2023 Data Compression Conference (DCC) | 979-8-3503-4795-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/DCC55655.2023.00039

On Stochastic Codebook Generation for Markov Sources

Ahmed Elshafiy, and Kenneth Rose
University of California, Santa Barbara
Santa Barbara, CA, 93117, USA
{a_elshafiy,rose}@ece.ucsb.edu

Abstract

This paper proposes an effective universal "on-the-fly" mechanism for stochastic codebook generation in lossy coding of Markov sources. Earlier work has shown that the rate-distortion bound can be asymptotically achieved by a "natural type selection" (NTS) mechanism that iteratively considers asymptotically long source strings (from an unknown distribution P) and regenerates the codebook from a distribution obtained within a maximum likelihood distribution estimation framework, based on observation of a set of K codewords that "d-match" (i.e., satisfy the distortion constraint for) a respective set of K independently generated source words. This result was later generalized, in a straightforward manner, to account for source memory, by considering the source as a vector source, i.e., a sequence of super-symbols from a corresponding super-alphabet. While ensuring asymptotic optimality, this extension suffered from a significant practical flaw: it requires asymptotically long vectors or super-symbols, hence exponentially large super-alphabet, in order to approach the rate-distortion bound, even for finite memory sources, e.g., Markov sources. Such exponentially large super-alphabet implies that even a single NTS iteration is intractable, thus compromising the promise of NTS to approach the rate-distortion function, in practice, for sources with memory. This work describes a considerably more efficient and tractable mechanism to achieve asymptotically optimal performance given a prescribed memory constraint, within a practical framework tailored to Markov sources. Specifically, the algorithm finds, asymptotically, the optimal codebook reproduction distribution, within a constrained set of distributions satisfying a prescribed Markovian property, e.g., of the same order as the source, which achieves the minimum per letter coding rate while maintaining a specified distortion level.

1 Introduction

Stochastic codebook design, based on source examples and string matching, has played a central role (with different flavors) in numerous applications in the areas of source coding, communications, machine learning, etc. Particularly influential were the seminal contributions of Lempel and Ziv in lossless coding, as evidenced by the numerous prevalent variants of the LZ77 and LZ78 algorithms [1, 2], which introduce stochastic codebook generation/adaptation, given source examples, as a powerful tool for lossless coding. Stochastic codebook generation mechanisms have been proposed for lossy coding as well, e.g., the gold-washing [3] and natural type selection [4, 5] algorithms. It is important to emphasize that optimizing the codebook reproduction distribution is fundamentally more difficult in the lossy coding setting. The lossless coding problem is "simpler" not only because perfect matching is less complex than matching with distortion, but more importantly because the optimal codebook generating distribution, which achieves the minimal coding rate, is exactly the source

This work is supported in part by NSF under grant CCF-1909423.

979-8-3503-4795-1/23/$31.00 ©2023 IEEE
DOI 10.1109/DCC55655.2023.00039

distribution. In other words, the problem is simply to learn the source distribution from examples. However, in lossy coding, the problem is vastly more difficult as the source distribution P and optimal codebook generating distribution Q^* are generally different, and more so in the high distortion regime [4, 6, 7].

Most relevant to the work presented herein, is the stochastic codebook generation and adaptation algorithm, known as "Natural Type Selection" (NTS), introduced in [4, 5] and further made practically tractable in [8, 9]. In this iterative algorithm, at each time step or iteration n, a sequence of K independently generated source words from an unknown source distribution, of length L, is encoded using a random codebook drawn from the current generating distribution Q_n. For each source word in the sequence, the first codeword in the codebook to satisfy a specified distortion constraint d, is recorded. Then, the codebook reproduction distribution is updated for iteration $n+1$, by estimating the *most likely distribution* to have generated the sequence of K d-matching codewords. In other words, the codebook reproduction distributions (or types) are naturally selected in response to source examples, and evolve through a sequence of "d-match" operations, hence the name natural type selection, with a nod to Darwin's theory of evolution. Consequently, it was shown that asymptotically in the statistical depth K, the number of iterations n, and the string length L, the sequence of codebook generating types $Q_1, Q_2, \ldots$ converges to the optimal reproduction distribution $Q^*_{P,d}$ that achieves the rate-distortion bound $R(P, d)$ for memoryless sources. This result was further extended to sources with memory in [9], by considering the source as a sequence of i.i.d. M-length *vectors* or super-symbols, i.e., neglecting inter-vector dependencies, for which the rate-distortion bound was achieved by a variant of the NTS algorithm, asymptotically as $M \to \infty$.

While ensuring asymptotic optimality, the NTS algorithm in [9] suffers from fundamental practical flaws. In order to converge to the optimal distribution that achieves the rate-distortion bound for sources with memory, the algorithm needs to encode source words that are composed of M-length vectors, each distributed according to the M-th order source joint distribution P_M, while sending M to infinity. Furthermore, even for finite-memory sources such as Markov sources of finite order, the algorithm nevertheless requires asymptotically large M (very long super-symbol vectors), within the codeword of L super-symbols, in order to achieve optimality. It is important to note that large M implies exponentially large cardinalities of both the source and code super-alphabet spaces, rendering intractable the main NTS operations such as d-search, maximum likelihood estimation and codebook regeneration. The requirement of asymptotically large M is also counter-intuitive, as it also applies to sources exhibiting modest memory, e.g., Markov sources where dependence on the past is fully captured by conditioning on a few past samples. In this paper, we propose to modify the NTS algorithm for Markov sources, such that the algorithm converges to the optimal distribution without sending M to infinity. Specifically, we restrict the generating codebook distribution to M-th order Markov distributions, which may in practice be chosen to be the same order as the source. Then, asymptotic convergence to the optimal constrained distribution, i.e., the M-th order Markov distribution that achieves the minimum per letter encoding rate amongst all codebook generating distributions of up to the same Markov order, is guaranteed.

2 Relevant Background on Random Coding

Let $\{X_u\}_{u=1}^{\infty}$ be a stationary ergodic source, where the source realizations are denoted as $x_u \in \mathcal{X}$. We assume that the source alphabet $\mathcal{X}$, and the reproduction alphabet $\mathcal{Y}$ are discrete spaces, equipped with their associated Borel σ-field $\mathcal{X}'$, and $\mathcal{Y}'$, respectively. Furthermore, let $\{\mathbf{X}_\ell\}_{\ell=1}^{\infty}$ and $\{\mathbf{Y}_\ell\}_{\ell=1}^{\infty}$ be a sequence of i.i.d. M-tuples source and reproduction vectors, where the realizations of source and reproductions vectors $\mathbf{x}_\ell \in \mathcal{X}^M$, and $\mathbf{y}_\ell \in \mathcal{Y}^M$, respectively. Let P_M denote the joint stationary distribution of source M-tuples $\mathbf{X}_\ell$. Define a random *source word* $\tilde{\mathbf{X}}$ and a random *codeword* $\tilde{\mathbf{Y}}$ as a concatenation of L i.i.d. random source and reproduction vectors, respectively, i.e., $\tilde{\mathbf{X}} = [\mathbf{X}_1, \ldots, \mathbf{X}_L]$, and $\tilde{\mathbf{Y}} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_L]$. Next, we define an arbitrary non-negative (measurable) scalar-valued distortion function $\rho : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$. The distortion between a realization of the source word $\tilde{\mathbf{x}}$ and a realization of the codeword $\tilde{\mathbf{y}}$, is assumed additive, and is, specifically, the average distortion over samples:

$$\rho(\mathbf{x}, \mathbf{y}) = \frac{1}{M} \sum_{m=1}^{M} \rho(x_m, y_m), \quad \rho(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = \frac{1}{L} \sum_{\ell=1}^{L} \rho(\mathbf{x}_\ell, \mathbf{y}_\ell). \tag{1}$$

For a scalar-valued fidelity constraint d, define a "d-match" event as the event that $\rho(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \leq d$. Suppose a random codebook $\mathcal{C}_L$ of infinite number of length-ML codewords $\Big(\tilde{\mathbf{Y}}(j), \text{ with } j \geq 1\Big)$ is generated such that, each codeword consists of L i.i.d. vectors as $Q_M = \{Q_M(\mathbf{y}) : \mathbf{y} \in \mathcal{Y}^M)\}$. We call Q_M the codebook reproduction distribution. Let $N_{M,L}$ be the index of the first codeword in $\mathcal{C}_L$ that d-matches the source word realization $\mathbf{x}$, i.e., $N_{M,L} = \inf\{j \geq 1 : \rho(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}(j)) \leq d\}$, with the convention that the infimum of an empty set is $+\infty$. Given a codebook reproduction distribution Q_M over $\mathcal{Y}^M$, we define,

$$D_{\min} \triangleq \mathbb{E}_{P_M}\left[\operatorname*{ess\,inf}_{\mathbf{y} \sim Q_M} \rho(\mathbf{X}, \mathbf{Y})\right], \quad D_{\text{av}} \triangleq \mathbb{E}_{P_M \times Q_M}\left[\rho(\mathbf{X}, \mathbf{Y})\right], \tag{2}$$

where $\operatorname{ess\,inf}_{\mathbf{y} \sim Q_M}(\cdot)$ denotes the essential infimum of a function, i.e.,

$$\operatorname*{ess\,inf}_{\mathbf{Y} \sim Q_M} \rho(\mathbf{x}, \mathbf{Y}) = \sup\{t \in \mathbb{R} : Q_M(\rho(\mathbf{x}, \mathbf{Y}) > t) = 1\}, \quad \text{for any } \mathbf{x} \in \mathcal{X}^M. \tag{3}$$

We will assume throughout this paper that D_{av} is finite, and $D_{\min} < D_{\text{av}} < \infty$. We will also restrict our attention to the non-trivial range of distortion levels $d \in (D_{\min}, D_{\text{av}})$. Then, Shannon's lossy coding theorem for scalar-valued distortion measures states: if a random codebook of length $\exp(L(R(P_M, d) + \epsilon))$ is generated using an optimal reproduction distribution $Q^*_{P_M,d}$, the probability of finding a codeword that d-matches an independently generated source word, drawn from the source distribution P_M, goes to one as L goes to infinity, wherein $R(P_M, d)$ is the *joint* (or M-th order) rate-distortion function, i.e., [10–12]

$$R(P_M, d) = \inf_{\substack{V : [V]_x = P_M, \\ \mathbb{E}_V(\rho(\mathbf{X}, \mathbf{Y})) \leq d}} I(\mathbf{X}, \mathbf{Y}). \tag{4}$$

Here, $I(\mathbf{X}, \mathbf{Y})$ is the mutual information between the M-tuples random vectors $\mathbf{X}$ and $\mathbf{Y}$, and the infimum is taken over all joint probability distributions V such that the

x-marginal of V, denoted $[V]_x$, is P_M and the expected distortion $\mathbb{E}_V(\rho(\mathbf{X},\mathbf{Y})) \leq d$. Let $V^*_{P_M,d}$ be the optimal joint distribution that realizes the infimum in (4), then the optimal codebook reproduction distribution $Q^*_{P_M,d}$ is the y-marginal of the optimal joint distribution $V^*_{P_M,d}$. However, if a random codebook is generated from distribution $Q_M \neq Q^*_{P_M,d}$, then the minimum encoding rate to guarantee a d-match in probability, as L goes to infinity, was effectively shown in [13], and extended to memoryless sources over abstract alphabets in [14], to be

$$R(P_M, Q_M, d) = \inf_{\substack{V:[V]_x=P_M,\\ \mathbb{E}_V(\rho(\mathbf{X},\mathbf{Y}))\leq d}} \mathcal{D}(V||P_M \times Q_M) = \inf_{Q'_M}\{I_{\min}(P_M||Q'_M, d) + \mathcal{D}(Q'_M||Q_M)\}, \quad (5)$$

where $\mathcal{D}(\cdot||\cdot)$ is the Kullback-Leibler (KL) divergence (or the relative entropy), and $I_{\min}(P_M||Q'_M, d)$ is the usual minimum mutual information but with an additional constraint on the output distribution, i.e.,

$$I_{\min}(P_M||Q'_M, d) = \inf_{\substack{V:[V]_x=P_M,\ [V]_y=Q'_M,\\ \mathbb{E}_V(\rho(\mathbf{X},\mathbf{Y}))\leq d}} I(\mathbf{X},\mathbf{Y}). \quad (6)$$

Here the infimum is taken over all joint distributions V of the random vectors $(\mathbf{X},\mathbf{Y})$, whose x-marginal, denoted by $[V]_x$, is P_M, and y-marginal, denoted by $[V]_y$, is Q'_M, and such that the expected distortion does not exceed d. In [15, Th. 2], it was shown that, under these assumptions for the memoryless case (for which extension to the sources with memory case is straight forward), $R(P_M, Q_M, d)$ is finite, strictly positive, and that the infimum in its L.H.S. definition of (5) is always achieved by some joint distribution $V^*_{P_M,Q_M,d}$. Moreover, since the set of V over which the infimum is taken is convex, from [16] it can be concluded that $V^*_{P_M,Q_M,d}$ is the unique minimizer. Hence, a unique minimizer to the R.H.S. of (5) also exists, and is denoted $Q^*_{P_M,Q_M,d}$. Next, we define the minimum coding rate per letter for stationary ergodic sources with memory required to guarantee a d-match with probability one asymptotically in L as [11] [12],

$$R(d) = \lim_{M\to\infty} M^{-1} R(P_M, d). \quad (7)$$

The limit in (7) exists for stationary ergodic sources, and for any M, $R(P_M, d)$ is an upper bound to $R(d)$ [17, Th. 9.8.1]. Consequently, the optimal codebook reproduction distribution that achieves $R(d)$ is $Q^*_d = \lim_{M\to\infty} Q^*_{P_M,d}$. Given a source with discrete input and reproduction alphabets, define a '*type*' of source or code vector as the fraction of occurrence of every letter in the alphabet as seen in the vector [18].

In this paper, we restrict our attention to stationary and ergodic sources with memory described by the Markovian property. The M-th order Markov source property implies that the current source sample distribution conditioned on the entire past is fully captured by conditioning only on the previous M-samples. This Markov source can be described by a state transition diagram, i.e., a Markov chain, with $|\mathcal{X}|^M$ states. Let $P_{j|i}$ be the source state transition probability, from state i to state j, where $i, j \in \mathcal{R} = \mathcal{X}^M$. Let $\mathbf{P}$ be the state transition probability matrix whose (i, j) element is $P_{j|i}$. Equivalently, let $P(X|\mathbf{x}) = \{P(x|\mathbf{x}) : x \in \mathcal{X}\}$ be the stationary source letter distribution conditioned on the M previous samples specified in the vector $\mathbf{x}$. Note that there exists a one-to-one mapping between the set $\{P_{j|i}, \forall(i, j) \in \mathcal{R}^2\}$ and the set $\{P(x|\mathbf{x}), \forall x \in \mathcal{X}, \forall \mathbf{x} \in \mathcal{X}^M\}$.

3 Natural Type Selection

This work builds on and expands the NTS random lossy codebook generation approach for discrete sources, which was originally proposed in [4], a tractable version for memoryless sources was proposed in [5, 8], and extension to sources with memory in [9]. Let the sequence of i.i.d. source vectors $\{\mathbf{x}_\ell\}_{\ell=1}^{\infty}$ be generated according to an unknown source distribution $P_M = \{P_M(\mathbf{x}) : \mathbf{x} \in \mathcal{X}^M\}$. Furthermore, let the codebook reproduction distribution be $Q_M = \{Q_M(\mathbf{y}) : \mathbf{y} \in \mathcal{Y}^M\}$. In [9], the authors showed that the Maximum Likelihood (ML) distribution that most likely generates a sequence of K codewords that respectively d-match a sequence of K independently generated source words, converges in probability to $Q^*_{P_M,Q_M,d}$ as the statistical depth $K \to \infty$ and string length $L \to \infty$. Note that $Q^*_{P_M,Q_M,d}$ is more efficient in coding the source than Q_M. This immediately suggests a recursive and iterative algorithm. Let n be the iteration index, and assume that the algorithm starts with a strictly positive initial codebook reproduction distribution denoted $Q_{0,M,L,K}$, over the entire reproduction space. At each iteration, the algorithm performs a sequence of K d-match events to a sequence of K independently generated source words. Afterward, the algorithm computes the ML codebook distribution that would generate the set of d-matching codewords. In other words, the next iteration's codebook reproduction distribution is naturally selected by the source through a sequence of d-match events, hence the name "natural type selection". Let $Q_{0,M,L,K}, Q_{1,M,L,K}, \ldots$ be the sequence of ML codebook reproduction distributions, it was shown in [9] that this sequence of distributions converges to the optimal codebook distribution $Q^*_{P_M,d}$ that achieves the M-th order rate-distortion function $R(P_M, d)$ in (4), i.e., $Q^*_{P_M,d} = \lim_{n\to\infty} \lim_{L\to\infty} \lim_{K\to\infty} Q_{n,M,L,K}$. In the next section, we introduce a variant of the NTS algorithm which is specialized for Markov sources and analyze the asymptotic optimality of its codebook generating distribution over all Markov distribution of the same order.

4 Proposed NTS Algorithm for Markov Sources

In order to take into account the M-th order Markovian property of the source, we restrict the codebook reproduction distribution to distributions with M-th order Markov property. Let $Q_{j|i}$ be the codebook distribution state transition probability from state i to state j, where $i \in \mathcal{S}, j \in \mathcal{S},$ and $\mathcal{S} = \mathcal{Y}^M$. Hence, let $\mathbf{Q}$ be the state transition probability matrix for which the entry in the i-th row and j-th column is $Q_{j|i}$. Let the random L-tuples source words and codewords $\mathbf{X} = [X_1, \ldots, X_L]$, and $\mathbf{Y} = [Y_1, \ldots, Y_L]$, be generated according to state transition matrices $\mathbf{P}$ and $\mathbf{Q}$, respectively. First, we introduce a variant of NTS algorithm for the above setup. At every NTS iteration with index n, the algorithm finds a set of d-matching codewords in the random codebook to a set of K independently generated source words. Let the realizations of the d-matching source and code sets be denoted as $\{\mathbf{x}(i_1), \ldots, \mathbf{x}(i_K)\}$, and $\{\mathbf{y}(j_1), \ldots, \mathbf{y}(j_K)\}$, where j_k is the index of the codeword that d-match the k-th source word in the codebook. Next, similar to before, the NTS algorithm finds the most likely (constrained) reproduction distribution to produce the set of d-matching codewords, where the distribution is constrained to have M-th order Markov property. Note that in the codebook training stage, we assume that each source word (and hence each codeword) among the K-size set is generated independently of the other source words. This condition is necessary to guarantee convergence to the desired optimal constrained distribution, as will be illustrated in Theorem 1. However, once the training is completed and the codebook distribution has converged, this condition can be relaxed. Furthermore, for a given codebook distribution, the encoder and the decoder "agree" on a given random codebook by synchronizing the random number generator seed.

Lemma 1 [19]: The ML estimate of the M-th order Markov process state transition probabilities underlying the codebook reproduction distribution, given a set of K d-matching codewords, is the *average* of the d-matching codewords' transitions, i.e.,

$$\mathbf{Q}_{n+1,M,L,K} = \mathbf{Q}^{\mathrm{ML}} = \left\{ Q_{j|i} : Q_{j|i} = \frac{\sum_{k=1}^{K} N(i \to j | \mathbf{y}(k))}{\sum_{k=1}^{K} \sum_{j' \in \mathcal{S}} N(i \to j' | \mathbf{y}(k))}, \quad \forall (i,j) \in \mathcal{S}^2 \right\}, \tag{8}$$

where k enumerates the d-matching events, and $N(j \to i|\mathbf{y}(j_k))$ is the number of transitions from state i to state j as seen in the k-th d-matching codeword $\mathbf{y}(j_k)$, whose index in the random codebook is j_k.

Thus, this algorithm yields a sequence of state transition matrices (8), or equivalently, conditional distributions $Q_{n+1,M,L,K}(Y|\mathbf{y}) = \{Q_{n+1,M,L,K}(y|\mathbf{y}) : y \in \mathcal{Y}\}$. We next quantify the asymptotic performance of the NTS variant tailored to Markov sources. Let the random codebook be generated according to a Markov process with conditional probabilities $Q(Y|\mathbf{y})$, $\forall \mathbf{y} \in \mathcal{Y}^M$. We start by transforming this variant of NTS algorithm into a dual set of NTS algorithms for *memoryless sources.*

Let the sets of K d-matching source words and codewords be concatenated into KL-length source and code blocks denoted as, $\mathbf{s} = [\mathbf{x}(i_1), \ldots, \mathbf{x}(i_K)]$, and $\mathbf{c} = [\mathbf{y}(j_1), \ldots, \mathbf{y}(j_K)]$, respectively. Next, let the source and code blocks be independently divided into sub-streams based on the previous source and code M-tuples, denoted as $\{\mathbf{s}_{\mathbf{x}}, \forall \mathbf{x} \in \mathcal{X}^M\}$, and $\{\mathbf{c}_{\mathbf{y}}, \forall \mathbf{y} \in \mathcal{Y}^M\}$. The first M letters in each source word and codeword are not assigned to any sub-stream, which is of negligible conse-

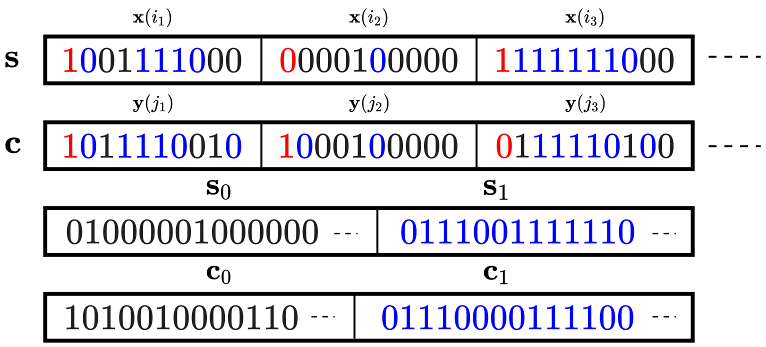

Figure 1: Division of the d-matching source and code blocks into i.i.d. sub-streams based on the previous sample.

quence asymptotically in L. Note that as we assume a time-invariant Markov source, each sub-stream sequence $\{\mathbf{s_x}\}$ is i.i.d. and drawn from $P(X|\mathbf{x})$. The set of d-match event $\{\rho(\mathbf{x}(i_k), \mathbf{y}(i_k)) \leq d, \ \ \forall k\}$, implies that $\rho(\mathbf{s}, \mathbf{c}) \leq d$, which is equivalent to a set of size $|\mathcal{X}^M \times \mathcal{Y}^M|$ events of distortion matches between the sub-streams $\{\mathbf{s_x}\}$ and $\{\mathbf{c_y}\}$, each with distortion level denoted as $d_{\mathbf{x},\mathbf{y}}$, such that, $\sum_{\mathbf{x},\mathbf{y}} \mathbb{M}_n(\mathbf{x}, \mathbf{y}) d_{\mathbf{x},\mathbf{y}} \leq d$. Here $\mathbb{M}_n(\mathbf{x}, \mathbf{y})$, is the empirical probability of mapping a letter in sub stream $\mathbf{s_x}$, to a letter in sub stream $\mathbf{c_y}$, at NTS iteration n, as seen by the code and source blocks $\mathbf{s}$, and $\mathbf{c}$, respectively. Fig 1 illustrates this for binary source and code blocks, which are formed by concatenating three L-length d-matching source and code words, with $L = 10$, where the source and codebook generating distributions are first order Markov, hence the number of Markov states is $|\mathcal{X}| = |\mathcal{Y}| = 2$, the distortion measure is Hamming, and the distortion constraint is $d = 1/3$. Samples in different i.i.d. sub-streams are assigned different colors. Samples whose immediate predecessor is a '0' are colored black, and samples following a '1' are colored blue. The i.i.d sub-streams $\mathbf{s_x}$ and $\mathbf{c_y}$ are formed by collecting all samples that follow the same letter, see Fig. 1.

Theorem 1: For an initial codebook generating Markov chain with strictly positive transition probabilities $Q(Y|\mathbf{y}) > 0$, $\forall \mathbf{y} \in \mathcal{S} = \mathcal{Y}^M$, and distortion measure satisfying $0 \leq D_{\min} < D_{\text{av}} < \infty$, the transition probabilities $Q(Y|\mathbf{y})$, of the recursive NTS algorithm for Markov sources, where each recursion involves collecting K d-matches, converge in probability and asymptotically, as $L \to \infty$, as follows,

$$\begin{aligned} &Q_{n+1,M,K}(Y|\mathbf{y}) \to \sum_{\mathbf{x} \in \mathcal{X}^M} \mathbb{M}_n^*(\mathbf{x}|\mathbf{y}) Q^* \left(P(X|\mathbf{x}), Q_{n,M,K}(Y|\mathbf{y}), d^*_{\mathbf{x},\mathbf{y}} \right), \\ &V^* \left(P(X|\mathbf{x}), Q(Y|\mathbf{y}), d^*_{\mathbf{x},\mathbf{y}} \right) \triangleq \arg \min_{V \in E_{\mathbf{x},\mathbf{y}}(d^*_{\mathbf{x},\mathbf{y}})} \mathcal{D} \left(V \middle\| P(X|\mathbf{x}) \times Q(Y|\mathbf{y}) \right), \\ &Q^* \left(P(X|\mathbf{x}), Q(Y|\mathbf{y}), d^*_{\mathbf{x},\mathbf{y}} \right) = \left[V^* \left(P(X|\mathbf{x}), Q(Y|\mathbf{y}), d^*_{\mathbf{x},\mathbf{y}} \right) \right]_y , \end{aligned} \tag{9}$$

where $Q_{n+1,M,K}(Y|\mathbf{y}) = \lim_{L \to \infty} Q_{n+1,M,L,K}(Y|\mathbf{y})$, and the set $E_{\mathbf{x},\mathbf{y}}(d_{\mathbf{x},\mathbf{y}})$ is defined as,

$$E_{\mathbf{x},\mathbf{y}}(d^*_{\mathbf{x},\mathbf{y}}) = \left\{ V : V = P' \circ W', P' = P(X|\mathbf{x}), \ \rho(P', W') \leq d^*_{\mathbf{x},\mathbf{y}} \right\}. \tag{10}$$

Here $\rho(P', W')$ is the average distortion computed over distributions, and the set of distortion levels $\{d^*_{\mathbf{x},\mathbf{y}}, \forall (\mathbf{x}, \mathbf{y}) \in \mathcal{X}^M \times \mathcal{Y}^M\}$, satisfies,

$$\frac{\partial}{\partial \delta} R(P(X|\mathbf{x}), Q(Y|\mathbf{y}), \delta) \Big|_{\delta = d^*_{\mathbf{x},\mathbf{y}}} = R'_{P,Q,d}, \ \forall (\mathbf{x}, \mathbf{y}), \quad \sum_{\mathbf{x},\mathbf{y}} \mathbb{M}_n^*(\mathbf{x}, \mathbf{y}) d^*_{\mathbf{x},\mathbf{y}} \leq d, \tag{11}$$

where $R'_{P,Q,d}$ is independent of the sub-stream pair $(\mathbf{x}, \mathbf{y})$. In other words, the distortion allocation to sub-stream pairs, $d^*_{\mathbf{x},\mathbf{y}}$, ensures they all maintain the same rate-distortion slope, given codebook generating distributions $\{Q(Y|\mathbf{y})\}$, while satisfying the overall average distortion constraint d.

Proof sketch[1]: We employ a variant of the conditional limit theorem [18] to establish that, conditioned on the rare event that the joint input-output distribution of a block of K concatenated respective source and codewords $(\mathbf{S}, \mathbf{C})$ belongs to a convex set of distributions that satisfy the distortion constraint d, the conditional distributions of this code block on $\mathcal{Y}$ converge in probability, as $L \to \infty$, to the distribution $\sum_{\mathbf{x} \in \mathcal{X}^M} \mathbb{M}_n^*(\mathbf{x}|\mathbf{y}) Q^* \left(P(X|\mathbf{x}), Q_{n,M,K}(Y|\mathbf{y}), d^*_{\mathbf{x},\mathbf{y}} \right)$. Next, by [11], the minimum

[1] While detailed theorem proofs are omitted for space constraints, see note at the end of this section for additional information

of the output-constrained rate is achieved by adding the output-constrained rate-distortion functions at points of equal slopes in all co-ordinates, implying (11). The theorem and proof are very closely related to the *Gibbs Conditioning Principle* of statistical mechanics (see [20] and references therein), which (roughly) states: Consider $\{X_1, \ldots, X_N\}$ i.i.d. random variables distributed over a Polish space with marginal distribution P_X and a measurable function $f : \mathcal{X} \to \mathbb{R}$. Under suitable conditions on P_X and $f(\cdot)$, and conditioned on the rare event that $\left\{\frac{1}{N}\sum_i f(X_i) \in [a-\delta, a+\delta]\right\}$, where $a \in \mathbb{R}$ and $\delta > 0$, the distribution of X_i converges in probability, as $N \to \infty$, to the distribution that minimizes the divergence $\mathcal{D}(\cdot||P_X)$ over all distributions that satisfy the constraint. Very similar arguments are used to prove Theorem 1.

Next, we look at the asymptotic convergence of the codebook reproduction conditional distributions as the number of iterations n goes to infinity.

Theorem 2: Given an initial codebook that is generated using a Markov process with strictly positive conditional distributions $Q(Y|\mathbf{y})$ for any state $\mathbf{y} \in \mathcal{S} = \mathcal{Y}^M$, the recursion in (8) achieves the minimum average coding rate over the cross product of all source-code sub streams, denoted as $\overline{R}(d)$, i.e.,

$$\overline{R}(d) = \min_{Q(Y|\mathbf{y})} \min_{\substack{\mathbb{M}(\mathbf{y}|\mathbf{x}) \\ d_{\mathbf{x},\mathbf{y}}, V_{\mathbf{x},\mathbf{y}}}} \sum_{\mathbf{x},\mathbf{y}} \mathbb{M}(\mathbf{x})\mathbb{M}(\mathbf{y}|\mathbf{x})\ \mathcal{D}\left(V_{\mathbf{x},\mathbf{y}}\ \big|\big|\ P(X|\mathbf{x}) \times Q(Y|\mathbf{y})\right). \tag{12}$$

and the set of optimization variables that achieves the minimum in (12), satisfies,

$$\frac{\partial}{\partial \delta} R\left(P(X|\mathbf{x}), Q^*(Y|\mathbf{y}), \delta\right)\Big|_{\delta = d^*_{\mathbf{x},\mathbf{y}}} = R'_{P,Q^*,d}, \ \sum_{\mathbf{x},\mathbf{y}} \mathbb{M}^*(\mathbf{x},\mathbf{y}) d^*_{\mathbf{x},\mathbf{y}} \leq d, \tag{13}$$

where $R'_{P,Q^*,d}$ is independent of the sub-stream pair $(\mathbf{x}, \mathbf{y})$.

Proof sketch: First, we show that the average encoding rate over the cross product of all i.i.d. source and code sub-streams can be written as double minimization over *convex* sets, due to the convexity of all constraints. Then, we invoke the Csiszar and Tusnady theorem of alternating minimization over convex sets [21] to show the convergence of the average encoding rate to its minimum, and consequently, the convergence of the conditional distributions to the distributions that achieve the minimum average encoding rate. Hence, this establishes that the NTS algorithm finds the conditional distributions that minimize the average encoding rate over the cross product of the sets of all i.i.d. source and code sub-streams $\{\mathbf{s_x} \times \mathbf{c_y}\}$ while maintaining the distortion level d, hence implying asymptotic optimality.

Note that the detailed proofs of Theorem 1 and Theorem 2 have been omitted here due to space limitations, but they largely follow from, and employ similar arguments as in optimality proofs for earlier NTS algorithms in our prior work [4, 8, 9], based on conditional limit theorems and alternating minimization over convex sets. Current detailed proofs can be found in an unpublished extended draft for a journal paper submission [22].

5 Toy Example

In this section, we illustrate the convergence behavior of the proposed NTS algorithm variant, which is tailored to Markov sources. We consider a first-order binary Markov

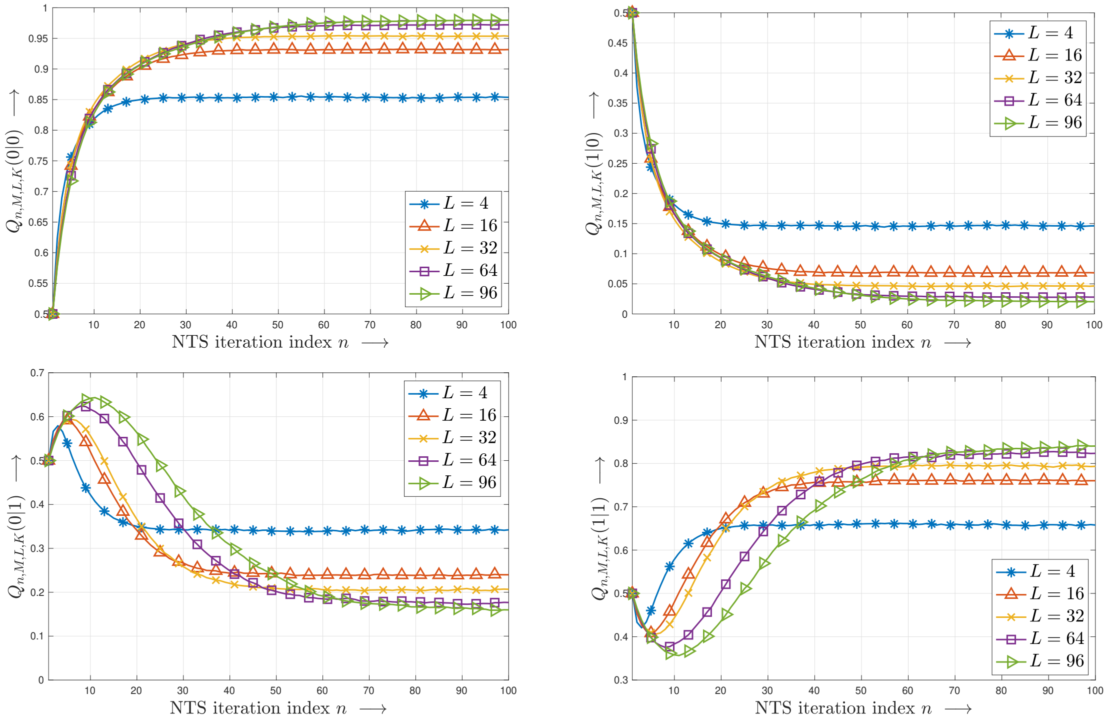

Figure 2: Evolution of state transition probabilities in the NTS codebook generating Markov chain versus iteration index n, for different source word lengths L, statistical depth is $K = 10^5$; operating on a binary Markov source with Hamming distortion constraint at $d = d_{\max} = 1/3$.

source with the transition probabilities $P(0|0) = 0.8, P(1|0) = 0.2, P(0|1) = 0.4$, and $P(1|1) = 0.6$, to be encoded under the Hamming distortion measure. To illustrate how the algorithm operates, we choose a distortion constraint for which one can guess the optimal solution: $d = d_{\max} = 1/3$. In Fig. 2, we depict the evolution of transition probabilities of the codebook reproduction distributions versus the number of NTS iterations n, for different values of source word length L . It is worth noting that as L increases, and for the given distortion level $d = d_{\max}$, the transition probabilities employed for codebook generation approach $Q^*(0|0) = 1$, which strongly favors the optimal codeword achieving $d_{\max}$, namely, the all zero codeword. Thus, NTS is converging to the optimal first-order Markov codebook generating distribution, without prior knowledge of the source distribution. Furthermore, it is important to note that, even for finite length L, the codebook generating distributions converge asymptotically in the statistical depth K, and the number of NTS iterations n.

6 Conclusion

This paper proposes a modified and more effective NTS approach for a stochastic generation of random codebook in the lossy coding settings, specifically when Markov sources are considered. Unlike the NTS approach in [9], the algorithm is not required to send the memory depth M to infinity in order to achieve the rate-distortion bound, which dramatically reduces the otherwise intractable complexity of the NTS algorithm and most importantly the search for d-match in the codebook, a central step in the iteration. It was further shown by Theorem 1 and Theorem 2, that the codebook generating distribution, that emerges from the proposed stochastic algorithm, converges to the optimal Markov codebook generating distribution of the prescribed order, asymptotically as $L \to \infty$, and $n \to \infty$.

References

[1] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. on Inf. Theory*, vol. 23, no. 3, pp. 337–343, 1977.

[2] J. Ziv and A. Lempel, "Compression of individual sequences via variable rate coding," *IEEE Trans. on Inf. Theory*, vol. IT-24, pp. 530–536, 1978.

[3] Z. Zhang and V. Wei, "An on-line universal lossy data compression algorithm via continuous codebook refinement—part i: Basic results," *IEEE Trans. on Inf. Theory*, vol. IT-42, pp. 803–821, 1996.

[4] R. Zamir and K. Rose, "Natural type selection in adaptive lossy compression," *IEEE Trans. on Inf. Theory*, vol. 47, pp. 99–110, 2001.

[5] Y. Kochman and R. Zamir, "Adaptive parametric vector quantization by natural type selection," in *Data Compression Conference (DCC)*, 2002.

[6] R. Zamir and K. Rose, "Towards lossy Lempel-Ziv: Natural type selection," in *Proc. of the Inf. Theory Workshop, Haifa, Israel*, June 1996, p. pp. 58.

[7] R. Zamir and K. Rose, "A type generation model for adaptive lossy compression," in *Proc. of ISIT97*, Ulm, Germany, June 1997, p. 186.

[8] A. Elshafiy, M. Namazi, and K. Rose, "On effective stochastic mechanisms for on-the-fly codebook regeneration," in *IEEE Int. Symposium on Inf. Theory (ISIT)*, 2020.

[9] A. Elshafiy, M. Namazi, R. Zamir, and K. Rose, "On-the-fly stochastic codebook re-generation for sources with memory," in *IEEE Inf. Theory Workshop (ITW)*, 2021.

[10] T. Berger, *Rate Distortion Theory*, Prentice-Hall, 1971.

[11] R. Gray, "Conditional rate-distortion theory," Tech. Rep. 6502-2, Stanford Electronics Lab, 1973.

[12] T. Berger, "Rate distortion theory for sources with abstract alphabets and memory," *Information and Control*, vol. 13, September 1968.

[13] E. Yang and J. Kieffer, "On the performance of data compression algorithms based upon string matching," *IEEE Trans. on Inf. Theory*, vol. 44, pp. 47–65, 1998.

[14] I. Kontoyiannis and R. Zamir, "Mismatched codebooks and the role of entropy coding in lossy data compression," *IEEE Tras. on Inform. Theory*, vol. 52, May 2006.

[15] A. Dembo and I. Kontoyiannis, "Source coding, large deviations, and approximate pattern matching," *IEEE Tras. on Inform. Theory*, vol. 48, June 2002.

[16] I. Csiszar, "I-Divergence Geometry of Probability Distributions and Minimization Problems," *Annals of Probability*, vol. 3, no. 1, pp. 146–158, 1975.

[17] R. Gallager, *Information Theory and Reliable Communication*, John Wiley and Sons, Inc., 1968.

[18] T. Cover and J. Thomas, *Elements of Inf. Theory*, Wiley-Interscience, 2006.

[19] M. Bartlett, "The frequency goodness of fit test for probability chains," *Mathematical Proc. of the Cambridge Philosophical Society*, vol. 47, no. 1, pp. 86–95, 1951.

[20] A. Dembo and O. Zeitouni, "Refinements of the gibbs conditioning principle," *Probability Theory and Related Fields*, vol. 104, pp. 1–14, 1996.

[21] I. Csiszar and G. Tusnady, "Information geometry and alternating minimization procedures," *Statist. Decision*, , no. 1, pp. 205–237, 1984.

[22] A. Elshafiy and K. Rose, "Asymptotically Optimal Stochastic Lossy Coding of Markov Sources," [Online]. Available on arXiv.org.