On-The-Fly Stochastic Codebook Re-generation for Sources with Memory

Ahmed Elshafiy*, Mahmoud Namazi*, Ram Zamir†, and Kenneth Rose*
*Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106, USA
Email: {a_elshafiy, mnamazi, rose}@ece.ucsb.edu
†Department of Electrical Engineering - Systems
Tel Aviv University, Tel Aviv, Israel
Email: zamir@eng.tau.ac.il

2020 IEEE Information Theory Workshop (ITW) | 978-1-7281-5962-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/ITW46852.2021.9457666

***Abstract*—This paper proposes a generalized stochastic mechanism for codebook generation in lossy coding settings for sources with memory. Earlier work has shown that the rate-distortion bound can be asymptotically achieved for discrete memoryless sources by a "natural type selection" (NTS) algorithm. In iteration n, the distribution that is most likely to produce the types of a sequence of K codewords of finite length ℓ that "d-match" a respective sequence of K source words of length ℓ, (i.e., which satisfy the distortion constraint), is used to regenerate the codebook for iteration $n+1$. The resulting sequence of codebook generating distributions converges to the optimal distribution Q^* that achieves the rate-distortion bound for the memoryless source, asymptotically in ℓ, K, and n. This work generalizes the NTS algorithm to account for sources with memory. The algorithm encodes $m\ell$-length source words consisting of ℓ vectors (or super-symbols) of length m. We show that for finite m and ℓ, the sequence of codebook reproduction distributions $Q_{0,m,\ell}, Q_{1,m,\ell}, \ldots$ (each computed after observing a sequence of K d-match events) converges to the optimal *achievable* distribution $Q^*_{m,\ell}$ (within a set of achievable distributions determined by m and ℓ), asymptotically in K and n. It is further shown that $Q^*_{m,\ell}$ converges to the optimal reproduction distribution Q^* that achieves the rate-distortion bound for sources with memory, asymptotically in m and ℓ.**

***Index Terms*—Random Codebook, String Matching, Rate-Distortion function, Natural Type Selection**

I. Introduction

Codebook generation and adaptation, based on codeword usage and source string matching, feature prominently in the source coding literature. In past decades, significant attention has been directed towards its successful application to optimal codebook generation in the lossless setting, following the seminal contributions of Lempel and Ziv [1]–[3]. The lossless setting is now relatively well understood. In lossless coding, achieving the optimal encoding rate performance is ultimately a matter of using the observed source samples to learn the source distribution, which by definition equals the optimal codebook reproduction distribution. For example, the Lempel-Ziv (LZ78) algorithm [3] generates a codebook which asymptotically consists (mostly) of typical source sequences, by string matching with source examples, thus achieving asymptotically optimal performance, without prior knowledge of the source statistics. However, finding the optimal codebook reproduction distribution in lossy coding is radically different, because "good" codewords in the rate-distortion sense are not statistically equivalent, and sometimes not even similar, to the source. For example, consider continuous-alphabet sources with continuous probability distribution P under the squared-error distortion measure. The optimal codebook reproduction distribution Q^* only approximates P at the low distortion regime (high resolution). As the fidelity constraint d is relaxed, Q^* increasingly deviates from P, often becomes discrete with successively decreasing cardinality, until it collapses on one point at $d = d_{\max}$ [4].

In [5], it was proposed that codebook adaptation in the lossy setting does not play the role of learning and matching the source statistics, but more generally that of "type-selection". In other words, the codebook adaptation algorithm ultimately estimates the optimal *reproduction type* for the source at a given distortion constraint d. Most relevant to this work is the iterative universally optimal NTS algorithm, originally proposed in [5], and practically improved in [6], [7]. Consider a discrete memoryless source with distribution P and a random codebook drawn from a strictly positive initial distribution. At NTS iteration n, the sequence of K source words, each of finite length ℓ, will favor a sequence of K codewords through independent respective distortion match events. Afterwards, a Maximum Likelihood (ML) estimation framework is employed to compute the new codebook reproduction distribution $Q_{n+1,\ell}$, which is the most likely to generate the set of the favoured codewords. It was shown that the sequence of codebook generating distributions $Q_{0,\ell}, Q_{1,\ell}, \ldots$ converges, asymptotically in K, n, and ℓ, to the optimal reproduction distribution $Q^*(P,d)$ that achieves the rate-distortion bound $R(P,d)$. In [6], a parametric set of codebook reproduction distributions $\mathcal{Q}$ was considered. It was shown that even for sources with memory, a variant of the NTS algorithm finds the distribution within $\mathcal{Q}$ that achieves the *first-order* rate-distortion bound (over the parametric distribution set $\mathcal{Q}$) asymptotically in ℓ and n. Hence, the NTS algorithm as it stands, guarantees optimal performance only for memoryless sources, while in practice, virtually all sources of interest are sources with memory.

This work generalizes the NTS codebook generation mechanism to be universally optimal for sources with memory. The remainder of this paper is organized as follows: Section II provides the relevant background, Section III summarizes the

This work is supported in part by NSF under grant CCF-1909423, and by BSF under grant 2018690.

original NTS algorithm, Section IV introduces the proposed effective and practical mechanism of codebook generation to capture the benefits of NTS for sources with memory, and establishes its asymptotic convergence to the rate-distortion bound achieving distribution. In Section V, the workings and convergence of the NTS algorithm are illustrated by a toy example. Conclusions are drawn in Section VI.

II. Relevant Background

Let $\mathcal{X}$ be the discrete source alphabet, and $\mathcal{Y}$ be the discrete reconstruction alphabet. Additionally, let $\{x_u\}_{u=1}^{\infty}$ be a stationary ergodic source, where $x_u \in \mathcal{X}$. Next, let $\{\mathbf{x}_i\}_{i=1}^{\infty}$ be a sequence of independent and identically distributed (i.i.d.) m-tuples (or source "super-symbols"), each obtained by drawing m successive symbols from the distribution underlying source $\{x_u\}$. Hence, let $P_m = \{P_m(\mathbf{x}) : \mathbf{x} \in \mathcal{X}^m\}$ be the vector source distribution. Define a source block (source word) that contain ℓ source vectors as $\tilde{\mathbf{x}} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\ell)$. Next, we define a vector-valued distortion measure $\rho = (\rho_1, \rho_2, \ldots, \rho_m)$, as in [8], where $\rho_r : \mathcal{X}^m \times \mathcal{Y}^m \to [0, \infty)$. The distortion between source block $\tilde{\mathbf{x}}$ and code block (codeword) $\tilde{\mathbf{y}} = (\mathbf{y}_1 \ldots, \mathbf{y}_\ell)$, with $\mathbf{y}_i \in \mathcal{Y}^m$, is assumed additive, and is specifically, the average distortion over super-symbols in the block:

$$\rho(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = \frac{1}{\ell} \sum_{i=1}^{\ell} \rho(\mathbf{x}_i, \mathbf{y}_i). \tag{1}$$

For a vector-valued fidelity constraint $\mathbf{d} = (d_1, d_2, \ldots, d_m)$, define a "$\mathbf{d}$-match" event as the event that $\rho_r(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \le d_r, \forall r \in \{1, 2, \ldots, m\}$. Suppose a random codebook is generated such that each codeword consists of ℓ i.i.d. vectors as $Q_m = \{Q_m(\mathbf{y}) : \mathbf{y} \in \mathcal{Y}^m)\}$. Additionally, let $N_{m,\ell}$ be the index of the first codeword to $\mathbf{d}$-match the source word, i.e.,

$$N_{m,\ell} \triangleq \inf\{i \ge 1 : \rho_r(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}_i) \le d_r, \forall r \in \{1, \ldots, m\}\}, \tag{2}$$

where $\tilde{\mathbf{y}}_i$ is the i-th codeword in the random codebook, and $\tilde{\mathbf{y}}_{N_{m,\ell}}$ is the $\mathbf{d}$-matching codeword. Then, the generalization of Shannon's lossy coding theorem to vector-valued distortion measures states: if a random codebook of length $\exp(\ell(R(P_m, \mathbf{d}) + \epsilon))$ is generated using an optimal reproduction distribution $Q_m^*(P_m, \mathbf{d})$, the probability of finding a codeword that $\mathbf{d}$-matches an independently generated source word, drawn from source distribution P_m, goes to one as ℓ goes to infinity, wherein $R(P_m, \mathbf{d})$ is the *joint* (or m-th order) rate-distortion function, i.e., [8]–[10]

$$R(P_m, \mathbf{d}) = \inf_{W_m : \rho_r(P_m, W_m) \le d_r, \forall r} I(P_m, W_m), \tag{3}$$

Here, $I(P_m, W_m)$ denotes the mutual information for a given source and channel (conditional) distributions, and $\rho_r(P_m, W_m)$, with $r \in \{1, 2, \ldots, m\}$, is r-th component of the vector-valued distortion function defined over the joint distribution, i.e.,

$$I(P_m, W_m) = \sum_{\mathbf{x} \in \mathcal{X}^m} \sum_{\mathbf{y} \in \mathcal{Y}^m} P_m(\mathbf{x}) W_m(\mathbf{y}|\mathbf{x}) \log \frac{W_m(\mathbf{y}|\mathbf{x})}{Q_m(\mathbf{y})}, \tag{4}$$

$$Q_m(\mathbf{y}) = \sum_{\mathbf{x}' \in \mathcal{X}^m} P_m(\mathbf{x}') W_m(\mathbf{y}|\mathbf{x}'), \tag{5}$$

$$\rho_r(P_m, W_m) = \sum_{\mathbf{x} \in \mathcal{X}^m} \sum_{\mathbf{y} \in \mathcal{Y}^m} P_m(\mathbf{x}) W_m(\mathbf{y}|\mathbf{x}) \rho_r(\mathbf{x}, \mathbf{y}). \tag{6}$$

Throughout the paper, the logarithm function is natural, i.e., taken over the base e. Let $W_m^* = \{W_m^*(\mathbf{y}|\mathbf{x}) : \mathbf{x} \in \mathcal{X}^m, \mathbf{y} \in \mathcal{Y}^m\}$ be the transition (conditional) distribution that achieves the infimum in (3), i.e., $R_m(P_m, \mathbf{d}) = I(P_m, W_m^*)$. Then, the optimal codebook reproduction distribution $Q^*(P_m, \mathbf{d})$ is obtained by marginalization:

$$Q_m^*(P_m, \mathbf{d}) = [P_m \circ W_m^*]_y = \left\{ \sum_{\mathbf{x} \in \mathcal{X}^m} P_m(\mathbf{x}) W_m^*(\mathbf{y}|\mathbf{x}) \right\}, \tag{7}$$

where $P_m \circ W_m^*$ denotes the joint distribution, and $[P_m \circ W_m^*]_y$ denotes the y-marginal of the joint distribution. Now suppose that an arbitrary but strictly positive distribution $Q_m \ne Q_m^*$ is used to generate random codewords consisting of ℓ vectors. Then the minimum coding rate to guarantee a $\mathbf{d}$-match in probability to an independently generated source word, as $\ell \to \infty$, is [11],

$$R(P_m, Q_m, \mathbf{d}) = \min_{Q'_m} \{I_{\min}(P_m || Q'_m, \mathbf{d}) + \mathcal{D}(Q'_m || Q_m)\}, \tag{8}$$

where $\mathcal{D}(\cdot||\cdot)$ denotes the Kullback-Leibler (KL) divergence function, $I_{\min}(P_m||Q_m, \mathbf{d})$ is the minimum mutual information with constrained output distribution Q_m, i.e.

$$I_{\min}(P_m||Q_m, \mathbf{d}) = \begin{cases} \min\limits_{W \in \mathcal{W}(P_m, Q_m, \mathbf{d})} I(P_m, W) & \text{if } \mathcal{W}(P_m, Q_m, \mathbf{d}) \ne \emptyset \\ \infty & \text{otherwise} \end{cases}, \tag{9}$$

$$\mathcal{W}(P_m, Q_m, \mathbf{d}) = \left\{ W : [P_m \circ W]_y = Q_m,\ \rho_r(P_m, W) \le d_r, \forall r \right\}. \tag{10}$$

Let $Q_m^*(P_m, Q_m, \mathbf{d})$ denote the distribution that achieves $R(P_m, Q_m, \mathbf{d})$, i.e., the distribution that achieves the minimum in (8).

III. Natural Type Selection

This work builds on and expands the NTS random lossy codebook generation approach for discrete memoryless sources, which was originally proposed in [5] and practically enhanced in [6], [7]. Consider the memoryless case for which m is set to one, and the source letters are generated according to $P_1 = \{P_1(x) : x \in \mathcal{X}\}$. In this case the distortion constraint vector $\mathbf{d}$ is one-dimensional, hence $\mathbf{d} = d_1 = d$. Additionally, let the memoryless codebook reproduction distribution be $Q_1 = \{Q_1(y) : y \in \mathcal{Y}\}$. In [5], the authors showed that the empirical type of the codeword that d-matches an independently generated source word, converges in probability to $Q_1^*(P_1, Q_1, d)$ as the string length $\ell \to \infty$. Note that $Q_1^*(P_1, Q_1, d)$ is more efficient in coding the source than Q_1. This immediately suggests a recursive and iterative algorithm. Let n be the iteration index, $N_{1,\ell}$ be the index of the first d-matching codeword, and $Q_{n,1,\ell}^{N_{1,\ell}}$ be the d-matching codeword type. The subscript "1" here and below stands for $m = 1$. Starting with a strictly positive initial codebook reproduction

distribution denoted $Q_{0,1,\ell}$, the type of the d-matching codeword at the current iteration is used to generate the codebook of the next iteration. In other words, the next iteration's codebook reproduction distribution is naturally selected by the source through a d-match event, hence the name "natural type selection". This recursion results in a sequence of reproduction distributions,

$$Q_{n,1,\ell} = Q_{n-1,1,\ell}^{N_{1,\ell}}, \tag{11}$$

$$Q_{n,1} = \lim_{\ell\to\infty} Q_{n,1,\ell} = Q_1^*(P_1, Q_{n-1,1}, d), \ n = 1, 2, \dots \tag{12}$$

Additionally, it is shown that the recursion in (12) converges to the optimal codebook distribution $Q_1^*(P_1, d)$ that achieves the rate-distortion function $R(P_1, d)$ in (3) for $m = 1$, i.e., [5]

$$Q_1^*(P_1, d) = \lim_{n\to\infty} \lim_{\ell\to\infty} Q_{n,1,\ell}, \tag{13}$$

$$R(P_1, d) = \lim_{n\to\infty} \lim_{\ell\to\infty} R(P_1, Q_{n,1,\ell}, d). \tag{14}$$

While ensuring optimality, the original NTS algorithm suffers from a fundamental practical flaw. In order to converge to the optimal distribution, first the string length is sent to infinity, and only then can NTS iterations be done. In other words, the limit as $n \to \infty$ assumes that the codeword length ℓ is already very large. However as the codeword length ℓ increases, the probability of finding a d-match decreases exponentially resulting in an intractable d-search complexity for large ℓ which is at the heart of the NTS iteration. Hence, in practice it is the reversed order of limits that would "make more sense". In [7], a practically-effective and asymptotically optimal NTS algorithm was devised, where a finite source word or codeword length ℓ is considered. Instead of naively using the restricted type of the d-matching finite-length codeword, i.e., the "favourite type" as the next iteration codebook reproduction distribution, an ML framework is leveraged to identify the *general* distribution that most likely generates a *set* of K codewords that respectively d-match an independently generated set of K source words. Lemma 1 of [7] shows that the codebook reproduction distribution obtained through the ML framework for iteration n, and $n_K = 1$, is computed as,

$$Q_{n+1,1,\ell,K} = \hat{Q}^{\mathrm{ML}}_{n+1,1,\ell,K} = \frac{1}{K}\sum_{i=1} Q_{n,1,\ell}(\tilde{\mathbf{y}}_{j(i)}), \tag{15}$$

where $\tilde{\mathbf{y}}_{j(i)}$ is the ℓ-length codeword, of index $j(i)$, that achieves a d-match event to the i-th source string, and $Q_{n,1,\ell}(\mathbf{y}_{j(i)})$ is its corresponding type. Next, Theorem 1 and Theorem 2 of [7] show that the recursive codebook reproduction distribution in (15) converges to the optimal codebook reproduction distribution $Q^*(P_1, d)$, in probability with the preferred (practically-effective) order of limits, i.e.,

$$Q^*(P_1, d) = \lim_{\ell\to\infty} \lim_{n\to\infty} \lim_{K\to\infty} Q_{n,1,\ell,K}. \tag{16}$$

Moreover, for finite string length ℓ, the codebook reproduction distribution converges to the optimal *achievable* distribution in the set of distributions constrained by the string length ℓ. While resolving central practical shortcomings of the original NTS algorithm, the resulting improved NTS mechanism is still restricted to memoryless, uncorrelated sources. The fact that virtually all practice sources exhibit time-dependencies motivates the current work. In this paper, we provide a generalization of the asymptotically-optimal NTS algorithm to the broader class of sources with memory.

TABLE I
SUMMARY OF THE NTS PARAMETERS DEFINITIONS.

n	NTS iteration index.
m	Memory depth or size of "super-symbol".
ℓ	Number of super symbols encoded together.
K	Statistical depth for ML codebook distribution estimation.

IV. GENERALIZED NTS ALGORITHM

First, we define the minimum coding rate per letter for stationary sources with memory required to guarantee a $\mathbf{d}$-match with probability one asymptotically in ℓ as [8], [10],

$$R(\mathbf{d}) = \lim_{m\to\infty} m^{-1} R(P_m, \mathbf{d}). \tag{17}$$

Consequently, the optimal codebook reproduction distribution that achieves $R(\mathbf{d})$ is obtained as,

$$Q^*(\mathbf{d}) = \lim_{m\to\infty} Q_m^*(P_m, \mathbf{d}). \tag{18}$$

Let $Q_{n,m,\ell}(\tilde{\mathbf{y}}_{j(i)}) = \left\{Q(\mathbf{y}) : Q(\mathbf{y}) = \frac{1}{\ell} N(\mathbf{y}|\tilde{\mathbf{y}}_{j(i)}), \mathbf{y} \in \mathcal{Y}^m\right\}$ be the *generalized type* of codeword $\tilde{\mathbf{y}}_{j(i)}$, where $N(\mathbf{y}|\tilde{\mathbf{y}}_{j(i)})$ is the number of occurrences of vector $\mathbf{y}$ in the codeword. The ML estimation framework of (15) is directly extendable to the general case ($m \geq 1$): Given a set of K codewords (each consisting of ℓ independent sub-blocks) that respectively $\mathbf{d}$-match an independently generated set of K source words (also consisting of ℓ independent sub-blocks), the ML codebook reproduction distribution equals the average of the generalized codeword types. So the generalized NTS algorithm yields a sequence of reproduction distributions:

$$Q_{n+1,m,\ell,K} = \hat{Q}^{\mathrm{ML}}_{n+1,m,\ell,K} = \frac{1}{K}\sum_{i=1}^{K} Q_{n,m,\ell}(\tilde{\mathbf{y}}_{j(i)}), \tag{19}$$

$$Q_{n,m,\ell} = \lim_{K\to\infty} Q_{n,m,\ell,K}, \quad n = 1, 2, \dots \tag{20}$$

NTS parameter definitions are listed in Table I. Next, we show that the generalized NTS algorithm (for sources with memory) converges to the optimal reproduction distribution that achieves $R(\mathbf{d})$, asymptotically as $K \to \infty$, $n \to \infty$, $m \to \infty$ and $\ell \to \infty$. We start by defining the set of all $\mathbf{d}$-matching joint types for a given (source word and codewor) block length $m\ell$, and vector length m as,

$$\begin{aligned}\mathcal{V}_{m,\ell}(\mathbf{d}) \triangleq \Big\{V : V = P' \circ W', P' \in \mathcal{P}_{m,\ell},\ Q' = [V]_y,\\ Q' \in \mathcal{Q}_{m,\ell}, \rho_r(P', W') \leq d_r, \forall r\Big\},\end{aligned} \tag{21}$$

where $\mathcal{P}_{m,\ell}$ and $\mathcal{Q}_{m,\ell}$ are the sets of all possible $m\ell$-length block types over the input alphabet $\mathcal{X}^m$, and the output alphabet $\mathcal{Y}^m$, respectively. Define the set $E_{m,\ell}(P_m, \mathbf{d})$ as,

$$E_{m,\ell}(P_m, \mathbf{d}) \triangleq \Big\{V : V \in \mathrm{Conv}(\mathcal{V}_{m,\ell}(\mathbf{d})),\ [V]_x = P_m\Big\}, \tag{22}$$

where $\mathrm{Conv}(\cdot)$ denotes the convex hull of the argument (set), and $[V]_x$ denotes the x-marginal of the joint distribution V. Obviously, $E_{m,\ell}(P_m, \mathbf{d})$ is a convex set. We can now state our first main result.

Theorem 1: The reproduction distribution of the generalized recursive NTS algorithm in (19) converges asymptotically, as $K \to \infty$, given a fixed block length $m\ell$, and a fixed vector length m, to the distribution $Q^*_{m,\ell}(P_m, Q_{n-1,m,\ell}, \mathbf{d})$ in probability, i.e.,

$$\begin{aligned} Q_{n,m,\ell} &= \lim_{K\to\infty} Q_{n,m,\ell,K} = Q^*_{m,\ell}(P_m, Q_{n-1,m,\ell}, \mathbf{d}) \\ V^*_{m,\ell}(P_m, Q_m, \mathbf{d}) &\triangleq \operatorname*{argmin}_{V\in E_{m,\ell}(P_m,\mathbf{d})} \mathcal{D}\left(V||P_m \times Q_m\right), \\ Q^*_{m,\ell}(P_m, Q_m, \mathbf{d}) &= \left[V^*_{m,\ell}(P_m, Q_m, \mathbf{d})\right]_y . \end{aligned} \tag{23}$$

Proof: Let $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{y}}_{j(i)}$, $i = 1, 2, \ldots, K$, be a sequence of $\mathbf{d}$-matching $m\ell$-length blocks that are generated with P_m and Q_m over discrete alphabets $\mathcal{X}^m$ and $\mathcal{Y}^m$, respectively. Let P_i be the i-th instantaneous generalized type of $\tilde{\mathbf{x}}_i$, Q_j be the j-th instantaneous type of $\tilde{\mathbf{y}}_{j(i)}$, and $W_{i,j}$ be the instantaneous channel, i.e., $[P_i \circ W_{i,j}]_y = Q_j$. Next we look at the concatenated blocks. Let $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ be the $Km\ell$-length blocks constructed by concatenating $\{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_K\}$ blocks of length $m\ell$, and $\left\{\tilde{\mathbf{y}}_{j(1)}, \ldots, \tilde{\mathbf{y}}_{j(K)}\right\}$ blocks of length $m\ell$, respectively. Additionally, define $E_{K,m,\ell}(P_m, \mathbf{d})$ as,

$$\begin{aligned} E_{K,m,\ell}(P_m, \mathbf{d}) \triangleq \Big\{ V : V = \frac{1}{K}\sum_{i=1}^{K} P_i \circ W_{i,j},\ P_i \in \mathcal{P}_{m,\ell}, \\ Q_j \in \mathcal{Q}_{m,\ell}, \rho_r(P_i, W_{i,j}) \le d_r, \forall r, \\ P_i(\mathbf{x}) = \frac{N(\mathbf{x}|\tilde{\mathbf{x}}_i)}{\ell}, \tilde{\mathbf{X}} \sim P_m \Big\}, \end{aligned} \tag{24}$$

where $N(\mathbf{x}|\tilde{\mathbf{x}}_i)$ is the number of occurrences of vector $\mathbf{x} \in \mathcal{X}^m$ in the block $\tilde{\mathbf{x}}_i$. Additionally, $\tilde{\mathbf{X}} \sim P_m$ indicates that the vectors contained in the block $\tilde{\mathbf{x}}_i$ are i.i.d., and generated according to P_m. Note that the types of the concatenated blocks $\bar{\mathbf{x}}$, and $\bar{\mathbf{y}}$ are equal to the average of the respective sets of types $\{P_i\}$, and $\{Q_j\}$. We will show that for any $\delta > 0$, and sufficiently large K,

$$\begin{aligned} \mathbb{P}(\mathcal{D}(Q_{\bar{\mathbf{y}}}||Q^*_{m,\ell}(P_m, Q_m, \mathbf{d})) > 3\delta | V_{\bar{\mathbf{x}},\bar{\mathbf{y}}} \in E_{K,m,\ell}(P_m, \mathbf{d})) \le \\ (K\ell+1)^{2|\mathcal{X}^m||\mathcal{Y}^m|} e^{-K\ell\delta}. \end{aligned} \tag{25}$$

In other words, if we condition on the event that the joint type of $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$, denoted as $V_{\bar{\mathbf{x}},\bar{\mathbf{y}}}$, belongs to $E_{K,m,\ell}(P_m, \mathbf{d})$, the type of $\bar{\mathbf{y}}$, denoted as $Q_{\bar{\mathbf{y}}}$, is with high probability close in the divergence-sense to $Q^*_{m,\ell}(P_m, Q_m, \mathbf{d})$. Since closeness in divergence also implies closeness in the $\mathcal{L}_1$ sense [12, Lemma 11.6.1], this establishes Theorem 1. We start by verifying that, as $K \to \infty$, by (22) and (24), $E_{K,m,\ell}(P_m, \mathbf{d})$ approaches $E_{m,\ell}(P_m, \mathbf{d})$, hence define,

$$D^* = \min_{V\in E_{m,\ell}(P_m,\mathbf{d})} \mathcal{D}(V||P_m \times Q_m). \tag{26}$$

Then following [7], [12, Th. 11.6.2], and generalizing to $m \ge 1$, we obtain,

$$\begin{aligned} &\mathbb{P}(\mathcal{D}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}}||P_m \times Q_m) > D^* + 3\delta, V_{\bar{\mathbf{x}},\bar{\mathbf{y}}} \in E_{K,m,\ell}(P_m, \mathbf{d})) \le \\ &(K\ell+1)^{|\mathcal{X}^m||\mathcal{Y}^m|} \exp\left(-K\ell\left(D^* + 3\delta\right)\right), \end{aligned} \tag{27}$$

$$\begin{aligned} \mathbb{P}(\mathcal{D}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}}||P_m \times Q_m) \le D^* + 2\delta, V_{\bar{\mathbf{x}},\bar{\mathbf{y}}} \in E_{K,m,\ell}(P_m, \mathbf{d})) \ge \\ \frac{\exp\left(-K\ell\left(D^* + 2\delta\right)\right)}{(K\ell+1)^{|\mathcal{X}^m||\mathcal{Y}^m|}}, \end{aligned} \tag{28}$$

Next, taking into account that the probability of one event is larger than or equal to the probability of the intersection, we have,

$$\mathbb{P}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}} \in E_{K,m,\ell}(P_m, \mathbf{d})) \ge \frac{\exp\left(-K\ell\left(D^* + 2\delta\right)\right)}{(K\ell+1)^{|\mathcal{X}^m||\mathcal{Y}^m|}}. \tag{29}$$

By Bayes' law we get,

$$\begin{aligned} \mathbb{P}(\mathcal{D}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}}||P_m \times Q_m) > D^* + 3\delta | V_{\bar{\mathbf{x}},\bar{\mathbf{y}}} \in E_{K,m,\ell}(P_m, \mathbf{d})) \le \\ (K\ell+1)^{2|\mathcal{X}^m||\mathcal{Y}^m|} \exp\left(-K\ell\delta\right). \end{aligned} \tag{30}$$

By the "Pythagorean" theorem [12, Th. 11.6.1], we have,

$$\mathcal{D}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}}||V^*_{m,\ell}) + \mathcal{D}(V^*_{m,\ell}||P_m \times Q_m) \le \mathcal{D}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}}||P_m \times Q_m), \tag{31}$$

where, for ease of notation, $V^*_{m,\ell} = V^*_{m,\ell}(P_m, Q_m, \mathbf{d})$. Hence, $\mathcal{D}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}}||P_m \times Q_m) \le D^* + 3\delta$ implies that,

$$\mathcal{D}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}}||V^*_{m,\ell}) \le 3\delta. \tag{32}$$

Finally, by the data processing inequality, we have,

$$\mathcal{D}(Q_{\bar{\mathbf{y}}}||Q^*(P_m, Q_m, \mathbf{d})) \le \mathcal{D}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}}||V^*_{m,\ell}) \tag{33}$$

since, both are the respective y-marginals of the joint types. Hence, Theorem 1 follows from (30) as desired. ∎

Theorem 2: Given a strictly positive initial distribution $Q_{0,m,\ell}$ over $\mathcal{Y}^m$, the recursion in (20) achieves,

$$i) \qquad Q_{n,m,\ell} \to Q^*_{m,\ell}(P_m, \mathbf{d}), \quad \text{as } n \to \infty , \tag{34}$$

$$ii) \begin{matrix} R_\ell(P_m, \mathbf{d}) \to R(P_m, \mathbf{d}) \\ Q^*_{m,\ell}(P_m, \mathbf{d}) \to Q^*_m(P_m, \mathbf{d}) \end{matrix}, \quad \text{as } \ell \to \infty , \tag{35}$$

where $Q^*_m(P_m, \mathbf{d})$ is the optimum reproduction distribution that achieves the rate distortion function $R(P_m, \mathbf{d})$ for a fixed vector length m, and $Q^*_{m,\ell}(P_m, \mathbf{d})$ is the optimum achievable reproduction distribution for finite block length $m\ell$ and vector length m, that would achieve $R_\ell(P_m, \mathbf{d})$, i.e.,

$$\begin{aligned} &R_\ell(P_m, Q_m, \mathbf{d}) \triangleq \min_{V\in E_{m,\ell}(P_m,\mathbf{d})} \mathcal{D}\left(V||P_m \times Q_m\right), \\ &\mathcal{W}_{m,\ell}(P_m, \mathbf{d}) \triangleq \{W : P_m \circ W = V, V \in E_{m,\ell}(P_m, \mathbf{d})\}, \\ &W^*_{m,\ell}(P_m, \mathbf{d}) = W^*_{m,\ell} \triangleq \arg \min_{W\in\mathcal{W}_{m,\ell}(P_m,\mathbf{d})} I(P_m, W), \\ &R_\ell(P_m, \mathbf{d}) \triangleq \min_{Q_m} R_\ell(P_m, Q_m, \mathbf{d}) = I(P_m, W^*_{m,\ell}), \\ &Q^*_{m,\ell}(P_m, \mathbf{d}) \triangleq \left[P_m \circ W^*_{m,\ell}(P_m, \mathbf{d})\right]_y . \end{aligned} \tag{36}$$

Proof: From (36), we can write $R_\ell(P_m, \mathbf{d})$ as a double minimization over *convex* sets, i.e.,

$$R_\ell(P_m, \mathbf{d}) = \min_{Q_m} \min_{V\in E_{m,\ell}(P_m,\mathbf{d})} \mathcal{D}\left(V||P_m \times Q_m\right). \tag{37}$$

Furthermore, it should be noted that for a fixed V, the reproduction distribution which minimizes $\mathcal{D}(V||P_m \times Q_m)$ is the y-marginal of V. On the other hand, for a fixed Q_m and distortion constraint $\mathbf{d}$, the joint distribution which minimizes $\mathcal{D}(V||P_m \times Q_m)$ over $E_{m,\ell}(P_m, \mathbf{d})$ will induce $Q^*_{m,\ell}(P_m, Q_m, \mathbf{d})$. By the results of Theorem 1, the recursion in (20) performs exactly this minimization over the convex sets, i.e.,

$$V^*_{m,\ell}(P_m, Q_{0,m,\ell}, \mathbf{d}) \to (P_m \times Q^*_{m,\ell}(P_m, Q_{0,m,\ell}, \mathbf{d})) \to V^*_{m,\ell}(P_m, Q_{1,m,\ell}, \mathbf{d}) \to (P_m \times Q^*_{m,\ell}(P_m, Q_{1,m,\ell}, \mathbf{d})) \dots$$

It should be noted that the distance in the alternating minimization is measured by divergence. Hence, by [13, Th. 3], the sequences of divergences and distributions will converge to the minimum divergence, i.e., $R_\ell(P_m, \mathbf{d})$, and the corresponding optimum reproduction distribution $Q^*_{m,\ell}(P_m, \mathbf{d})$. Next, to show part $ii)$ of Theorem 2 stated in (35), first verify that the minimum coding rate with constrained reproduction distribution Q_m, denoted as $R(P_m, Q_m, \mathbf{d})$, is written as [5],

$$R(P_m, Q_m, \mathbf{d}) = \min_{W:\rho_r(P_m,W)\le d_r,\forall r} I(P_m, W) + \mathcal{D}([P_m \circ W]_y \,||Q_m), \quad (38)$$

$$R(P_m,Q_m,\mathbf{d}) = \min_{W:\rho_r(P_m,W)\le d_r,\forall r} \mathcal{D}(P_m \circ W||P_m \times Q_m), \quad (39)$$

Hence, $R(\mathbf{d})$ follows from (39) as,

$$R(\mathbf{d}) = \lim_{m\to\infty} \frac{1}{m} R(P_m, \mathbf{d}) = \lim_{m\to\infty} \frac{1}{m} \min_{Q_m} \min_{W:\rho_r(P_m,W)\le d_r,\forall r} \mathcal{D}(P_m \circ W||P_m \times Q_m). \quad (40)$$

Now, as $\ell \to \infty$, it is straight forward to show that,

$$E_{m,\ell}(P_m, \mathbf{d}) \to \{V : V = P_m \circ W', \rho_r(P_m, W') \le d_r, \forall r\}. \quad (41)$$

Consequently, as $\ell \to \infty$, and from (36), (37), and (40),

$$R_\ell(P_m,Q_m,\mathbf{d}) \to \min_{W:\rho_r(P_m,W)\le d_r,\forall r} \mathcal{D}\left(P_m \circ W||P_m \times Q_m\right), \quad (42)$$

$$R_\ell(P_m, \mathbf{d}) \to R(P_m, \mathbf{d}). \quad (43)$$

Thus, by the definition of $Q^*_{m,\ell}(P_m, \mathbf{d})$ in (36), and (7), part $ii)$ of Theorem 2 follows. ■

V. Toy Example

A binary-alphabet stationary Markov chain source is considered, i.e., $\mathcal{X} = \mathcal{Y} = \{0, 1\}$. Let the source transition probabilities be $P(0|1) = P(0|0) = 0.7$, and $P(1|0) = P(1|1) = 0.3$ (This is also a simple Gilbert-Elliott model). In the simulation environment, the vector length is set to two, i.e., $m = 2$, which corresponds to source distribution $P_2 = \{0.49, 0.21, 0.21, 0.09\}$. For this setting, the average Hamming distortion function is considered, i.e., the distortion between source vector $\mathbf{x}$ and code vector $\mathbf{y}$ is the frequency of positions at which the corresponding source and code letters differ. Two cases of distortion constraints are considered $(\forall r \in \{1, 2, \dots, m\})$: $i)$ $d_r = d_{\max} = 0.3$, and $ii)$ $d_r = 0.25$. Note that $Q^*_2(P_2, \mathbf{d}_{\max}) = \{1, 0, 0, 0\}$, i.e., the optimal reproduction distribution collapses onto one point that corresponds to the vector $\mathbf{y} = [0, 0]$. The generalized NTS algorithm is run with $K = 10^5$, and different values of ℓ. The evolution of the codebook reproduction distribution of the vector $\mathbf{y} = [0, 0]$ is plotted across the NTS iteration index n in Fig 1. For $\ell = 1$, and the considered distortion constraints, the only codeword that can $\mathbf{d}$-match the source word is identical to it. Hence, the codebook reproduction distribution in (20) immediately converges to the source distribution P_2. As ℓ increases, the set of all possible $\mathbf{d}$-matching joint types $\mathcal{V}_{2,\ell}(\mathbf{d})$ expands, and it can be observed that the codebook reproduction distribution $Q^*_{2,\ell}(P_2, \mathbf{d})$ approaches $Q^*_2(P_2, \mathbf{d})$ as derived in Theorem 2.

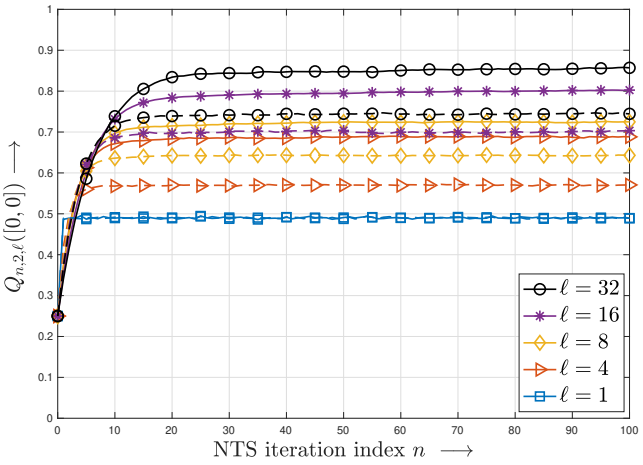

Fig. 1. Convergence of the codebook reproduction distributions for $d_r = 0.3$ (solid lines), or $d_r = 0.25$ (dashed lines), and different values of ℓ.

VI. Conclusion

This paper generalizes the asymptotically-optimal iterative NTS lossy codebook generating algorithm to sources with memory. The codeword type is generalized to be the frequency of occurrence of *vectors* with length m in a codeword of length $m\ell$. Similar to [7], an ML framework is leveraged to identify the next iteration codebook reproduction distribution after observing sequence of K $\mathbf{d}$-matching events. We show that for finite vector length m, and codeword length $m\ell$, the codebook reproduction distribution converges to the optimal *achievable* distribution $Q^*_{m,\ell}$ in probability as $K \to \infty$ and $n \to \infty$. Additionally, we show that asymptotically in m and ℓ, the codebook reproduction distribution converges to the optimal distribution Q^* that achieves the rate-distortion bound for sources with memory.

References

[1] A. Lempel and J. Ziv, "On the complexity of finite sequences," *IEEE Trans. on Inf. Theory*, vol. IT-22, No. 1, pp. 75–81, 1976.

[2] 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.

[3] ——, "Compression of individual sequences via variable rate coding," *IEEE Trans. on Inf. Theory*, vol. IT-24, pp. 530–536, 1978.

[4] K. Rose, "A mapping approach to rate-distortion computation and analysis," *IEEE Trans. on Inform. Theory*, vol. 40, Nov. 1994.

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

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

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

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

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

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

[11] E. H. 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.

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

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