence there is no “hole in the middle”. It should be noted that our proposed generalization of the DUDE will not be effective with reasonable computational complexity unless the fidelity boosting sequence depends on the original clean signal in a sequential manner, such as a sliding-window. An example of a non-sequential dependence is a fidelity boosting sequence generated by an arbitrary linear block code.

In order to show that the optimality results of [12] carry over to this case, consider a channel Π, with input \( \hat{x}_i = (x_i, y_i) \), and output \( \tilde{z}_i = (z_i, y_i) \), where \( z_i \) is the output of the original channel Π, when the input is \( x_i \). Note that this channel does not disturb the second component of its input vector \( (x_i, y_i) \). As shown in the next result, since the newly defined channel Π inherits its invertibility from the original channel Π, the results of [12] concerning asymptotic optimality of DUDE can be applied to this case as well.

**Theorem 3.1:** Provided that \( t_n |X|^{2t_n} = o(n/ \log n) \), \( \forall x, y \)

\[
\lim_{n \to \infty} \left[ \frac{1}{n - 2t_n} \sum_{i = t_n + 1}^{n - t_n} \lambda(x_i, \hat{x}_i) \right] = \lim_{n \to \infty} D_{t_n, m_n}(x^n, y^n, Z^n), \quad \text{a.s.,}
\]

where,

\[
D_{t_n, m_n}(x^n, y^n, z^n) = \min_{f : \mathcal{X}^{2t_n+1} \times \mathcal{Z}^{2m_n+1} \to \mathcal{X}} \lambda(x_i, f(y_i, z_i + m_n))
\]

and \( \hat{x}_i \) is the output of the denoiser in (5) and (6) with parameters \( l_n, m_n \), and \( t_n = \max\{l_n, m_n\} \).

**Remark 1:** Here the class of decompressors is restricted to sliding-window decoders of finite-window length on both noisy data and FB sequence. Theorem 3.1 states that in the semi-stochastic setting where both the source and the FB sequence are individual sequences, with probability one, the asymptotic accumulated loss of the ΠDUDE decoder is no more than the loss incurred by the best decoder of the same order in this class.

**Remark 2:** Although Theorem 3.1 is stated for the semi-stochastic setting, as in [12], there is a counterpart in the stochastic setting where the source and FB sequences are jointly stationary processes.

**IV. SLIDING-WINDOW WYNER-ZIV CODING**

The majority of achievability proofs in the information theory literature are based on the idea of random block coding. Shannon pioneered this technique for proving his coding theorems for both lossy compression and channel coding. In rate-distortion theory, besides the conventional block codes, sliding block codes were introduced in 1975 by Gray, Neuhoff, and Ornstein [11] and independently by Marton [25], and shown to achieve the (block-coding) rate-distortion function in [11]. ΠDUDE encoders apply a function with a finite number of arguments to the source sequence, outputting another sequence that has lower entropy, but resembles the original sequence as much as the designer desires.
In the rest of the section, we show that sliding block WZ coding achieves the Wyner-Ziv rate-distortion function for stationary sources.

A. Block coding

A WZ block code of length $n$ and rate $R$ consists of encoding and decoding mappings, $f_n$ and $g_n$ respectively, which are defined as follows:

$$f_n : X^n \rightarrow \{1, 2, \ldots, [2^{nR}]\},$$
$$g_n : Z^n \times \{1, 2, \ldots, [2^{nR}]\} \rightarrow \hat{X}^n.$$  

The performance of such code is defined as the expected average distortion per symbol between the source and reconstruction sequences, i.e.

$$E[\rho_n(X^n, \hat{X}^n)] \triangleq \frac{1}{n} E \left[ \sum_{i=1}^{n} \lambda(X_i, \hat{X}_i) \right],$$

where $\hat{X}^n = g_n(Z^n, f_n(X^n))$.

The rate distortion pair $(R, D)$ is said to be achievable if for any given $\epsilon > 0$, there exists $f_n$ and $g_n$, such that

$$E[\rho_n(X^n, g_n(Z^n, f_n(X^n)))] \leq D + \epsilon,$$

for all sufficiently large $n$. For a given source $X$, and memoryless channel described by transition matrix $\Pi$, the infimum of all achievable distortions at rate $R$ is called $D_{X, \Pi}$, i.e.

$$D_{X, \Pi}(R) = \inf \{D : (R, D) \text{ is achievable} \}.$$  

More explicitly, $D_{X, \Pi}(R)$ is the distortion-rate function of our WZ coding setting.

B. Sliding-Block WZ compression

An extension of the idea of SB rate distortion coding is SB WZ coding. In this section, using the techniques of [11], we show that in WZ coding any performance that is achievable by block codes is also achievable by SB codes.

A WZ SB code consists of two time-invariant encoding and decoding mappings $f$ and $g$. The encoding mapping $f$ with constraint length of $2k + 1$ maps every $2k + 1$ source symbols into a symbol of $\mathcal{Y}$ which is the alphabet of the FB sequence; In other words

$$f : \mathcal{X}^{2k+1} \rightarrow \mathcal{Y}. \quad (8)$$

This encoder moves over the source sequence and generates the FB sequence $Y$ by letting

$$Y_i = f(X_{i-k}^{i+k}). \quad (9)$$

On the other hand, the decoding mapping $g$ with the constraint length of $\max\{2l + 1, 2m + 1\}$ maps a block of length $2l + 1$ of the noise corrupted signal and a block of length $2m + 1$ of $\mathcal{Y}$ sequence to a reconstruction symbol, i.e.

$$g : \mathcal{Z}^{2l+1} \times \mathcal{Y}^{2m+1} \rightarrow \hat{X}.$$  

The decoder slides over the noisy and the FB sequences in a synchronous manner, and generates the reconstruction sequence by letting

$$\hat{X}_i = g(Z_{i-l}^{i+l}, Y_{i-m}^{i+m}). \quad (11)$$

The following theorem states that SB-WZ codes perform at least as well as WZ block codes.

**Theorem 4.1:** Let $(R, D)$ be an interior point in the (block) WZ rate-distortion region of a jointly stationary processes $X$ and $Z$ representing the source and FB sequences respectively. For any given $\epsilon_1 > 0$, there exists a SB-WZ encoder $f : \mathcal{X}^{2k+1} \rightarrow \mathcal{Y}$, where $\log |\mathcal{Y}| \geq R$, and a SB decoder $g$ with parameters $l$ and $m$, such that

1. $E[\lambda(X_i, g(Z_{i-l}^{i+l}, Y_{i-m}^{i+m}))] \leq D + \epsilon_1$, where $Y_i = f(X_{i-k}^{i+k}),$
2. $H(Y) = \lim_{n \to \infty} \frac{1}{n} H(Y_1, \ldots, Y_n) \leq R - \epsilon_2$, for some $\epsilon_2 > 0$.

**Proof:**

The complete proof is presented in Appendix A; A sketch of the main idea follows. The proof is an extension of the proof given in [11] for showing that SB codes achieve the same performance of block codes in the rate-distortion problem, which in our scenario corresponds to the case where the decoder has only access to the FB sequence and there is no channel output.

Since $(R, D)$ is assumed to be an interior point of the achievable region in the $R-D$ plane, it is possible to find a point $(R_1, D)$ such that $R_1 < R$, but still the new point is an interior point of the achievable region. Since $(R_1, D)$ is an interior point, there exists a block WZ encoder/decoder, $(f_n, g_n)$, of rate $R_1$ and block length $n$, and average expected distortion less than $D + \epsilon$, for any $\epsilon > 0$. Instead of considering the initial point of $(R, D)$, we consider this new point with $R_1 < R$ because, according to the theorem, our goal is to show that there exists a SB encoder resulting in a FB sequence with entropy rate lower than $R$. In order to achieve this goal we follow techniques similar to those in [11]. To derive a SB code from a block code, the most challenging part is dividing the source sequence into blocks of fixed length such that it is possible to apply the block code to these sub-blocks. This is a demanding task first because we are looking for a stationary SB code, and second because the decoder is also a SB decoder which should be able to discriminate between different coded blocks concatenated by the encoder. The main tool in our proof, as in [11], is the Rohlin-Kakutani theorem of ergodic theory. This theorem enables us to define a SB encoder which finds blocks of length $n$ to apply the block code to them, and puts a tag sequence of negligible length $[n\epsilon]$ after each encoded block. This tag sequence is not included in any of the codewords of the WZ block coder $(f_n, g_n)$ (the existence of such tag sequence is shown in the proof). This would enable the decoder to discriminate between different coded blocks, while letting the encoder to generate a stationary FB sequence. The rest of the proof is devoted to showing that the SB code defined in this way would satisfy our desired constraints.

To conclude this section, note that the two-step achievability proof of Wyner and Ziv in [11] for proving their WZ theorem (rate-distortion with side information) is extended in [20] to devise a method which is used to prove a SB source coding theorems (theorems of Berger, Kaspi and Tung [27], [28], [29]) for a general finite-alphabet ergodic multiterminal source. The
focus in [26] is on multiterminal sources for which we can no longer use the method used in [11] to derive SB codes because the \(tag\) sequences in the coded version received by different terminal are not synchronized. In our case, since we have only one terminal to code, and the side information is just the output of the DMC due to the source, it is still possible to use the method used in [11].

V. WYNER-ZIV DUDE

In Section III we introduced the \(\text{FB-DUDE}\), a natural extension of the DUDE algorithm to the case where in addition to the noisy signal the denoiser has access to encoded side information. As described in Section III, this extension could easily be obtained by considering a larger context for denoising each symbol which comes from working on both signals simultaneously. Then Theorem 3.1 expressed the asymptotic optimality of the \(\text{FB-DUDE}\) denoiser. Section IV introduced \(\text{SB-WZ}\) coding, and in Theorem 4.1 it was shown that using \(\text{SB-WZ}\) codes instead of \(\text{WZ}\) block codes incurs no loss of optimality. Motivated by the results established so far, in this section, we propose a new \(\text{WZ}\) coding scheme, and prove its asymptotic optimality.

For any given block length \(n\), let \(f_n^*\) and \(g_n^*\) denote the encoder and the decoder of the scheme respectively. The scheme has a number of parameters, namely \(l_n, k_n, m_n\) and \(\delta\), that their meaning is made explicit in the following description of the algorithm.

1) Encoder: For a given source sequence \(x^n\) define \(S(x^n, k_n, R)\) to be the set of all \(\text{SB}\) mappings of window length \(2k_n + 1\) with the property that their output is a sequence whose Lempel-Ziv compressed version \(\text{LZ}(\cdot)\) is not longer than \(nR\), i.e.
\[
S(x^n, k_n, R) \triangleq \left\{ f : X^{2k_n+1} \rightarrow Y : \frac{1}{n} \text{LZ}(f(x^n)) \leq R \right\}. \tag{12}
\]

Note that \(f(x^n)\) is assumed to be equal to \(y^n\), where \(y_i = f(x_{i-k_n}^{i+k_n})\) for \(k_n + 1 \leq i \leq n - k_n\), and \(y_i = 0\) otherwise. For each \(f \in S\), and integers \(l_n\) and \(m_n\) define
\[
V(f, l_n, m_n) = \min_{g \circ f} \mathbb{E} \left[ \sum_{i=k_n}^{n-k_n} \lambda \left( x_i, g(Z_i^{l_n}, y_i^{m_n}) \right) \right], \tag{13}
\]
where the minimization is over all decoding mappings \(g : Z^{2l_n+1} \times Y^{2m_n+1} \rightarrow X\). Let \(f^*(l_n, m_n)\) be the mapping in \(S\) that minimizes \(V(f, l_n, m_n)\), i.e.
\[
f^*(l_n, m_n) = \arg \min_{f \in S} V(f, l_n, m_n), \tag{14}
\]
and also let \(g^*\) be the decoder mapping corresponding to \(f^*\) that achieves \(V(f^*, l_n, m_n)\). Then, the FB encoded sequence is the LZ compression of \(f_n^*(x^n)\) which is sent to the decoder.

2) Decoder: Upon obtaining \(f_n^*(x^n)\) with an LZ decompressor, the decoder employs the \(\text{FB-DUDE}\) described in Section III, i.e. the reconstructed signal is \(X^n, k_n, m_n \circ Z^n, f^*(x^n)\).

The main result of this paper is the following theorem, which shows that the described \(\text{WZ-DUDE}\) coding algorithm is asymptotically optimal.

**Theorem 5.1:** Let \(k_n, l_n, \) and \(m_n\) increase without bound with \(n\), but sufficiently slowly that \(t_n |X|^l_n = o(n/\log n)\), where \(t_n = \max \{l_n, m_n\}\). Then, for any \(R \geq 0\), and any stationary ergodic source \(X\),
\[
\lim_{n \to \infty} \mathbb{E} \left[ \rho_n \left( X^n, g_n^*(Z^n, f_n^*(X^n)) \right) \right] = D_{\mathbf{x}, \mathbf{\Pi}}(R). \tag{15}
\]

**Proof:** The full proof can be found in Appendix B. A brief outline of the proof is as follows. The first step is using Theorem 4.1 to find a \(\text{SB-WZ}\) code with mappings \(f\) and \(g\) which results in a final expected distortion less that \(D + \frac{\epsilon_2}{2}\), and a FB sequence of entropy rate less than \(R - \epsilon_2\) (\(\epsilon_2\) goes to zero as \(\epsilon_1\) does). The second step uses the fact that for any stationary ergodic process, the LZ coding algorithm is an asymptotically optimal lossless compression scheme. Therefore, by choosing sufficiently large block length, the difference between the bit per symbol resulting from LZ compression of the FB sequence, and its entropy rate could be made sufficiently small. The third step is using the asymptotic optimality of \(\text{FB-DUDE}\) decoding algorithm which guarantees that by choosing decoding window lengths properly, there is no loss in using \(\text{FB-DUDE}\) decoder instead of any other possible sliding-window decoder.

Note that the only part of our scheme of questionable practicality is its encoding which requires listing all mappings of some finite window length which generate a FB sequence with LZ description length less than some fixed value depending on block length and coding rate. This is a huge number, e.g. for the binary case there are \(2^{2k+1}\) mappings having a window length of \(2k+1\) (for \(k = 1\) there are 256 mappings). Therefore, we cannot implement the algorithm as described, but as shown in Section IV, this new scheme inspires pragmatic universal \(\text{WZ}\) coding schemes attaining good performance.

Finally, it is worth mentioning the relationship between our encoding algorithm and the Yang-Kieffer fixed-rate lossy coding algorithm described in [40]. For block length \(n\), the encoder constructs a codebook \(C_n\) consisting of all reconstruction blocks having LZ description length less than \(nR\). The \(x^n\) is represented by the nearest codeword \(\hat{x}^n\) in \(C_n\). Yang and Kieffer [40] show that for any stationary ergodic source, this conceptually simple (but not implementable) scheme achieves the rate-distortion function as \(n\) goes to infinity. In our case, we construct our codebook in a similar way, but since the encoder knows that the decoder has access to the output of the DMC as well, the codeword that results in minimum average expected loss is chosen (the expectation is taken over all possible channel outputs for the best possible \(\text{SB}\) decoder).

VI. PRAGMATIC APPROACHES AND EXPERIMENTAL RESULTS

As mentioned earlier, the demanding aspect of the \(\text{WZ-DUDE}\) algorithm is finding the optimal mapping \(f^*\). In all of the following cases, instead of looking for the optimal mapping, we choose a not-necessarily optimal mapping along with the
WZ-DUDE decoder. Furthermore, in all of the following cases, the distortion measure is Hamming distortion, i.e., for \( x, \hat{x} \in \mathcal{X} \)

\[
\lambda(x, \hat{x}) = \begin{cases} 
0, & x = \hat{x}; \\
1, & x \neq \hat{x}.
\end{cases}
\]

**A. Binary Image with BSC**

In this experiment, instead of looking for the optimal mapping, we use a lossy JPEG encoder. Since except for the encoding of the DC component, JPEG works on \( 8 \times 8 \) blocks separately, it can be considered as a SB encoder of window length 1 working on the super-alphabets formed by \( 8 \times 8 \) binary blocks. Fig. 2 and Fig. 3 show the original binary image and its noise-corrupted version under a binary symmetric channel with crossover probability 0.15. Fig. 4 shows the JPEG encoded image which requires 0.22 bit per pixel (b.p.p.) after JPEG lossless compression, compared to 0.6 b.p.p. required by the original image. The average Hamming distortion between the original image and the decompressed one is 0.0556. Fig. 5 shows the result of denoising the noise corrupted image with DUDE ignoring the FB sequence. In this case the resulting average distortion is 0.0635. On the other hand, Fig. 6 shows the result of denoising the noisy signal when the FB sequence is also taken into account. The decoder/denoiser in this case is WZ-DUDE with parameters \( l = 1 \) and \( m = 1 \). The final average distortion between the reconstructed image and the original image is 0.0407.

**B. Text with erasure channel**

In this section we consider the case where our source is an English text document, and the DMC is a memoryless erasure channel that erases each symbol with probability \( \epsilon \). To construct the FB sequence, we use a method which is similar to the first run of the DUDE algorithm in which it tries to estimate \( P_{Z^i \mid Z_{\leq i}}(\cdot \mid \cdot) \), to estimate \( P_{X_{\leq i} \mid X_{> i}}(\cdot \mid \cdot) \). For a given window length of \( 2k + 1 \), the encoder generates the count matrix \( r_{\text{enc}} \) as follows

\[
r_{\text{enc}}(x^n, a_k, b_k)[\beta] = |\{i : x_{i-k}^i = a_k, x_{i+k}^{i+1} = b_k, x_i = \beta\}|.
\]

(17)

For each left and right contexts \( a_k \) and \( b_k \), the vector \( r_{\text{enc}}(x^n, a_k, b_k) \) is a \( 1 \times |\mathcal{X}| \) vector with \( \beta \)-th component being equal to the number of times the \( \beta \)-th element of \( \mathcal{X} \) have appeared in \( x^n \) sequence with its right and left contexts being equal to \( a_k \) and \( b_k \) respectively. Therefore, from the count vector \( r_{\text{enc}}(x^n, x_{i-k}^i, x_{i+1}^{i+1}) \) corresponding to the right and left contexts of \( x_i \), the MAP estimation of \( x_i \) is

\[
\arg \max_{\beta \in \mathcal{X}} r_{\text{enc}}(x^n, x_{i-k}^i, x_{i+1}^{i+1})[\beta],
\]

(18)

which is the symbol in \( \mathcal{X} \) is the most frequent symbol in \( x^n \) among those with the same right and left contexts of \( z_i \). Similarly, for given right and left contexts, we can rank all the symbols in \( \mathcal{X} \) according to their repetition frequency in our text within the given contexts. Now for a FB alphabet cardinality of \( N \), define \( \mathcal{Y} = \{1, \ldots, N\} \). The encoding function \( f \) is as follows

\[
f(x_{i-k}^{i+1}) = \begin{cases} 
\ell, & \text{if } x_i = \beta, \text{ where } r_{\text{enc}}(x^n, x_{i-k}^{i+1})[\beta] \text{ is the } \ell \text{-th largest element, and } \ell < N; \\
N, & \text{otherwise}.
\end{cases}
\]

(19)

After constructing the sequence \( y^n \) by sliding \( f \) over the original text, the LZ description of the resulting sequence...
is transmitted to the decoder. As mentioned in [12], the DUDE denoising rule for an erasure channel is equivalent to a majority-vote of the context counts, i.e. replacing each erasure with the most frequent symbol with the same context. WZ-DUDE decodes the erased symbol \( x_i \) by first computing \( r_{\text{dec}}(z^{i-1}_n, z^{i+k}_n) \). For moderate values of \( e \), one would expect \( r_{\text{enc}} \) and \( r_{\text{dec}} \) to rank the symbols similarly. Therefore, based on \( r_{\text{dec}} \) count vector, and \( y_i \), \( \hat{x}_i \) is the source alphabet corresponding to the \( y_i \)-th largest element of \( r_{\text{dec}}(z^{i-1}_n, z^{i+k}_n) \). Note that in this case, the window length of the SB encoder and decoder should be the same, otherwise \( y^n \) does not help the decoder.

Fig. 7 shows the percentage of erased symbols that are recovered by our WZ-DUDE decoder for different values of \( N \). For our experiments we have used the English translation of Don Quixote de La Mancha, by Miguel de Cervantes Saavedra (1547-1616)\footnote{The text consists of approximately 2.3 \times 10^6 characters. The channel is assumed to have erasure probability 0.1. In Fig. 7, \( N = 1 \), corresponds to the case where there is no FB available to the decoder, or in other words, it corresponds to the performance of the DUDE. As it can be observed, for \( N = 1 \) using larger context size \( k \) improves performance; As \( N \) increases, \( k = 1 \) outperforms \( k = 2 \). In addition both curves seem to eventually saturate as \( N \) increases. Although one might expect that increasing \( N \) would always improve the performance, and tending it to \( |X| \), one should be able to recover all erased symbols, we see in Fig. 8, this does not hold for our scheme. The reason is that, once one of the symbols in the context of an erased symbol is erased, the decoder is not able to construct the count vector \( r_{\text{dec}}(z^{i-1}_n, z^{i+k}_n) \), which is crucial in interpreting the FB sequence. In such cases we let \( \hat{x}_i \) to be the space character which has the largest frequency in the text. Therefore, the best error-correction performance that can be achieved by our scheme is upper bounded by the probability that none of the symbols in the context of an erased symbol are erased, which is equal to \((1-e)^2k\). In our example, for \( k = 1 \) the upper bound is \( 0.9^2 = 0.81 \), and for \( k = 2 \), it is \( 0.9^4 \approx 0.66 \), which coincides with our curves.

To illustrate the performance of the algorithm, a small excerpt of length 154 of the original text, its noise-corrupted version, and the outputs of its DUDE and WZ-DUDE decoded versions, for different values of \( k \), and \( N \) are presented.

- **Clean text:**
  
  ... methodising with rare patience and judgment what had been previously brought to light, he left, as the saying is, no stone unturned under which anything ...

- **Erasure-corrupted source:** (12 erasures)
  
  ...*et*odising with ra*e pati*nce and judgment what had been previously brought to ligh*, he l*ft* as the sayin* is, no stone untu*ned under which any*hin*...

- **DUDE denoiser with no FB sequence**, \( k = 1 \) (7 errors + 1 erasure)
  
  ...et*odising with rave patience and judgment what had been previously brought to light, he left as the saying is, no stone unturned under which any*thing ...

- **DUDE denoiser**, \( k = 1 \), \( N = 2 \), \( R = 0.16 \) b.p.s. (2 errors)

- **DUDE denoiser with no FB**, \( k = 2 \) (3 errors)
...methodising with rate patience and judgment what had been previously brought to light, he lefts as the saying is, no stone unturned under which anything ...

- WZ-DUDE denoiser; \( k = 2, N = 2, R = 0.12 \) b.p.s. (2 errors)
  
  ...methodising with race patience and judgment what had been previously brought to light, he left, as the saying is, no stone unturned under which anything ...

where

\[ f(l, r, \alpha, \beta) \triangleq \lim_{e \to 0} \frac{1}{e} \mathbb{P}(\hat{X}_i \neq X_i) = q, \]

and

\[ \lim_{e \to 0} \frac{1}{e} \mathbb{P}(\hat{X}_i \neq X_i) = 0.5. \]
relatively small, the erased symbol can be recovered correctly with high probability. On the other hand, as \( e \) increases, the probability of having two consecutive erased symbols, which are harder to recover, increases as well.

Now consider the WZ setup, where in addition to the output of the BEC, the decoder has access to a FB sequence of rate \( R \) designed by the encoder to improve the decoder’s performance. For generating this FB sequence we again use DUDE counts. For a fixed \( k \), the encoder first forms the count matrix consisting of \( r_{\text{enc}}(a^k, b^k) \) for all \( 2^{2k} \) possible right and left contexts, each of length \( k \). Then the FB sequence is

\[
Y_i = \begin{cases} 
0, & \text{if } r_{\text{enc}}(X_{i-k}^{i-1}, X_{i+k}^{i+1})[X_i] \geq r_{\text{enc}}(X_{i-k}^{i-1}, X_{i+k}^{i+1})[1-X_i] \\
1, & \text{otherwise.} 
\end{cases}
\]  

(30)

In other words, \( Y_i \) is 1 whenever \( X_i \) is different from what it is predicted to be from its context. As mentioned above, the DUDE decision rule for the BEC is majority-vote decoding. In the case of a binary source instead of text, if there are erased symbols in the context of an erased bit, we do not simply let \( X_i \) equal to some pre-fixed symbol. When in addition to \( Z_i \) some other bits of \( Z_{i-k}^{i+k} \) are erased, the decoder’s count vector \( r_{\text{dec}}(Z_{i-k}^{i-1}, Z_{i+k}^{i+1}) \) is the average of count vectors corresponding to all possible binary contexts coinciding with \( Z_{i}^{i+k} \) at the non-erased positions. If \( k_e \) bits out of \( 2k \) are erased, then there exist \( 2^{k_e} \) such contexts that agree with original context in the \( 2k-k_e \) non-erased bits. Let \( b = \arg \max_{\beta \in \{0,1\}} r_{\text{dec}}(Z_{i-k}^{i-1}, Z_{i+k}^{i+1})[\beta] \), then

\[
\hat{X}_i = \begin{cases} 
b, & \text{if } Y_i = 0; \\
1-b, & \text{if } Y_i = 1. 
\end{cases}
\]

(31)

Consider again the BSMS with \( q = 0.25 \) passed through a BEC with \( e = 0.1 \). From Fig. 8, an optimal denoiser that only has access to the output of BEC will decode at least 25.13% of the erased bits wrongly. On the other hand, the DUDE denoiser decodes 25.44% of erased bits erroneously, which is almost equal to the performance of the optimal non-universal denoiser which knows the statistics of the source. Now assume that the encoder also sends to the decoder the FB sequence \( Y^n \) constructed as described in (30). From our simulation results, for this case, the entropy of the FB sequence is around \( R = 0.3 \), and applying the described WZ-DUDE decoder to \( (Y^n, Z^n) \) reduces the probability of error to 19.3%.

VII. CONCLUSION AND FUTURE DIRECTIONS

This paper deals with WZ coding of a source with unknown statistics; a new WZ coding algorithm, WZ-DUDE, was presented and its asymptotic optimality was established. In order to optimize the scheme one would list all possible mappings that have a certain property and look for the one that gives minimum expected loss. However, we saw that even a simple encoding mapping, namely an off-the-shelf lossy compressor, achieves considerable improvement compared to the case where either the FB sequence or the noisy signal are not present at the decoder.

The original DUDE is tailored to discrete-alphabet sources going through a DMC, and making it applicable to continuous alphabet sources entails more than a trivial extension, which has been accomplished in [31] and [32]. As mentioned in Section III, since our fbDUDE decoder is a special case of the original DUDE algorithm, one would expect that by following the same methods used in [31], [32], it might be possible to devise a decoder which works on continuous data. The non-trivial part is finding a proper encoder. In this case it is not possible to list all SB encoders of some finite block length, because there are an infinite number of them even for a window length of one. One simple solution is to look into all mappings which map to a quantized version of the source alphabet. How to choose this quantized alphabet, and whether this would result in a scheme that asymptotically achieves the performance bounds is a question that requires further study. Finding a sequential version of the described scheme, where the decoder is subject to a delay constraint is another interesting open avenue. Adapting our WZ-DUDE algorithm to
perform effectively with non-stationary data is another open avenue. For example, often real data is more accurately modeled as a piecewise stationary source. In the recent work [33], the sDüDE denoising algorithm is described which, unlike DüDE, tries to compete with a genie-aided SB denoiser that can switch between SB denoisers up to $m$ times, where $m$ is sub-linear in the block length $n$. When the clean data is emitted by a piecewise stationary process, the sDüDE algorithm achieves the optimum distribution-dependent performance.

**APPENDIX A: PROOF OF THEOREM 4.1**

The proof is an extension of the proof given in [34] which is for the case where there is no FB sequence. Let $X = \{X_i; \forall i \in \mathbb{N}^+\}$ be a stochastic process defined on a probability space $(X, \Sigma, P)$, where $\Sigma$ denotes the $\sigma$-algebra generated by cylinder sets, and $P$ is a probability measure defined on it. The shift operator $T : X^\infty \rightarrow X^\infty$ is defined by

$$(Tx)_n = x_{n+1}, \quad x \in X^\infty, n \geq 1.$$ 

Let $\mathcal{X}$ and $\hat{X}$ denote the source and reconstruction alphabets respectively, which are both assumed to be finite.

Since $(R, D)$ is assumed to be an interior point of the achievable region in the $R-D$ plane, there exists $\delta_0 > 0$, such that $(R - \delta, D)$ is also an interior point for any $0 < \delta \leq \delta_0$. Define $R_1 \triangleq R - \delta$. Since $(R_1, D)$ is an achievable point, for any given $\epsilon > 0$, there exists a block WZ encoder/decoder, $(f_n, g_n)$ of rate $R_1$ and block length $n$, which is sufficiently large based on $\epsilon$, and average expected distortion less than $D + \epsilon$. Assume that among our infinite choices, we pick a WZ code whose block length $n$ is large enough such that

$$\max \left\{ \frac{1}{\sqrt{n}}, \frac{\log n}{n \log |Y|} \right\} < \epsilon. \quad (A-1)$$

This constraint will be useful in our future analysis.

In order to prove that there exists a SB code satisfying the constraints given in Theorem 4.1, the given block code $(f_n, g_n)$ should somehow be embedded in the SB encoder/decoder mappings. For defining a SB code based on a block code, the natural question is how to define blocks in the infinite length source sequence. Moreover, after finding a way for distinguishing blocks in the input sequence, the next problem is how the decoder is going to detect the coded blocks in the infinite length received FB sequence. To answer the first question, as usually done in the literature, we resort to the Rohlin-Kakutani (RK) Theorem of ergodic theory [35]:

**Theorem 7.1 (Rohlin-Kakutani Theorem):** Given the ergodic source $[A, \mu, U]$, integers $L$ and $n \leq L$, and $\epsilon > 0$, there exists an event $F$ (called the base) such that

1. $F, TF, \ldots, T^{L-1}F$ are disjoint,
2. $P \left( \bigcup_{i=0}^{L-1} T^iF \right) \geq 1 - \epsilon,$
3. $P(S(a^n)|F) = P(S(a^n))$, where $S(a^n) = \{x : x^n = a^n\}$.

This theorem states that for any given $L$, and any $n$ less than $L$, there exists a base event $F$, such that the base and its $L$ disjoint shifts, basically cover the entire space, i.e., any given sequence $X$ with high probability belongs to $T^3F$ for some $0 \leq i \leq L - 1$. The last property states that the probability distribution of the $n$-tuples is the same both in the base and in the whole space.

For a given $\epsilon > 0$, $n$, the block length of the block encoder/decoder $(f_n, g_n)$, and $L_n \triangleq n + \lceil n\epsilon \rceil$, let $F$ be the base event given by the RK theorem for these parameters. Define $G$ to be everything in the event space which is not included in $\bigcup_{i=0}^{L_n+1} T^iF$. Note that by the RK theorem $P(G) \leq \epsilon$. To show the existence of a finite length SB encoder, we first prove the existence of an infinite length SB encoder, $f^{(\infty)}$, and then show that it can be truncated appropriately such that the resulting finite window length code also satisfies our desired properties.

Note that $f^{(\infty)}$ maps every infinite length sequence $x$ into a symbol in the FB sequence alphabet $Y$, and defines the FB sequence as $y_i = f^{(\infty)}(T^i x)$. As mentioned earlier, one problem is enabling the decoder to discriminate between the encoded blocks embedded in the FB sequence. One simple solution is requiring the SB encoder to interject a pre-defined synchronization sequence, which is not contained in any of the codewords, between the encoded blocks. Let $s = 1_{[ne]}$, where $1_r$ is a vector of length $r$ with all of its elements equal to 1, denote the synchronization block. From now on, symbols 0 and 1 denote two arbitrary distinct symbols in $Y$. The following lemma shows that as long as $|Y| > 2^{R - \epsilon_n}$, it is possible to construct a codebook of $2^{nR}$ distinct codewords none of them containing $s$.

**Lemma 7.2:** If $|Y| > 2^{R - \epsilon_n}$, where $\epsilon_n = \log(1 - n|Y|^{-\lceil ne \rceil})$, it is possible to find a codebook $C \subset Y^n$ with $2^{nR}$ codewords such that $s = 1_r$ is not contained in any of them.

**Proof:** Let $N_s$ denote the number of sequences in $Y^n$ that contain $s$ as part of them. There are $n - \lceil ne \rceil + 1$ positions that might be the start of the $s$. For each of them it is possible to construct $|Y|^{n - \lceil ne \rceil}$ sequences that contain $s$ starting at that certain position. Therefore, $N_s$ is upper-bounded as follows,

$$N_s < (n - \lceil ne \rceil + 1)|Y|^{n - \lceil ne \rceil}. \quad (A-2)$$

On the other hand, if $|Y|^{n - N_s} > 2^{nR}$, then it is possible to choose $2^{nR}$ codewords as desired. Combining this with (A-2), it is sufficient to have

$$|Y|^{n - \lceil ne \rceil} > n|Y|^{n - \lceil ne \rceil}, \quad (A-3)$$

or

$$|Y|^{n - 1 - \lceil ne \rceil} > 2^{nR}, \quad (A-4)$$

or

$$\log |Y| > R - \epsilon_n, \quad (A-5)$$

where $\epsilon_n$ is as defined in the statement of the lemma.

Therefore, using the previous lemma, it is possible to construct a codebook $C$ with $2^{nR}$ codewords, such that none of them contain $s$. Further, we can assume that the codewords in $C$ are chosen such that the first and the last symbols of all of them are equal to 0. Note that if $|Y|$ satisfies (A-5), then the number of codewords that satisfy the requirement of Lemma 7.2 is exponentially more than $2^{nR}$. Therefore, it is possible to choose such a codebook. This assumption makes sure that for any $y^n \in C$, the synchronization sequence can uniquely be detected in $y_1, \ldots, y_n, s_1, \ldots, s_{\lceil ne \rceil}$ and also in $s_1, \ldots, s_{\lceil ne \rceil}, y_1, \ldots, y_n$ with no ambiguity.
Now each codeword in the codebook \( \mathcal{C} \) can be mapped into a unique codeword in \( \hat{\mathcal{C}} \). The role of each element in \( \mathcal{C} \) in the coding is then played by the corresponding vector in \( \hat{\mathcal{C}} \) that it is mapped to. Since in the WZ coding, the codebook is only an indexing of the input blocks, such a mapping only acts as a renaming of the vectors in the codebook, and does not have any other effect.

Now we define an infinite length encoder \( f^{(\infty)} \) based on the partitioning of the event space given by the RK Theorem as follows.

1. \( x \in T^i, F \), for some \( 0 \leq i \leq n - 1 \); let \( f^{(\infty)}(x) = \hat{y}_i \) where \([y_0, y_1, \ldots, y_{n-1}] = f_n(x_{-i-1}^{-i}).\)
2. \( x \in T^n F \), for some \( n \leq i \leq L_n - 1 \); let \( f^{(\infty)}(x) = s[i - n + 1], \)
3. \( x \in G \); let \( f^{(\infty)}(x) = y_0 \), where \( y_0 \) is an element in \( \mathcal{Y} \) which is not used in \( s \).

After defining the SB encoder, we can define the SB decoder, \( g \), which generates the reconstruction process as \( \hat{X}_i = g(Z_{i-M}^{i-M}, Y_{i-M}^{i-M}) \) with \( M = 2(\lfloor n \rfloor + \lceil ne \rceil) + 1 \). The decoder \( g \) searches the block \( Y_{i+1} \) for a synchronization sequence. At most there will be one such sequence. If it detects one string \( s \), which starts at position \( i + r, 1 \leq r \leq n + 1 \), then it lets \( \hat{X}_i = U_{n-r+1} \), where \([\hat{U}_1, \hat{U}_2, \ldots, \hat{U}_n] = g_n(Z_{i+r-n}^{i+r-n}).\)

In order to compute the expected average distortion between the source and reconstruction sequences, note that since the original process and its reconstruction are jointly stationary, the average expected distortion between them is equal to

\[
E\left[ \lambda(X_0, \hat{X}_0) \right] = \sum_{i=0}^{n-1} E\left[ \lambda(X_0, \hat{X}_0) | T^i F \right] P(T^i F)
\]

By the stationarity of the source,

\[
P\left( \bigcup_{i=n}^{L_n-1} T^i F \right) = [nc] P(F).
\]

And \( P\left( \bigcup_{i=n}^{L_n-1} T^i F \right) \leq [nc] P(F) \). Therefore,

\[
P\left( \bigcup_{i=n}^{L_n-1} T^i F \right) = [nc] P\left( \bigcup_{i=0}^{L_n-1} T^i F \right)
\]

where (a) follows from (A-1). Moreover, from the RK Theorem, \( P(G) \) \( < \epsilon \), which together with (A-8) shows that

\[
E\lambda(X_0, \hat{X}_0) < \sum_{i=0}^{n-1} E\left[ \lambda(X_0, \hat{X}_0) | T^i F \right] P(T^i F) + 2\lambda_{\text{max}} \epsilon.
\]  

(A-9) For bounding the first term in (A-9), note that

\[
\sum_{i=0}^{n-1} E\left[ \lambda(X_0, \hat{X}_0) | T^i F \right] P(T^i F) = \sum_{i=0}^{n-1} E\left[ \lambda(X_0, \hat{X}_0) | F \right] P(F)
\]

\[
< E\left[ \rho_n(X_n, \hat{X}_n) | F \right]
\]

where the last line follows from the fact that \( P(F) < \frac{1}{n} \). On the other hand, by the RK theorem, \( E\left[ \rho_n(X_n, \hat{X}_n) | F \right] = E\left[ \rho_n(X_n, g_n(Z_n, f_n(X_n))) \right] \). Consequently, combining all of the previous results,

\[
E\left[ \lambda(X_0, \hat{X}_0) \right] < E\left[ \rho_n(X_n, g_n(Z_n, f_n(X_n))) \right] + 2\lambda_{\text{max}} \epsilon,
\]

(A-10)

So far, we have shown the existence of a SB code which generates a reconstruction sequence within maximum distance of \( D + (1 + 3\lambda_{\text{max}}) \epsilon \). Moreover, from the RK Theorem, \( P(G) \) \( < \epsilon \), which together with (A-8) shows that

\[
E\lambda(X_0, \hat{X}_0) < \sum_{i=0}^{n-1} E\left[ \lambda(X_0, \hat{X}_0) | T^i F \right] P(T^i F) + 2\lambda_{\text{max}} \epsilon.
\]  

(A-9)
where for \( y, \tilde{y} \in \mathcal{Y} \), \( y \oplus \tilde{y} = 0 \) if \( y = \tilde{y} \) and 1 otherwise, also for \( 0 \leq \alpha \leq 1 \), \( h_b(\alpha) = -\alpha \log \alpha - (1 - \alpha) \log(1 - \alpha) \). Therefore, letting \( m \) grow to infinity, we conclude that
\[
H(\mathbf{Y}) \leq H(\hat{\mathbf{Y}}) + h_b(\sigma). \tag{A-15}
\]

Now we turn to bounding \( H(\hat{\mathbf{Y}}) \). Let \( \{\theta_i\} \) denote a sequence defined as follows,
\[
\theta_i = \begin{cases} 
  j, & \hat{Y}_i \text{ is the } j^{th} \text{ letter of a codeword}, \\
  0, & \text{otherwise}.
\end{cases} \tag{A-16}
\]
Then the entropy rate of the generated FB process \( \hat{\mathbf{Y}} \) can be upper-bounded as follows
\[
H(\hat{\mathbf{Y}}) = \lim_{m \to \infty} \frac{1}{m} H(\hat{\mathbf{Y}}^m),
\leq \lim_{m \to \infty} \frac{1}{m} H(\hat{\mathbf{Y}}^m, \theta^m),
= \lim_{m \to \infty} \frac{1}{m} \left[ H(\theta^m) + H(\hat{\mathbf{Y}}^m | \theta^m) \right],
= \lim_{m \to \infty} \frac{1}{m} \left[ H(\theta^m) + H(\hat{\mathbf{Y}}_{m-1}^m, \theta^m) \right],
\leq \lim_{m \to \infty} \left[ \frac{1}{m} H(\theta^m) + H(\hat{\mathbf{Y}}_{m-1}^m, \theta^m) \right], \tag{A-17}
\]
where
\[
H(\hat{\mathbf{Y}}_{m-1}^m, \theta_m) = \sum_{j=0}^{L_n} H(\hat{\mathbf{Y}}_{m-1}^m, \theta_m = j) P(\theta_m = j),
\leq \epsilon \log |\mathcal{Y}| + \frac{1}{n} \sum_{j=1}^{n} H(\hat{\mathbf{Y}}_{m-1}^m, \theta_m = j) P(\theta_m = j),
\leq \epsilon \log |\mathcal{Y}| + \frac{1}{n} \sum_{j=1}^{n} H(\hat{\mathbf{Y}}_{m-1}^m, \theta_m = j),
= \epsilon \log |\mathcal{Y}| + \frac{1}{n} \sum_{j=1}^{n} H(\tilde{\mathbf{f}}_m(\mathbf{X}^n)),
\leq \epsilon \log |\mathcal{Y}| + R_1, \tag{A-18}
\]
where \((a)\) follows from the facts that, for \( n + 1 \leq j \leq L_n \), \( H(\hat{\mathbf{Y}}_{m-1}^m, \theta_m = j) = 0 \), and \( P(\theta_m = j) < \epsilon \). Moreover, we show that the entropy rate of the \( \{\theta_i\} \) process can be made arbitrary small:
\[
\lim_{m \to \infty} \frac{1}{m} H(\theta^m) = \lim_{m \to \infty} H(\theta_m | \theta_{m-1}),
\leq \lim_{m \to \infty} H(\theta_m | \theta_{m-1}),
= \sum_{j=0}^{L_n} P(\theta_0 = j) H(\theta_1 | \theta_0 = j), \tag{A-19}
\]
where the last step is a result of stationarity. By the definition of the \( \{\theta_i\} \) sequence, \( H(\theta_1 | \theta_0 = j) = 0 \), for \( 1 \leq j \leq n - 1 \), \( P(\theta_0 = 0) = P(G) \leq \epsilon \), and \( P(\theta_0 = L_n) = P(F) \leq \frac{1}{n} \). Given \( \theta_0 = 0, \theta_1 \) can either be zero or one, therefore \( H(\theta_1 | \theta_0 = 0) \leq 1 \). Similarly, conditioned on \( \theta_0 = L_n, \theta_1 \) can only be zero or one, and \( H(\theta_1 | \theta_0 = n) \) computes the uncertainty that one has in determining whether a sequence belongs to \( \bigcup_{i=0}^{L_n} T^i F \) or not when it is known that \( \mathbf{X} \in T^{-L_n} F \). Since conditioning can only reduce entropy, and \( P(G) < \epsilon \), it follows that \( H(\theta_1 | \theta_0 = L_n) < h_b(\epsilon) \). Consequently,
\[
\lim_{m \to \infty} \frac{1}{m} H(\theta^m) = P(\theta_0 = 0) H(\theta_1 | \theta_0 = 0),
\leq \frac{1}{n} h_b(\epsilon).
\tag{A-20}
\]

Combining (A-15), (A-17) and (A-20), it follows that
\[
H(\mathbf{Y}) \leq R_1 + (1 + \log |\mathcal{Y}|) \epsilon + \frac{1}{n} h_b(\epsilon) + h_b(\sigma), \tag{A-21}
\]
where as defined before
\[
\sigma = \frac{\epsilon}{2(n + |n\epsilon|)} + 1.
\]

Note that \( (1 + \log |\mathcal{Y}|) \epsilon + \frac{1}{n} h_b(\epsilon) + h_b(\sigma) \) goes to zero as \( \epsilon \) goes to zero. Therefore, there exists \( \epsilon' > 0 \), such that \( \frac{1}{n} + \log |\mathcal{Y}| \epsilon + \frac{1}{n} h_b(\epsilon) + h_b(\sigma) < \frac{\epsilon}{2} \), for any \( \epsilon < \epsilon' \). By definition \( R_1 = R - \delta \). Consequently, by choosing \( \epsilon < \min(\epsilon', \epsilon'') \), where \( \epsilon'' \doteq \frac{2 \delta}{1 + 3 \log |\mathcal{Y}|} \), and \( \epsilon_2 \doteq \frac{\epsilon}{2} > 0 \), we get a SB encoder \( f^{(k)} \) and SB decoder \( g \) generating FB and reconstruction sequences satisfying
1) \( E[\lambda(X_0, \hat{X}_0)] < D + \epsilon_1 \),
2) \( H(\hat{\mathbf{Y}}) < R - \epsilon_2 \).

**APPENDIX B: PROOF OF THEOREM 5.1**

First, we prove that for any given \( \epsilon > 0 \), there exists \( N_\epsilon > 0 \) such that for \( n > N_\epsilon \),
\[
E[\rho_n (X^n \cdot \mathbf{g}_n(Z^n, T_n^n(X^n)))] < D_{X_\Pi}(R) + \epsilon. \tag{B-1}
\]
By definition, \( D_{X_\Pi}(R) \) denotes the infimum of all distortions achievable by WZ coding of source \( X \) at rate \( R \) when the DMC is described by \( \Pi \). Therefore, for any \( \epsilon > 0 \), \( (D + \frac{\epsilon}{2}, R) \) would be an interior point of the rate-distortion region. Hence, by theorem [4.1] for \( \epsilon_1 = \frac{\epsilon}{2} > 0 \), there exist some \( \epsilon_2 > 0 \), and a sliding-block WZ code with mappings \( f \) and \( g \), each one having a finite window length, such that
1) \( E[\lambda(X_i, g(Z_{i+k}^{i+k} | Y_{i-k}^{i-1}, Y_{i-1}^{i+m}))] \leq D + \frac{\epsilon}{2} \), where \( Y_i = f(X_{i+k}^{i+k}) \),
2) \( H(Y_i) = \lim_{n \to \infty} \frac{1}{n} H(Y_1, \ldots, Y_n) \leq R - \epsilon_2 \), for some \( \epsilon_2 > 0 \).

On the other hand, the FB process \( \{Y_i\} \) generated by sliding windowing a stationary ergodic process \( \{X_i\} \) with a time invariant mapping \( f \), is also a stationary ergodic process. Consequently, since for any stationary ergodic process Lempel-Ziv coding algorithm is an asymptotically optimal lossless compression scheme [27], for any given \( \sigma > 0 \), there exists \( N_\sigma > 0 \), such that for \( n > N_\sigma \),
\[
\frac{1}{n} \text{LZ}(Y_1, \ldots, Y_n) \leq H(\mathbf{Y}) + \sigma. \tag{B-2}
\]
Letting \( \sigma = \frac{\epsilon_2}{2} \), and choosing \( n \) greater than the corresponding \( N_\sigma \), yields
\[
\frac{1}{n} \text{LZ}(Y_1, \ldots, Y_n) \leq R - \epsilon_2 + \sigma,
\leq R - \frac{\epsilon_2}{2},
< R.
\]
Therefore, for any given $\epsilon > 0$, and any source output sequence, by choosing the block length $n$ sufficiently large, the mapping $f$ would belong to $S(x^n, k_n, R)$. On the other hand, since for any individual source sequence $x^n$, $f^*$ is the mapping in $S$ that defines the FB sequence minimizing the expected distortion, it follows that

$$ V(f^*, l, m) < V(f, l, m). \quad (B-3) $$

Moreover, since $V(f^*, l, m)$ is the minimum accumulated loss attainable by the mappings in $S(n, l, m)$, when the decoder is constrained to be a sliding window decoder with parameters $l$ and $m$, it is in turns less than the expected distortion obtained by the specific mapping $g$ given by Theorem 4.1 i.e.

$$ E \left[ \sum_{i=k+1}^{n-k} \lambda \left( x_i, g^* \left( Z_{i-l}^{i+l}, y_{i-m}^{i+m} \right) \right) \right] \leq E \left[ \sum_{i=k+1}^{n-k} \lambda \left( x_i, g \left( Z_{i-l}^{i+l}, y_{i-m}^{i+m} \right) \right) \right] + \frac{\epsilon}{4}, \quad (B-4) $$

where $y_i = f(x_{i-k}^{i+k})$ and $\hat{y}_i = f^*(x_{i-k}^{i+k})$.

The final step is applying Theorem 3.1 which is the asymptotic optimality of WZ DUDE algorithm in the semi-stochastic setting. From this result, when the parameter $l$ and $m$ are such that $t_n = \max\{l, m\} = o(n/\log n)$, the difference between the performance of the WZ DUDE decoding algorithm and the optimal sliding-window decoder of the same order goes to zero as the block length goes to infinity. In other words, for any given $\epsilon > 0$, there exists $N'_\epsilon > 0$, such that for $n > N'_\epsilon$,

$$ \frac{1}{n-2k} E \left[ \sum_{i=k+1}^{n-k} \lambda \left( x_i, \hat{x}_i \right) \right] \leq \frac{1}{n-2k} E \left[ \sum_{i=k+1}^{n-k} \lambda \left( x_i, g^* \left( Z_{i-l}^{i+l}, y_{i-m}^{i+m} \right) \right) \right] + \frac{\epsilon}{4}, \quad (B-5) $$

where $\hat{x}^n = g^*_n \left( Z^n, f^*_n(X^n) \right)$. Note that the only uncertainty in (B-5) is due to the channel noise, and the source and FB sequence are assumed to be individual sequences. Combining (B-4) and (B-5), it follows that with probability one

$$ \frac{n}{n-2k} E \left[ \rho_n \left( x^n, g^*_n \left( Z^n, f^*_n(X^n) \right) \right) \right] \leq \frac{1}{n-2k} E \left[ \lambda \left( x_i, g \left( Z_{i-l}^{i+l}, y_{i-m}^{i+m} \right) \right) \right] + \frac{\epsilon}{4}. \quad (B-6) $$

On the other hand, since $\{(X_i, Y_i)\}_{i=1}^{\infty}$ is also a stationary ergodic process with super-alphabet $\mathcal{X} \times \mathcal{Y}$, by the ergodic theory, with probability one,

$$ \lim_{n \to \infty} \frac{1}{n-2k} \sum_{i=k+1}^{n-k} E \left[ \lambda \left( x_i, g \left( Z_{i-l}^{i+l}, y_{i-m}^{i+m} \right) \right) \right] \leq D_{X, Y} \left( R \right) + \frac{\epsilon}{2}. $$

This means that with probability one, there exists $N''_\epsilon > 0$, such that for $n > N''_\epsilon$,

$$ \frac{1}{n-2k} \sum_{i=k+1}^{n-k} E \left[ \lambda \left( x_i, g \left( Z_{i-l}^{i+l}, y_{i-m}^{i+m} \right) \right) \right] \leq D_{X, Y} \left( R \right) + \frac{\epsilon}{2} + \frac{\epsilon}{4}. \quad (B-7) $$

Finally, combining (B-6) and (B-7), and taking $n > N_\epsilon$, where $N_\epsilon = \max\{N'_\epsilon, N''_\epsilon\}$, yields the desired result as follows

$$ \frac{n}{n-2k} E \left[ \rho_n \left( X^n, g^*_n \left( Z^n, f^*_n(X^n) \right) \right) \right] \leq D_{X, Y} \left( R \right) + \epsilon. \quad (B-8) $$

ACKNOWLEDGMENT

The authors would like to thank Erik Ordentlich and Pascal Vontobel for the helpful discussions. They also acknowledge Stanford Graduate Fellowship supporting the first author, NSF Grant CCR-0312839, National Science Foundation, Theoretical Foundations Research Grants CCF-0635154 and CCF-0728445 for supporting the second author, and NSF CAREER grant for supporting the third author. The authors also would like to thank the anonymous reviewers for their helpful comments.

REFERENCES


