Near Optimal Lossy Source Coding and Compression-Based Denoising via Markov Chain Monte Carlo

Shirin Jalali* and Tsachy Weissman*†,
*Department of Electrical Engineering, Stanford University, Stanford, CA 94305, {shjalali, tsachy}@stanford.edu
† Department of Electrical Engineering, Technion, Haifa 32000, Israel

***Abstract*— We propose an implementable new universal lossy source coding algorithm. The new algorithm utilizes two well-known tools from statistical physics and computer science: Gibbs sampling and simulated annealing. In order to code a source sequence x^n, the encoder initializes the reconstruction block as $\hat{x}^n = x^n$, and then at each iteration uniformly at random chooses one of the symbols of $\hat{x}^n$, and updates it. This updating is based on some conditional probability distribution which depends on a parameter β representing inverse temperature, an integer parameter $k = o(\log n)$ representing context length, and the original source sequence. At the end of this process, the encoder outputs the Lempel-Ziv description of $\hat{x}^n$, which the decoder deciphers perfectly, and sets as its reconstruction. The complexity of the proposed algorithm in each iteration is linear in k and independent of n. We prove that, for any stationary ergodic source, the algorithm achieves the optimal rate-distortion performance asymptotically in the limits of large number of iterations, β, and n. We also show how our approach carries over to such problems as universal Wyner-Ziv coding and compression-based denoising.**

I. INTRODUCTION

Consider the basic setup of lossy coding of a stationary ergodic source $\mathbf{X} = \{X_i : i > 1\}$. Each source output block of length n, X^n, is mapped to an index $f(X^n) \in \{1, 2, \ldots, nR\}$. The index is sent to the decoder which decodes it to a reconstruction block $\hat{X}^n = g(f(X^n))$. The performance of a coding scheme $\mathcal{C} = (f, g, n, R)$ is measured by its average expected distortion between source and reconstruction blocks, i.e.

$$D = Ed_n(X^n, \hat{X}^n) \triangleq \frac{1}{n}\sum_{i=1}^{n} Ed(X_i, \hat{X}_i), \qquad (1)$$

where $d : \mathcal{X} \times \mathcal{X} \to R^+$ is a single-letter distortion measure. Here $\mathcal{X}$ and $\hat{\mathcal{X}}$ denote the source and reconstruction alphabets respectively. For any $D \geq 0$, the minimum rate that is achievable (cf. [2] for exact definition of achievability) is characterized as [1]

$$R(\mathbf{X}, D) = \lim_{n\to\infty} \min_{p(\hat{X}^n|X^n): Ed_n(X^n,\hat{X}^n)\leq D} \frac{1}{n} I(X^n; \hat{X}^n). \qquad (2)$$

For the case of lossless compression, i.e. $D = 0$ (assuming a non-degenerate distortion measure), we know that the minimum required rate is the entropy rate of the source, i.e. $R(\mathbf{X}, 0) = H(\mathbf{X}) \triangleq \lim_{k\to\infty} H(X_0|X_{-k}^{-1})$. Moreover, there are known implementable *universal* schemes, such as Lempel-Ziv coding [3] and arithmetic coding [4], that are able to describe any stationary ergodic source at rates as close as desired to the entropy rate of the source without any error. In contrast to the situation for lossless compression, for $D > 0$, neither the explicit solution of (2) is known for a general source (even not for a first-order Markov source), nor are there known practical schemes that universally achieve the rate-distortion curve. In recent years, there has been considerable progress towards designing universal lossy compressor, especially in trying to tune one of the existing universal lossless coders to work in the lossy case as well [5], [6], [7]. In [5], a lossy version of Lempel-Ziv algorithm at fixed distortion is rendered, and is shown to be optimal for memoryless sources. On the other hand, for the non-universal setting, specifically the case of lossy compression of an i.i.d. source with a known distribution, there is an ongoing progress towards designing codes that get very close to the optimal performance [14], [15], [16], [17].

In this paper, we present a new *universal* lossy source coding algorithm that, in addition to being asymptotically *optimal* is also implementable. Our algorithm borrows two well-known tools from statistical physics and computer science, namely Markov Chain Monte Carlo (MCMC) methods, and simulated annealing [10],[11]. MCMC methods refer to a class of algorithms that are designed to generate samples of a given distribution through generating a Markov chain having the desired distribution as its stationary distribution. MCMC methods include a large number of algorithms; For our application, we use Gibbs sampler [9] also known as the *heat bath* algorithm, which is well-suited to the case where the desired distribution is hard to compute, but the conditional distributions of each variable given the rest are easy to work out.

The second required tool is simulated annealing which is a well-known method in statistical physics. Its goal is to find the minimum of a function $f_{\min} \triangleq \min f(s)$ along with the minimizing state $s_{\min}$ over a set of possibly huge number of states S. In order to do simulated annealing, a sequence of probability distributions $p_1, p_2, \ldots$ corresponding to the temperatures $T_1 > T_2 > \ldots$, where $T_i \to 0$ as $i \to \infty$, and a sequence of positive integers N_1, N_2, ..., are considered. For the N_1 first steps, the algorithm runs one of the relevant MCMC methods in an attempt to sample from

978-1-4244-2247-0/08/$25.00 ©2008 IEEE.

distribution p_1. Then, for the N_2 next steps, the algorithm, using the output of the previous part as the initial point, aims to sample from p_2, and so on. The probability distributions are designed such that: 1) their output, with high probability, is the minimizing state $s_{\min}$, or one of the states close to it, 2) the probability of getting the minimizing state increases as the temperature drops. The probability distribution that satisfies these characteristics, and is almost always used, is the Boltzman distribution $p_\beta(s) \propto e^{-\beta f(s)}$, where $\beta \propto \frac{1}{T}$. It can be proved that using Boltzman distribution, if the temperature drops slowly enough, the probability of ultimately getting the minimizing state as the output of the algorithm approaches one [9]. Simulated annealing has been suggested before in the context of lossy compression, either as a way for approximating the rate distortion function (i.e., the optimization problem involving minimization of the mutual information) or as a method for designing the codebook in vector quantization [12],[13], as an alternative to the conventional generalized Lloyd algorithm (GLA) [8]. In contrast, in this paper we use the simulated annealing approach to obtain a particular reconstruction sequence, rather than a whole codebook.

Let us briefly describe how the new algorithm codes a source sequence x^n. First, to each reconstruction block y^n, it assigns an *energy*, $E(y^n)$, which is a linear combination of its conditional empirical entropy, to be defined formally in the next section, and its distance from the source sequence x^n. Then, it assumes a Boltzman probability distribution over the reconstruction blocks as $p(y^n) \propto e^{-\beta E(y^n)}$, for some $\beta > 0$, and tries to generate $\hat{x}^n$ from this distribution using Gibbs sampling [9]. As we will show, for β large enough, with high probability the reconstruction block of our algorithm would satisfy $E(\hat{x}^n) \approx \min E(y^n)$. The encoder will output $LZ(\hat{x}^n)$, which is the Lempel-Ziv [3] description of $\hat{x}^n$. The decoder, upon receiving $LZ(\hat{x}^n)$, reconstructs $\hat{x}^n$ perfectly.

In this paper, instead of working at a fixed rate or at a fixed distortion, we are fixing the slope. A fixed slope rate-distortion scheme, for a fixed slope $s < 0$, looks for the coding scheme that minimizes $R - s \cdot D$, where as usual R and D denote the rate and the average expected distortion respectively. In comparison to a given coding scheme of rate R and expected distortion D, for any $0 < \delta < R - R(\mathbf{X}, D)$, there exists a code which works at rate $R(\mathbf{X}, D) + \delta$ and has the same average expected distortion, and consequently a lower cost. Therefore, it follows that any point that is optimal in the fixed-slope setup corresponds to a point on the rate-distortion curve.

The organization of the paper is as follows. In Section II we set up the notation, and Section III describes the count matrix and empirical conditional entropy of a sequence. Section IV describes an exhaustive search scheme for fixed-slope lossy compression which universally achieves the rate-distortion curve for any stationary ergodic source. Section V described our new universal MCMC-based lossy coder. Section VI describes the application of the method used in Section V to universal Wyner-Ziv coding, and Section VII puts forward its application to universal compression-based denoising. Section VIII concludes the paper and discusses some future directions. The proofs of all stated results will appear in the full version of the paper.

II. Notation

Let $\mathcal{X}$ and $\hat{\mathcal{X}}$ denote the source and reconstructed signals alphabets respectively. In the section on WZ coding, let $\mathcal{Z}$ denote the noisy channel output alphabet. In this paper, for simplicity, we restrict attention to the case where $\mathcal{X} = \hat{\mathcal{X}} = \mathcal{Z} = \{\alpha_1, \ldots, \alpha_N\}$, though our derivations and results carry over directly to general finite alphabets. Bold low case symbols, e.g. $\mathbf{x}, \mathbf{y}, \mathbf{z}$, denote individual sequences. The discrete memoryless channel is described by its transition matrix $\mathbf{\Pi}$, where $\mathbf{\Pi}(i,j)$ denotes the probability of getting α_j at the output of the channel when the input is α_i.

Let $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$ be the loss function (fidelity criterion) which measures the loss incurred in denoising (decoding) a symbol α_i to another symbol α_j, which will be represented by a $N \times N$ matrix, $\mathbf{\Lambda} : \{d(\alpha_i, \alpha_j)\}$. Moreover, let $d_m = \max_{i,j} d(\alpha_i, \alpha_j)$, and note that $d_m < \infty$, since the alphabets are finite. The normalized cumulative loss between a source sequence x^n and reconstructed sequence $\hat{x}^n$, is denoted by $d_n(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i)$.

Let π_i and $\mathbf{d}_j$ denote the i-th column and the j-th column of $\mathbf{\Pi}$ and $\mathbf{\Lambda}$ matrices respectively, i.e. $\mathbf{\Pi} = [\pi_1 | \ldots | \pi_N]$ and $\mathbf{\Lambda} = [\mathbf{d}_1 | \ldots | \mathbf{d}_N]$.

For two N-dimensional vectors $\mathbf{u}$ and $\mathbf{v}$, $\mathbf{u} \odot \mathbf{v}$ denotes the N-dimensional vector that results from componentwise multiplication of $\mathbf{u}$ and $\mathbf{v}$, i.e. $\mathbf{u} \odot \mathbf{v}[i] = u_i v_i$.

III. Counts and Empirical Conditional Entropy

Let $H_k(y^n)$ denote the conditional empirical entropy of order k induced by y^n, i.e.

$$H_k(y^n) = H(Y_{k+1} | Y^k), \tag{3}$$

where Y^{k+1} on the right hand side of (3) is distributed according to

$$P(Y^{k+1} = u^{k+1}) = \frac{1}{n} \left| \left\{ 1 \le i \le n : y_{i-k}^i = u^{k+1} \right\} \right|, \tag{4}$$

where in (4), and throughout we assume a cyclic convention whereby $y_i \triangleq y_{n+i}$ for $i \le 0$. We introduce the count notation $\mathbf{m}_k(y^n, u^k)$, which is a column vector counting the number of appearances of the different symbols in y^n with the left context u^k. More explicitly, $\mathbf{m}_k(y^n, u^k)$ is a column vector whose y-th component is given by

$$\mathbf{m}_k(y^n, u^k)[y] = \left| \left\{ 1 \le i \le n : y_{i-k}^i = u^k y \right\} \right|, \tag{5}$$

where $u^k y$ denotes the $k+1$-tuple obtained by concatenating u^k with the symbol y. We let $\mathbf{m}_k(y^n, \cdot)$ denote the $|\mathcal{Y}| \times |\mathcal{Y}|^k$ matrix whose columns are given by $\mathbf{m}_k(y^n, u^k)$, for the $|\mathcal{Y}|^k$ values of u^k lexicographically ordered. Note that, with the count notation, the conditional empirical entropy in (3) can be expressed as

$$H_k(y^n) = \frac{1}{n} \sum_{u^k} \mathcal{H}\left(\mathbf{m}_k(y^n, u^k)\right) \mathbf{1}^T \mathbf{m}_k(y^n, u^k), \tag{6}$$

where $\mathbf{1}$ denotes the all-ones column vector of length $|\mathcal{Y}|$, and for a vector $v = (v_1, \ldots, v_\ell)^T$ with non-negative components, we let $\mathcal{H}(v)$ denote the entropy of the random variable whose probability mass function (PMF) is proportional to v. Formally,

$$\mathcal{H}(v) = \begin{cases} \sum_{i=1}^{\ell} \frac{v_i}{\|v\|_1} \log \frac{\|v\|_1}{v_i} & \text{if } v \neq (0,\ldots,0)^T \\ 0 & \text{if } v = (0,\ldots,0)^T. \end{cases} \quad (7)$$

IV. An exhaustive search scheme for fixed-slope compression

Consider the following scheme for lossy source coding at fixed slope $s \leq 0$. For each source sequence x^n let the reconstruction block $\hat{x}^n$ be

$$\hat{x}^n = \arg\min_{y^n} \left[H_k(y^n) - s \cdot d(x^n, y^n) \right]. \quad (8)$$

The encoder, after computing $\hat{x}^n$, losslessly conveys it to the decoder using LZ compression. Let k grow slowly enough with n so that

$$\limsup_{n\to\infty} \max_{y^n} \left[\frac{1}{n} \ell_{LZ}(y^n) - H_k(y^n) \right] \leq 0, \quad (9)$$

where $\ell_{LZ}(y^n)$ denotes the length of the LZ representation of y^n. Note that Ziv's inequality guarantees that if $k = k_n = o(\log n)$ then (9) holds. We can prove the following:

Theorem 1: Let $\mathbf{X}$ be a stationary and ergodic source, let $R(\mathbf{X}, D)$ denote its rate distortion function, and let $\hat{X}^n$ denote the reconstruction using the above scheme on X^n. Then

$$E\left[\frac{1}{n} \ell_{LZ}(\hat{X}^n) - s \cdot d(X^n, \hat{X}^n) \right] \overset{n\to\infty}{\longrightarrow} \min_{D \geq 0} \left[R(\mathbf{X}, D) - s \cdot D \right]. \quad (10)$$

V. Universal lossy coding via MCMC

In this section, we will show how simulated annealing Gibbs sampling enables us to get close to the performance of the impractical exhaustive search coding algorithm described in the previous section. Throughout this section we fix the slope $s \leq 0$.

Associate with each reconstruction sequence y^n the *energy*

$$E(y^n) \triangleq n \left[H_k(y^n) - s \cdot d(x^n, y^n) \right] = \sum_{u^k} \mathcal{H}\left(\mathbf{m}_k(y^n, u^k) \right) \mathbf{1}^T \mathbf{m}_k(y^n, u^k) - s \cdot \sum_{i=1}^{n} d(x_i, y_i).$$

The *Boltzmann distribution* can now be defined as the PMF on $\mathcal{Y}^n$ given by

$$p_\beta(y^n) = \frac{1}{Z(\beta)} \exp\{-\beta E(y^n)\}, \quad (11)$$

where $Z(\beta)$ is the normalization constant (partition function). Note that, though this dependence is suppressed in the notation for simplicity, $E(y^n)$, and therefore also p_β and $Z(\beta)$ depend on x^n and s, which are fixed until further notice. When β is large and $Y^n \sim p_\beta$, then with high probability

$$H_k(Y^n) - s \cdot d(x^n, Y^n) \approx \min_{y^n} \left[H_k(y^n) - s \cdot d(x^n, y^n) \right]. \quad (12)$$

Thus, using a sample from the Boltzmann distribution p_β, for large β, as the reconstruction sequence, would yield performance close to that of an exhaustive search scheme that would use the achiever of the minimum in (12). Unfortunately, it is hard to sample from the Boltzmann distribution directly, since the partition function is hard to evaluate. We can, however, get approximate samples via MCMC, as we describe next.

As mentioned earlier, Gibbs sampler [9] is useful in cases where one is interested in sampling from a probability distribution which is hard to compute, but the conditional distribution of each variable given the rest can be calculated more easily. In our case, since the partition function is not accessible, $p_\beta(y^n)$ can not be computed directly, but the conditional probability under p_β of Y_i given the other variables $Y^{n\setminus i} \triangleq \{Y_n : n \neq i\}$ can be shown, following straightforward arithmetic, to be expressible as

$$p_\beta(Y_i = y_i | Y^{n\setminus i} = y^{n\setminus i}) = \frac{1}{\sum_y \exp\{-\beta \left[n \Delta H_k(y^{i-1} y y_{i+1}^n, y_i) - s \cdot \Delta d(y, y_i, x_i) \right]\}}, \quad (13)$$

where $\Delta H_k(y^{i-1} y y_{i+1}^n, y_i)$ and $\Delta d(y^{i-1} y y_{i+1}^n, y_i, x_i)$ are defined to be $\Delta H_k(y^{i-1} y y_{i+1}^n, y_i) \triangleq H_k(y^{i-1} y y_{i+1}^n) - H_k(y^n)$, and $\Delta d(y, y_i, x_i) \triangleq d(y, x_i) - d(y_i, x_i)$, respectively. Evidently, $p_\beta(Y_i = y_i | Y^{n\setminus i} = y^{n\setminus i})$ depends on y^n only through $\{H_k(y^{i-1} y y_{i+1}^n) - H_k(y^n)\}_{y\in\mathcal{Y}}$ and $\{d(x_i, y)\}_{y\in\mathcal{Y}}$. In turn, $\{H_k(y^{i-1} y y_{i+1}^n) - H_k(y^n)\}_{y\in\mathcal{Y}}$ depends on y^n only through $\{\mathbf{m}_k(y^{i-1} y y_{i+1}^n, \cdot)\}_{y\in\mathcal{Y}}$.

Note that, given $\mathbf{m}_k(y^n, \cdot)$, the number of operations required to obtain $\mathbf{m}_k(y^{i-1} y y_{i+1}^n, \cdot)$, for any $y \in \mathcal{Y}$ is linear in k, since the number of contexts whose counts are affected by a change of one component in y^n is no larger than $2k+2$. I.e., letting $\mathcal{S}_i(y^n, y)$ denote the set of contexts whose counts are affected when the ith component of y^n is flipped from y_i to y, we have $|\mathcal{S}_i(y^n, y)| \leq 2k+2$. Further, since

$$n[H_k(y^{i-1} y y_{i+1}^n) - H_k(y^n)] = \sum_{u^k \in \mathcal{S}_i(y^n, y)} \mathcal{H}\left(\mathbf{m}_k(y^{i-1} y y_{i+1}^n, u^k) \right) \mathbf{1}^T \mathbf{m}_k(y^{i-1} y y_{i+1}^n, u^k) - \mathcal{H}\left(\mathbf{m}_k(y^n, u^k) \right) \mathbf{1}^T \mathbf{m}_k(y^n, u^k), \quad (14)$$

it follows that, given $\mathbf{m}_k(y^n, \cdot)$ and $H_k(y^n)$, the number of operations required to compute $\mathbf{m}_k(y^{i-1} y y_{i+1}^n, \cdot)$ and $H_k(y^{i-1} y y_{i+1}^n)$ is linear in k (and independent of n).

Now consider the following algorithm based on Gibbs sampling method for generating samples from p_β, and let $\hat{X}_{s,\beta,r}^n(X^n)$ denote its (random) outcome when taking $k = k_n$ to be a deterministic sequence satisfying $k_n = o(\log n)$, applied to the source sequence X^n as input.[1]

Theorem 2: Let $\mathbf{X}$ be a stationary and ergodic source. Then

$$\lim_{n\to\infty} \lim_{\beta\to\infty} \lim_{r\to\infty} E\left[\frac{1}{n} \ell_{LZ}\left(\hat{X}_{s,\beta,r}^n(X^n) \right) - s \cdot d(X^n, \hat{X}^n) \right] = \min_{D\geq 0} \left[R(\mathbf{X}, D) - s \cdot D \right]. \quad (15)$$

[1]Here and throughout it is implicit that the randomness used in the algorithms is independent of the source.

Algorithm: Universal lossy coder based on Gibbs sampler
Input: x^n, k, s, β, r
Output: a reconstruction sequence $\hat{x}^n$

1) Set $y^n = x^n$
2) Run one pass over y^n to compute $\mathbf{m}_k(y^n, \cdot)$ and $H_k(y^n)$
3) For $t = 1$ to $t = r$:
   a) Draw an integer $i \in \{1, \ldots, n\}$ uniformly at random
   b) For each $y \in \mathcal{Y}$ compute $p_\beta(Y_i = y | Y^{n\setminus i} = y^{(t-1),n\setminus i})$ given in (13)
   c) Update y^n by replacing its ith component y_i by y drawn from the PMF computed in Part (b)
   d) Update the values of $\mathbf{m}_k(y^n, \cdot)$ and $H_k(y^n)$
4) Set $\hat{x}^n = y^n$

VI. UNIVERSAL WYNER-ZIV CODING

Consider the basic setup shown in Fig. 1, where a source block x^n with unknown statistics drives a known discrete memoryless channel (DMC), and a decoder which receives the noisy channel output Z^n in addition to a fidelity-boosting (FB) sequence y^n which is describable to the decoder losslessly with no more than nR bits. The goal is to minimize the distortion between the source and the reconstructed signals by optimally designing the encoder and decoder. This is a non-conventional viewpoint of the rate-distortion coding with decoder side information commonly known as Wyner-Ziv (WZ) compression after the seminal paper [18]. Adopting this point of view, a new provably asymptotically optimal universal WZ coder is proposed in [22]. As mentioned in [22], the problematic part of this algorithm is the encoder which requires an exhaustive search over the space of all possible FB sequences with Lempel-Ziv (LZ) description length less than nR. The size of this set increases exponentially with the block length. In this section, we extend the ideas of the previous sections to the WZ coding setting, and put forward an implementable WZ encoder to be replaced by the encoder proposed in [22] without compromising the optimality or universality of the results.

A. An Exhaustive Search Scheme for Fixed-Slope WZ Compression

Consider the following scheme for WZ coding at fixed slope $s \leq 0$. Given x^n, the encoder constructs the FB sequence Y^n as follows

$$\begin{aligned} Y^n &= Y^n(x^n, \tilde{Z}^n) \\ &= \arg\min_{y^n} \left[H_k(y^n) - s \cdot d(x^n, \hat{X}^{n,\ell,m}(y^n, \tilde{Z}^n)) \right], \end{aligned} \quad (16)$$

where $\tilde{Z}^n$ is a simulated output of the channel Π for the input x^n generated by the encoder, and $\hat{X}^{n,\ell,m}(y^n, \tilde{Z}^n)$ is the output of a WZ-DUDE decoder with parameters ℓ and m as defined in (22) [22].

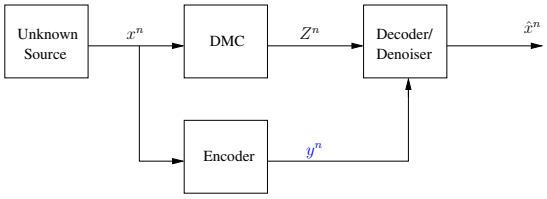

Fig. 1. The basic setup of WZ coding problem

Then the encoder sends the LZ description of Y^n to the decoder. Then the decoder receives the LZ description of Y^n and the output of the DMC channel Z^n. It first reconstructs Y^n, and then employs the WZ-DUDE decoder described in (22) to Y^n and Z^n to reconstruct $\hat{X}^n = \hat{X}^{n,\ell,m}(Y^n, Z^n)$.

Theorem 3: Let $\mathbf{X}$ be a stationary and ergodic source, and $R_\Pi(\mathbf{X}, D)$ denote its associated WZ rate distortion function. Let Y^n and $\hat{X}^n$ refer to the FB and reconstructed signals of the above WZ coding scheme respectively, then

$$\lim_{n\to\infty} \left[LZ(Y^n) - s \cdot d(X^n, \hat{X}^n) \right] = R_{\mathbf{\Pi}}(\mathbf{X}, D) - s \cdot D. \quad (17)$$

In the next section, inspired by the MCMC-based lossy quantizer described earlier, we put forward an implementable MCMC-based WZ encoder to be replaced with the exhaustive search encoder of the WZ coding algorithm just described. We show that asymptotically the new encoder incurs no loss compared to the original one.

B. Count vectors and DUDE operator

For each $u_{-\ell}^{\ell}$ and v_{-m}^{m}, let ϕ be $\mathcal{Z}$-dimensional with u_0-th component defined as

$$\begin{aligned} &\phi_{\ell,m}(z^n, y^n, u_{-\ell}^{-1}, u_1^{\ell}, v_{-m}^{m})[u_0] = \\ &|\{i : z_i = u_0, z_{i-\ell}^{i-1} = u_{-\ell}^{-1}, z_{i+1}^{i+\ell} = u_1^{\ell}, y_{i-m}^{i+m} = u_{-m}^{m}\}|. \end{aligned} \quad (18)$$

Likewise, let

$$\begin{aligned} &\tau_{\ell,m}(z^n, y^n, x^n, u_{-\ell}^{\ell}, v_{-m}^{m})[x] = \\ &\qquad |\{i : x_i = x, z_{i-\ell}^{i+\ell} = u_{-\ell}^{\ell}, y_{i-m}^{i+m} = u_{-m}^{m}\}|. \end{aligned} \quad (19)$$

Note that τ presents a more refined count vector compared to ϕ , and they are related as follows

$$\begin{aligned} &\phi_{\ell,m}(z^n, y^n, u_{-\ell}^{-1}, u_1^{\ell}, v_{-m}^{m})[u_0] = \\ &\qquad \mathbf{1}^T \tau_{\ell,m}(z^n, y^n, x^n, u_{-\ell}^{-1} u_0 u_1^{\ell}, v_{-m}^{m}), \end{aligned} \quad (20)$$

where $\mathbf{1}$ represents a $\mathcal{X}$-dimensional vector of all ones.

For a $\mathcal{Z}$-dimensional vector θ define the DUDE operator as

$$\hat{X}_{\text{DUDE}}(\theta, z) = \arg\min_{\hat{x}} \theta^T \mathbf{\Pi}^{-1} [\mathbf{d}_{\hat{x}} \odot \pi_z]. \quad (21)$$

To interpret the definition of (21), note that when θ corresponds to the count vector obtained by a DUDE denoiser [20] such that for $\alpha \in \mathcal{Z}$, $\theta(\alpha)$ is the number of appearances α within some specific left and right contexts along the noisy signal, then (21) represents the DUDE decision rule for denoising symbol z in the noisy signal having the same left/right contexts.

$$\hat{X}^{n,l,m}(z^n, y^n)[i] = \arg\min_{\hat{x}\in\mathcal{A}} \mathbf{r}^T(z^n, y^n, z_{i-l}^{i-1}, z_{i+1}^{i+l}, y_{i-m}^{i+m})\mathbf{\Pi}^{-1}[d_{\hat{x}} \odot \pi_{z_i}], \tag{22}$$

where, for $\beta \in \mathcal{Z}$, and $t = \max\{l, m\}$

$$\mathbf{r}(z^n, y^n, a^l, b^l, c_{-m}^m)[\beta] = |\left\{t+1 \le i \le n-t : z_{i-l}^{i-1} = a^l, z_{i+1}^{i+l} = b^l, y_{i-m}^{i+m} = c_{-m}^m\right\}| \tag{23}$$

C. MCMC-based WZ encoder

Like in the lossy compression case, in order to solve the minimization required to find the FB sequence by the exhaustive search WZ coder, we resort to simulated annealing [10]. Again we assign some *energy* to each possible FB sequence as follows

$$E(y^n) \triangleq H_k(y^n) - s \cdot d(x^n, \hat{X}^{n,\ell,m}(y^n, \tilde{Z}^n)) \tag{24}$$

$$= \frac{1}{n}\sum_{u^k} \mathcal{H}(m_k(y^n, u^u))\mathbf{1}^T m_k(y^n, u^u) - s \cdot \frac{1}{n} \sum_{u_{-\ell}^{\ell}, v_{-m}^{m}} \tau_{\ell,m}(z^n, y^n, x^n, u_{-\ell}^{\ell}, v_{-m}^{m})^T \mathbf{d}_{\hat{X}_{\mathsf{DUDE}}(\phi_{\ell,m}(z^n, y^n, u_{-\ell}^{-1}, u_1^{\ell}, v_{-m}^{m})), u_0)}, \tag{25}$$

and then assume the Boltzman probability distribution over the space of FB sequences as $p_\beta(y^n) \propto e^{-\beta E(y^n)}$. The advantage of decomposing the energy function as in (25) is that now we can run simulated annealing Gibbs sampling algorithm in a bid for eventually sampling from the defined Boltzman distribution for large values of β. The energy of the output FB sequence is going to be with high probability very close to the energy of the minimizer of $E(y^n)$ which in turn is what the exhaustive search WZ coder finds at its first step. From (25), in order to compute the change in $E(y^n)$ resulting from updating the value of some y_i chosen at random, as required by the Gibbs sampler, we should update the value of the count vectors m_k, $\tau_{m,\ell}$ and $\phi_{m,\ell}$ which at most requires a number of calculations linear in $\max(k, \ell, m)$. After finding the FB sequence through this process, then the rest of the procedure is similar to the universal WZ coder described in [22], or the exhaustive search algorithm described in the previous section. As it will be proved in the full version of this paper, the described WZ data compression algorithm features asymptotic universal optimality results analogous to the result of Theorem 2 for the MCMC-based lossy compression scheme described in Section V.

VII. Optimal denoising via MCMC-based lossy coding

Consider the problem of denoising a stationary ergodic source $\mathbf{X}$ with unknown distribution corrupted by additive white noise $\mathbf{N}$. Compression-based denoising algorithms have been proposed before by a number of researchers[2]. The idea of using a universal lossy compressor in denoising was first proposed in [21], and then was refined in [23] to result in a universal denoising algorithm. In this section, we show how our new MCMC-based lossy encoder enables the denoising algorithm proposed in [23] to lead to an implementable universally optimal denoiser. We will compare between the DUDE [20], and this new MCMC compression-based approach to universal denoising in future work

In [23], it is shown how a universally optimal lossy coder tuned to the right distortion measure and distortion level combined with some simple "post-processing" results in a universally optimal denoiser. In the sequel first we briefly go over this compression-based denoiser described in [23], and then show how our lossy coder can be embedded in it.

Throughout this section we assume that the source, noise, and reconstruction alphabets are $\mathcal{M}$-ary alphabet $\mathcal{A} = \{0, 1, \ldots, M-1\}$, and the noise is assumed to be additive modulo-M and $P_N(a) > 0$ for any $a \in \mathcal{A}$, i.e. $Z_i = X_i + N_i$.

As mentioned earlier, in the denoising scheme outlined in [23], first the denoiser lossily compresses' the noisy signal appropriately, and partly removes the additive noise. Consider a sequence of *good* lossy coders characterized by encoder/decoder pairs $(\mathrm{E}_n, \mathrm{D}_n)$ of block length n working at distortion level $H(N)$ under the difference distortion measure defined as

$$\rho(x, y) = \log \frac{1}{P_N(x-y)}. \tag{26}$$

By *good*, it is meant that for any stationary ergodic source $\mathbf{X}$, as n grows, the rate distortion performance of the sequence of codes converges to a point on the rate-distortion curve. The next step is the simple "post-processing" which works as follows. For a fixed m, define the following count vector over the noisy signal Z^n and its quantized version $Y^n = \mathrm{D}_n(\mathrm{E}_n(Z^n))$,

$$\hat{Q}^{2m+1}[Z^n, Y^n](z^{2m+1}, y) \triangleq \frac{1}{n}|\{1 \le i \le n : (Z_{i-k}^{i+k}, Y_i) = (z^{2m+1}, y)\}|. \tag{27}$$

After constructing these count vectors, the denoiser output is generated through the "post-processing" or the "derandomization" process as follows

$$\hat{X}_i = \arg\min_{\hat{x}\in\mathcal{A}} \sum_{y\in\mathcal{A}} \hat{Q}^{2m+1}[Z^n, Y^n](z^{2m+1}, y) d(\hat{x}, y), \tag{28}$$

where $d(\cdot,\cdot)$ is the original loss function under which the performance of the denoiser is to be measured. The described denoiser is shown to be universally optimal [23], and the basic theoretical justification of this is that the rate-distortion function of the noisy signal $\mathbf{Z}$ under the difference distortion measure satisfies the Shannon lower bound with equality, and

[2]Refer to the references mentioned in [23] for the literature on compression-based denoisers.

it is proved in [23] that for such sources [3] for a fixed k, the k-th order empirical joint distribution between the source and reconstructed blocks defined as

$$\hat{Q}^k[X^n, Y^n](x^k, y^k) \triangleq \frac{1}{n}|\{1 \leq i \leq n : (X_i^{i+k-1}, Y_i^{i+k-1}) = (x^k, y^k)\}|, \quad (29)$$

resulting from a sequence of *good* codes converge to P_{X^k,Y^k} in distribution, i.e. $\hat{Q}^k[X^n, Y^n] \stackrel{d}{\Rightarrow} P_{X^k,Y^k}$, where P_{X^k,Y^k} is the unique joint distribution that achieves the k-th order rate-distortion function of the source. In the case of quantizing the noisy signal under the distortion measure defined in (26), at level $H(N)$ P_{X^k,Y^k} is the k-th order joint distribution between the source and noisy signal. Hence, the count vector $\hat{Q}^{2m+1}[Z^n, Y^n](z^{2m+1}, y)$ defined in (27) asymptotically converges to $P_{X_i|Z^n}$ which is what all universally optimal denoisers would base its decision on. After estimating $P_{X_i|Z^n}$, the post-processing step is just making the optimal Bayesian decision at each position.

The main ingredient of the described denoiser is a universal lossy compressor. Note that the MCMC-based lossy compressor described in Section V is applicable to any distortion measure. The main problem is choosing the parameter s corresponding to the distortion level of interest. To find the right slope, we run the quantization MCMC-based part of the algorithm independently from two different initial points s_1 and s_2. After convergence of the two runs we compute the average distortion between the noisy signal and its quantized versions. Then assuming a linear approximation, we find the value of s that would have resulted in the desired distortion, and then run the algorithm again from this start point, and again computed the average distortion, and then find a better estimate of s from the observations so far. After a few repetitions of this process, we have a reasonable estimate of the desired s [19]. Note that for finding s it is not necessary to work with the whole noisy signal, and one can consider only a long enough section of data first, and find s from it, and then run the MCMC-based denoising algorithm on the whole noisy signal with the estimated parameter s.

VIII. CONCLUSIONS AND FUTURE WORK

In this paper, a new implementable universal lossy source coding algorithm based on simulated annealing Gibbs sampling was proposed, and it was shown that it is capable of getting arbitrarily close to the rate-distortion curve of any stationary ergodic source. For coding a source sequence x^n, the algorithm starts from some initial reconstruction block, and updates one of its coordinates at each iteration. The algorithm can be viewed as a process of systematically introducing 'noise' into the original source block, but in a biased direction that results in a decrease of its description complexity. We further developed the application of this new method to universal WZ coding and universal denoising.

[3] In fact it is shown in [23] that this is true for a large class of sources including i.i.d sources and those satisfying the Shannon lower bound with equality.

The convergence rate of the algorithm and the effect of different parameters on it is a topic for further study. As an example, one might wonder how the convergence rate of the algorithm is affected by choosing an initial point other than the source output block itself. Although our theoretical result on universal asymptotic optimality remains intact for any initial starting point, in practice the choice of the starting point might significantly impact the number of iterations required.

REFERENCES

[1] C. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec*, part 4, pp. 142-163, 1959.

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

[3] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. on Inf. Theory*, 24(5):530-536, Sep. 9178.

[4] I. H. Witten, R. M. Neal, and J. G. Cleary, "Arithmetic coding for data compression", *Commun. Assoc. Comp. Mach.*, VOL. 30, NO. 6, pp. 520-540, 1987.

[5] I. Kontoyiannis, "An implementable lossy version of the Lempel Ziv algorithm-Part I: Optimality for memoryless sources," *IEEE Trans. Inform. Theory*, Vol. 45, pp. 2293-2305, Nov. 1999.

[6] E. Yang, Z. Zhang, and T. Berger, "Fixed-Slope Universal Lossy Data Compression," , *IEEE Trans. Inform. Theory*, VOL. 43, NO. 5, pp. 1465-1476, Sep. 1997.

[7] E. H. Yang and J. Kieffer, "Simple universal lossy data compression schemes derived from the Lempel-Ziv algorithm," *IEEE Trans. Inform. Theory*, 1996.

[8] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, Vol. COM-28, pp. 8495, Jan. 1980.

[9] S. Geman and D. Geman. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images,". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721-741, 1984.

[10] S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi, "Optimization by Simulated Annealing", *Science*, Vol. 220, NO. 4598, pp. 671-680, 1983.

[11] V. Cerny, "A thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm," *Journal of Optimization Theory and Applications*, 45:41-51, 1985.

[12] K. Rose, "Deterministic annealing for clustering, compression,classification, regression, and related optimization problems," *Proceedings of the IEEE*, Vol. 86, Issue 11, pp. 2210 - 2239, Nov 1998.

[13] J. Vaisey and A. Gersho, "Simulated annealing and codebook design," *Proc. ICASSP*, pp. 1176-1179, 1988.

[14] E. Maneva and M. J. Wainwright, "Lossy source encoding via messagepassing, and decimation over generalized codewords of LDGM codes," *IEEE Int. Symp. on Inform. Theory*, Adelaide, Austrailia, Sept. 2005.

[15] A. Gupta and S. Verdú, "Nonlinear sparse-graph codes for lossy compression," submitted to: *IEEE Trans. on Inform. Theory.*

[16] J. Rissanen and I. Tabus, "Rate-distortion without random codebooks," Workshop on Information Theory and Applications (ITA), UCSD, 2006.

[17] A. Gupta, S. Verdú, T. Weissman, "Linear-Time Near-Optimal Lossy Compression", submitted to: *IEEE Int. Symp. on Inform. Thoery*, Toronto, Canada, 2008

[18] A. Wyner, and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, Vol. IT-22,pp. 1-10, Jan. 1976.

[19] K. Ramchandran and M. Vetterli, "Best wavelet packet bases in a rate-distortion sense", *IEEE Trans. on Image Process*, Vol. 2, no. 2, pp. 160-175, April 1993.

[20] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú and M. Weinberger, "Universal Discrete Denoising: Known Channel," *IEEE Trans. Inform. Theory*, Vol. 51, no. 1, pp. 5-28, January 2005.

[21] D. Donoho, (2002, Jan.) The Kolmogorov Sampler. [Online]. Available: http://www-stat.stanford.edu/ donoho/reports.html

[22] S. Jalali, S. Verdú, T. Weissman "A Universal Wyner-Ziv Scheme for Discrete Sources," *Proceedings of the IEEE Intl. Symp. on Inform. Theory*, Nice, France, July 2007.

[23] T. Weissman, and E. Ordentlich, "The Empirical Distribution of Rate-Constrained Source Codes", *IEEE Trans. on Inform. Theory*, Vol. 51 , no. 11 , pp. 3718- 3733, Nov. 2005