An Iterative Scheme for Near Optimal and Universal Lossy Compression

Shirin Jalali[1], Andrea Montanari[1], and Tsachy Weissman[1,2]

[1]Department of Electrical Engineering, Stanford University
[2]Department of Electrical Engineering, Technion

***Abstract*—We present a new lossy compression algorithm for discrete sources. The encoder assigns a certain cost to each reconstruction sequence, finds the sequence that minimizes the cost, and describes it losslessly to the decoder via a universal lossless compressor. The cost of a sequence is defined as a linear combination of its empirical probabilities of some order $k+1$ and its distortion relative to the source sequence. The linear structure of the cost in the empirical count matrix allows the encoder to employ a Viterbi-like algorithm for obtaining the minimizing reconstruction sequence simply. We identify a choice of coefficients for the linear combination in the cost function which ensures that the algorithm universally achieves the optimum rate-distortion performance of any Markov source in the limit of large n, provided k is increased as $o(\log n)$. Finding the optimal coefficients is complex and requires solving a non-convex optimization problem. As a detour, we propose a simple heuristic iterative procedure, and demonstrate its efficiency through our experimental results.**

I. Introduction

Let $\mathbf{X}=\{X_i : i \geq 1\}$ represent a discrete-valued stationary ergodic process with unknown statistics, and consider the problem of compressing $\mathbf{X}$ at rate R such that the incurred distortion is minimized. Let $\mathcal{X}$ and $\hat{\mathcal{X}}$ denote finite source and reconstruction alphabets respectively. The performance of the described coding scheme is measured by its average expected distortion between source and reconstruction blocks, i.e.

$$D = \mathrm{E}\, d_n(X^n, \hat{X}^n) \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\, d(X_i, \hat{X}_i), \qquad (1)$$

where $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ is a single-letter distortion measure. For any $R \geq 0$, the minimum achievable distortion (cf. [4] for exact definition of achievability) is characterized as [1], [2], [3]

$$D(\mathbf{X}, R) = \lim_{n\to\infty} \min_{p(\hat{X}^n|X^n):I(X^n;\hat{X}^n)\leq R} \mathrm{E}\, d_n(X^n, \hat{X}^n). \quad (2)$$

A sequence of codes at rate R is called universal if for every stationary ergodic source $\mathbf{X}$ its asymptotic performance converges to $D(\mathbf{X}, R)$, i.e.,

$$\limsup_{n\to\infty} \mathrm{E}\, d_n(X^n, \hat{X}^n) \leq D(\mathbf{X}, R). \qquad (3)$$

For lossless compression where the source is to be recovered without any errors, there already exist well-known implementable universal schemes such as Lempel-Ziv coding [5] or arithmetic coding [6]. In contrast to the situation of lossless compression, for $D > 0$, there are no well-known practical schemes that universally achieve the rate-distortion curve. In recent years, there has been progress towards designing universal lossy compressor especially in trying to tune some of the existing universal lossless coders to work in the lossy case as well [7], [8], [9]. Also, there are many other well-known non-optimal practically appealing methods for lossy compression some of which are comprehensively described in [10].

In a recent work [11], a new implementable algorithm for lossy compression of discrete-valued stationary ergodic sources has been proposed. Instead of fixing rate (or distortion) and minimizing distortion (or rate), the new algorithm fixes the Lagrangian coefficient α, and minimizes $R+\alpha D$. This is done by assigning energy $\mathcal{E}(y^n)$ representing $R+\alpha D$ to each possible reconstruction sequence and finding the sequence that minimizes the cost by simulated annealing. It is shown that using a universal lossless compressor to describe the reconstruction sequence resulting from this process to the decoder results in a scheme which is universal in the limit of many iterations and large block length. The drawback of the proposed scheme is that although its computational complexity per iteration is independent of the block length n and linear in a parameter $k = o(\log n)$, there is no useful bound on the number of iterations required for convergence.

In this paper, inspired by the previous method, we propose yet another approach for lossy compression of discrete Markov sources which universally achieves optimum rate-distortion performance for any discrete Markov source. We start by assigning the same cost that was defined for each possible reconstruction sequence in [11]. The cost of each sequence is a linear combination of two terms: its empirical conditional entropy and its distance to the source sequence to be coded. We show that there exists proper linear approximation of the first term such that minimizing the linearized cost results in the same performance as minimizing the original cost. But the advantage is that minimizing the modified cost can be done via Viterbi algorithm in lieu of simulated annealing which was used for minimizing the original cost. The problem is that finding the optimal coefficients is hard and requires solving a non-convex optimization problem. To resolve this problem, we propose an iterative approach which starts by estimating the coefficients from the input sequence, and runs the Viterbi algorithm. Then the coefficients are re-estimated from the

978-1-4244-4536-3/09/$25.00 ©2009 IEEE

output sequence. Effectiveness of this approach is shown by simulation results.

The organization of the paper is as follows. In Section II, we set up the notation, and define the count matrix and empirical conditional entropy of a sequence. Section III describes a new coding scheme for fixed-slope lossy compression which universally achieves the rate-distortion curve for any discrete Markov source and IV describes how to compute the coefficients required by the algorithm outlined in the previous section. Section V explains how Viterbi algorithm can be used for implementing the coding scheme described in Section III, and suggests a heuristic iterative approach to detour the problem of finding the optimal coefficients. Section VI presents some simulations results, and finally, Section VII concludes the paper with a discussion of some future directions. All the proofs will appear in the full version.

II. Notations and required definitions

Let matrix $\mathbf{m}(y^n) \in \mathbb{R}^{|\hat{\mathcal{X}}|} \times \mathbb{R}^{|\hat{\mathcal{X}}|^k}$ represent $(k+1)^{\text{th}}$ order empirical count of y^n defined as

$$m_{\beta,\mathbf{b}}(y^n) = \frac{1}{n} \left| \left\{ 1 \le i \le n : y_{i-k}^{i-1} = \mathbf{b}, y_i = \beta \right\} \right|. \quad (4)$$

In (4), and throughout we assume a cyclic convention whereby $y_i \triangleq y_{n+i}$ for $i \le 0$. Let $H_k(y^n)$ denote the conditional empirical entropy of order k induced by y^n, i.e.

$$H_k(y^n) = H(Y_{k+1}|Y^k), \quad (5)$$

where Y^{k+1} on the right hand side of (5) is distributed according to

$$\mathrm{P}(Y^{k+1} = u^{k+1}) = m_{u_{k+1},u^k}(y^n). \quad (6)$$

The conditional empirical entropy in (5) can be expressed as a function of $\mathbf{m}[y^n]$ as follows

$$H_k(y^n) = \frac{1}{n} \sum_{u^k} \mathcal{H}\left(m_{\cdot,u^k}(y^n)\right) \mathbf{1}^T m_{\cdot,u^k}(y^n), \quad (7)$$

where $\mathbf{1}$ and $m_{\cdot,u^k}(y^n)$ denote the all-ones column vector of length $|\hat{\mathcal{X}}|$, and the column in $\mathbf{m}(y^n)$ corresponding to u^k respectively. For a vector $\mathbf{v} = (v_1, \ldots, v_\ell)^T$ with non-negative components, we let $\mathcal{H}(\mathbf{v})$ denote the entropy of the random variable whose probability mass function (pmf) is proportional to $\mathbf{v}$. Formally,

$$\mathcal{H}(\mathbf{v}) = \begin{cases} \sum_{i=1}^{\ell} \frac{v_i}{\|\mathbf{v}\|_1} \log \frac{\|\mathbf{v}\|_1}{v_i} & \text{if } \mathbf{v} \neq (0,\ldots,0)^T \\ 0 & \text{if } \mathbf{v} = (0,\ldots,0)^T. \end{cases} \quad (8)$$

III. Linearized cost function

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

$$\hat{x}^n = \arg \min_{y^n \in \hat{\mathcal{X}}^n} \left[H_k(y^n) + \alpha \cdot d_n(x^n, y^n) \right]. \quad (9)$$

The encoder, after computing $\hat{x}^n$, losslessly conveys it to the decoder using LZ compression. Let k grow with n as $k = k_n = o(\log n)$.

Theorem 1: [11] 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 for coding X^n. Then

$$\mathbb{E}\left[\frac{1}{n} \ell_{\mathsf{LZ}}(\hat{X}^n) + \alpha d_n(X^n, \hat{X}^n) \right] \overset{n\to\infty}{\longrightarrow} \min_{D \ge 0} \left[R(\mathbf{X}, D) + \alpha D \right]. \quad (10)$$

In other words, conveying the reconstruction sequence to the decoder via universal lossless compression (selection of LZ algorithm here is for concreteness, but other universal lossless methods can be used as well) achieves optimum fixed-slope rate-distortion performance universally.

As proposed in [11], the exhaustive search required by this algorithm can be tackled through simulated annealing Gibbs sampling. Here assuming the source is a discrete Markov source, we propose another method for finding a sequence achieving the minimum in (9).

Before describing the new scheme, consider the problems (P1) and (P2) described below.

$$\text{(P1)}: \quad \min_{y^n} \ \left[H_k(\mathbf{m}(y^n)) + \alpha d_n(x^n, y^n) \right], \quad (11)$$

and

$$\text{(P2)}: \quad \min_{y^n} \ \left[\sum_{\beta \in \hat{\mathcal{X}}} \sum_{\mathbf{b} \in \hat{\mathcal{X}}^k} \left[\lambda_{\beta,\mathbf{b}} m_{\beta,\mathbf{b}}(y^n) + \alpha d_n(x^n, y^n) \right] \right]. \quad (12)$$

Comparing (P1) with (9) reveals that it is the optimization required by the exhaustive search coding scheme described before. The question is whether it is possible to choose a set of coefficients $\{\lambda_{\beta,\mathbf{b}}\}_{\beta,\mathbf{b}}$, $\beta \in \hat{\mathcal{X}}$ and $\mathbf{b} \in \hat{\mathcal{X}}^k$, such that (P1) and (P2) have the same set of minimizers or at least, the set of minimizers of (P2) is a subset of minimizers of (P1). The reason for asking this question is that if this is true, instead of solving (P1) one can solve (P2), which, as we describe in Section V,is straightforward using Viterbi algorithm.

Let S_1 and S_2 denote the set of minimizers of (P1) and (P2). Consider some $z^n \in S_1$, and let $\mathbf{m}_n^* = \mathbf{m}(z^n)$. Since $H(\mathbf{m})$ is concave in $\mathbf{m}$, for any empirical count matrix $\mathbf{m}$, i.e.,

$$\sum_{\beta \in \hat{\mathcal{X}}} \sum_{\mathbf{b} \in \hat{\mathcal{X}}^k} m_{\beta,\mathbf{b}} = 1,$$

and $m_{\beta,\mathbf{b}} \ge 0$ for any $b \in \hat{\mathcal{X}}$ and $\mathbf{b} \in \hat{\mathcal{X}}^k$, we have

$$H(\mathbf{m}) \le H(\mathbf{m}_n^*) + \sum_{\beta,\mathbf{b}} \frac{\partial}{\partial m_{\beta,\mathbf{b}}} H(\mathbf{m})|_{\mathbf{m}_n^*} (m_{\beta,\mathbf{b}} - m^*_{\beta,\mathbf{b}}).$$

Now assume that in (P2), the coefficients are chosen as follows

$$\lambda_{\beta,\mathbf{b}} = \frac{\partial}{\partial m_{\beta,\mathbf{b}}} H(\mathbf{m})|_{\mathbf{m}_n^*}. \quad (13)$$

Lemma 1: (P1) and (P2) have the same minimum value, if the coefficients are chosen according to (13). Moreover, if all the sequences in S_1 have the same type, $S_1 = S_2$.

This shows that if we knew the optimal type $\mathbf{m}_n^*$, then we could compute the optimal coefficients via (13), and solve (P2) instead of $(P1)$. The problem is that $\mathbf{m}_n^*$ is not known to the encoder (since knowledge of m_n^* requires solving (P1) which is the problem we are trying to avoid). In the next section, we describe a method for approximating $\mathbf{m}_n^*$, and hence the coefficients $\{\lambda_{\beta,\mathbf{b}}\}$.

IV. How to choose the coefficients?

For a given stationary ergodic source $\mathbf{X}$, and for any given count matrix $\mathbf{m}$ define $D(\mathbf{m})$ to be the minimum average expected distortion among all processes $\mathbf{Y}$ that are jointly stationary ergodic with $\mathbf{X}$ and their $(k+1)^{\text{th}}$ order stationary distribution is according to $\mathbf{m}$. $D(\mathbf{m})$ can equivalently be defined as

$$D(\mathbf{m}) = \lim_{k_1\to\infty} \min_{p(x^{k_1},y^{k_1})\in\mathcal{M}^{(k_1)}} \mathrm{E}_p\, d(x^{k_1},y^{k_1}), \tag{14}$$

where $\mathcal{M}^{(k_1)}$ is the set of all jointly stationary distributions $p(x^{k_1},y^{k_1})$ of (X^{k_1},Y^{k_1}) with marginal distributions with respect to x coinciding with the k_1^{th} order distribution of $\mathbf{X}$ process, and with marginal distributions with respect to y coinciding with $\mathbf{m}$, i.e., having the $(k+1)^{\text{th}}$ order marginal distribution described by $\mathbf{m}$.

Lemma 2: If the source is ℓ^{th} order Markov, then

$$D(\mathbf{m}) = \min_{p(x^{k_1},y^{k_1})\in\mathcal{M}^{(k_1)}} \mathrm{E}_p\, d(x^{k_1},y^{k_1}), \tag{15}$$

where $k_1 = \max(\ell, k+1)$.

Now consider the following optimization problem,

$$\begin{aligned} \min \quad & H(\mathbf{m}) + \alpha D(\mathbf{m}) \\ \text{s.t.} \quad & \mathbf{m}\in\mathcal{M}^{(k_1)}, \end{aligned} \tag{16}$$

for $k_1 = k+1$. By Lemma 2, an equivalent representation of the above optimization problem is

$$\begin{aligned} \min \quad & H(\mathbf{m}) + \alpha \sum_{\substack{\beta,\beta'\\ \mathbf{b},\mathbf{b}'}} d_{k_1}(\beta'\mathbf{b}',\beta\mathbf{b}) p_x(\beta'\mathbf{b}') q_{y|x}(\beta\mathbf{b}|\beta'\mathbf{b}') \\ \text{s.t.} \quad & 0 \le q_{y|x}(\beta\mathbf{b}|\beta'\mathbf{b}') \le 1, \quad \forall\, \beta,\beta',\mathbf{b},\mathbf{b}' \\ & \sum_{\beta,\mathbf{b}} q_{y|x}(\beta\mathbf{b}|\beta'\mathbf{b}') = 1, \quad \forall\, \beta',\mathbf{b}', \\ & \sum_{\beta,\beta'} p_x(\beta'\mathbf{b}') q_{y|x}(\beta\mathbf{b}|\beta'\mathbf{b}') = \\ & \qquad \sum_{\beta,\beta'} p_x(\mathbf{b}'\beta') q_{y|x}(\mathbf{b}\beta|\mathbf{b}'\beta'), \quad \forall \mathbf{b},\mathbf{b}', \\ & m_{\beta,\mathbf{b}} \quad = \sum_{\beta'\mathbf{b}'} p_x(\beta'\mathbf{b}') q_{y|x}(\beta\mathbf{b}|\beta'\mathbf{b}'), \quad \forall\, \beta,\mathbf{b}. \end{aligned} \tag{17}$$

Note that the variables in (17) are conditional distributions $q(y^{k_1}|x^{k_1})$, but we are only interested in the $\mathbf{m}$ that they induce.

Lemma 3: If for each n, (P1) has a unique minimizing type $\mathbf{m}_n^*$, then

$$\|\mathbf{m}_n^* - \hat{\mathbf{m}}_n^*\|_{\text{TV}} \to 0, \qquad \text{a.s.}, \tag{18}$$

where $\hat{\mathbf{m}}_n^*$ is the solution of (17).

Remark: In (17), the only dependence on n is through k_1.

Therefore, if the encoder knew the distribution of the source, it could solve (17), find a good approximation of $\mathbf{m}_n^*$, and then use (13) to compute the coefficients required by (P2). The problem is that the encoder does not have this information, and only knows that the source is Markov (but does not know its order). To overcome its lack of information, a reasonable step is to use empirical distribution of the source instead of the true unknown distribution in (17). For $a^{k_1}\in\mathcal{X}^{k_1}$, define the k_1^{th} order empirical distribution of the source as

$$\hat{p}^{(k_1)}(a^{k_1}) \triangleq \frac{|\{i : (x_{i-k_1},\ldots,x_{i-1}) = a^{k_1}\}|}{n}. \tag{19}$$

The following lemma shows that for $k_1 = o(\log n)$, $\hat{p}^{(k_1)}$ converges to the actual k_1^{th} order distribution of the source, and therefore can be considered as a good approximation for it.

Lemma 4: For $k_1 = o(\log n)$, and any stationary ergodic Markov source,

$$\|\hat{p}^{(k_1)} - p^{(k_1)}\|_{\text{TV}} \to 0 \quad \text{a.s.}, \tag{20}$$

where $p^{(k_1)}$ is the true k_1^{th} order distribution of the Markov source.

Assume x^n is generated by a discrete Markov source, and let $\hat{p}^{(k_1)}$ be the empirical distribution defined by equation (19). Consider the optimization problem given in (17), and replace $p_x(\beta'\mathbf{b}')$ with $\hat{p}_x^{(k_1)}(\beta'\mathbf{b}')$, for all β' and $\mathbf{b}'$, and let $\tilde{\mathbf{m}}_n^*$ denote the output of the this alternative optimization problem.

Lemma 5: For $k_1 = k_1(n) = o(\log n)$,

$$\|\tilde{\mathbf{m}}_n^* - \hat{\mathbf{m}}_n^*\|_{\text{TV}} \to 0, \quad a.s.$$

Let $\{\lambda_{\beta,\mathbf{b}}(n)\}_{\beta,\mathbf{b}}$ denote the optimal values of the coefficients defined at $\mathbf{m}_n^*$ (as given in (13)), and let $\{\hat{\lambda}_{\beta,\mathbf{b}}(n)\}_{\beta,\mathbf{b}}$ be coefficients computed at $\tilde{\mathbf{m}}_n^*$, then

Lemma 6:

$$\max_{\beta,\mathbf{b}} |\lambda_{\beta,\mathbf{b}}(n) - \hat{\lambda}_{\beta,\mathbf{b}}(n)| \to 0 \text{ as } n\to\infty. \tag{21}$$

These results suggest that for computing the coefficients, we can solve the approximate version of optimization problem given in (17) (whose complexity can be controlled with the rate of increase of k_1), and then substitute the result in (13) to obtain the approximate coefficients. After that (P2) defined by these coefficients can be solved using the Viterbi algorithm in a way that will be detailed in the next section. The succession of lemmas detailed in the previous sections then allow us to prove the following theorem.

Theorem 2: Let $\mathbf{X}$ be a stationary and ergodic Markov source, and $R(\mathbf{X},D)$ denote its rate distortion function. Let $\hat{X}^n$ be the reconstruction sequence obtained using the above scheme for coding X^n taking k_1 to increase without bound as $o(\log n)$. Then

$$\mathbb{E}\left[\frac{1}{n}H_{k_1}(\mathbf{m}(\hat{X}^n)) + \alpha d_n(X^n,\hat{X}^n)\right] \overset{n\to\infty}{\longrightarrow} \min_{D\ge 0}\left[R(\mathbf{X},D) + \alpha D\right]. \tag{22}$$

Remark: Theorem 2 implies the fixed-slope universality of the scheme which does the lossless compression of the reconstruction by first describing its count matrix (costing a number of bits which is negligible for large n) and then doing the conditional entropy coding.

V. Iterative Viterbi coder

After finding the optimal coefficients, $\{\lambda_{\mathbf{b},\beta}\}$, as described in Section III, Encoder needs to solve (P2). In this section, we show how this can be done efficiently via Viterbi algorithm. Moreover, we propose a heuristic iterative approach for finding the coefficients.

Note that the linearized cost can be written as

$$\sum_{\substack{\mathbf{b}\in\hat{\mathcal{X}}^k\\ \beta\in\hat{\mathcal{X}}}} [\lambda_{\beta,\mathbf{b}} m_{\beta,\mathbf{b}}(y^n) + \alpha d_n(x^n, y^n)] = \frac{1}{n}\sum_{i=1}^{n}\left[\lambda_{y_i, y_{i-k}^{i-1}} + \alpha d(x_i, y_i)\right]. \quad (23)$$

The advantage of this alternative representation is that, as it will be described, instead of using simulated annealing, we can find the sequence that minimizes (23) via Viterbi algorithm. Viterbi algorithm is a dynamic programming optimization method which finds the path of minimum weight in a Trellis diagram efficiently. For $i = k+1, \ldots, n$, let $s_i := y_{i-k}^i$ be the state at time i, and $\mathcal{S}$ be the set of all $|\hat{\mathcal{X}}|^{k+1}$ possible states. From this definition, s_i is a function of s_{i-1} and y_i, i.e. $s_i = g(s_{i-1}, y_i)$, for some $g : \mathcal{S}\times\hat{\mathcal{X}} \to \mathcal{S}$. This representation leads to a Trellis diagram corresponding to the evolution of the states $\{s_i\}_{i=k+1}^n$ in which each state has $|\hat{\mathcal{X}}|$ states leading to it and $|\hat{\mathcal{X}}|$ states branching from it. Now to each edge in this graph, we assign a weight which depends on the tail state and the input symbol, i.e., to the edge $e = (s', s)$ connecting states s' and $s = b^{k+1}$ at time i we assign weight $w_i(e)$ as

$$w_i(e) := \lambda_{b_{k+1}, b^k} + \alpha d(x_i, b_{k+1}). \quad (24)$$

Note that these edge weights depend on time through the input sequence x^n. In this representation, there is a 1-to-1 correspondence between sequences $y^n \in \hat{\mathcal{X}}^n$, and sequences of states $\{s_i\}_{i=k+1}^n$, and minimizing (23) is equivalent to finding the path of minimum weight in the corresponding Trellis diagram, i.e., the path $\{s_i\}_{i=k+1}^n$ that minimizes $\sum_{i=k+1}^n w_i(e_i)$, where $e_i = (s_{i-1}, s_i)$. Solving this minimization can readily be done by Viterbi algorithm which can be described as follows. For each state s, let $\mathcal{L}(s)$ be the $|\hat{\mathcal{X}}|$ states leading to it, and for any $i > 1$, define

$$C_i(s) := \min_{s'\in\mathcal{L}(s)} [w_i((s', s)) + C_{i-1}(s')]. \quad (25)$$

For $i = 1$ and $s = b^{k+1}$, let $C_1(s) := \lambda_{b_{k+1}, b^k} + \alpha d_{k+1}(x^{k+1}, b^{k+1})$. Using this procedure, each state s at each time j has a path of length $j-k-1$ which is the minimum path among all the possible paths between the states from time $i = k+1$ to $i = j$ such that $s_j = s$. After computing $\{C_i(s)\}$ for all $s \in \mathcal{S}$ and all $i \in \{k+1, \ldots, n\}$, at time $i = n$, let

$$s^* = \arg\min_{s\in\mathcal{S}} C_n(s). \quad (26)$$

It is not hard to see that the path leading to s^* is the path of minimum weight among all possible paths. Note that the computational complexity of this procedure is linear in n but exponential in k because the number of states increases exponentially with k.

Therefore, given the coefficients $\{\lambda_{\mathbf{b},\beta}\}$, solving (P2) is straightforward using Viterbi algorithm. The problem is finding an approximation of the optimal coefficients. The procedure outlined in Section III for finding the coefficients involves solving a non-convex optimization. To bypass this process, an alternative heuristic method is as follows: Start with $\mathbf{m}(x^n)$. Compute the coefficients from (13) at $\mathbf{m}(x^n)$. Employ Viterbi algorithm to solve (P2) at the computed coefficients. Let y^n denote the output sequence. Compute $\mathbf{m}(y^n)$, and recalculate the coefficients using (13) at $\mathbf{m}(y^n)$. Again, use Viterbi algorithm to solve (P2) at the updated coefficients. Iterate.

The effectiveness of this approach is discussed in the next section through some simulations.

VI. Simulation results

As a first example, consider an i.i.d. $\mathrm{Bern}(q)$ source with $q = 0.25$. Fig. 1 shows the rate-distortion curve corresponding this source, i.e., $R(D) = h(q) - h(D)$, where $h(p) = -p\log p - (1-p)\log(1-p)$, and the point on the curve corresponding to $\alpha = 2$. Moreover, the succession of points derived from the iterative approach after $I = 10$ iterations of the Viterbi encoder toward the optimal point is shown in the figure. Each point is the average performance of $L = 50$ independent runs of the algorithm. Fig. 2 shows how the iterative approach reduces the cost $H_k(\hat{x}^n) + \alpha d_n(x^n, \hat{x}^n)$ after $I = 4$ iterations. It can be observed that increasing α results in less reduction. One explanation is the following: larger values of α correspond to smaller values of desired distortion. Therefore, since x^n and optimal y^n are close together, so are $\mathbf{m}(x^n)$ and $\mathbf{m}(y^n)$. Hence, in such cases, there is not much gain in repeating the process of estimating the coefficients since they are close to the optimal from the beginning.

As a next example, consider a binary symmetric Markov source (BSMS) with transition probability $q = 0.2$. Fig. 3 compares the average performance of the Viterbi encoder against $R(D)$ and Shannon lower bound for different values of α. Note that the rate-distortion of this source is not known in general except up to $D = 0.0159$. Beyond that there are upper and lower bounds on $R(D)$, the most famous of which is Shannon lower bound: $R(D) \geq h(q) - h(D)$. The average is taken over $L = 20$ runs of the algorithm. The coefficients are computed at $\mathbf{m}(x^n)$. As described before, for large values of α, or equivalently small values of D, $\mathbf{m}(x^n)$ is a reasonable approximation of $\mathbf{m}^*$. But as α decreases, it becomes a poor approximation. Fig. 4 shows the reduction in the cost obtained by using the heuristic iterative approach. Note that

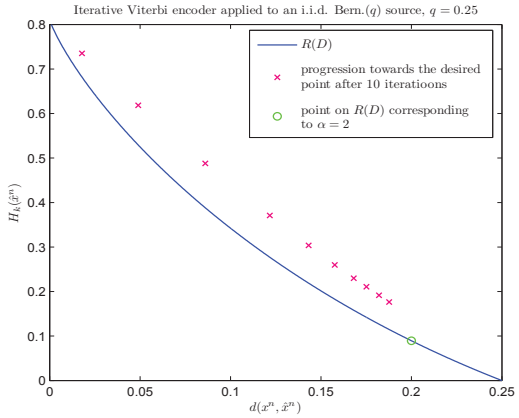

Fig. 1. Iterative Viterbi lossy coder with $I = 10$ iterations. [Bern(q) i.i.d. source, $q = 0.25$, $\alpha = 2$, $n = 20000$, $k = 8$, and $L = 50$]

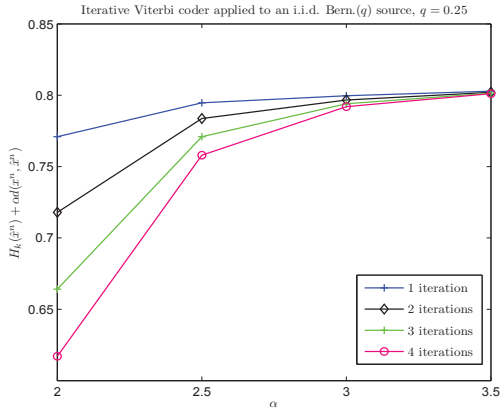

Fig. 2. Reduction in the final cost after $I = 4$ iterations versus α. [Bern(q) i.i.d. source, $q = 0.25$, $n = 20000$, $k = 8$, and $L = 50$]

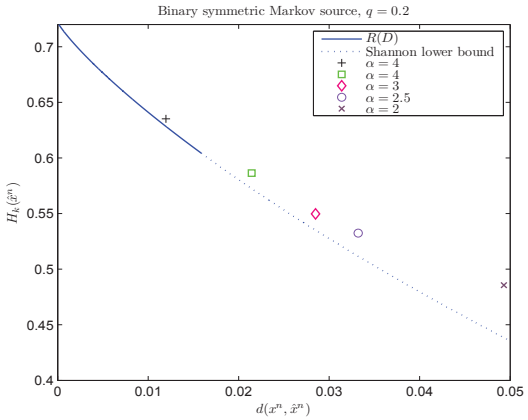

Fig. 3. Viterbi encoder with the coefficients computed at $\mathbf{m}(x^n)$. [BSMS(q), $q = 0.2$, $n = 20000$, $k = 9$, and $L = 20$]

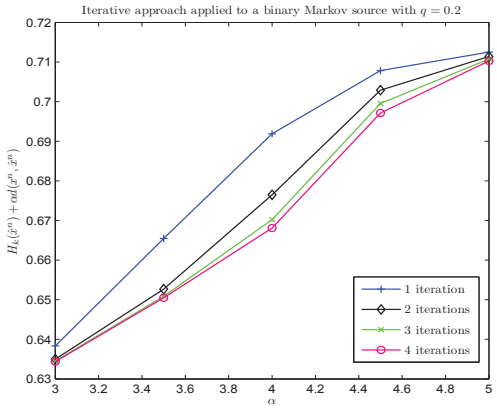

Fig. 4. Demonstrating the reduction in the cost, $H_k(\hat{x}^n) + \alpha d_n(x^n, \hat{x}^n)$, resulting from the iterative approach. [BSMS(q), $q = 0.2$, $n = 20000$, $k = 8$, and $L = 20$]

for large and small values of α the gain is not significant, but for moderate values there is a significant gain in using the iterations. There are different explanations for the same effect observed at the two extreme cases of small and large values of α. While for large values, the iterations does not help because the initial point is already a reasonable approximation of the optimal value, for the other case, iterations does not help, because $\mathbf{m}(x^n)$ is a too poor approximation of $\mathbf{m}^*$ and leads the algorithm into a local minima which it cannot escape through more iterations.

VII. Conclusions and Current Directions

In this paper, a new method for universal fixed-slope lossy compression of discrete Markov sources was proposed. The new method achieves the rate-distortion curve for any discrete Markov source. Extending the algorithm to work on any stationary ergodic source is under current investigation. We believe that in fact the same algorithm works for the general class of stationary ergodic sources, and only the proof should be extended to work in this case as well.

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] R.G. Gallager, "Information Theory and Reliable Communication," New York, NY: John Wiley & Sons, 1968.

[3] T. Berger, *Rate-distortion theory: A mathematical basis for data compression*, Englewood Cliffs, NJ: Prentice-Hall, 1971.

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

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

[6] 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.

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

[8] E. Yang, Z. Zhang, and T. Berger, "Fixed-slope universal lossy data compression," , *IEEE Trans. on Inform. Theory*, vol. 43, no. 5, pp. 1465-1476, Sep. 1997.

[9] E. H. Yang and J. Kieffer, "Simple universal lossy data compression schemes derived from the Lempel-Ziv algorithm," *IEEE Trans. on Inform. Theory*, vol. 42, no. 1, pp. 239-245, 1996.

[10] A. Gersho, R.M. Gray, *Vector Quantization and Signal Compression* Springer, 1992.

[11] S. Jalali, T. Weissman, "Lossy coding via Markov chain Monte Carlo," IEEE International Symposium on Information Theory, Toronto, Canada, 2008.