Universal Lossless Compression-based Denoising

Han-I Su and Tsachy Weissman
Department of Electrical Engineering
Stanford University
Stanford, CA 94305, USA
Email: {hanisu, tsachy}@stanford.edu

Abstract—**In a discrete denoising problem, if the denoiser knows the clean source distribution, the Bayes optimal denoiser is the Bayes response of the posterior distribution of the source given the noisy observations. However, in many applications the source distribution is unknown. We consider the Bayes response based on the approximate posterior distribution induced by a universal lossless compression code. Motivated by this approach, we present the empirical conditional entropy-based denoiser. Simulations show that when the source alphabet is small, the proposed denoiser achieves the performance of the Universal Discrete DEnoiser (DUDE). Furthermore, if the alphabet size increases, the proposed denoiser degrades more gracefully than the DUDE.**

I. Introduction

The setting of a denoising problem is shown in Fig. 1. The clean source sequence may be a Markov process or texts. The source is corrupted by a memoryless noisy channel, and the denoiser observes the noisy channel output sequence and reconstructs an estimation sequence of the source. The denoising performance is measured by a per-symbol loss function. If the denoiser computes the estimation sequence causally, the denoising problem is called a filtering problem. There are several well-developed approaches to denoising, such as linear filtering [1], [2], nonlinear denoising [3], and the forward-backward recursions for hidden Markov processes (see [4]). Most of these techniques assume that the denoiser has complete knowledge about the source distribution. However, this is not always realistic.

In a causal filtering problem, if the source distribution is unknown, a two-step denoising approach is to first estimate the source distribution and then reconstruct the source based on the estimated source distribution. For example, in [5] a universal lossless compression code induces an approximation of the source distribution, and the approximation yields a universal predictor. The universal predictor then induces a universal filter [6]. For the noncausal denoising problem, the Discrete Universal DEnoiser (DUDE) in [7] first estimates the posterior distribution of the source sequence given the observed noisy sequence and then compute the Bayes response of the approximate posterior distribution. The DUDE asymptotically achieves the optimal performance of genie-aided stationary (sliding-window) denoisers which know the clean source sequence. In [8], the classical unidirectional context-tree models [9] are generalized to multi-directional settings for universal denoising. In [10], bidirectional models are established from unidirectional models and achieve better denoising performance than the DUDE does in some applications.

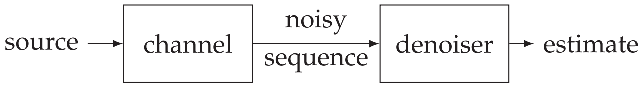

Fig. 1. Discrete denoising problem

In this paper, we approximate the unknown source distribution via a universal lossless compression code and apply it to universal lossless compression-based denoising. However, the computational complexity of this approach is in general too high. We then introduce the empirical conditional entropy-based denoiser to reduce the computational complexity. Our simulation results show that the proposed denoiser is competitive with the DUDE for denoising binary first-order Markov processes. For denoising problems with larger source alphabets, such as randomly corrupted English texts, the empirical conditional entropy-based denoiser has better performance than the DUDE.

In the next section, we formally define the discrete denoising problem. In Section III, we present the empirical conditional entropy-based denoiser. Simulations in Section V compares the empirical conditional entropy-based denoiser with the DUDE. We conclude this paper in Section VI.

II. Problem Formulation and Preliminaries

We consider the discrete denoising problem in Fig. 1. The source sequence is an n-block stochastic sequence $X^n \triangleq (X_1, X_2, \ldots, X_n)$ (or deterministic sequence x^n), where X_i (or x_i) is in a finite alphabet $\mathcal{X}$. The source sequence is corrupted by a discrete memoryless channel $\Pi \in \mathbb{R}^{\mathcal{X}\times\mathcal{Z}}$, where $\mathcal{Z}$ is a finite alphabet of noisy symbols Z_i, and $\Pi(x,z) = P_{Z|X=x}(z)$. We assume that the channel matrix Π has full row rank and is available to the denoiser. The i-th component of the *n-block denoiser* $\hat{X}_i$ is

978-1-4244-7892-7/10/$26.00 ©2010 IEEE

a deterministic mapping from $\mathcal{Z}^n \triangleq \mathcal{Z} \times \mathcal{Z} \times \ldots \times \mathcal{Z}$ to the finite reconstruction alphabet $\hat{\mathcal{X}}$ for $i = 1, 2, \ldots, n$. Given a loss function $\Lambda \in \mathbb{R}^{\mathcal{X} \times \hat{\mathcal{X}}}$, where $\Lambda(x, \hat{x}) : \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$, *the expected per-symbol loss* is

$$\mathrm{E}\, L_{\hat{X}^n}(x^n, Z^n) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} \Lambda(X_i, \hat{X}_i(Z^n))\right).$$

We recall the Bayes optimal denoiser [7] minimizing the expected per-symbol loss in the following theorem.

Theorem 1 (see [7]): For the discrete denoising problem with channel matrix Π, loss matrix Λ, and arbitrarily distributed source X^n, the Bayes optimal denoiser $\hat{X}^n_{\text{opt}}$ which minimizes $\mathrm{E}\, L_{\hat{X}^n}(X^n, Z^n)$ is

$$\hat{X}_{\text{opt},i}(z^n) = \hat{X}_{\text{Bayes}}(P_{X_i|z^n}) = \Phi(\Pi, \Lambda, P_{Z_i|z^{n\backslash i}}, z_i). \quad (1)$$

The vector $P_{X_i|z^n}$ is a column vector in $R^{\mathcal{X}}$ with the x element $P_{X_i|z^n}(x) = \mathrm{P}(X_i = x | Z^n = z^n)$. The vector $P_{Z_i|z^{n\backslash i}}$ is defined in a similar way, where $z^{n\backslash i} = (z_1^{i-1}, z_{i+1}^n)$. The functions $\hat{X}_{\text{Bayes}}$ and Φ are defined as

$$\hat{X}_{\text{Bayes}}(P_{X_i|z^n}) = \underset{\hat{x}\in\hat{\mathcal{X}}}{\arg\min}\left(\lambda_{\hat{x}}^T P_{X_i|z^n}\right),$$

$$\Phi(\Pi, \Lambda, P_{Z_i|z^{n\backslash i}}, z_i) = \hat{X}_{\text{Bayes}}\left((\Pi\Pi^T)^{-1}\Pi P_{Z_i|z^{n\backslash i}} \odot \pi_{z_i}\right),$$

where $\lambda_{\hat{x}}$ is the $\hat{x}$ column of Λ, π_{z_i} is the z_i column of Π, and operator $\odot$ is component-wise multiplication. Note that $\hat{X}_{\text{Bayes}}(v)$ is invariant to positive scaling of v.

We also need the following definitions of universal lossless compression codes and universal probability assignments.

Definition 1: A compression code c_n is a mapping from $\mathcal{X}^n$ to $\{0,1\}^*$, the set of all finite binary strings. We denote the length of codeword $c_n(x^n)$ as $l_n(x^n)$. A compression code is *non-singular* if $c_n(x^n) \neq c_n(\tilde{x}^n)$ for all $x^n \neq \tilde{x}^n$, and a compression code is *lossless* if its extension is non-singular. A lossless compression code is *universal* if $\mathrm{E}((1/n)l_n(X^n)) \to \mathbb{H}(\mathbf{X})$ as $n \to \infty$ for all stationary processes $\mathbf{X}$, where $\mathbb{H}(\mathbf{X})$ is the entropy rate of the process $\mathbf{X}$.

Definition 2: A *probability assignment* Q is a set of conditional probabilities $\{\{Q_{X_i|x^{i-1}}\}_{x^{i-1}\in\mathcal{X}^{i-1}}\}_{i\geq 1}$. A probability assignment is *universal* if

$$\lim_{n\to\infty}\frac{1}{n}D(P_{X^n}\|Q_{X^n}) = 0$$

for all stationary processes $\mathbf{X}$ with distribution P, denoted as $\mathbf{X} \sim P$, where $Q_{X^n}(x^n) = \prod_{i=1}^n Q_{X_i|x_{i-1}}(x_i)$.

The following theorem shows that a universal lossless compression code induces a universal probability assignment.

Theorem 2 (see [11]): Given a universal lossless compression code c_n and the corresponding length function l_n, the probability assignment induced by

$$Q_{X^n}(x^n) = \frac{2^{-l_n(x^n)}}{\sum_{\tilde{x}^n} 2^{-l_n(\tilde{x}^n)}} \quad (2)$$

is universal.

III. Universal Lossless Compression-based Denoiser

In a discrete denoising problem, if the denoiser knows the source distribution, it also knows the noisy source distribution. Thus, the denoiser can compute the conditional probability $P_{Z_i|z^{n\backslash i}}$ and then find the Bayes optimal denoiser in (1). However, the computational complexity may grow too fast as n increases. Furthermore, in many practical scenarios, the source distribution is unknown. Thus, algorithms using approximations of the conditional probability $P_{Z_i|z^{n\backslash i}}$ are considered. For example, the DUDE of order k in [7] uses the empirical two-sided context counts

$$m_k\left(z^n, u_{-k}^{-1}, u_1^k\right)[u] = \\ |\{k+1 \leq i \leq n-k : z_{i-k}^{i+k} = (u_{-k}^{-1}, u, u_1^k)\}|$$

for $u \in \mathcal{Z}$ as an (unnormalized) estimate of the distribution of the noisy sequence to compute the Bayes response

$$\hat{X}^{(k)}_{\text{DUDE},i}(z^n) = \Phi\left(\Pi, \Lambda, m_k\left(z^n, z_{i-k}^{i-1}, z_{i+1}^{i+k}\right), z_i\right).$$

The DUDE is a *sliding-window* denoiser of order $2k+1$, that is, if $z_{i-k}^{i+k} = z_{j-k}^{j+k}$, then $\hat{X}_{\text{DUDE},i}(z^n) = \hat{X}_{\text{DUDE},j}(z^n)$.

According to Theorem 2, the probability assignment induced by a universal lossless compression code is also a good approximation of the distribution of a sequence in the Kullback-Leiber divergence sense. Thus, for any universal lossless compression code employed on the noisy process with length function l_n, we consider the *universal lossless compression-based denoiser*

$$\hat{X}_{\text{cmpr},i}(z^n) = \Phi(\Pi, \Lambda, Q_{Z_i|z^{n\backslash i}}, z_i), \quad (3)$$

where $Q_{Z_i|z^{n\backslash i}}$ is computed based on $Q_{Z^n}(z^n)$ as given in (2). In general, a universal lossless compression-based denoisers is not a sliding-window denoiser and may have high computational complexity. For example, if the Lempel-Ziv compression in [12] is used, the best known computation complexity is $\Theta(n^2)$: at each location i, computing $l_n(z^{i-1}zz_{i+1}^n)$ for all $z \in \mathcal{Z}$ requires incremental parsing of the $(n-i+1)$-block sequence (z, z_{i+1}^n). For context-tree weighting mothods, there is also no known efficient algorithm to update the probability estimate for the sequence (z^{i-1}, z, z_{i+1}^n) from the estimate for the sequence z^n, cf. discussions in [8], [10].

Motivated by the complexity improvement of the Yang-Kieffer lossy compression code [13] in [14], [15] by replacing the Lempel-Ziv length function with the empirical conditional entropy, we consider the following denoiser.

Definition 3: The *empirical conditional entropy of order k* of the *n*-block sequence z^n is

$$H_k(z^n) = \frac{1}{n-k}\sum_{u^k\in\mathcal{Z}^k}\|c_k(z^n,u^k)\|_1 H(c_k(z^n,u^k)) \qquad (4)$$

where

$$c_k(z^n,u^k)[u] = |\{k+1\le i\le n : z_{i-k}^{i-1}=u^k,\ z_i=u\}|, \text{ and}$$

$$H(c_k(z^n,u^k)) = \sum_{u\in\mathcal{Z}} \frac{c_k(z^n,u^k)[u]}{\|c_k(z^n,u^k)\|_1}\log\frac{\|c_k(z^n,u^k)\|_1}{c_k(z^n,u^k)[u]}.$$

The *empirical conditional entropy-based denoiser of order k*, denoted as $\hat{X}^{(k),n}_{\text{emp}}$, is defined as

$$\hat{X}^{(k)}_{\text{emp},i}(z^n) = \Phi(\Pi,\Lambda,Q^{(k)}_{Z_i|z^{n\setminus i}}, z_i), \qquad (5)$$

where

$$Q^{(k)}_{Z_i|z^{n\setminus i}}(z_i) = \frac{2^{-nH_k(z^n)}}{\sum_{\tilde{z}\in\mathcal{Z}}2^{-nH_k(z^{i-1}\tilde{z}z^n_{i+1})}}. \qquad (6)$$

Note that the conditional probability in (6) is induced by the probabiilty

$$Q^{(k)}_{Z^n}(z^n) = \frac{2^{-l_n^{(k)}(z^n)}}{\sum_{\tilde{z}^n\in\mathcal{Z}^n}2^{-l_n^{(k)}(\tilde{z}^n)}}, \qquad (7)$$

which, in turn, is induced by the bona fide length function

$$l_n^{(k)}(z^n) = nH_k(z^n) + |\mathcal{Z}|^{k+1}\log n.$$

In the following theorem, we show that the empirical conditional entropy induces a universal probability assignment provided context length k is increased sufficiently slowly with data length n.

Theorem 3: The probability assignment $Q = \{\{Q^{(k)}_{Z_i|z^{i-1}}\}_{z^{i-1}}\}_{i\ge1}$ given by (6) is universal provided $k=k_n$ such that $k_n\to\infty$ as $n\to\infty$ and $k_n\le c\log_{|\mathcal{Z}|} n$ for $c<1$.

Proof: For any stationary process $\mathbf{Z}\sim P$,

$$\limsup_{k\to\infty} \mathrm{E}\, H_k(Z^n) \le \limsup_{k\to\infty} H(Z_0|Z_{-k}^{-1}) = \mathbb{H}(\mathbf{Z}).$$

Now we consider

$$\begin{aligned}
0 &\le \frac{1}{n}D(P_{Z^n}\|Q^{(k_n)}_{Z^n}) = \sum_{z^n\in\mathcal{Z}^n} P_{Z^n}(z^n)\log\frac{P_{Z^n}(z^n)}{Q^{(k)}_{Z^n}(z^n)}\\
&\le \frac{1}{n}\sum_{z^n\in\mathcal{Z}^n}\Big(P_{Z^n}(z^n)\log P_{Z^n}(z^n)\\
&\qquad +nP_{Z^n}(z^n)H_{k_n}(Z^n) + P_{Z^n}(z^n)|\mathcal{Z}|^{k_n+1}\log n\Big)\\
&\le -\frac{1}{n}H(Z^n) + \mathrm{E}\,H_{k_n}(Z^n) + |\mathcal{Z}|\frac{n^c\log n}{n},
\end{aligned}$$

where the second inequality follows by the Kraft's inequality, and the last inequality follows by assumptions. Taking $n\to\infty$, we obtain

$$0\le \lim_{n\to\infty}\frac{1}{n}D(P_{Z^n}\|Q^{(k_n)}_{Z^n}) \le -\mathbb{H}(\mathbf{Z})+\mathbb{H}(\mathbf{Z}) = 0.$$

Thus, the probability assignment Q is universal. ■

The empirical conditional entropy-based denoising process consists of two passes. In the first pass, at location $k+1\le i\le n$, the context count $c_k(z^n,z_{i-k}^{i-1})[z_i]$ and the associated term in empirical conditional entropy $H_k(z^n)$ in (4) are updated. At the end of the first pass, we obtain the empirical context counts $c_k(z^n,u^k)[u_{k+1}]$ for all $u^{k+1}\in\mathcal{Z}^{k+1}$ and the empirical conditional entropy $H_k(z^n)$. In order to compute $Q^{(k)}_{Z_i|z^{n\setminus i}}(z_k)$ in (6), we need to compute $H_k(z^{i-1}\tilde{z}z^n_{i+1})$ for all $\tilde{z}$ in the second pass. At location $k+1\le i\le n$, we first remove context counts associated with $(k+1)$-block sequences $z^i_{i-k}, z^{i+1}_{i-k+1},\ldots,z^{i+k}_i$ and update the corresponding terms in $H_k(z^n)$ in (4). Then we flip z_i to $\tilde{z}$ and denote the new sequence as $\tilde{z}^n$. Now we add context counts associated with these new $(k+1)$-block sequences $\tilde{z}^i_{i-k}, \tilde{z}^{i+1}_{i-k+1},\ldots,\tilde{z}^{i+k}_i$ and update the empirical conditional entropy to obtain $H_k(z^{i-1}\tilde{z}z^n_{i+1})$. We repeat this process for all $\tilde{z}\in\mathcal{Z}\setminus\{z_i\}$ and compute the posterior distribution in (6) and the estimate in (5). Note that if $z^{i+k}_{i-k}=z^{j+k}_{j-k}$, then the second pass processes at locations i and j are identical and thus $\hat{x}_i=\hat{x}_j$. Therefore, a empirical conditional entropy-based denoiser is a sliding-window denoiser of order $2k+1$. Furthermore, the computational complexity of the empirical conditional entropy-based denoiser is $O(kn)$.

IV. Context Length Selection

The empirical conditional entropy induces an universal probability assignment if context length k grows with data length n at the rate given in Theorem 3. For particular finite source and noisy sequences x^n and z^n, the optimal context length for the empirical conditional entropy-based denoiser is

$$k^*(x^n,z^n) = \arg\min_k L_{\hat{X}^{n,(k)}_{\text{emp}}}(x^n,z^n).$$

However, the denoiser only observes the noisy sequence and cannot find the optimal context length. We thus consider the heuristic method based on the compressibility of the noisy sequence z^n. For each fixed context length k, we obtain the empirical conditional entropy $H_k(z^n)$ after the first pass of empirical conditional entropy-based denoising. The denoiser then selects context length $\hat{k}$ that minimizes the compression code length of the noisy sequence z^n, that is,

$$\begin{aligned}
\hat{k}(z^n) &= \arg\min_k l_n^{(k)}(z^n)\\
&= \arg\min_k \Big(nH_k(z^n) + |\mathcal{Z}|^{k+1}\log n\Big).
\end{aligned}$$

Since the probability $P_{Z^n}(z^n)$ is essentially given by $2^{-l_n^{(k)}(z^n)}$, the context length $\hat{k}(z^n)$ maximizes the likelihood of the noisy sequence over all possible context lengths. Note that, in [7], the heuristic method for choosing the context length is based on the compressibility of the estimation sequence, instead of the noisy sequence. In the next section, we will show that in our experimental data the context length $\hat{k}$ is close to the optimal (genie-aided) context length k^* and, as a consequence, the denoising performance achieved when using $\hat{k}$ is close to that based on the genie-aided k^*.

V. Simulation Results

We compare the performance of the empirical conditional entropy-based denoiser and the DUDE for two settings, the first-order Markov binary process source with binary symmetric channel (BSC) and the corrupted English texts.

A. Binary sequence denoising

In the binary case, we assume that $\mathcal{X} = \mathcal{Z} = \hat{\mathcal{X}} = \{0, 1\}$ and the source $\mathbf{X}$ is a first-order Markov process with transition matrix

$$\begin{bmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{bmatrix}.$$

The channel is BSC with parameter δ, and the loss function is Hamming loss, that is,

$$\Pi = \begin{bmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{bmatrix} \text{ and } \Lambda = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

It can be shown that

$$\Phi(\Pi, \Lambda, v, z) = \begin{cases} z & \text{if } \dfrac{v(z)}{v(1-z)} \geq \dfrac{2\delta(1-\delta)}{\delta^2 + (1-\delta)^2}, \\ 1-z & \text{otherwise.} \end{cases}$$

In this case, if the denoiser knows the source distribution, the Bayes optimal denoiser can be computed by the forward-backward recursions (see [4]).

We consider denoising an individual sequence of length $n = 10^6$. The context lengths k for the empirical conditional entropy-based denoiser and the DUDE are chosen to minimize the per-symbol loss. Table I shows the simulation results. Since we only consider an individual source sequence rather than averaging the per-symbol loss over multiple source realizations, the Bayes optimal denoiser that minimizes the expected per-symbol loss can have some per-symbol losses greater than the BSC parameter δ, which is the per-symbol loss when $\hat{x}_i = 0$ for all $1 \leq i \leq n$. The performances of the empirical conditional entropy-based denoiser and the DUDE are almost the same as the performance of the Bayes optimal denoiser. In general, the empirical conditional entropy-based algorithm requires longer context length to achieve the performance of the DUDE.

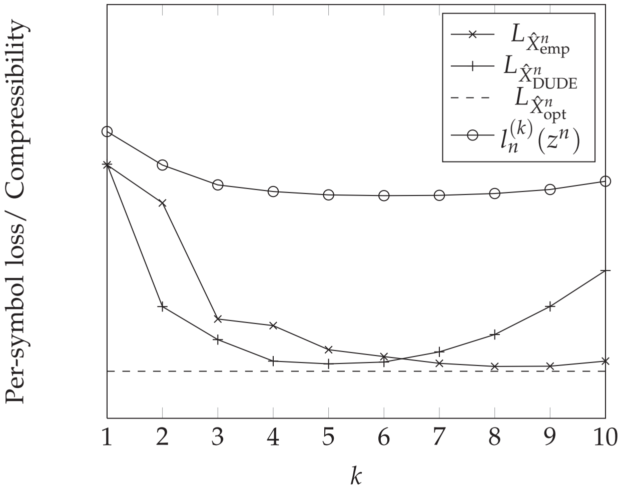

Fig. 2. Per-symbol losses and compressibility under different context lengths for the binary case with parameters $\alpha = 0.01$ and $\delta = 0.1$

In Fig. 2, we compare the per-symbol losses of the empirical conditional entropy and the DUDE and the properly-scaled compression code length of the noisy sequence under different context lengths for $\alpha = 0.01$ and $\delta = 0.1$. We also plot the per-symbol loss of the optimal denoiser as a reference. For the particular source and noisy sequences, the optimal context length for the empirical conditional entropy-based denoiser is $k^* = 8$. If we use the heuristic method in Section IV, we obtain the context length $\hat{k} = 6$.

B. Text Denoising

The English texts are obtained from Project Gutenberg (http://www.gutenberg.org). We remove punctuations and convert uppercase letters into lowercase letters. Thus, the source alphabet $\mathcal{X} = \{\text{'a'}, \text{'b'}, \ldots, \text{'z'}, \text{space}\}$. The text is corrupted by the QWERTY keyboard channel. The non-space letter is corrupted with probability δ. If a letter is corrupted, it is flipped to one of its neighbor uniformly at random. For example, the letter 'w' has 3 neighbors 'e', 'q', and 's'. Thus,

$$\Pi(\text{'w'}, z) = \begin{cases} 1-\delta & \text{if } z = \text{'w'}, \\ \delta/3 & \text{if } z = \text{'q', 'e', 's'}, \\ 0 & \text{otherwise.} \end{cases}$$

The loss function is Hamming loss, that is,

$$\Lambda(x, \hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x}, \\ 1 & \text{otherwise.} \end{cases}$$

In this case, we cannot compute the Bayes optimal denoiser in (1) since the source statistics are unknown. We only compare the empirical conditional entropy-based denoiser with the DUDE. Since for context lengths greater than 3, the counts of most contexts are zero, the context length k for both denoisers are optimized over

TABLE I
Per-symbol loss for denoising a binary sequence with optimal context length shown in the brackets

	$\delta = 0.01$			$\delta = 0.10$		
α	emp	DUDE	opt	emp	DUDE	opt
0.01	0.0714 δ [5]	0.0683 δ [2]	0.0681 δ	0.0627 δ [8]	0.0663 δ [5]	0.0567 δ
0.05	0.4313 δ [3]	0.4324 δ [3]	0.4313 δ	0.2971 δ [6]	0.3000 δ [5]	0.2953 δ
0.10	1.0202 δ [5]	1.0203 δ [2]	1.0203 δ	0.5577 δ [4]	0.5566 δ [4]	0.5568 δ
0.15	1.0076 δ [6]	1.0076 δ [3]	1.0076 δ	0.7518 δ [5]	0.7521 δ [3]	0.7521 δ
0.20	0.9972 δ [7]	0.9971 δ [5]	0.9972 δ	0.9253 δ [4]	0.9254 δ [3]	0.9254 δ

TABLE II
Per-symbol loss for denoising corrupted texts

length	errors	emp	DUDE
992050	40235	20169	25394
2230536	90325	41593	44457
3099626	126763	56422	61651

1, 2, and 3. Table II shows the simulation results for corruption probability $\delta = 0.05$. The empirical conditional entropy-based denoiser corrects more errors than the DUDE does. This can be explained by the following observation. We consider context length $k = 2$ for both denoisers. At location i, the estimate $\hat{X}_{\text{DUDE},i}(z^n)$ depends on the vector $m(z^n, z_{i-k}^{i-k}, z_{i+1}^{i+k})$ and the symbol z_i, while the estimate $\hat{X}_{\text{emp},i}(z^n)$ depends on vectors $c_k(z^n, z_{i-k}^{i-1}), \ldots, c_k(z^n, z_i^{i+k-1})$ and the symbol z_i. Since

$$c_k(z^n, z_{i-k}^{i-1})[z_i] = \sum_{u^k \in \mathcal{Z}^k} m(z^n, z_{i-k}^{i-1}, u^k)[z_i],$$

$m(z^n, z_{i-k}^{i-1}, z_{i+1}^{i+k})[z_i]$ may be much smaller than $c_k(z^n, z_{i-k}^{i-1})[z_i]$ when the alphabet size $|\mathcal{Z}|$ or the context length k is large. However, the empirical distribution is close to the distribution of the noisy sequence only when the number of counts are large enough. In this case of denoising corrupted texts, the alphabet size is 27, which is much larger than 2, the alphabet size of the binary case. Thus, the performance of DUDE is worse than the performance of the empirical conditional entropy-based denoiser.

VI. Conclusions

It has been shown in the literature that a universal lossless compression code induces a universal probability assignment, and a universal probability assignment directly yields an asymptotically optimal universal predictor as well as a universal filter (causal denoiser). In this work, we use the universal probability assignment to approximate the posterior distribution of the clean source given the noisy sequence, and we present the universal lossless compression-based denoiser. We then specialize and take a close look at this approach when employing the empirical conditional entropy-based denoiser, which is a sliding-window denoiser and has lower computational complexity than the denoiser that would be induced from Lempel-Ziv or context tree weighting compression. The simulation results show that the empirical conditional entropy-based denoiser has the same performance as the DUDE for denoising binary sequences corrupted by binary symmetric channels. In correction of corrupted English texts, the sizes of the source and reconstruction alphabets are large, and the empirical conditional entropy-based denoiser corrects more errors than the DUDE does.

We are currently working on comparing the convergence rate of the empirical conditional entropy-based denoiser and the DUDE, and on proving that universal lossless compression-based denoisers are universal, i.e., achieve the optimum (Bayes) denoising performance on any stationary source.

References

[1] N. Wiener, *The Extrapolation, Interpolation and Smoothing of Stationary Time Series*. Wiley, 1949.

[2] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME: Journal of Basic Engineering*, pp. 35–45, Mar. 1960.

[3] D. Donoho, "De-noising by soft-thresholding," *Information Theory, IEEE Transactions on*, vol. 41, no. 3, pp. 613–627, May 1995.

[4] Y. Ephraim and N. Merhav, "Hidden markov processes," *Information Theory, IEEE Transactions on*, vol. 48, no. 6, pp. 1518–1569, Jun 2002.

[5] B. Y. Ryabko, "Prediction of random sequences and universal coding," *Probl. Peredachi Inf.*, pp. 3–14, 1988.

[6] T. Weissman, E. Ordentlich, M. Weinberger, A. Somekh-Baruch, and N. Merhav, "Universal filtering via prediction," *Information Theory, IEEE Transactions on*, vol. 53, no. 4, pp. 1253–1264, April 2007.

[7] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. Weinberger, "Universal discrete denoising: known channel," *Information Theory, IEEE Transactions on*, vol. 51, no. 1, pp. 5–28, Jan. 2005.

[8] E. Ordentlich, M. Weinberger, and T. Weissman, "Multi-directional context sets with applications to universal denoising and compression," in *Information Theory, 2005. ISIT 2005. Proceedings. International Symposium on*, Sept. 2005, pp. 1270–1274.

[9] F. Willems, Y. Shtarkov, and T. Tjalkens, "The context-tree weighting method: basic properties," *Information Theory, IEEE Transactions on*, vol. 41, no. 3, pp. 653–664, May 1995.

[10] J. Yu and S. Verdu, "Schemes for bidirectional modeling of discrete stationary sources," *Information Theory, IEEE Transactions on*, vol. 52, no. 11, pp. 4789–4807, Nov. 2006.

[11] B. Ryabko, "Compression-based methods for nonparametric density estimation, on-line prediction, regression and classification for time series," in *Information Theory Workshop, 2008. ITW '08. IEEE*, May 2008, pp. 271–275.

[12] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *Information Theory, IEEE Transactions on*, vol. 24, no. 5, pp. 530–536, Sep 1978.

[13] E. hui Yang and J. Kieffer, "Simple universal lossy data compression schemes derived from the lempel-ziv algorithm," *Information Theory, IEEE Transactions on*, vol. 42, no. 1, pp. 239–245, Jan 1996.

[14] S. Jalali and T. Weissman, "Rate-distortion via markov chain monte carlo," in *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, July 2008, pp. 852–856.

[15] S. Jalali, A. Montanari, and T. Weissman, "An implementable scheme for universal lossy compression of discrete markov sources," *Data Compression Conference*, vol. 0, pp. 292–301, 2009.