Lossy compression of discrete sources via Viterbi algorithm

Shirin Jalali, Andrea Montanari and Tsachy Weissman

Abstract

We present a new lossy compressor for discrete-valued sources. For coding a sequence x^n, the encoder starts by assigning a certain cost to each possible reconstruction sequence. It then finds the one that minimizes this cost and describes it losslessly to the decoder via a universal lossless compressor. The cost of each sequence is a linear combination of its distance from the sequence x^n and a linear function of its k^{th} order empirical distribution. The structure of the cost function allows the encoder to employ the Viterbi algorithm to recover the minimizer of the cost. We identify a choice of the coefficients comprising the linear function of the empirical distribution used in the cost function which ensures that the algorithm universally achieves the optimum rate-distortion performance of any stationary ergodic source in the limit of large n, provided that k diverges as $o(\log n)$. Iterative techniques for approximating the coefficients, which alleviate the computational burden of finding the optimal coefficients, are proposed and studied.

I. Introduction

Consider the problem of universal lossy compression of stationary ergodic sources described as follows. Let $\mathbf{X} = \{X_i; \forall\, i \in \mathbb{N}^+\}$ be a stochastic process and let $\mathcal{X}$ denote its alphabet which is assumed discrete and finite throughout this paper. Consider a family of source codes $\{\mathcal{C}_n\}_{n\geq 1}$. Each code $\mathcal{C}_n$ in this family consists of an encoder f_n and a decoder g_n such that

$$f_n : \mathcal{X}^n \to \{0,1\}^*, \tag{1}$$

and

$$g_n : \{0,1\}^* \to \hat{\mathcal{X}}^n, \tag{2}$$

where $\hat{\mathcal{X}}$ denotes the reconstruction alphabet which also is assumed to be finite and in most cases is equal to $\mathcal{X}$. $\{0,1\}^*$ denotes the set of all finite length binary sequences. The encoder f_n maps each source block X^n to a binary sequence of finite length, and the decoder g_n maps the coded bits back to the signal space as $\hat{X}^n = g_n(f_n(X^n))$. Let $l_n(f_n(X^n))$ denote the length of the binary sequence assigned to sequence X^n by the encoder f_n. The performance of each code in this family is measured by the expected rate and the expected average distortion it induces. For a given source $\mathbf{X}$ and coding scheme $\mathcal{C}_n$, the expected rate R_n, and expected average distortion D_n, of $\mathcal{C}_n$ in coding the process $\mathbf{X}$ are defined as follows:

$$R_n = \mathrm{E}[\frac{1}{n} l_n(f_n(X^n))], \tag{3}$$

and

$$D_n = \mathrm{E}[d_n(X^n, \hat{X}^n)] \triangleq \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} d(X_i, \hat{X}_i)\right], \tag{4}$$

where $\hat{X}^n = g_n(f_n(X^n))$, and $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$ is a per-letter distortion measure.

For a given process and any rate $R \geq 0$, the minimum achievable distortion (cf. [1] for exact definition of achievability) is characterized as [2], [3], [4]

$$D(R, \mathbf{X}) = \lim_{n\to\infty} \min_{p(\hat{X}^n|X^n): I(X^n;\hat{X}^n)\leq R} \mathrm{E}[d_n(X^n, \hat{X}^n)]. \tag{5}$$

Similarly, for any distortion $D > 0$, define $R(D, \mathbf{X})$ to denote the minimum required rate for achieving distortion D, i.e.,

$$R(D, \mathbf{X}) = \min_{D(r,\mathbf{X})\leq D} r.$$

Universal lossy compression codes are usually defined in the literature in one of the following modes [5]:

I. Fixed-rate: A family of lossy compression codes $\{\mathcal{C}_n\}$ is called fixed-rate universal, if for every stationary ergodic process $\mathbf{X}$, $R_n \leq R$, $\forall n \geq 1$, and

$$\limsup_n D_n = D(R, \mathbf{X}).$$

II. Fixed-distortion: A family of lossy compression codes $\{\mathcal{C}_n\}$ is called fixed-distortion universal, if for every stationary ergodic process $\mathbf{X}$, $D_n \leq D$, $\forall n \geq 1$, and

$$\limsup_n R_n = R(D, \mathbf{X}).$$

III. Fixed-slope: A family of lossy compression codes $\{\mathcal{C}_n\}$ is called fixed-slope universal, if there exists $\alpha > 0$, such that for every stationary ergodic process $\mathbf{X}$

$$\limsup_n [R_n + \alpha D_n] = \min_{D\geq 0} [R(D, \mathbf{X}) + \alpha D].$$

Existence of universal lossy compression codes for all these paradigms has already been established in the literature a long time ago [6], [7], [8], [9], [10], [11]. The remaining challenging step is to design universal lossy compression algorithms that are implementable and appealing from a practical viewpoint.

A. Related prior work

Unlike lossless compression, where there exists a number of well-known universal algorithms which are also attractive from a practical perspective (cf. Lempel-Ziv algorithm [12] or arithmetic coding algorithm [13]), in lossy compression, despite all the progress in recent years, no such algorithm is yet known. In this section, we briefly review some of the related literature on universal lossy compression with the main emphasis on the progress towards the design of practically appealing algorithms.

There have been different approaches towards designing universal lossy compression algorithms. Among them the one with longest history is that of tuning the well-known universal lossless compression algorithms to work

for the lossy case as well. For instance, Cheung and Wei [14] extended the move-to-front transform to the case where the reconstruction is not required to perfectly match the original sequence. One basic tool used in LZ-type compression algorithms, is the idea of string-matching, and hence there have been many attempts to find optimal approximate string-matching. Morita and Kobayashi [15] proposed a lossy version of LZW algorithm, and Steinberg and Gutman [16] suggested a fixed-database lossy compression algorithms based on string-matching. Although the extensions could all be implemented efficiently, they were later proved to be sub-optimal by Yang and Kieffer [17], even for memoryless sources. Another related example, is the work by Luczak and Szpankowski which proposes another suboptimal compression algorithm which again uses the ideas of approximate pattern matching [18]. For some other related work see [19] [20][21].

Another well-studied approach to lossy compression is Trellis coded quantization [22] and more generally vector quantization (c.f. [23], [24] and the references therein). Codes of this type are usually designed for a given distributions encountered in a specific application. For example, such codes are used in image compression (JPEG) or video compression (MPEG). Nevertheless, there have been attempts at extending such codes to more general settings. For instance Kasner, Marcellin, and Hunt proposed universal Trellis coded quantization which is used in the JPEG2000 standard [25].

There has been a lot of progress in recent years in designing *non-universal* lossy compression algorithms of discrete memoryless sources. Some examples of the recent work in this area are as follows. Wainwright and Maneva [26] proposed a lossy compression algorithm based on message-passing ideas. The effectiveness of the scheme was shown by simulations. Gupta and Verdú proposed an algorithm based on non-linear sparse-graph codes [27]. Another algorithm with near linear complexity is suggested by Gupta, Verdú and Weissman in [28]. The algorithm is based on a 'divide and conquer' strategy. It breaks the source sequence into sub-blocks and codes the subsequences separately using a random codebook. Finally, the capacity-achieving polar codes proposed by Arikan [29] for channel coding are shown to be optimal for lossy compression of binary-symmetric memoryless sources in [30].

The idea of fixed-slope universal lossy compression was first suggested by Yang, Zhang and Berger in [5]. They proposed a generic fixed-slope universal algorithm which leads to specific coding algorithms based on different universal lossless compression algorithms. Although the constructed algorithms are all universal, they involve computationally demanding minimizations, and hence are impractical. In [5], the authors considered lowering the search complexity by choosing appropriate lossless codes which allow to replace the required exhaustive search by a low-complexity sequential search scheme that approximates the solution of the required minimization. However, these schemes only find an approximation of the optimal solution.

In a recent work [31], a new implementable algorithm for fixed-slope lossy compression of discrete sources was proposed. Although the algorithm involves a minimization which resembles a specific realization of the generic cost proposed in [5], it is somewhat different. The reason is that the cost used in [31] cannot be derived directly from a lossless compression algorithm. The advantage of the new cost function is that it lends itself to rather naturally Gibbs simulated annealing in that the computational effort involved in each iteration is modest. It was shown that

using a universal lossless compressor to describe the reconstruction sequence found by the annealing 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 $k_n = o(\log n)$, there is no useful bound on the number of iterations required for convergence.

In this paper, motivated by the algorithm proposed in [31], we propose another approach to fixed-slope lossy compression of discrete sources. We start by making a linear approximation of the cost used in [31]. The cost assigned to each possible reconstruction sequence consists of a linear combination of two terms: a linear function of its empirical distribution plus its distance to (distortion from) the source sequence. We show that there exists proper coefficients such that minimizing the linearized cost function results in the same performance as would minimizing the original cost. The advantage of the modified cost is that its minimizer can be found simply using the Viterbi algorithm.

B. Organization of this paper

The organization of the paper is as follows. In Section II, the count matrix of a sequence and its empirical conditional entropy is introduced and some of their properties are studied. Section III reviews the fixed-slope universal lossy compression algorithm used in [31]. Section IV describes a new coding scheme for fixed-slope lossy compression derived by replacing part of the cost used in the mentioned exhaustive-search algorithm by a linear function. We prove that using appropriate coefficients for the linear function, the performance of the two algorithms remains the same. In Section V, a method for approximating these optimal coefficients is presented. This method, along with the result of the previous section, gives rise to a fixed-slope universal lossy compression algorithm that achieves the rate-distortion performance for any discrete stationary ergodic source. The advantage of this modified cost is discussed in Section VI where we show that the minimizer of the new cost can be found using the Viterbi algorithm. The method introduced for approximating the coefficients is computationally demanding, and hence is impractical. Therefore, in Section VII, we discuss a low-complexity iterative detour for approximating the coefficients. Section VIII presents some simulations results and, finally, Section IX concludes the paper with a discussion of some future directions.

II. Conditional empirical entropy and its properties

For any $y^n \in \mathcal{Y}^n$, let the $|\mathcal{Y}| \times |\mathcal{Y}|^k$ matrix $\mathbf{m}(y^n)$ denote its $(k+1)^{\text{th}}$ order empirical distribution[1]. For $\mathbf{b} = (b_1, \ldots, b_k) \in \mathcal{Y}^k$, and $\beta \in \mathcal{Y}$, the element in the β^{th} row and the $\mathbf{b}^{\text{th}}$ column of the matrix $\mathbf{m}$, $m_{\beta,\mathbf{b}}$, is defined as

$$m_{\beta,\mathbf{b}}(y^n) \triangleq \frac{1}{n-k} \left| \left\{ k+1 \le i \le n : y_{i-k}^{i-1} = \mathbf{b}, y_i = \beta] \right\} \right|. \tag{6}$$

[1]For any set $\mathcal{A}$, $|\mathcal{A}|$ denotes its size.

Based on the distribution induced by $\mathbf{m}(y^n)$, define the k^{th} order conditional empirical entropy of y^n, $H_k(y^n)$, as

$$H_k(y^n) \triangleq H(Z_{k+1}|Z^k), \tag{7}$$

where Z^{k+1} is assumed to be distributed according to $\mathbf{m}$, i.e.,

$$\mathrm{P}\left(Z^{k+1} = [b_1, \ldots, b_k, \beta] = [\mathbf{b}, \beta]\right) = m_{\beta,\mathbf{b}}(y^n). \tag{8}$$

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,

$$\mathcal{H}(\mathbf{v}) = \begin{cases} \sum_{i=1}^{\ell} \frac{v_i}{\|\mathbf{v}\|_1} \log \frac{\|\mathbf{v}\|_1}{v_i} & \text{if } \mathbf{v} \neq (0, \ldots, 0)^T \\ 0 & \text{if } \mathbf{v} = (0, \ldots, 0)^T, \end{cases} \tag{9}$$

where $0 \log(0) = 0$ by convention. With this notation, the conditional empirical entropy $H_k(y^n)$ defined in (7) is readily seen to be expressible in terms of $\mathbf{m}(y^n)$ as

$$H_k(y^n) \triangleq H(\mathbf{m}(y^n)) \triangleq \sum_{\mathbf{b}} \mathcal{H}\left(\mathbf{m}_{\cdot,\mathbf{b}}\right) \sum_{\beta \in \mathcal{Y}} m_{\beta,\mathbf{b}}, \tag{10}$$

where $\mathbf{m}_{\cdot,\mathbf{b}}$ denotes the column of $\mathbf{m}$ indexed by $\mathbf{b}$.

Remark 1: Note that $H_k(\cdot)$ has a discrete domain, while the domain of $H(\cdot)$ is continuous and consists of all $|\mathcal{Y}| \times |\mathcal{Y}|^k$ matrices with positive real entries adding up to one. In other words,

$$H_k : \mathcal{Y}^n \to [0, 1], \tag{11}$$

but

$$H : [0, 1]^{|\mathcal{Y}|} \times [0, 1]^{|\mathcal{Y}|^k} \to [0, 1]. \tag{12}$$

Conditional empirical entropy of sequences, $H_k(\cdot)$, plays key role in our results. Hence, in the following two subsections, we focus on this function, and study some of its properties.

A. Concavity

We prove that like the standard entropy function, conditional empirical entropy is also a concave function. By definition

$$H(\mathbf{m}) = \sum_{\mathbf{b} \in \mathcal{Y}^k} \left(\sum_{\beta \in \mathcal{Y}} m_{\beta,\mathbf{b}}\right) \mathcal{H}(\mathbf{m}_{\cdot,\mathbf{b}}), \tag{13}$$

where $\mathcal{H}(\cdot)$ is defined in (9). We need to show that for any $\theta \in [0, 1]$, and matrices $\mathbf{m}^{(1)}$ and $\mathbf{m}^{(2)}$ with non-negative components adding up to one,

$$\theta H(\mathbf{m}^{(1)}) + \bar{\theta} H(\mathbf{m}^{(2)}) \leq H(\theta \mathbf{m}^{(1)} + \bar{\theta} \mathbf{m}^{(2)}), \tag{14}$$

where $\bar{\theta} = 1 - \theta$. From the concavity of entropy function $\mathcal{H}$, it follows that

$$\begin{aligned}
&\theta(\sum_{\beta\in\mathcal{Y}} m^{(1)}_{\beta,\mathbf{b}})\mathcal{H}(\mathbf{m}^{(1)}_{\cdot,\mathbf{b}}) + \bar{\theta}(\sum_{\beta\in\mathcal{Y}} m^{(2)}_{\beta,\mathbf{b}})\mathcal{H}(\mathbf{m}^{(2)}_{\cdot,\mathbf{b}}) \\
&= (\theta(\sum_{\beta\in\mathcal{Y}} m^{(1)}_{\beta,\mathbf{b}}) + \bar{\theta}(\sum_{\beta\in\mathcal{Y}} m^{(2)}_{\beta,\mathbf{b}})) \sum_{i\in\{1,2\}} \frac{\theta_i(\sum_{\beta\in\mathcal{Y}} m^{(i)}_{\beta,\mathbf{b}})}{(\theta(\sum_{\beta\in\mathcal{Y}} m^{(1)}_{\beta,\mathbf{b}}) + \bar{\theta}(\sum_{\beta\in\mathcal{Y}} m^{(2)}_{\beta,\mathbf{b}}))}\mathcal{H}(\mathbf{m}^{(i)}_{\cdot,\mathbf{b}}) \\
&\leq (\theta(\sum_{\beta\in\mathcal{Y}} m^{(1)}_{\beta,\mathbf{b}}) + \bar{\theta}(\sum_{\beta\in\mathcal{Y}} m^{(2)}_{\beta,\mathbf{b}}))\mathcal{H}(\theta\mathbf{m}^{(1)}_{\cdot,\mathbf{b}} + \bar{\theta}\mathbf{m}^{(2)}_{\cdot,\mathbf{b}}), \qquad (15)
\end{aligned}$$

where $\theta_1 \triangleq 1 - \theta_2 \triangleq \theta$. Summing up both sides of (15) over all $\mathbf{b} \in \mathcal{Y}^k$ yields the desired result.

B. Stationarity condition

Let $p(y^{k+1})$ be a given pmf defined on $\mathcal{Y}^{k+1}$. Under what condition(s) does there exist a a stationary process with its $(k+1)^{\text{th}}$ order distribution equal to p?

Lemma 1: The necessary and sufficient condition for $\{p(y^{k+1})\}_{y^{k+1}\in\mathcal{Y}^{k+1}}$ to represent the $(k+1)^{\text{th}}$ order marginal distribution of a stationary process is

$$\sum_{\beta\in\mathcal{Y}} p(\beta, y^k) = \sum_{\beta\in\mathcal{Y}} p(y^k, \beta),\ \forall\ y^k \in \mathcal{Y}^k. \qquad (16)$$

Proof:

i. Necessity: The necessity of (16) is just a direct result of the stationarity of the process. If $p(y^{k+1})$ is to represent the $(k+1)^{\text{th}}$ order marginal distribution of a stationary process $\mathbf{Y} = \{Y_i\}$, then it should be consistent with the k^{th} order marginal distribution. Hence, (16) should hold.

ii. Sufficiency: In order to prove the sufficiency, we assume that (16) holds, and build a stationary process with $(k+1)^{\text{th}}$ order marginal distribution equal to $p(y^{k+1})$. Let $\mathbf{Y} = \{Y_i\}_i$ be a Markov chain of order k whose transition probabilities are defined as

$$\mathrm{P}(Y_{k+1} = y_{k+1}|Y^k = y^k) \triangleq q(y_{k+1}|y^k) \triangleq \frac{p(y^{k+1})}{p(y^k)}, \qquad (17)$$

where

$$p(y^k) \triangleq \sum_{\beta\in\mathcal{Y}} p(\beta, y^k) = \sum_{\beta\in\mathcal{Y}} p(y^k, \beta).$$

Now, given (16), it is easy to check that $p(y^{k+1})$ is the $(k+1)^{\text{th}}$ order stationary distribution of the defined Markov chain. Therefore, $\mathbf{Y}$ is a stationary process with the desired marginal distribution.

■

Throughout the paper, we refer to the condition stated in (16) as the *stationarity condition*.

Corollary 1: For any $|\mathcal{Y}| \times |\mathcal{Y}|^k$ matrix $\mathbf{m}$ corresponding to the $(k+1)^{\text{th}}$ order empirical distribution of some $y^n \in \mathcal{Y}^n$, there exists a stationary process whose marginal distribution coincides with $\mathbf{m}$.

Proof: From Lemma 1, we only need to show that (16) holds, i.e.,

$$\sum_{\beta\in\mathcal{Y}} m_{\beta,\mathbf{b}} = \sum_{\beta\in\mathcal{Y}} m_{b_k,[\beta,b_1\ldots,b_{k-1}]}, \forall\ \mathbf{b} \in \mathcal{Y}^k, \qquad (18)$$

which obviously holds because both sides of (18) are equal to $|\{i : y_{i+1}^{i+k} = \mathbf{b}\}|/(n-k)$. ■

III. Exhaustive search algorithm

Consider the following lossy source coding algorithm. Given $\alpha > 0$, for encoding sequence $x^n \in \mathcal{X}^n$, find

$$\hat{x}^n = \arg\min_{y^n \in \hat{\mathcal{X}}^n} [H_k(y^n) + \alpha d_n(x^n, y^n)], \tag{19}$$

and describe $\hat{x}^n$ using the Lempel-Ziv coding algorithm. As proved before [5], [31], the described algorithm is a universal lossy compression algorithm. That is, for any stationary ergodic source $\mathbf{X}$,

$$\frac{1}{n}\ell_{\mathrm{LZ}}(\hat{X}^n) + \alpha d_n(X^n, \hat{X}^n) \to \min[R(D, \mathbf{X}) + \alpha D], \quad \text{a.s.}, \tag{20}$$

where X^n is generated by the source $\mathbf{X}$, and $\hat{X}^n$ denotes the minimizer of (19) for the input X^n. Here ℓ_{LZ} denotes the length of the codeword assigned to $\hat{X}^n$ by the Lempel-Ziv algorithm [12]. Clearly, given the size of the search space, this is not an implementable algorithm. An approach for approximating the solution of (19) using Markov chain Monte Carlo methods has been suggested in [31]. One problem with the MCMC-based algorithms is that no useful bound is yet known on the required number of iterations. Moreover, the performance of the algorithm depends on the cooling process chosen. There exist cooling schedules with guaranteed convergence, but they are very slow, and usually not used in practice. On the other hand, if we use faster cooling processes, there is a risk of getting stuck in a local minima and missing the optimum solution. The goal of this paper is to propose a new approach for approximating the solution of (19). This new approach, as we show later, suggests a new implementable algorithm for lossy compression. The main idea here is using linear approximation of the conditional entropy function, $H(\mathbf{m})$, at some point $\mathbf{m}_0$, and proving that if $\mathbf{m}_0$ is chosen correctly, then while we have reduced the exhaustive search algorithm to the Viterbi algorithm, we have not changed its performance.

IV. Linearized cost function

Consider the problems (P1) and (P2) described by (21) and (22) respectively, where (P1) corresponds to the optimization required by the exhaustive search lossy compression scheme described in (19), and (P2) involves a similar optimization problem. The difference between (P1) and (P2) is that the term corresponding to conditional empirical entropy in (P1), which is a highly non-linear function of $\mathbf{m}$, is replaced by a linear function of $\mathbf{m}$.

$$\text{(P1)}: \quad \min_{y^n} \; \left[H(\mathbf{m}(y^n)) + \alpha d_n(x^n, y^n)\right], \tag{21}$$

and

$$\text{(P2)}: \quad \min_{y^n} \; \left[\sum_{\beta}\sum_{\mathbf{b}} \lambda_{\beta,\mathbf{b}} m_{\beta,\mathbf{b}}(y^n) + \alpha d_n(x^n, y^n)\right], \tag{22}$$

where $\{\lambda_{\beta,\mathbf{b}}\}_{\beta,\mathbf{b}}$ are a set of real-valued coefficients. In this section we are interested in answering the following question:

Is it possible to choose the 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 the minimizers of (P1)?

The reason we are interested in answering this question is that if the answer is affirmative, then instead of solving (P1) one can solve (P2), which we describe in Section VI can be done efficiently via the Viterbi algorithm.

Let $\mathcal{S}_1$ and $\mathcal{S}_2$ denote the set of minimizers of (P1) and (P2) respectively. Consider some $z^n \in \mathcal{S}_1$, and let $\mathbf{m}_n^* = \mathbf{m}(z^n)$, and let the coefficients used in (P2)

$$\begin{aligned}\lambda_{\beta,\mathbf{b}} &= \frac{\partial}{\partial m_{\beta,\mathbf{b}}} H(\mathbf{m})\Big|_{\mathbf{m}_n^*} \\ &= \log\left(\frac{\sum_{\beta'} m^*_{\beta',\mathbf{b}}}{m^*_{\beta,\mathbf{b}}}\right). \end{aligned} \tag{23}$$

Theorem 1: If the coefficients used in (P2) are chosen according to (23), then the minimum values of (P1) and (P2) will be the same. Moreover,

$$\mathcal{S}_2 \subset \mathcal{S}_1$$

and contains all the sequences $w^n \in \mathcal{S}_1$ with $\mathbf{m}(w^n) = \mathbf{m}_n^*$.

Proof: Since, as proved earlier, $H(\mathbf{m})$ is concave in $\mathbf{m}$, for any empirical count matrix $\mathbf{m}$, we have

$$H(\mathbf{m}) \le H(\mathbf{m}^*) + \sum_{\beta,\mathbf{b}} \frac{\partial}{\partial m_{\beta,\mathbf{b}}} H(\mathbf{m})\Big|_{\mathbf{m}_n^*} (m_{\beta,\mathbf{b}} - m^*_{\beta,\mathbf{b}}) \tag{24}$$

$$= H(\mathbf{m}^*) + \sum_{\beta,\mathbf{b}} \lambda_{\beta,\mathbf{b}} (m_{\beta,\mathbf{b}} - m^*_{\beta,\mathbf{b}}) \tag{25}$$

$$\triangleq \hat{H}(\mathbf{m}). \tag{26}$$

Adding a constant to the both sides of (26), we conclude that for any $y^n \in \hat{\mathcal{X}}^n$,

$$H(\mathbf{m}(y^n)) + \alpha d_n(x^n, y^n) \le \hat{H}(\mathbf{m}(y^n)) + \alpha d_n(x^n, y^n). \tag{27}$$

Taking the minimum of both sides of (27) yields

$$\min_{y^n}[H(\mathbf{m}(y^n)) + \alpha d_n(x^n, y^n)] \le \min_{y^n}[\hat{H}(\mathbf{m}(y^n)) + \alpha d_n(x^n, y^n)] \tag{28}$$

$$\le \hat{H}(\mathbf{m}(z^n)) + \alpha d_n(x^n, z^n) \tag{29}$$

$$= H(\mathbf{m}(z^n)) + \alpha d_n(x^n, z^n) \tag{30}$$

$$= \min_{y^n}[H(\mathbf{m}(y^n)) + \alpha d_n(x^n, y^n)], \tag{31}$$

because $z^n \in \mathcal{S}_1$. Therefore,

$$\min_{y^n}[H(\mathbf{m}(y^n)) + \alpha d_n(x^n, y^n)] = \min_{y^n}[\hat{H}(\mathbf{m}(y^n)) + \alpha d_n(x^n, y^n)], \tag{32}$$

i.e., (P1) and (P2) have the same minimum values.

For any sequence w^n with $\mathbf{m}(w^n) \neq \mathbf{m}_n^*$, by strict concavity of $H(\mathbf{m})$,

$$\hat{H}(\mathbf{m}(w^n)) + \alpha d_n(x^n, w^n) > H(\mathbf{m}(w^n)) + \alpha d_n(x^n, w^n), \tag{33}$$

$$\ge \min_{y^n}[H_k(y^n) + \alpha d_n(x^n, y^n)]. \tag{34}$$

Hence, the empirical count matrices of all the sequences in $\mathcal{S}_2$, i.e., all the minimizers of (P2) for the selected coefficients, are equal to $\mathbf{m}_n^*$.

Let $w^n \in \mathcal{S}_2$. We prove that $w^n \in \mathcal{S}_1$ as well. As we just proved, $\mathbf{m}(w^n) = \mathbf{m}(z^n) = \mathbf{m}_n^*$. Moreover, since both z^n and w^n belong to $\mathcal{S}_2$,

$$\begin{aligned}\min_{y^n}[\hat{H}(\mathbf{m}(y^n)) + \alpha d_n(x^n, y^n)] &= \hat{H}(\mathbf{m}(w^n)) + \alpha d_n(x^n, w^n) \\ &= \hat{H}(\mathbf{m}(z^n)) + \alpha d_n(x^n, z^n). \end{aligned} \tag{35}$$

Therefore, $d_n(x^n, w^n) = d_n(x^n, z^n)$, and consequently,

$$\begin{aligned} H_k(w^n) + \alpha d_n(w^n, x^n) &= H_k(z^n) + \alpha d_n(z^n, x^n), \\ &= \min_{y^n}[H_k(y^n) + \alpha d_n(y^n, x^n)], \end{aligned} \tag{36}$$

which proves that $w^n \in \mathcal{S}_1$, and concludes the proof. ■

Theorem 1 states that if the optimal type $\mathbf{m}_n^*$ is known, then the desired coefficients can be computed according to (23), and solving (P2) instead of (P1) using the computed coefficients finds a minimizer of (P1). In Section VI, we describe how (P2) can be solved efficiently using Viterbi algorithm for a given set of coefficients. The problem of course is that the optimal type $\mathbf{m}_n^*$ required for computing the desired coefficients is not known to the encoder (since knowledge of $\mathbf{m}_n^*$ seems to require solving (P1) which is the problem we are trying to avoid). In Section V, we introduce another optimization problem whose solution is a good approximation of $\mathbf{m}_n^*$, and hence of the desired coefficients $\{\lambda_{\beta,\mathbf{b}}\}$ when substituting in (23).

V. Computing the coefficients

As mentioned in the previous section, there exists a set of coefficients for which (P1) and (P2) have the same value. However, computing the desired coefficients requires the knowledge of $\mathbf{m}_n^*$ which is not available without solving (P1). In order to alleviate this issue, in this section we introduce another optimization problem that gives an asymptotically tight approximation of $\mathbf{m}_n^*$, and therefore a reasonable approximation of the set of coefficients.

For a given sequence x^n and a given order k, let $\mathcal{M}^{(k)} = \mathcal{M}^{(k)}(x^n)$ be the set of all jointly stationary probability distributions on $(X^k, \hat{X}^k)$ (in the sense of Lemma 1) such that their marginal distributions with respect to X coincides with the k^{th} order empirical distribution induced by x^n defined as follows

$$\begin{aligned}\hat{p}^{(k)}_{[x^n]}(a^k) &\triangleq \frac{|\{k+1 \le i \le n : (x_{i-k}, \ldots, x_{i-1}) = a^k\}|}{n-k}, \\ &= \frac{1}{n-k}\sum_{i=k+1}^{n} \mathbb{1}_{x_{i-k}^{i-1} = a^k}, \end{aligned} \tag{37}$$

where $a^k \in \mathcal{X}^k$. More specifically a distribution $p^{(k)}$ in $\mathcal{M}^{(k)}$ should satisfy the following two constraints:

1) Stationarity condition: as described in Section II-B, for any $a^{k-1} \in \mathcal{X}^{k-1}$ and $b^{k-1} \in \hat{\mathcal{X}}^{k-1}$,

$$\sum_{a_k \in \mathcal{X}, b_k \in \hat{\mathcal{X}}} p^{(k)}(a^k, b^k) = \sum_{a_k \in \mathcal{X}, b_k \in \hat{\mathcal{X}}} p^{(k)}(a_k a^{k-1}, b_k b^{k-1}). \tag{38}$$

2) Consistency: for each $a^k \in \mathcal{X}^k$,

$$\sum_{b^k \in \hat{\mathcal{X}}^k} p^{(k)}(a^k, b^k) = \hat{p}^{(k)}_{[x^n]}(a^k). \tag{39}$$

For given x^n, k and $\ell > k$, consider the following optimization problem

$$\begin{aligned} \min \quad & H(\hat{X}_{k+1}|\hat{X}^k) + \alpha \, \mathrm{E}\, d(X_1, \hat{X}_1) \\ \text{s.t.} \quad & (X^\ell, \hat{X}^\ell) \sim p^{(\ell)} \\ & p^{(\ell)} \in \mathcal{M}^{(\ell)}. \end{aligned} \tag{40}$$

Remark 2: Note that the rate-distortion function of a stationary ergodic process $\mathbf{X}$ has the following representation [32]:

$$\begin{aligned} R(D, \mathbf{X}) &= \inf\{\bar{H}(\hat{\mathbf{X}}) : \ (\mathbf{X}, \hat{\mathbf{X}}) \ \text{ jointly stationry and ergodic, and } \mathrm{E}\, d(X_0, \hat{X}_0) \leq D\}, \\ &= \inf_{k \geq 1} \inf\{H(\hat{X}_{k+1}|\hat{X}^k) : (\mathbf{X}, \hat{\mathbf{X}}) \text{ jointly stationary and ergodic, and } \mathrm{E}\, d(X_0, \hat{X}_0) \leq D\}, \end{aligned} \tag{41}$$

where $\bar{H}(\hat{\mathbf{X}})$ denotes the entropy rate of the stationary ergodic process $\hat{\mathbf{X}}$, i.e.,

$$\bar{H}(\hat{\mathbf{X}}) \triangleq \lim_{n \to \infty} H(\hat{X}_{n+1}|\hat{X}^n). \tag{42}$$

This representation gives the motivating intuition behind the optimization described in (40). It shows that (40) is basically performing the search required by (41).

Using the properties of the set $\mathcal{M}^{(\ell)}$, and the definition of conditional empirical entropy, (40) can be written more explicitly as

$$\begin{aligned} \min \quad & H(\mathbf{m}) + \alpha \sum_{a \in \mathcal{X}, b \in \hat{\mathcal{X}}} d(a,b) q(a,b) \\ \text{s.t.} \quad & 0 \leq p^{(\ell)}(a^\ell, b^\ell) \leq 1, \quad \forall \, \alpha^\ell \in \mathcal{X}^\ell, b^\ell \in \hat{\mathcal{X}}^\ell, \\ & \sum_{a^\ell, b^\ell} p^{(\ell)}(a^\ell, b^\ell) = 1, \quad \forall \, a^\ell \in \mathcal{X}^\ell, b^\ell \in \hat{\mathcal{X}}^\ell, \\ & \sum_{a_\ell \in \mathcal{X}, b_\ell \in \hat{\mathcal{X}}} p^{(\ell)}(a^\ell, b^\ell) = \sum_{a_\ell \in \mathcal{X}, b_\ell \in \hat{\mathcal{X}}} p^{(\ell)}(a_\ell a^{\ell-1}, b_k b^{\ell-1}), \\ & \qquad\qquad \forall \, a^{\ell-1} \in \mathcal{X}^{\ell-1}, b^{\ell-1} \in \hat{\mathcal{X}}^{\ell-1}, \\ & \sum_{b^\ell \in \hat{\mathcal{X}}^\ell} p^{(\ell)}(a^\ell, b^\ell) = \hat{p}^{(\ell)}_{[x^n]}(a^\ell) \quad \forall \, a^\ell \in \mathcal{X}^\ell, \\ & q(a,b) = \sum_{a^{\ell-1} \in \mathcal{X}^{\ell-1}, b^{\ell-1} \in \hat{\mathcal{X}}^{\ell-1}} p^{(\ell)}(a a^{\ell-1}, b b^{\ell-1}) \\ & m_{\beta, \mathbf{b}} = \sum_{a^\ell \in \mathcal{X}^\ell, b^{\ell-k} \in \hat{\mathcal{X}}^{\ell-k}} p^{(\ell)}(a^\ell, \mathbf{b} \beta b^{\ell-k}), \quad \forall \, \beta, \mathbf{b}. \end{aligned} \tag{43}$$

Note that the optimization in (43) is done over the joint distributions $p^{(\ell)}$ of $(X^\ell, \hat{X}^\ell)$. Let $\hat{\mathcal{P}}^*_n$ denote the set of minimizers of (43), and $\hat{\mathcal{S}}^*_n$ be their $(k+1)^{\text{th}}$ order marginalized versions with respect to $\hat{X}$. Let $\{\hat{\lambda}_{\beta, \mathbf{b}}\}_{\beta, \mathbf{b}}$ be the coefficients evaluated at some $\hat{\mathbf{m}}^*_n \in \hat{\mathcal{S}}^*_n$ using (23). Let $\mathbf{X}$ be a stationary ergodic source, and $R(\mathbf{X}, D)$ denote its rate distortion function. Finally, let $\hat{X}^n$ be the reconstruction sequence obtained by solving (P2) (recall (22)) at the evaluated coefficients.

Theorem 2: If $k = k_n = o(\log n)$, $\ell = \ell_n = o(n^{1/4})$ and $k = o(\ell)$ such that $k_n, \ell_n \to \infty$, as $n \to \infty$, then for any stationary ergodic source

$$H_k(\hat{X}^n) + \alpha d_n(X^n, \hat{X}^n) \stackrel{n\to\infty}{\longrightarrow} \min_{D\geq 0} \left[R(\mathbf{X}, D) + \alpha D\right], \quad \text{a.s.} \tag{44}$$

The proof of Theorem 2 is presented in Appendix A.

Remark 3: 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.

Remark 4: Note that all the constraints in (43) are linear, and the cost is a concave function. Hence, overall, we have a concave minimization problem (of dimension $|\mathcal{X}|^\ell|\hat{\mathcal{X}}|^\ell + |\hat{\mathcal{X}}|^{k+1} + |\mathcal{X}||\hat{\mathcal{X}}|$).

VI. VITERBI CODER

In this section, we show how, for a given set of coefficients, $\{\lambda_{\mathbf{b},\beta}\}$, (P2) can be solved efficiently via the Viterbi algorithm [33], [34].

Note that the linearized cost used in (P2) can also be written as

$$\sum_{\substack{\mathbf{b}\in\hat{\mathcal{X}}^k \\ \beta\in\hat{\mathcal{X}}}} [\lambda_{\beta,\mathbf{b}} m_{\beta,\mathbf{b}}(y^n) + \alpha d_n(x^n, y^n)] = \frac{1}{n}\sum_{i=1}^{n} \left[\lambda_{y_i, y_{i-k}^{i-1}} + \alpha d(x_i, y_i)\right]. \tag{45}$$

The advantage of this alternative representation is that, as we will describe, instead of using simulated annealing, we can find the sequence that exactly minimizes (45) via the Viterbi algorithm, which is a dynamic programming optimization method for finding the path of minimum weight in a Trellis diagram efficiently. For $i = k+1, \ldots, n$, let

$$s_i \triangleq y_{i-k}^i \tag{46}$$

to be the state at time i, and define $\mathcal{S}$ to be the set of all $|\hat{\mathcal{X}}|^{k+1}$ possible states. From this definition, the state at time i, s_i, is determined by the state at time $i-1$, s_{i-1}, and y_i. In other words, $s_i = g(s_{i-1}, y_i)$, 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. To the edge $e = (s', s)$ connecting states s' and $s = b^{k+1}$ at stage i, we assign the weight $w_i(e)$ defined as

$$w_i(e) := \lambda_{b_{k+1}, b^k} + \alpha d(x_i, b_{k+1}). \tag{47}$$

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 (45) 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 the 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

$$C_i(s) := \min_{s' \in \mathcal{L}(s)} [w_i((s', s)) + C_{i-1}(s')]. \tag{48}$$

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

$$s^* = \arg\min_{s \in \mathcal{S}} C_n(s). \tag{49}$$

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 the Viterbi algorithm. The problem is finding an approximation of the optimal coefficients. The procedure outlined in Section IV for finding the coefficients involves solving a concave minimization problem of dimension that becomes intractable even for moderate values of n. To bypass this process, an alternative heuristic method is proposed in the next section. The effectiveness of this approach is discussed in the next section through some simulations.

VII. Approximating the optimal coefficients

As we discussed in Section IV, having known the optimal coefficients, solving (P2) which can be done using the Viterbi algorithm is equivalent to solving (P1) which has exponential complexity in n. However, the problem is finding such desired coefficients. In Section V, it was proposed that for finding a good approximation of these coefficients, one method is to solve (43) and find $\hat{\mathbf{m}}^*$. Then an approximation of the coefficients $\{\lambda_{\beta,\mathbf{b}}\}$ can be made via (23) by evaluating the partial derivatives of $H(\mathbf{m})$ at $\hat{\mathbf{m}}^*$. But solving (43) requires solving a concave minimization problem of dimension which is demanding for even moderate values of n. Therefore, in this section, we consider a detour with moderate computational complexity.

First, assume that the desired distortion is small, or equivalently α is large. In that case, the distance between the original sequence x^n and its quantized version $\hat{x}^n$ should be small. Therefore, their types, i.e., their $(k+1)^{\text{th}}$ order empirical distributions, are close. Hence, the coefficients computed based on $\mathbf{m}(x^n)$ provide a reasonable approximation of the coefficients derived from $\mathbf{m}^*$. This implies that if our desired distortion is small, one possibility is to compute the type of the input sequence, and evaluate the coefficients at $\mathbf{m}(x^n)$.

In the case where the desired distortion is not very small, we can use an iterative approach as follows. Start with $\mathbf{m}(x^n)$. Compute the coefficients from (23) at $\mathbf{m}(x^n)$. Employ Viterbi algorithm to solve (P2) at the computed coefficients. Let $\hat{x}^n$ denote the output sequence. Compute $\mathbf{m}(\hat{x}^n)$, and recalculate the coefficients using (23) at $\mathbf{m}(\hat{x}^n)$. Again, use Viterbi algorithm to solve (P2) at the updated coefficients. Iterate.

For a conditional empirical distribution matrix $\mathbf{m}$, define its coefficient matrix as $\Lambda(\mathbf{m})$, where $\lambda_{\beta,\mathbf{b}}$ is defined as (23). For two matrices A and B of the same dimensions, define the scalar product of A and B as

$$A \odot B \triangleq \sum_{i,j} A_{i,j} B_{i,j}.$$

Now succinctly, the iterative approach can be described as follows. For $t = 0$, let $y^{n,(0)} = x^n$. For $t = 1, 2, \ldots$

$$\Lambda^{(t)} = \Lambda(\mathbf{m}(y^{n,(t-1)})),$$

$$y^{n,(t)} = \arg\min_{z^n \in \hat{\mathcal{X}}^n} [\Lambda^{(t)} \odot \mathbf{m}(z^n) + \alpha d(x^n, z^n)].$$

Stop as soon as $y^{n,(t)} = y^{n,(t-1)}$.

For a given sequence x^n, and slope α, assign to each sequence $y^n \in \hat{\mathcal{X}}^n$ the energy

$$\mathcal{E}(y^n) = H_k(y^n) + \alpha d(x^n, y^n). \tag{50}$$

As mentioned before, the goal is to find the sequence with minimum energy. Theorem 3 below gives some justification on how the described approach serves this purpose. It shows that, through the iterations, the energy level of the output is decreasing at each step. Moreover, since the number of energy levels is finite, it proves that the algorithm converges in a finite number of iterations.

Theorem 3: For the described iterative algorithm, at each $t \geq 1$,

$$\mathcal{E}(y^{n,(t+1)}) \leq \mathcal{E}(y^{n,(t)}). \tag{51}$$

Proof: For the ease of notations, let $\hat{x}^n = y^{n,(t)}$, $\hat{\mathbf{m}} = \mathbf{m}(\hat{x}^n)$, and $\hat{\Lambda} = \Lambda(\hat{\mathbf{m}})$. Similarly, let $\tilde{x}^n = y^{n,(t+1)}$, $\tilde{\mathbf{m}} = \mathbf{m}(\tilde{x}^n)$, and $\tilde{\Lambda} = \Lambda(\tilde{\mathbf{m}})$. From the concavity of $H(\mathbf{m})$ in $\mathbf{m}$,

$$H(\tilde{\mathbf{m}}) \leq H(\hat{\mathbf{m}}) + \hat{\Lambda} \odot (\tilde{\mathbf{m}} - \hat{\mathbf{m}}), \tag{52}$$

where $A \odot B$ with A and B two matrices of the same dimensions is equal to $\sum_{i,j} a_{i,j} b_{i,j}$. On the other hand

$$\begin{aligned} \hat{\Lambda} \odot \hat{\mathbf{m}} &= \sum_{\beta,\mathbf{b}} \hat{\lambda}_{\beta,\mathbf{b}} \hat{m}_{\beta,\mathbf{b}} \\ &= \sum_{\beta,\mathbf{b}} \hat{m}_{\beta,\mathbf{b}} \log \left(\frac{\sum_{\beta' \in \mathcal{X}} \hat{m}_{\beta',\mathbf{b}}}{\hat{m}_{\beta,\mathbf{b}}} \right) \qquad (53) \\ &= H(\hat{\mathbf{m}}). \qquad (54) \end{aligned}$$

Therefore, combining (52) and (53) yields

$$H(\tilde{\mathbf{m}}) \leq \hat{\Lambda} \odot \tilde{\mathbf{m}}. \tag{55}$$

Adding a constant term to the both sides of (56), we get

$$\mathcal{E}(\tilde{x}^n) = H(\tilde{\mathbf{m}}) + \alpha d(x^n, \tilde{x}^n) \leq \hat{\Lambda} \odot \tilde{\mathbf{m}} + \alpha d(x^n, \tilde{x}^n). \tag{56}$$

But, since $\tilde{x}^n$ is assumed to be a minimizer of (P2) for the computed coefficients,

$$\begin{aligned}\hat{\Lambda} \odot \tilde{\mathbf{m}} + \alpha d(x^n, \tilde{x}^n) &\leq \hat{\Lambda} \odot \hat{\mathbf{m}} + \alpha d(x^n, \hat{x}^n) \\ &= H(\hat{\mathbf{m}}) + \alpha d(x^n, \hat{x}^n) \\ &= \mathcal{E}(\hat{x}^n) \end{aligned} \tag{57}$$

Therefore, combining (56) and (57) yields the desired result, i.e.,

$$\mathcal{E}(\tilde{x}^n) \leq \mathcal{E}(\hat{x}^n). \tag{58}$$

■

Remark 5: In the described iterative algorithm, for any slope α, we assumed that the algorithm starts at $y^{n,(0)} = x^n$. However, as mentioned earlier, only for large values of α, $\mathbf{m}(x^n)$ provides a reasonable approximation of the desired type $\mathbf{m}_n^*$. Hence, in order to address this issue, we can slightly modify the algorithm as follows. The idea is that instead of starting at $y^{n,(0)} = x^n$ for all values of α, we can gradually decrease the slope to our desired value, and use the final output of each step as the initial point for the next step. More explicitly, for any given α_0, start from some large slope, $\alpha_{\max}$, (corresponding to very low distortion). Run the previous iterative algorithm and find $\hat{x}^n(\alpha_{\max})$. Pick some integer N_α, and define

$$\Delta\alpha \triangleq \frac{\alpha_{\max} - \alpha_0}{N_\alpha}.$$

Again run the iterative algorithm, but this time at $\alpha = \alpha_{\max} - \Delta\alpha$. Now, instead of starting from $y^{n,(0)} = x^n$, initialize $y^{n,(0)} = \hat{x}^n(\alpha_{\max})$. Repeat this process N_α times. I.e, At the r^{th} step, $r = 1, \ldots, N_\alpha$, run the algorithm at $\alpha = \alpha_{\max} - r\Delta\alpha$, and initialize $y^{n,(0)} = \hat{x}^n(\alpha_{\max} - (r-1)\Delta\alpha)$. At the final step $\alpha = \alpha_0$, and we have a reasonable quantized version of x^n for initialization.

To gain further insight on $(P2)$, for the coefficients matrix $\Lambda = \{\lambda_{\beta,\mathbf{b}}\}_{\beta,\mathbf{b}}$, define

$$\begin{aligned}\phi(\Lambda) &= \min_{y^n \in \hat{\mathcal{X}}^n} \left[\sum_{\beta,\mathbf{b}} \lambda_{\beta,\mathbf{b}} m_{\beta,\mathbf{b}}(y^n) + \alpha d_n(x^n, y^n) \right] \\ &= \min_{y^n \in \hat{\mathcal{X}}^n} \left[\Lambda \odot \mathbf{m}(y^n) + \alpha d_n(x^n, y^n) \right]. \end{aligned} \tag{59}$$

Since $\phi(\Lambda)$ is the minimum of multiple affine functions of Λ, it is a concave function. To each sequence $y^n \in \hat{\mathcal{X}}^n$, assign a coefficient matrix $\Lambda = [\lambda_{\beta,\mathbf{b}}]$ as

$$\lambda_{\beta,\mathbf{b}} = \left. \frac{\partial H(\mathbf{m})}{\partial m_{\beta,\mathbf{b}}} \right|_{\mathbf{m}(y^n)}. \tag{60}$$

Let $\mathcal{L}_d$ be the set of all such coefficient matrices. Similarly to each possible conditional distribution matrix $\mathbf{m}$ on $\hat{\mathcal{X}}^{k+1}$ which satisfies the stationarity condition defined in Section II-B, assign a coefficients matrix Λ defined according to (60). Let $\mathcal{L}_c$ be the set of coefficient matrices calculated at $(k+1)^{\text{th}}$ order stationary distributions on $\hat{\mathcal{X}}^{k+1}$. Note that while $\mathcal{L}_d$ is a discrete set (consisting of no more than $|\mathcal{Y}|^n$ elements), $\mathcal{L}_c$ is continuous.

For a sequence x^n, let

$$\hat{x}^n = \arg\min_{y^n \in \hat{\mathcal{X}}^n} \mathcal{E}(y^n),$$

and

$$\Lambda^* \triangleq \Lambda(\hat{x}^n).$$

Note that Λ^* is the optimal coefficients matrix required for replacing (P1) with (P2).

Lemma 2:

$$\Lambda^* = \arg\min_{\Lambda \in \mathcal{L}_d} \phi(\Lambda). \tag{61}$$

Proof: As shown before,

$$f(\hat{\Lambda}) = \mathcal{E}(\hat{x}^n). \tag{62}$$

On the other hand, if $\tilde{x}^n$ is the minimizer of $\sum_{\beta,\mathbf{b}} \lambda_{\beta,\mathbf{b}} m_{\beta,\mathbf{b}}(y^n) + \alpha d_n(x^n, y^n)$ for some $\Lambda \in \mathcal{L}_d$, then, as shown in the proof of Theorem 3,

$$H(\tilde{\mathbf{m}}) \leq \Lambda \odot \tilde{\mathbf{m}}. \tag{63}$$

Therefore, adding $d(x^n, \tilde{x}^n)$ to both sides of (63) yields

$$\mathcal{E}(\tilde{x}^n) \leq \phi(\Lambda). \tag{64}$$

But, by assumption,

$$\mathcal{E}(\hat{x}^n) \leq \mathcal{E}(\tilde{x}^n). \tag{65}$$

Combining (62), (64) and (65) yields the desired result. ■

Remark 6: Note that

$$\begin{aligned} &\min_{\Lambda \in \mathcal{L}_c} \min_{y^n} \left(\sum_{\beta,\mathbf{b}} \lambda_{\beta,\mathbf{b}} m_{\beta,\mathbf{b}}(y^n) + \alpha d_n(x^n, y^n) \right) \\ &= \min_{y^n} \left[\min_{\Lambda \in \mathcal{L}_c} \left(\sum_{\beta,\mathbf{b}} \lambda_{\beta,\mathbf{b}} m_{\beta,\mathbf{b}}(y^n) \right) + \alpha d_n(x^n, y^n) \right]. \end{aligned} \tag{66}$$

But $H(\mathbf{m}(y^n)) \leq \sum_{\beta,\mathbf{b}} \lambda_{\beta,\mathbf{b}} m_{\beta,\mathbf{b}}(y^n)$, for any $\Lambda \in \mathcal{L}_c$, and the lower bound is achieved at $\Lambda(y^n)$. Therefore,

$$\min_{\Lambda \in \mathcal{L}_c} \min_{y^n} \left(\sum_{\beta,\mathbf{b}} \lambda_{\beta,\mathbf{b}} m_{\beta,\mathbf{b}}(y^n) + \alpha d_n(x^n, y^n) \right) = \min_{y^n} (H_k(y^n) + \alpha d(x^n, y^n)). \tag{67}$$

Hence, we can replace $\mathcal{L}_d$ by $\mathcal{L}_c$ in (61), and still get the same result. This transform converts the discrete optimization stated in (61), which can be solved by exhaustive search, to an optimization over a continuous function of relativley low dimentions.

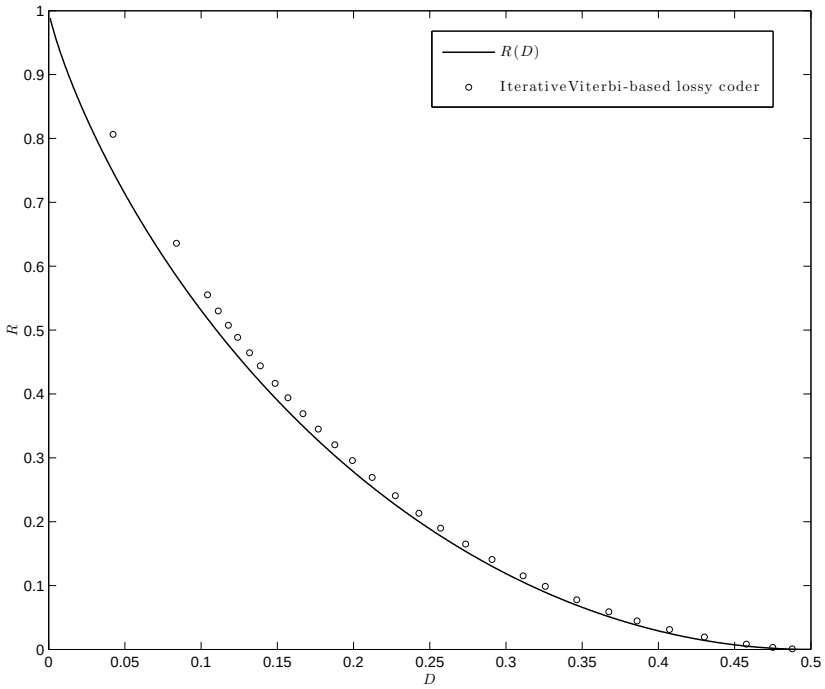

Fig. 1. Average performance of the iterative Viterbi-based lossy coder applied to an i.i.d. $\mathrm{Bern}(0.5)$ source. ($n = 10^4$, $k = 8$, $\alpha = (3, 2.9, \ldots, 0.1)$, and $L = 50$)

VIII. Simulation results

As the first example, consider an i.i.d. $\mathrm{Bern}(p)$ source with $p = 0.5$. Fig. 1 shows the performance of the iterative algorithm described in Section VII slightly modified, as suggested in Remark 5. The simulations parameters are as follows: $n = 10^4$, $k = 8$, and $\alpha = (3, 2.9, \ldots, 0.1)$. Each point corresponds to the average performance over $L = 50$ independent source realizations. As mentioned in Section VII, the iterative algorithm continues until there is no decrease in the cost. Fig. 2 shows the average, minimum and maximum number of required iterations before convergence versus α. Again, the number of trials are $L = 50$. It can be observed that the number of iterations in this case is always below 60, which, given the size of the search space, i.e, 2^n, shows fast convergence.

The next example involves a binary symmetric Markov source (BSMS) with transition probability $q = 0.2$. Fig. 3 compares the average performance of the Viterbi encoder against upper and lower bounds on $R(D)$ [35]. The reason for only comparing the performance of the algorithm against bounds on $R(D)$ in this case is that the rate-distortion function of a Markov source is not known, except for a low-distortion region. For low distortions, the Shannon lower bound is tight [36]. More explicitly, for $D \leq D_c \approx 0.0159$,

$$R(D) = H_b(q) - H_b(D),$$

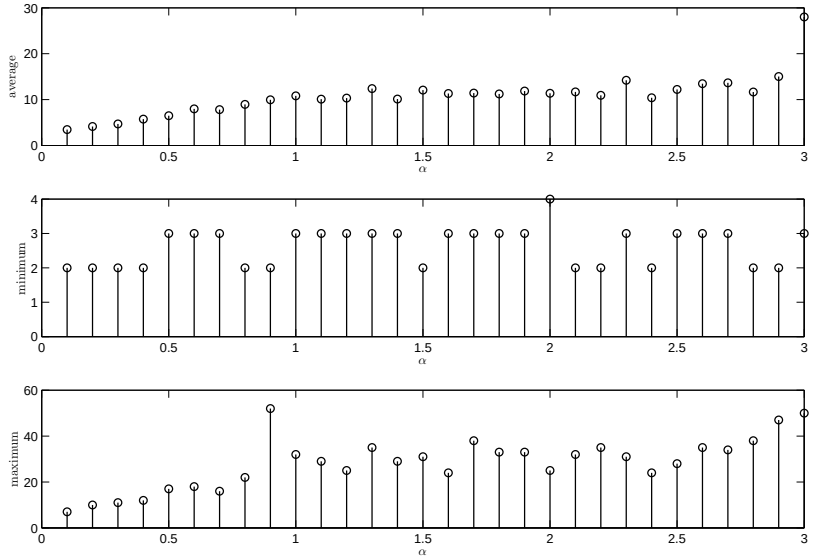

Fig. 2. From top to bottom: average, minimum and maximum number of iterations before convergence. (i.i.d. $\mathrm{Bern}(0.5)$ source, $n = 10^4$, $k = 8$, $\alpha = (3, 2.9, \ldots, 0.1)$, and $L = 50$)

where $H_b(\epsilon) \triangleq \mathcal{H}(\epsilon, 1-\epsilon)$. For $D > D_c$, $R(D) > H_b(q) - H_b(D)$.

A comparison with the memoryless case (Fig. 1) seems to suggest that the problem is less with how quickly (in n) we are converging to the exhaustive search performance scheme of (19) than with how quickly the convergence in (44) is taking place, which is source dependent and not at our control.

Fig. 4 shows the average number of iterations before convergence versus α. It can be observed that the average is always below 15. To give some examples on how the energy is decreasing, Fig. 5 and Fig. 6 show the energy decay through iterations for $\alpha = 1.6$ and $\alpha = 1$ respectively.

Remark 7: Similar to [31], here in the figures we are using $H_k(\hat{x}^n)$ as the rate, while in fact it is not a true length function. The reason is that as explained in [31], by Ziv inequality [37], if $k = o(\log(n))$, then for any $\epsilon > 0$, there exits $N_\epsilon \in \mathbb{N}$ such that for any $n > N_\epsilon$ and any sequence $= (y_1, y_2, \ldots)$,

$$\left[\frac{1}{n}\ell_{\mathsf{LZ}}(y^n) - H_k(y^n)\right] \leq \epsilon. \tag{68}$$

IX. Conclusions

In this paper, a new approach to for fixed-slope lossy compression of discrete sources is proposed. The core ingredient is the use of the Viterbi algorithm, which is a dynamic programing algorithm. It enables the encoder to find the reconstruction sequence with minimum cost. The encoder first assigns some weights to different contexts of length k, i.e, subsequences of length $k+1$, that appear within the reconstruction sequence. Then, the overall cost assigned to each possible reconstruction sequence is the sum of the weights of different contexts multiplied by

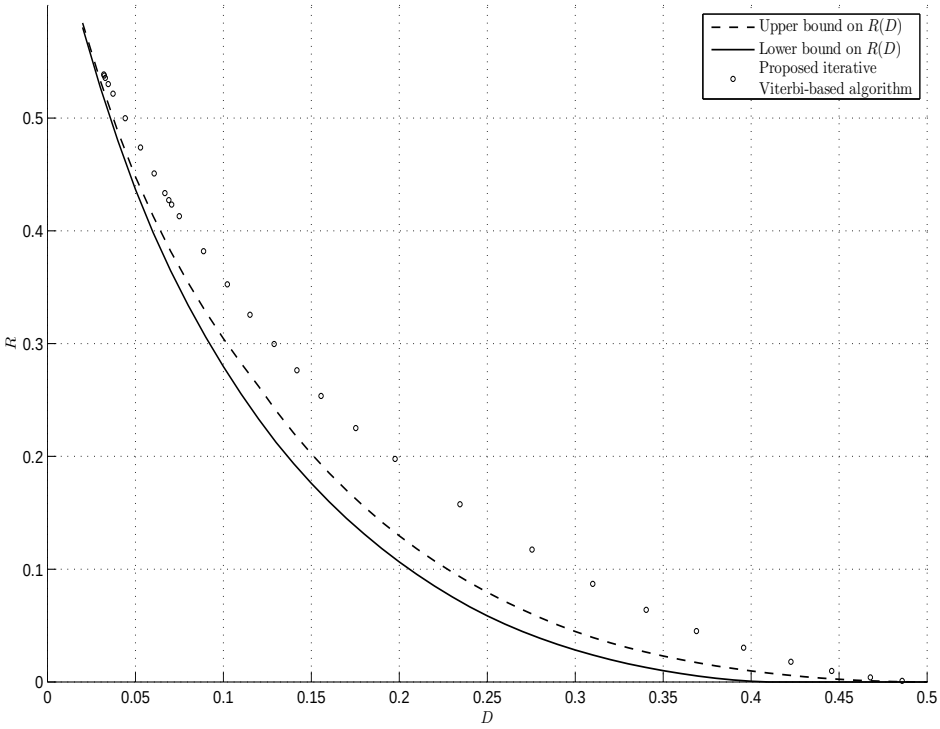

Fig. 3. Average performance of the iterative Viterbi-based lossy coder applied to a BSMS with $q = 0.2$ source. ($n = 25 \times 10^3$, $k = 8$, $\alpha = 3 : -0.1 : 0.1$ and $L = 50$)

their number of appearances in the sequence, plus some constant times the distance between the original sequence and the candidate reconstruction sequence. From this definition, it turns out that the state of the Viterbi algorithm at time t is the last k symbols observed plus the current symbol in the sequence, i.e, $(y_{t-k}, \ldots, y_t)$. Therefore, the Trellis has overall $|\hat{\mathcal{X}}|^{k+1}$ different states, corresponding to $|\hat{\mathcal{X}}|^{k+1}$ different possible contexts of length k. Hence for coding a sequence of length n, the computational complexity of the Viterbi algorithm will be of the order of $O(n2^{k+1})$. We prove that there exists a set of optimal coefficients for which the described algorithm will achieve the rate-distortion performance for any stationary ergodic process. The problem is finding those weights. We provide an optimization problem whose solution can be used to find an asymptotically tight approximation of the optimal coefficients resulting in an overall scheme which is universal with respect to the class of stationary ergodic sources. However, solving this optimization problem is computationally demanding, and in fact infeasible in practice for even moderate blocklengths. In order to overcome this problem, we propose an iterative approach for approximating the optimal coefficients. This approach is partially justified by a guarantee of convergance to at least a local minimum.

In the described iterative approach, the algorithm starts at a large slope (corresponding to a small distortion) and

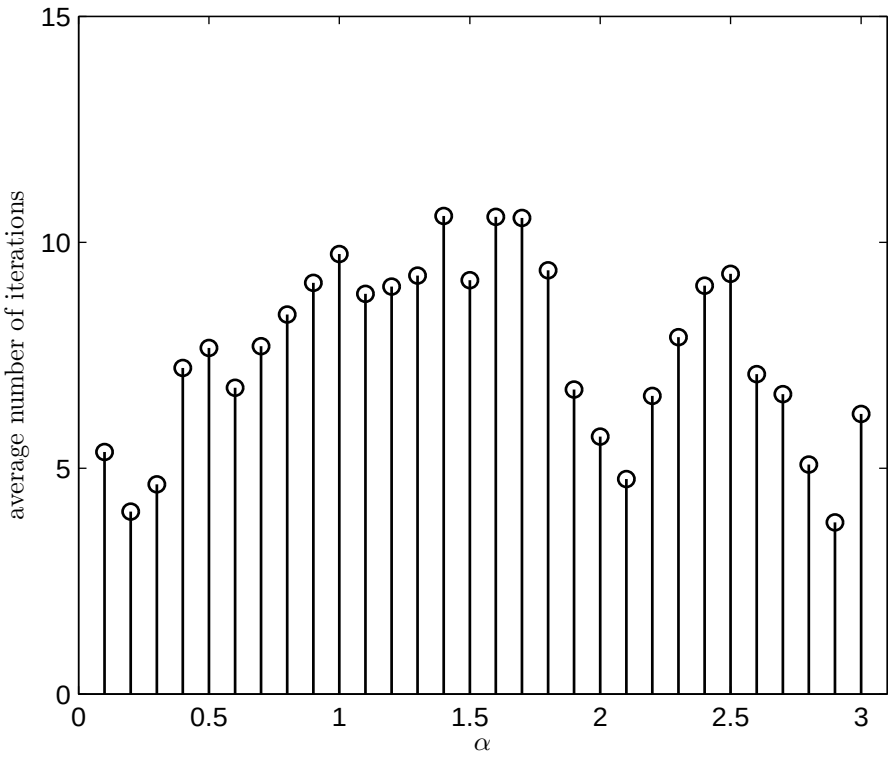

Fig. 4. Average number of iterations before convergence.(BSMS with $q = 0.2$, $n = 25 \times 10^3$, $k = 8$, $\alpha = 3 : -0.1 : 0.1$ and $L = 50$)

gradually decreases the slope until it hits the desired value. At each slope, the algorithm runs the Viterbi algorithm iteratively until it converges. An interesting possible next step is to explore whether there exisits a sequence of slopes converging to the desired value in a small number of steps (e.g. of $o(n)$) for which we can guarantee convergence of the algorithm to the global minimum at the end of the porcess. Existance of such sequence of slopes implies a universal lossy compression algorithm with moderate computatioal complexity.

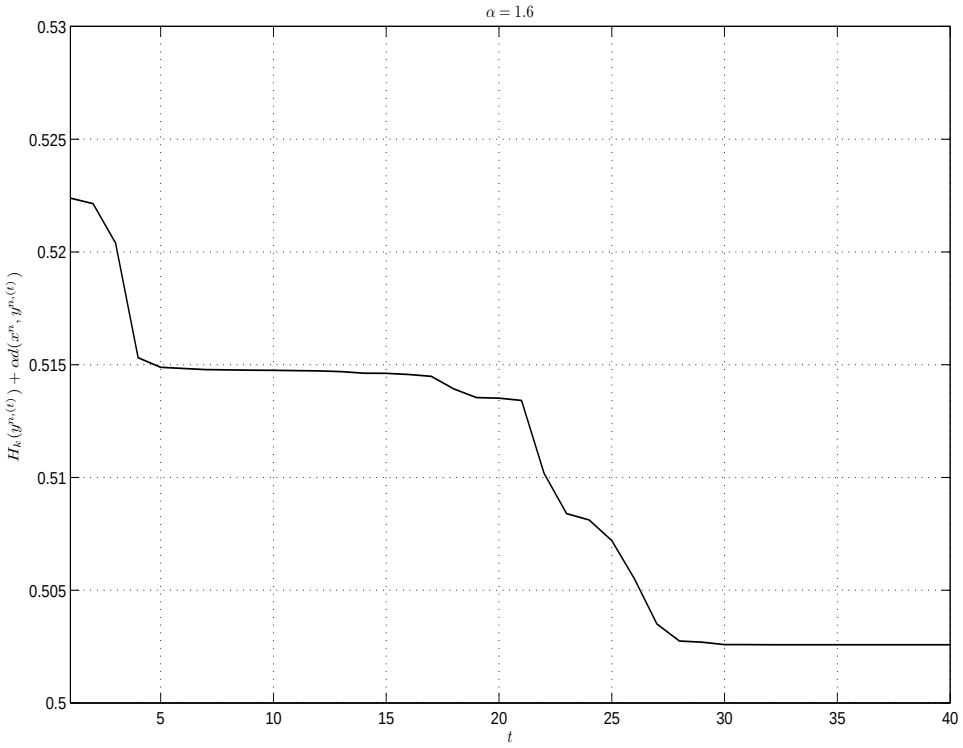

Fig. 5. Energy decay through the iterations for $\alpha = 1.6$. (BSMS with $q = 0.2$, $n = 25 \times 10^3$ and $k = 8$)

APPENDIX A: PROOF OF THEOREM 2

Proof: By rearranging the terms, the cost that is to be minimized in (P1) can alternatively be represented as follows

$$\begin{aligned}
H_k(y^n) + \alpha d_n(x^n, y^n) &= H_k(\mathbf{m}(y^n)) + \alpha \frac{1}{n} \sum_{i=1}^{n-k} d(x_i, y_i), \\
&= H_k(\mathbf{m}(y^n)) + \alpha \frac{1}{n} \sum_{i=k+1}^{n} d(x_i, y_i) \sum_{a \in \mathcal{X}, b \in \hat{\mathcal{X}}} \mathbb{1}_{x_i = a, y_i = b}, \\
&= H_k(\mathbf{m}(y^n)) + \alpha \frac{1}{n} \sum_{i=k+1}^{n} \sum_{a \in \mathcal{X}, b \in \hat{\mathcal{X}}} d(a, b) \mathbb{1}_{x_i = a, y_i = b}, \\
&= H_k(\mathbf{m}(y^n)) + \alpha \sum_{a \in \mathcal{X}, b \in \hat{\mathcal{X}}} d(a, b) \frac{1}{n} \sum_{i=k+1}^{n} \mathbb{1}_{x_i = a, y_i = b}, \\
&= H_k(\mathbf{m}(y^n)) + \alpha \sum_{a \in \mathcal{X}, b \in \hat{\mathcal{X}}} d(a, b) \hat{p}^{(1)}_{[x^n, y^n]}(a, b), \\
&= H_{\hat{p}^{(k+1)}_{[y^n]}}(Y_{k+1} | Y^k) + \alpha \, \mathrm{E}_{\hat{p}^{(1)}_{[x^n, y^n]}(\cdot, \cdot)} \, d(X_1, Y_1). \qquad \text{(A-1)}
\end{aligned}$$

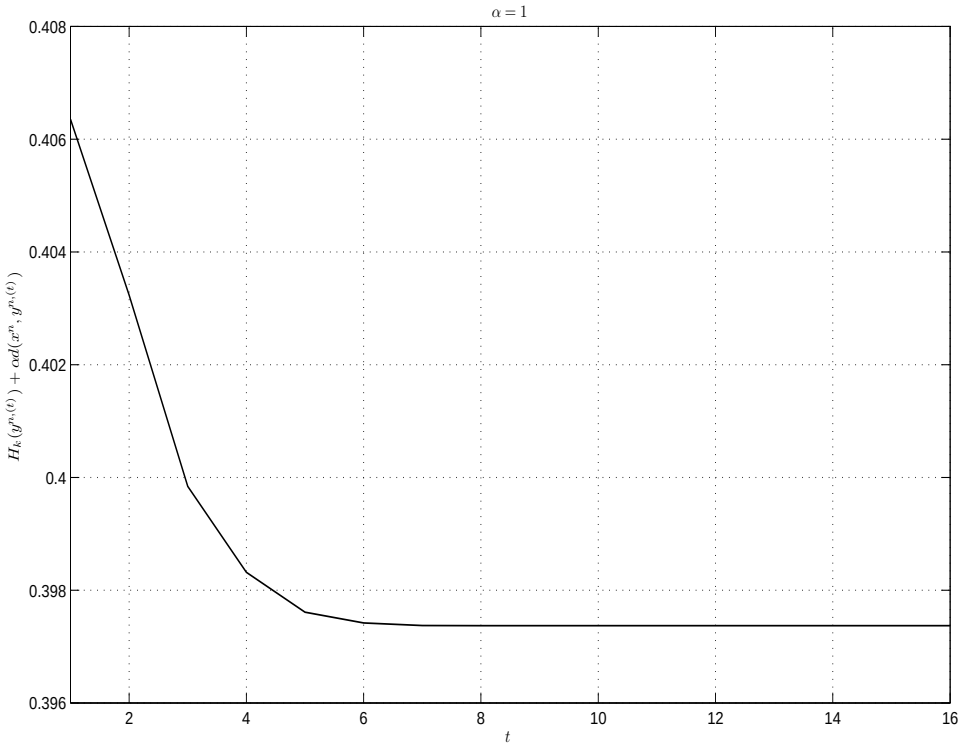

Fig. 6. Energy decay through the iterations for $\alpha = 1$. (BSMS with $q = 0.2$, $n = 25 \times 10^3$ and $k = 8$)

This new representation reveals the close connection between (P1) and (40). Although the costs we are trying to minimize in the two problems are equal, there is a fundamental difference between them: (P1) is a discrete optimization problem, while the optimization space in (40) is continuous.

Let $\mathcal{E}_n^*$ and $\mathcal{P}_n^*$ be the sets of minimizers of (P1), and joint empirical distributions of order ℓ, $\hat{p}^{(\ell)}_{[x^n,y^n]}(\cdot,\cdot)$, induced by them respectively. Also let $\mathcal{S}_n^*$ be the set of marginalized distributions of order $k+1$ in $\mathcal{P}_n^*$ with respect to Y. Finally, let C_n^* and $\hat{C}_n^*$ be the minimum values achieved by (P1) and (43) respectively.

In order to make the proof more tractable, we break it down into several steps as follows.

1) Let $y^n \in \mathcal{E}_n^*$, and $\hat{p}^{(\ell)}_{[x^n,y^n]}(\cdot,\cdot)$ be the induced joint empirical distribution. It is easy to check that $\hat{p}^{(\ell)}_{[x^n,y^n]}(\cdot,\cdot)$ satisfies all the constraints mentioned in (43). The only condition that might need some thought is the stationarity constraint, which also holds because

$$\sum_{a_\ell \in \mathcal{X}, b_\ell \in \hat{\mathcal{X}}} \hat{p}^{(\ell)}_{[x^n,y^n]}(a^\ell, b^\ell) = \frac{1}{n-\ell} \left| \left\{ i : x_{i-\ell+1}^{i-1} = a^{\ell-1}, y_{i-\ell+1}^{i-1} = b^{\ell-1} \right\} \right|,$$
$$= \sum_{a_\ell \in \mathcal{X}, b_\ell \in \hat{\mathcal{X}}} \hat{p}^{(\ell)}_{[x^n,y^n]}(a_\ell a^{\ell-1}, b_k b^{\ell-1}). \tag{A-2}$$

Therefore, since $\hat{C}_n^*$ is the minimum of (43), we have

$$\begin{aligned}\hat{C}_n^* &\leq H_k(\mathbf{m}(y^n)) + \alpha \, \mathrm{E}_{\hat{p}^{(1)}_{[x^n,y^n]}(\cdot,\cdot)}(X_{k+1}, Y_{k+1}) \\ &= H_k(\mathbf{m}(y^n)) + \alpha d_n(x^n, y^n) \\ &= C_n^*. \end{aligned} \tag{A-3}$$

2) Let $p^{*(\ell)} \in \hat{\mathcal{P}}_n^*$. Based on this joint probability distribution and x^n, we construct a reconstruction sequence $\tilde{X}^n$ as follows: divide x^n into $r = \lceil \frac{n}{\ell} \rceil$ consecutive blocks:

$$x^{\ell}, x_{\ell+1}^{2\ell}, \ldots, x_{(r-2)\ell+1}^{(r-1)\ell}, x_{(r-1)\ell+1}^{n},$$

where except for possibly the last block, the other blocks have length ℓ. The new sequence is constructed as follows

$$\tilde{X}^{\ell}, \tilde{X}_{\ell+1}^{2\ell}, \ldots, \tilde{X}_{(r-2)\ell+1}^{(r-1)\ell}, \tilde{X}_{(r-1)\ell+1}^{n},$$

where for $i = 1, \ldots, r-1$, $\tilde{X}_{(i-1)\ell+1}^{i\ell}$ is a sample from the conditional distribution $p^{*(\ell)}(\hat{X}^{\ell} | X^{\ell} = x_{(i-1)\ell+1}^{i\ell})$, and $\tilde{X}_{(r-1)\ell+1}^{n} \sim p^{*(\ell)}(\hat{X}_{(r-1)\ell+1}^{n} | X_{(r-1)\ell+1}^{n} = x_{(r-1)\ell+1}^{n})$.

3) Assume that $\mathbf{x} = \{x_i\}_{i=1}^{\infty}$ is a given individual sequence. For each n, let $p^{*(k+1)}$ be the $(k+1)^{\text{th}}$ order marginalized version of the solution of (43) on $\hat{\mathcal{X}}^{(k+1)}$. Moreover, let $\tilde{X}^n$ be the constructed as described in the previous item, and $\hat{p}_{[\tilde{X}^n]}^{(k+1)}$ be the $(k+1)^{\text{th}}$ order empirical distribution induced by $\tilde{X}^n$. We now prove that

$$\| p^{*(k+1)} - \hat{p}_{[\tilde{X}^n]}^{(k+1)} \|_1 \to 0, \quad \text{a.s.}, \tag{A-4}$$

where the randomization in (A-4) is only in the generation of $\tilde{X}^n$.

Remark 8: Since $p^{*(\ell)}$ satisfies *stationarity condition*, its $(k+1)^{\text{th}}$ order marginalized distribution, $p^{*(k+1)}$, is well-defined and can be computed with respect to any of the $(k+1)$ consecutive positions in $1, \ldots, \ell$. In other words for $a^{k+1} \in \hat{\mathcal{X}}^{k+1}$,

$$p^{*(k+1)}(a^{k+1}) = \sum_{b^{\ell-k-1} \in \hat{\mathcal{X}}^n} p^{*(k+1)}(b^j a^{k+1} b_{j+1}^{\ell-k-1}), \tag{A-5}$$

for any $j \in \{0, \ldots, \ell-k-1\}$, and the result does not depend on the choice of j.

In order to show that the difference between $\hat{p}_{[\tilde{X}^n]}^{(k+1)}(a^{k+1})$ and $p^{*(k+1)}(a^{k+1})$ is going to zero almost surely, we decompose $\hat{p}_{[\tilde{X}^n]}^{(k+1)}(a^{k+1})$ into the average of $\ell - k$ terms each of which is converging to $p^{*(k+1)}(a^{k+1})$. Then using the union bound we get the desired result which is the convergence of $\hat{p}_{[\tilde{X}^n]}^{(k+1)}(a^{k+1})$ to $p^{*(k+1)}(a^{k+1})$.

For $a^{k+1} \in \hat{\mathcal{X}}^{k+1}$,

$$\begin{aligned}
&\left|\hat{p}^{(k+1)}_{[\tilde{X}^n]}(a^{k+1}) - p^{*(k+1)}(a^{k+1})\right| \\
&= \left|\frac{1}{n-k-1}\sum_{i=k+1}^{n} \mathbb{1}_{\tilde{X}^i_{i-k}=a^{k+1}} - p^{*(k+1)}(a^{k+1})\right| \\
&= \left|\frac{1}{n-k-1}\sum_{j=0}^{\ell-k-1}\sum_{i=1}^{r-1} \mathbb{1}_{\tilde{X}^{i\ell-j}_{i\ell-j-k}=a^{k+1}} + \delta_\ell - p^{*(k+1)}(a^{k+1})\right|, \\
&= \left|\frac{r}{n-k-1}\sum_{j=0}^{\ell-k-1}\left[\frac{1}{r}\sum_{i=1}^{r-1} \mathbb{1}_{\tilde{X}^{i\ell-j}_{i\ell-j-k}=a^{k+1}}\right] + \delta_\ell - p^{*(k+1)}(a^{k+1})\right|, \\
&= \left|\frac{1}{\ell-k-1}\sum_{j=0}^{\ell-k-1}\left[\frac{1}{r}\sum_{i=1}^{r-1} \mathbb{1}_{\tilde{X}^{i\ell-j}_{i\ell-j-k}=a^{k+1}}\right] + \delta'_{n,\ell} - p^{*(k+1)}(a^{k+1})\right|, \qquad \text{(A-6)}
\end{aligned}$$

where δ_ℓ accounts for the edge effects between the blocks, and $\delta'_{n,\ell}$ is defined such that $\delta'_{n,\ell} - \delta_\ell$ takes care of the effect of replacing $\frac{r}{n-k-1}$ with $\frac{1}{\ell-k-1}$. Therefore, $0 \le \delta_\ell \le \frac{4k}{\ell}$, $\delta_\ell \to 0$ as $\ell \to \infty$, and finally $|\delta'_{n,\ell} - \delta_\ell| = o(k/\ell)$.

The interesting point about the new representation is that it decomposes a sequence of correlated random variables, $\{\mathbb{1}_{\tilde{X}^i_{i-k}=a^{k+1}}\}_{i=k+1}^n$, into $\ell-k$ sub-sequences where each of them is an independent process. For achieving this some counts that lie between two blocks are ignored, i.e., if $\mathbb{1}_{\tilde{X}^i_{i-k}=a^{k+1}}$ is such that it depends on more than one block of the form $\tilde{X}^{i\ell}_{(i-1)\ell+1}$, we ignore it. The effect of such ignored counts will be no more than δ_r which goes to zero as $k, \ell \to \infty$ because the Theorem requires $k = o(\ell)$. More specifically in (A-6), $\{\mathbb{1}_{\tilde{X}^{i\ell-j}_{i\ell-j-k}=a^{k+1}}\}_{i=1}^r$ is a sequence of independent random variables for each $j \in \{0, \ldots, \ell-k-1\}$. For n large enough, $|\delta'_{n,\ell}| < \epsilon/2$. Therefore, by Hoeffding inequality [38], and the union bound,

$$\begin{aligned}
&\mathrm{P}\left(\left|\hat{p}^{(k+1)}_{[\tilde{X}^n]}(a^{k+1}) - p^{*(k+1)}(a^{k+1})\right| > \epsilon\right), \\
&\le \mathrm{P}\left(\left|\frac{1}{\ell-k-1}\sum_{j=0}^{\ell-k-1}\left[\frac{1}{r}\sum_{i=1}^{r-1} \mathbb{1}_{\tilde{X}^{i\ell-j}_{i\ell-j-k}=a^{k+1}} - p^{*(k+1)}(a^{k+1})\right]\right| > \frac{\epsilon}{2}\right), \\
&\le \mathrm{P}\left(\frac{1}{\ell-k-1}\sum_{j=0}^{\ell-k-1}\left|\frac{1}{r}\sum_{i=1}^{r-1} \mathbb{1}_{\tilde{X}^{i\ell-j}_{i\ell-j-k}=a^{k+1}} - p^{*(k+1)}(a^{k+1})\right| > \frac{\epsilon}{2}\right), \\
&\le \sum_{j=0}^{\ell-k-1}\mathrm{P}\left(\left|\frac{1}{r}\sum_{i=1}^{r-1} \mathbb{1}_{\tilde{X}^{i\ell-j}_{i\ell-j-k}=a^{k+1}} - p^{*(k+1)}(a^{k+1})\right| > \frac{\epsilon}{2}\right), \\
&\le 2(\ell-k)e^{-r\epsilon^2/2}. \qquad \text{(A-7)}
\end{aligned}$$

Now again by applying the union bound,

$$\begin{aligned}
&\mathrm{P}\left(\|\hat{p}^{(k+1)}_{[\tilde{X}^n]} - p^{*(k+1)}\|_1 > \epsilon\right) \\
&\qquad \le \sum_{a^{k+1}\in\hat{\mathcal{X}}^{k+1}} \mathrm{P}\left(\left|\hat{p}^{(k+1)}_{[\tilde{X}^n]}(a^{k+1}) - p^{*(k+1)}(a^{k+1})\right| > \frac{\epsilon}{|\hat{\mathcal{X}}|^{k+1}}\right), \\
&\qquad \le |\hat{\mathcal{X}}|^{k+1} 2(\ell-k) e^{-\frac{n\epsilon^2}{2\ell|\hat{\mathcal{X}}|^{2(k+1)}}}. \qquad \text{(A-8)}
\end{aligned}$$

Our choices of $k = k_n = o(\log n)$, $\ell = \ell_n = o(n^{1/4})$, $k = o(\ell)$, and $k_n, \ell_n \to \infty$, as $n \to \infty$ now guarantee that the right hand side of (A-8) is summable on n which together with Borel-Cantelli Lemma yields the desired result of (A-4).

4) Using similar steps as above we can prove that

$$\|q^* - \hat{q}^{(1)}_{[x^n,\tilde{X}^n]}\| \to 0, \quad \text{a.s.} \qquad \text{(A-9)}$$

Again we first prove that $|q^*(a,b) - \hat{q}^{(1)}_{[x^n,\tilde{X}^n]}(a,b)| \to 0$ for each $a \in \mathcal{X}$ and $b \in \hat{\mathcal{X}}$. For doing this we again need to decompose

$$\{\mathbb{1}_{x_i=a,\tilde{X}_i=b}\}_{i=1}^n$$

into ℓ sub-sequences each of which is a sequence of independent random variables, and then apply Hoeffding inequality plus the union bound. Finally we apply the union bound again in addition to the Borel-Cantelli Lemma to get our desired result.

5) Combing the results of the last two parts, and the fact that $H_k(\mathbf{m})$ and $\mathrm{E}_q\, d(X,Y)$ are bounded continuous functions of $\mathbf{m}$ and $q(\cdot,\cdot)$ respectively, we conclude that

$$\begin{aligned}
H_k(\tilde{X}^n) + \alpha d_n(x^n,\tilde{X}^n) &= H_{\hat{p}^{(k+1)}_{[\tilde{X}^n]}}(Y_{k+1}|Y^k) + \alpha\, \mathrm{E}_{\hat{q}^{(1)}_{[x^n,\tilde{X}^n]}} d(X_1,Y_1) \\
&= H_{p^{*(k+1)}}(Y_{k+1}|Y^k) + \alpha\, \mathrm{E}_{q^*}\, d(X_1,Y_1) + \epsilon_n \\
&= \hat{C}^*_n + \epsilon_n, \qquad \text{(A-10)}
\end{aligned}$$

where $\epsilon_n \to 0$ with probability 1.

6) Since C^*_n is the minimum of (P1), we have

$$\begin{aligned}
C^*_n &\le H_k(\tilde{X}^n) + \alpha d_n(x^n,\tilde{X}^n), \\
&= \hat{C}^*_n + \epsilon_n. \qquad \text{(A-11)}
\end{aligned}$$

On the other hand, as shown in (A-3), $\hat{C}^*_n \le C^*_n$. Therefore,

$$|C^*_n - \hat{C}^*_n| \to 0 \qquad \text{(A-12)}$$

as $n \to \infty$.

7) For a given set of coefficients $\boldsymbol{\lambda} = \{\lambda_{\beta,\mathbf{b}}\}_{\beta,\mathbf{b}}$ computed at some $\mathbf{m}$ according to (23), define

$$f(\boldsymbol{\lambda}) = \min_{y^n\in\hat{\mathcal{X}}^n} \left[\sum_{\beta,\mathbf{b}} \lambda_{\beta,\mathbf{b}} m_{\beta,\mathbf{b}}(y^n) + \alpha d_n(x^n,y^n)\right]. \qquad \text{(A-13)}$$

It is easy to check that f is continuous, and bounded by $1+\alpha$. Therefore, since $\boldsymbol{\lambda}$ is in turns a continuous function of $\mathbf{m}$, and as proved in (A-4),

$$\|p^{*(k+1)} - \hat{p}^{(k+1)}_{[\tilde{X}^n]}\|_1 \to 0,$$

we conclude that,

$$|f(\boldsymbol{\lambda}^*) - f(\hat{\boldsymbol{\lambda}})| \to 0, \tag{A-14}$$

where $\boldsymbol{\lambda}^*$ and $\hat{\boldsymbol{\lambda}}$ are the coefficients computed at $p^{*(k+1)}$ and $\hat{p}^{(k+1)}_{[\tilde{X}^n]}$ respectively.

8) Let $\bar{X}^n$ be the output of (P2) when the coefficients are computed at $\mathbf{m}(\tilde{X}^n)$. Then, from Theorem 3,

$$\begin{aligned} H_k(\bar{X}^n) + \alpha d_n(x^n, \bar{X}^n) &\le H_k(\tilde{X}^n) + \alpha d_n(x^n, \tilde{X}^n) \\ &= \hat{C}^*_n + \epsilon_n. \end{aligned} \tag{A-15}$$

Since, $\epsilon_n \to 0$, this shows that haven computed the coefficients at $\mathbf{m}(\tilde{X}^n)$, we would get a universal lossy compressor. But instead, we want to compute the coefficients at $\mathbf{m}^*$. From (A-14), the difference between the performances of these two algorithms goes to zero. Therefore, we finally get our desired result which is

$$\left[H_k(\hat{X}^n) + \alpha d_n(X^n, \hat{X}^n)\right] \overset{n\to\infty}{\longrightarrow} \min_{D\ge 0} \left[R(\mathbf{X}, D) + \alpha D\right], \quad \text{a.s.} \tag{A-16}$$

■

REFERENCES

[1] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, New York, 2nd edition, 2006.

[2] C. Shannon. Coding theorems for a discrete source with fidelity criterion. In R. Machol, editor, *Information and Decision Processes*, pages 93–126. McGraw-Hill, 1960.

[3] R.G. Gallager. *Information Theory and Reliable Communication*. NY: John Wiley, 1968.

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

[5] En hui Yang, Z. Zhang, and T. Berger. Fixed-slope universal lossy data compression. *Information Theory, IEEE Transactions on*, 43(5):1465–1476, Sep 1997.

[6] D. J. Sakrison. The rate of a class of random processes. *Information Theory, IEEE Transactions on*, 16:10–16, Jan. 1970.

[7] J. Ziv. Coding of sources with unknown statistics part ii: Distortion relative to a fidelity criterion. *Information Theory, IEEE Transactions on*, 18:389–394, May 1972.

[8] D. L. Neuhoff, R. M. Gray, and L.D. Davisson. Fixed rate universal block source coding with a fidelity criterion. *Information Theory, IEEE Transactions on*, 21:511–523, May 1972.

[9] D. L. Neuhoff and P. L. Shields. Fixed-rate universal codes for Markov sources. *Information Theory, IEEE Transactions on*, 24:360–367, May 1978.

[10] J. Ziv. Distortion-rate theory for individual sequences. *Information Theory, IEEE Transactions on*, 24:137–143, Jan. 1980.

[11] R. Garcia-Munoz and D. L. Neuhoff. Strong universal source coding subject to a rate-distortion constraint. *Information Theory, IEEE Transactions on*, 28:285295, Mar. 1982.

[12] J. Ziv and A. Lempel. Compression of individual sequences via variable-rate coding. *Information Theory, IEEE Transactions on*, 24(5):530–536, Sep 1978.

[13] I. H. Witten, R. M. Neal, , and J. G. Cleary. Arithmetic coding for data compression. *Commun. Assoc. Comp. Mach.*, 30(6):520–540, 1987.

[14] K. Cheung and V. K. Wei. A locally adaptive source coding scheme. *Proc. Bilkent Conf on New Trends in Communication, Control, and Signal Processing*, pages 1473–1482, 1990.

[15] H. Morita and K. Kobayashi. An extension of LZW coding algorithm to source coding subject to a fidelity criterion. In *In Proc. 4th Joint Swedish-Soviet Int. Workshop on Information Theory*, page 105109, Gotland, Sweden, 1989.

[16] Y. Steinberg and M. Gutman. An algorithm for source coding subject to a fidelity criterion based on string matching. *Information Theory, IEEE Transactions on*, 39:877886, Mar. 1993.

[17] En hui Yang and J.C. Kieffer. On the performance of data compression algorithms based upon string matching. *Information Theory, IEEE Transactions on*, 44(1):47 –65, jan 1998.

[18] T. Luczak and T. Szpankowski. A suboptimal lossy data compression based on approximate pattern matching. *Information Theory, IEEE Transactions on*, 43:14391451, Sep. 1997.

[19] R. Zamir and K. Rose. Natural type selection in adaptive lossy compression. *Information Theory, IEEE Transactions on*, 47(1):99 –111, jan 2001.

[20] W. Szpankowski Atallah, Y. Génin. Pattern matching image compression: algorithmic and empirical results. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21:618627, Sept. 1999.

[21] Amir Dembo and Ioannis Kontoyiannis. The asymptotics of waiting times between stationary processes, allowing distortion. *The Annals of Applied Probability*, 9(2):413–429, May 1999.

[22] M. W. Marcellin and T. Fischer. Trellis coded quantization of memoryless and Gauss-Markov sources. *IEEE Trans. on Comm.*, 38(1):82–93, jan 1990.

[23] T. Berger and J.D. Gibson. Lossy source coding. *Information Theory, IEEE Transactions on*, 44(6):2690–2723, Sep 1998.

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

[25] J.H. Kasner, M.W. Marcellin, and B.R. Hunt. Universal trellis coded quantization. *Image Processing, IEEE Transactions on*, 8(12):1677 –1687, dec 1999.

[26] M.J. Wainwright and E. Maneva. Lossy source encoding via message-passing and decimation over generalized codewords of LDGM codes. In *Proc. IEEE Int. Symp. Inform. Theory*, pages 1493–1497, Sept. 2005.

[27] A. Gupta and S. Verdú. Nonlinear sparse-graph codes for lossy compression. *Information Theory, IEEE Transactions on*, 55(5):1961 –1975, may 2009.

[28] A. Gupta, S. S. Verdú, and T. Weissman. Rate-distortion in near-linear time. In *Proc. IEEE Int. Symp. Inform. Theory*, pages 847–851, Toronto, Canada, July 2008.

[29] E. Arikan. Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *arXiv:0807.3917*.

[30] S. Babu Korada and R. Urbanke. Polar codes are optimal for lossy source coding. *arXiv:0903.0307*.

[31] S. Jalali and T. Weissman. Rate-distortion via Markov chain Monte Carlo. *arXiv:0808.4156v2*.

[32] R. Gray, D. Neuhoff, and J. Omura. Process definitions of distortion-rate functions and source coding theorems. *Information Theory, IEEE Transactions on*, 21(5):524–532, Sep 1975.

[33] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260 – 269, apr 1967.

[34] Jr. Forney, G.D. The Viterbi algorithm. *Proceedings of the IEEE*, 61(3):268 – 278, march 1973.

[35] S. Jalali and T. Weissman. New bounds on the rate-distortion function of a binary Markov source. In *Proc. IEEE Int. Symp. Inform. Theory*, Nice, France, July 2007.

[36] R. Gray. Rate distortion functions for finite-state finite-alphabet markov sources. *Information Theory, IEEE Transactions on*, 17(2):127–134, Mar 1971.

[37] E. Plotnik, M.J. Weinberger, and J. Ziv. Upper bounds on the probability of sequences emitted by finite-state sources and on the redundancy of the Lempel-Ziv algorithm. *Information Theory, IEEE Transactions on*, 38(1):66–72, Jan 1992.

[38] W. Hoeffding. Probability inequalities for sums of bounded random vaiables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.