Stochastic Codebook Regeneration for Sequential Compression of Continuous Alphabet Sources

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

2021 IEEE International Symposium on Information Theory (ISIT) | 978-1-5386-8209-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISIT45174.2021.9517877

***Abstract*—This paper proposes an effective and asymptotically optimal framework for stochastic, adaptive codebook regeneration for sequential ("on the fly") lossy coding of continuous alphabet sources. Earlier work has shown that the rate-distortion bound can be asymptotically achieved for discrete alphabet sources, by a "natural type selection" (NTS) algorithm. At each iteration n, a maximum-likelihood framework is used to estimate the reproduction distribution most likely to generate the empirical types of a sequence of K length-ℓ codewords that respectively "d-match" (i.e., are within distortion d from) a sequence of K length-ℓ source words. The reproduction distribution estimated at iteration n is used to regenerate the codebook for iteration $n+1$. The sequence of reproduction distributions was shown to converge, asymptotically in K, n, and ℓ, to the optimal distribution that achieves the rate-distortion bound for discrete alphabet sources. This work generalizes the NTS framework to handle sources over more general (e.g., continuous) alphabet spaces, which often preclude a natural interpretation of the concept of "type". We show, for continuous alphabet sources and fixed block length ℓ, that as $K \to \infty$ and $n \to \infty$, the sequence of estimated reproduction distributions converges, in the weak convergence sense, to a distribution that achieves the rate-distortion bound, albeit for an auxiliary distortion measure introduced as subterfuge to effectively impose a maximum distortion constraint over K blocks. Leveraging this result, we establish that the sequence of reproduction distributions converges, asymptotically in ℓ, to the optimal codebook reproduction distribution Q^* that achieves the rate-distortion bound, with respect to the original distortion measure.**

I. INTRODUCTION

Random codebook generation is a cornerstone of the theory of source coding and serves to establish its performance bounds. It also has had a significant impact on practical source coding, especially in the context of lossless coding, most notably due to the seminal contributions of Lempel and Ziv [1] [2] [3], and extensive consequent contributions by others. For example, in LZ78 [3], a tree of codewords is grown, as source strings are encoded, in a way that ensures that the frequency of typical source sequences among codewords in the tree, asymptotically approaches one, without recourse to any prior knowledge of the source statistics.

In the lossy coding setting, stochastic codebook generation is fundamentally more challenging. Recall that in lossless coding the optimal codebook must exhibit the same statistics as the source, and hence the underlying problem is effectively that of *learning* the source statistics from source examples, and generating a codebook that exhibits such statistics. However, in lossy coding, the optimal codebook-generating distribution generally differs from the source distribution, and considerably so at high distortion levels. Hence, it is not enough to simply "mimic" the source, and thus, finding the optimal codebook reproduction distribution represents a significant challenge. This observation was recognized in [4], where codebook adaptation in the lossy settings (for discrete alphabet, memoryless sources) is viewed as a sequential process of "type selection" rather than learning and matching the source statistics. Ultimately, the codebook adaptation algorithm estimates the optimal *reproduction type* for the source at a given distortion constraint d. Most relevant to this work is the early NTS algorithm, originally proposed for discrete-alphabet memoryless sources in [4], practically improved (enabling most iterations to be performed at short block lengths) in [5], and most recently generalized to sources with memory in [6].

Consider a discrete memoryless source drawn from distribution P. At each iteration n, NTS algorithm generates a codebook of independent codewords from distribution Q_n. A set of K independently generated (ℓ-length) source words get "d-matched" by a respective set of K codewords in the codebook. A Maximum Likelihood (ML) framework is employed to find a new codebook reproduction distribution, Q_{n+1}, which is most likely to have generated the sequence of d-matching codewords. NTS algorithm then uses Q_{n+1} to regenerate a new codebook for iteration $n+1$. It was shown that the sequence of codebook generating distributions $\{Q_n\}$ converges, asymptotically in K, n, and ℓ, to the optimal reproduction distribution $Q^*(P,d)$ that achieves the rate-distortion bound $R(P,d)$. The recent extension to accommodate discrete sources with memory [6], also establishes asymptotic optimality, but details are omitted here for brevity.

While earlier NTS work was focused on discrete alphabet sources, the prevalence of continuous sources in practical compression applications provides strong motivation for this paper. It is important to emphasize that the standard concept of types, which was the cornerstone of the earlier NTS work on discrete alphabet sources, and was specifically instrumental to showing asymptotic convergence to the reconstruction distribution that achieves the rate-distortion bound, does not apply to continuous alphabet sources. Hence, the generalization of the NTS algorithm to accommodate continuous alphabet sources is not straightforward and is, in fact, fundamentally more challenging. Important advances were made in [7], which studied abstract alphabet spaces in the random codebook coding context of plain and entropy-constrained quantization, and further generalized the conditional limit theorem, which is at the heart of the NTS algorithm, to stationary ergodic sources with abstract alphabet spaces. The current work generalizes the NTS stochastic codebook generation algorithm and establishes its asymptotic optimality for continuous sources and sources over abstract alphabet spaces (specifically, Polish spaces). Theorems 1 and 2 establish that, at fixed block length ℓ, the codebook reproduction probability measure converges asymptotically, in the weak convergence sense, as $K \to \infty$, $n \to \infty$, to the optimal probability measure that achieves the rate-distortion function, albeit for an auxiliary distortion measure, designed to capture a maximum distortion constraint over a set of blocks. Theorem 3 proceeds to establish that the marginal probability measure of the codebook reproduction distribution converges weakly to the optimal distribution $Q^*(P,d)$, as $\ell \to \infty$ and $n \to \infty$.

The remainder of the paper is organized as follows: Section II provides background; Section III reviews NTS algorithms for discrete memoryless sources; The main results for continuous (and abstract) alphabet sources are covered in Section IV; with conclusions in Section V.

II. Relevant Background

Let $\{X_u\}_{u\geq 1}$ be a memoryless source, whose samples are drawn from alphabet $\mathcal{X}$ according to probability distribution P. Denote the source realizations, $x_u \in \mathcal{X}$. We assume that the alphabet $\mathcal{X}$ is a complete separable metric space (often called Polish space), equipped with its associated Borel σ-field $\mathcal{X}'$. Similarly, we assume that the reproduction alphabet $\mathcal{Y}$ is also a Polish space equipped with its associated Borel σ-field $\mathcal{Y}'$. We denote the ℓ-length source random vector $\mathbf{X}_1^\ell = (X_1, X_2, \ldots, X_\ell)$, and its realizations $\mathbf{x}_1^\ell \in \mathcal{X}^\ell$. We define an arbitrary non-negative (measurable) single-letter distortion function $\rho : \mathcal{X}\times\mathcal{Y} \to [0,\infty)$. The distortion between source vector $\mathbf{x}_1^\ell$ and codevector $\mathbf{y}_1^\ell = (y_1, y_2, \ldots, y_\ell) \in \mathcal{Y}^\ell$, is assumed additive:

$$\rho(\mathbf{x}_1^\ell, \mathbf{y}_1^\ell) = \frac{1}{\ell}\sum_{i=1}^{\ell} \rho(x_i, y_y). \tag{1}$$

Given fidelity constraint d, we define a "d-match" event as the event that $\rho(\mathbf{x}_1^\ell, \mathbf{y}_1^\ell) \leq d$ is satisfied. Suppose a random codebook $\mathcal{C}_\ell$ of infinite i.i.d. codewords $\mathbf{Y}_1^\ell(i)$, with $i \geq 1$, such that the letters of any codeword $\mathbf{Y}_1^\ell(i)$ are i.i.d. and drawn from distribution Q over the reproduction alphabet $\mathcal{Y}$. We call Q the codebook reproduction distribution. Let N_ℓ be the index of the first codeword in $\mathcal{C}_\ell$ that d-matches the source word realization $\mathbf{x}_1^\ell$, i.e.,

$$N_\ell = \inf\{i \geq 1 : \rho(\mathbf{x}_1^\ell, \mathbf{y}_1^\ell(i)) \leq d\}, \tag{2}$$

with the convention that the infimum of an empty set is $+\infty$. Shannon's lossy source coding theorem states: if a random codebook $\mathcal{C}_\ell$ of size $\exp(\ell(R(P,d)+\epsilon))$ is generated using an optimal codebook reproduction distribution $Q^*_{P,d}$, then the probability of finding a d-match in the codebook to an independently generated source vector via P goes to one as ℓ goes to infinity, wherein $R(P,d)$ is the rate-distortion function, i.e., [8]

$$R(P,d) = \inf_{\substack{V:[V]_x = P,\\ \mathbb{E}_V(\rho(X,Y))\leq d}} I(X,Y). \tag{3}$$

Here, $I(X,Y)$ is the mutual information between random variables X and Y, and the infimum is taken over all joint probability measure V such that the x-marginal of V, denoted $[V]_x$, is P and the expected distortion $\mathbb{E}_V(\rho(X,Y)) \leq d$. Let $V^*_{P,d}$ be the optimal joint distribution that realizes the infimum in (3), then the optimal codebook reproduction distribution $Q^*_{P,d}$ is the y-marginal of the optimal joint distribution $V^*_{P,d}$. However, if a random codebook is generated from distribution $Q \neq Q^*_{P,d}$, then the minimum encoding rate to guarantee a d-match in probability, as ℓ goes to infinity, was effectively shown in [9], and extended to abstract alphabets in [7], to be

$$R(P,Q,d) = \inf_{\substack{V:[V]_x = P,\\ \mathbb{E}_V(\rho(X,Y))\leq d}} \mathcal{D}(V||P\times Q), \tag{4}$$

$$R(P,Q,d) = \inf_{Q'}\{I_{\min}(P||Q',d) + \mathcal{D}(Q'||Q)\}, \tag{5}$$

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

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

Here the infimum is taken over all joint distributions V, whose x-marginal is P, and y-marginal is Q', and such that the expected distortion does not exceed d. In [10, Th. 2], it was shown that, under these assumptions, $R(P,Q,d)$ is finite, strictly positive, and that the infimum in its definition in (4) is always achieved by some joint distribution $V^*_{P,Q,d}$. Moreover, since the set of V over which the infimum is taken is convex, from [11] it can be concluded that $V^*_{P,Q,d}$ is the unique minimizer. Hence, a unique minimizer to (5) also exists, i.e.,

$$Q^*_{P,Q,d} = \arg\min_{Q'}\{I_{\min}(P||Q',d) + \mathcal{D}(Q'||Q)\}. \tag{7}$$

III. Natural Type Selection

The results summarized in the next section build on and expand the NTS random lossy codebook generation approach for discrete memoryless sources, which was originally proposed in [4] and practically enhanced in [5], [12]. Let the source be memoryless and drawn from *discrete* alphabet $\tilde{\mathcal{X}}$ according to $\tilde{P} = \{\tilde{P}(x) : x \in \tilde{\mathcal{X}}\}$. Additionally, let $\tilde{\mathcal{Y}}$ be the *discrete* codebook reproduction alphabet, and consider a codebook generated according to distribution $\tilde{Q} = \{\tilde{Q}(y) : y \in \tilde{\mathcal{Y}}\}$. In [4], it was shown that the empirical type of the codeword that d-matches an independently generated source word, converges in probability to $Q^*_{P,Q,d}$ as the string length ℓ goes to infinity. Note that $Q^*_{\tilde{P},\tilde{Q},d}$ is more efficient in coding the source than $\tilde{Q}$. This immediately suggests an iterative codebook generation algorithm. Let n be the iteration index, N_ℓ be the first codeword index that d-matches the source, and $Q^{N_\ell}_{n,\ell}$

be the d-matching code-word type. Starting with a strictly positive initial codebook reproduction distribution denoted $Q_{0,\ell}$, 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" (with a nod to Darwin's theory of evolution). This recursion results in a sequence of reproduction distributions,

$$Q_{n,\ell} = Q^{N_\ell}_{n-1,\ell}, \tag{8}$$

$$Q_n = \lim_{\ell\to\infty} Q_{n,\ell} = Q^*_{P,Q_{n-1},d},\ n = 1, 2, \ldots \tag{9}$$

Additionally, it is shown that the recursion in (9) converges to the optimal codebook distribution $Q^*_{\tilde{P},d}$ that achieves the rate-distortion function $R(P, d)$ in (3), i.e., [4]

$$Q^*(\tilde{P}, d) = \lim_{n\to\infty} \lim_{\ell\to\infty} Q_{n,\ell}, \tag{10}$$

$$R(\tilde{P}, d) = \lim_{n\to\infty} \lim_{\ell\to\infty} R(\tilde{P}, Q_{n,\ell}, d). \tag{11}$$

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 run. In other words, the limit as $n \to \infty$ assumes that the string length ℓ is already very large. Unfortunately, the probability of finding a d-match decreases exponentially with the string length, resulting in an intractable d-search complexity even at the early NTS iterations. Hence, in practice it is the reversed order of limits that would be desirable. This shortcoming was eliminated in [5], by devising a practically-effective and asymptotically optimal NTS algorithm, where a finite string length ℓ is considered. Instead of "naively" using the restricted type of the first d-matching codeword, i.e., the "favourite type" as the next iteration's codebook reproduction distribution, an ML framework is leveraged to identify the *general* distribution that most likely generates a set of d-matching codewords for a set of independently generated source words. Let the size of the d-matching codewords set be K. Lemma 1 of [5] shows that the codebook reproduction distribution obtained through the ML framework for iteration n, is computed as,

$$Q_{n+1,\ell,K} = \hat{Q}^{\mathrm{ML}}_{n+1,\ell,K} = \frac{1}{K}\sum_{i=1}^{K} Q_{\mathbf{y}_1^\ell(j(i))}, \tag{12}$$

where $\mathbf{y}_1^\ell(j(i))$ is the ℓ-length code-word, of index $j(i)$, that achieves a d-match event to the i-th source string, and $Q_{\mathbf{y}_1^\ell(j(i))}$ is its corresponding type. Next, Theorem 1 and Theorem 2 of [5] establish that the recursive codebook reproduction distribution in (12) tends to the optimum codebook reproduction distribution $Q^*_{\tilde{P},d}$ in probability through a practically-effective order of limits, i.e.,

$$Q^*_{\tilde{P},d} = \lim_{\ell\to\infty} \lim_{n\to\infty} \lim_{K\to\infty} Q_{n,\ell,K}. \tag{13}$$

Additionally, for a fixed string length ℓ, the codebook reproduction distribution converges to the optimal *achievable* distribution in a set defined by the string length ℓ. While resolving the main practical shortcoming of the original NTS algorithm, the main limitation is the fact that NTS is still restricted to discrete alphabets, and the importance of lossy coding for continuous alphabet sources provides strong motivation for this work and the results provided next.

TABLE I
SUMMARY OF THE NTS PARAMETERS DEFINITIONS.

n	NTS iteration index.
ℓ	Number of super symbols encoded together.
K	Statistical depth for ML codebook distribution estimation.

IV. MAIN RESULTS: GENERALIZATION TO CONTINUOUS ALPHABET SOURCES

This section generalizes the NTS algorithm to abstract alphabet memoryless sources. It is worth noting that this generalization is *fundamentally challenging* because the concept of types, which is the heart of earlier work NTS theorems in [4]–[6] that establish the asymptotic algorithm convergence behavior, does not immediately apply to continuous alphabets sources. Let n denote the iteration index, and let $Q_{n,\ell}$ be the codebook reproduction distribution (probability measure) for generating ℓ-length codewords in $\mathcal{Y}^\ell$. A summary of the parameter definitions is given in Table I. The algorithm considers a sequence of d-match events for a sequence of independently generated source words (or vectors) $\mathbf{x}_1^\ell(1)$, $\mathbf{x}_1^\ell(2)$, …, $\mathbf{x}_1^\ell(K)$. Let $\mathbf{y}_1^\ell(j(1))$, $\mathbf{y}_1^\ell(j(2))$, …, $\mathbf{y}_1^\ell(j(K))$ be the sequence of codewords, each generated according to $Q_{n,\ell}$, which d-matches the respective sequence of source words. The ML estimation framework in (12) can be extended to the general case of abstract alphabets. Hence, the ML codebook reproduction distribution that would have generated these d-matching code-words is equal to the average of the codeword empirical distributions, i.e.,

$$Q_{n+1,\ell,K} = \frac{1}{K}\sum_{i=1}^{K} Q_{\mathbf{y}_1^\ell(j(i))}, \tag{14}$$

$$Q_{n+1,\ell} = \lim_{K\to\infty} Q_{n+1,\ell,K}, \tag{15}$$

$$Q_{n+1} = \lim_{\ell\to\infty} Q_{n+1,\ell}, \tag{16}$$

where $Q_{\mathbf{y}_1^\ell(j(i))}$ is the empirical distribution of the d-matching codeword $\mathbf{y}_1^\ell(j(i))$ on $\mathcal{Y}^\ell$,i.e.,

$$Q_{n+1,\ell,K} = \frac{1}{K}\sum_{i=1}^{K} \delta_{\mathbf{y}_1^\ell(j(i))}, \tag{17}$$

with $\delta_{\mathbf{y}_1^\ell(j(i))}$ denoting a Dirac delta function located at $\mathbf{y}_1^\ell(j(i))$. Next consider a sequence of K concatenated (not necessarily sequentially generated) source and code vectors, i.e., let the $K\ell$-length source and code blocks be denotes as $\overline{\mathbf{x}} = [\mathbf{x}_1^\ell(i_1)\ \mathbf{x}_1^\ell(i_2)\ \ldots \mathbf{x}_1^\ell(i_K)]$, and $\overline{\mathbf{y}} = [\mathbf{y}_1^\ell(j_1)\ \mathbf{y}_1^\ell(j_2)\ \ldots \mathbf{y}_1^\ell(j_K)]$, respectively. Define the following auxiliary distortion measure $\left(\overline{\rho}_d : \mathcal{X}^\ell \times \mathcal{Y}^\ell \to \{0, 1\}\right)$, which is additive across the K ℓ-length blocks, i.e.,

$$\overline{\rho}_d\left(\mathbf{x}_1^\ell(i_r), \mathbf{y}_1^\ell(j_r)\right) = \begin{cases} 0 & \text{if } \rho\left(\mathbf{x}_1^\ell(i_r), \mathbf{y}_1^\ell(j_r)\right) \le d \\ 1 & \text{if } \rho\left(\mathbf{x}_1^\ell(i_r), \mathbf{y}_1^\ell(j_r)\right) > d \end{cases} \tag{18}$$

$$\overline{\rho}_d(\overline{\mathbf{x}}, \overline{\mathbf{y}}) = \frac{1}{K}\sum_{i=1}^{K} \overline{\rho}_d\left(\mathbf{x}_1^\ell(i_r), \mathbf{y}_1^\ell(j_r)\right), \tag{19}$$

Note that by setting $\overline{\rho}_d(\overline{\mathbf{x}},\overline{\mathbf{y}})=0$, we impose a requirement of maximum distortion d per block, over the K blocks. Thus, the auxiliary distortion measure $\overline{\rho}_d$ is a subterfuge to impose maximum distortion while maintaining the additive property over the K blocks.

Theorem 1: For a memoryless source $\{X_u\}_{u\geq 1}$, the probability measure $Q_{n,\ell,K}$ on $\mathcal{Y}^\ell$ converges as K goes to infinity, to the optimal distribution $\overline{Q}^*_{P^\ell,Q^\ell,d'}$ that achieves the bound $R(P^\ell,Q^\ell,d')$, for the auxiliary distortion measure $\overline{\rho}_d$, with the extreme distortion constraint $d'=0$, i.e.,

$$Q_{n,\ell,K} \Longrightarrow \overline{Q}^*_{P^\ell,Q^\ell,d'},\ d'=0, \quad \text{as} \quad K\to\infty, \tag{20}$$

where "$\Longrightarrow$" denotes weak convergence of probability measures.

Proof: Let $\mathbf{x}_1^\ell(i)$ and $\mathbf{y}_1^\ell(j(i))$, $i=1,2,\ldots,K$, be a sequence of d-matching ℓ-length words that are generated with the product probability measure P^ℓ over $\mathcal{X}^\ell$, and $Q_{n-1,\ell}$ over $\mathcal{Y}^\ell$, respectively. For the ease of notations, denote $Q_{n-1,\ell}$ as Q^ℓ. In other words, $\rho(\mathbf{x}_1^\ell(i),\mathbf{y}_1^\ell(j(i)))\leq d, \forall i\in\{1,2,\ldots,K\}$. Now let us consider the realizations of the concatenated source and code vectors $\overline{\mathbf{x}}=[\mathbf{x}_1^\ell(1)\ \ \mathbf{x}_1^\ell(2)\ \ \ldots\mathbf{x}_1^\ell(K)]$, and $\overline{\mathbf{y}}=[\mathbf{y}_1^\ell(j(1))\ \ \mathbf{y}_1^\ell(j(2))\ \ \ldots\mathbf{y}_1^\ell(j(K))]$. The set of conditions $\rho\left(\mathbf{x}_1^\ell(i),\mathbf{y}_1^\ell(j(i))\right)\leq d,\forall i\in\{1,2,\ldots,K\}$ implies that $\overline{\rho}_d(\overline{\mathbf{x}},\overline{\mathbf{y}})=0$, or in other words, the auxiliary distortion function $\overline{\rho}_d$ is satisfied between $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ with *zero* distortion constraint. Next define $\overline{Q}^*_{P^\ell,Q^\ell,d'}$ as the optimal distribution that minimize the coding rate for a given codebook reproduction distribution Q^ℓ, and for the auxiliary distortion measure $\overline{\rho}_d$, as in (7), i.e.,

$$\overline{Q}^*_{P^\ell,Q^\ell,d'}=\arg\inf_{Q'}\left\{I_{\min}(P^\ell||Q',d')+D(Q'||Q^\ell)\right\}, \tag{21}$$

$$I_{\min}(P^\ell||Q',d')=\inf_{\substack{V:[V]_x=P^\ell,\ [V]_y=Q',\\ \mathbb{E}_V(\overline{\rho}(\mathbf{X}_1^\ell,\mathbf{Y}_1^\ell))\leq d'}} I(\mathbf{X}_1^\ell,\mathbf{Y}_1^\ell), \tag{22}$$

Now consider the codewords' empirical distributions obtained by the following algorithms:

1) At iteration n, the algorithm *independently* finds a sequence of d'-matching codewords $\mathbf{y}_1^\ell(j(i))$ of length ℓ (considering distortion measure $\overline{\rho}_d$) to a respective sequence of independently generated source-words $\mathbf{x}_1^\ell(i)$, with $i\in 1,2,\ldots,K$. We say a source word "d'-matches" a codeword if $\overline{\rho}_d\left(\mathbf{x}_1^\ell(i),\mathbf{y}_1^\ell(j(i))\right)\leq d'$. The codewords $\mathbf{y}_1^\ell(j(i))$ belongs to random codebook $\mathcal{C}_\ell$ generated by distribution $Q_{n-1,\ell}$. Let $\overline{\mathbf{y}}=[\mathbf{y}_1^\ell(j(1)),\ldots,\mathbf{y}_1^\ell(j(K))]$ be a realization of the concatenated d'-matching code vectors and let $Q_{\overline{\mathbf{Y}}}$ be its empirical distribution on $\mathcal{Y}^\ell$. Note that by definition $Q_{\overline{\mathbf{Y}}}=Q_{n,\ell,K}$ in (17).
2) At iteration n, the algorithm finds a *single* long d'-matching codeword $\mathbf{y}_1^{K\ell}$ of length $K\ell$ (considering distortion measure $\overline{\rho}_d$) to a respective long source-word $\mathbf{x}_1^{K\ell}$, such that every respective ℓ-length source word and codeword *jointly* d'-matches. In other words, if $\mathbf{x}_1^{K\ell}=[\mathbf{x}_1^\ell(1),\ldots,\mathbf{x}_1^\ell(K)]$, and $\mathbf{y}_1^{K\ell}=[\mathbf{y}_1^\ell(1),\ldots,\mathbf{y}_1^\ell(K)]$, then $\overline{\rho}_d(\mathbf{x}_1^\ell(i),\mathbf{y}_1^\ell(i))\leq d',\ \forall i\in\{1,\ldots,K\}$. The d'-matching codeword $\mathbf{y}_1^{K\ell}$ belongs to random codebook $\mathcal{C}_{K\ell}$ generated by K-th product probability measure of $Q_{n-1,\ell}$, i.e., $Q_{n-1,\ell}^K$ on $\mathcal{Y}^{K\ell}$. Let $Q_{\overline{\mathbf{Y}}'}$ be the marginal empirical distribution of the d'-matching codeword on $\mathcal{Y}$.

By the independent generation of every ℓ-length part of the source blocks, code blocks, and the definition of the distortion measure $\overline{\rho}_d$ for $d'=0$, we show that for every (measurable) set $E\subset\mathcal{Y}^\ell$, $Q_{\overline{\mathbf{Y}}}(E)=Q_{\overline{\mathbf{Y}}'}(E)$ for $d'=0$. In view of (18) and (19) we have,

$$\mathbb{P}\left(\overline{\rho}_d(\overline{\mathbf{X}},\overline{\mathbf{Y}})=0\ \middle|\ \overline{\mathbf{X}}=\overline{\mathbf{x}}\right)= \prod_{i=1}^{K}\mathbb{P}\left(\overline{\rho}_d\left(\mathbf{X}_1^\ell(i),\mathbf{Y}_1^\ell(j(i))\right)=0\ \middle|\ \mathbf{X}_1^\ell(i)=\mathbf{x}_1^\ell(i)\right). \tag{23}$$

$$\mathbb{P}\left(\overline{\rho}_d(\overline{\mathbf{X}},\overline{\mathbf{Y}})=0\ \middle|\ \overline{\mathbf{X}}=\overline{\mathbf{x}}\right)= \prod_{i=1}^{K}\mathbb{P}\left(\rho\left(\mathbf{X}_1^\ell(i),\mathbf{Y}_1^\ell(j(i))\right)\leq d\ \middle|\ \mathbf{X}_1^\ell(i)=\mathbf{x}_1^\ell(i)\right). \tag{24}$$

Hence, the d'-match event (for $\overline{\rho}_d$ with $d'=0$) between the $K\ell$-length random source word $\mathbf{X}_1^{K\ell}$ and codeword $\mathbf{Y}_1^{K\ell}$ implies a sequence of d'-match events for every ℓ-length part. This implies that for any (measurable) set $E\subset\mathcal{Y}^\ell$, we have,

$$\mathbb{P}\left(Q_{\overline{\mathbf{Y}}}(E)=q\ \middle|\ \overline{\rho}_d\left(\mathbf{X}_1^{K\ell},\mathbf{Y}_1^{K\ell}\right)=0,\ \mathbf{X}_1^{K\ell}=\mathbf{x}_1^{K\ell}\right)= \mathbb{P}\left(Q_{\overline{\mathbf{Y}}}(E)=q\ \middle|\ \overline{\rho}_d\left(\mathbf{X}_1^\ell(i),\mathbf{Y}_1^\ell(i)\right)=0,\ \mathbf{X}_1^\ell(i)=\mathbf{x}_1^\ell(i)\right) \tag{25}$$

This, together with the independent generation of source words and codewords for every ℓ-length part, immediately shows that for every (measurable) $E\subset\mathcal{Y}^\ell$, $Q_{\overline{\mathbf{Y}}}(E)=Q_{\overline{\mathbf{Y}}'}(E)$.

Next by [7, Th. 3], for every (measurable) $E\subset\mathcal{Y}^\ell$, the probability,

$$\mathbb{P}\left(\left|\hat{Q}_{\overline{\mathbf{Y}}}(E)-\overline{Q}^*_{P^\ell,Q^\ell,d'}(E)\right|>\delta\ \middle|\ \overline{\rho}_d(\overline{\mathbf{X}},\overline{\mathbf{Y}})=0,\overline{\mathbf{X}}=\overline{\mathbf{x}}\right)\to 0, \tag{26}$$

as $K\to\infty$, exponentially fast, where $\hat{Q}_{\overline{\mathbf{Y}}}$ is the empirical distribution of the random concatenated code vector $\overline{\mathbf{Y}}$ on $\mathcal{Y}^\ell$. Thus conditioning on the $\mathbb{P}$-almost every realization $\overline{\mathbf{x}}$ (as $K\to\infty$) and the d'-match event $\overline{\rho}_d(\overline{\mathbf{X}},\overline{\mathbf{Y}})=0$, the probability that the difference between empirical distributions $Q_{\overline{\mathbf{Y}}'}$ and $\overline{Q}^*_{P^\ell,Q^\ell,d'}$ over any (measurable) $E\subset\mathcal{Y}^\ell$ is larger than δ, with $\delta>0$, goes to zero asymptotically in K. Note that the effective length of the concatenated vectors is $\overline{\ell}\triangleq K\ell$, hence sending $K\to\infty$, obviously implies that $\overline{\ell}\to\infty$. Furthermore, by [7], we have,

$$\mathbb{P}\left(\left|\hat{Q}_{\overline{\mathbf{Y}}}(E)-\overline{Q}^*_{P^\ell,Q^\ell,d'}(E)\right|>\delta\ \middle|\ \overline{\rho}_d(\overline{\mathbf{X}},\overline{\mathbf{Y}})=0,\overline{\mathbf{X}}=\overline{\mathbf{x}}\right) =\mathbb{P}\left(\left|Q_{\overline{\mathbf{Y}}'}(E)-\overline{Q}^*_{P^\ell,Q^\ell,d'}(E)\right|>\delta\ \middle|\ \overline{\mathbf{X}}=\overline{\mathbf{x}}\right). \tag{27}$$

This together with Borel-Cantelli lemma, we conclude that for any measurable set $E\subset\mathcal{Y}^\ell$,

$$Q_{\overline{\mathbf{Y}}'}(E)\to\overline{Q}^*_{P^\ell,Q^\ell,d'}(E), \quad \text{as } K\to\infty, \quad \text{w.p. 1.} \tag{28}$$

Since $\mathcal{Y}^\ell$ is a Polish space, then there exists a countable convergence determining class $\mathcal{E}=\{E_i\}\subset\mathcal{Y}^\ell$. Therefore with probability one we have,

$$Q_{\overline{\mathbf{Y}}'}(E_i)\to\overline{Q}^*_{P^\ell,Q^\ell,d'}(E_i), \quad \text{as } K\to\infty, \quad \forall i, \tag{29}$$

which subsequently implies Theorem 1. $\square$

Theorem 2: For an initial distribution $Q_{0,\ell}$ having strictly positive density everywhere over $\mathcal{Y}^\ell$, the recursion in (15) achieves,

$$Q_{n,\ell} \Longrightarrow \overline{Q}^*_{P^\ell,d'}, \quad \text{for } d'=0, \quad \text{as } n \to \infty \ , \tag{30}$$

where $\overline{Q}^*_{P^\ell,d'}$ is the optimal reproduction distribution that achieves the rate distortion function $R(P^\ell, d')$ for the auxiliary distortion measure $\overline{\rho}_d$, i.e.,

$$R(P^\ell,d') = \inf_{Q^\ell} \inf_{\substack{V^\ell:\left[V^\ell\right]_x=P^\ell,\\ \mathbb{E}_{V^\ell}\left(\overline{\rho}_d(\mathbf{X}_1^\ell,\mathbf{Y}_1^\ell)\right)\le d'}} \mathcal{D}\left(V^\ell || P^\ell \times Q^\ell\right), \tag{31}$$

where the inner infimum is taken over all joint distributions V^ℓ of the random vectors $(\mathbf{X}_1^\ell, \mathbf{Y}_1^\ell)$ such that the x-marginal of V^ℓ is P^ℓ, and the expected distortion $\mathbb{E}_{V^\ell}(\overline{\rho}_d\left(\mathbf{X}_1^\ell, \mathbf{Y}_1^\ell\right)) \le d'$.

Proof: It is straightforward to verify that the sets of joint distributions $\{P^\ell \times Q^\ell \ : \ \text{any } Q^\ell\}$, and $\{V^\ell : \left[V^\ell\right]_x = P^\ell, \ \mathbb{E}_{V^\ell}(\overline{\rho}_d\left(\mathbf{X}_1^\ell, \mathbf{Y}_1^\ell\right)) \le d'\}$ are convex sets. Furthermore, it should be noted that for a fixed V^ℓ, the reproduction distribution which minimizes $\mathcal{D}(V^\ell||P^\ell \times Q^\ell)$ is the y-marginal of V^ℓ on $\mathcal{Y}^\ell$. On the other hand, for a fixed Q^ℓ and distortion constraint d', the joint distribution which minimizes $\mathcal{D}(V^\ell||P^\ell \times Q^\ell)$ over $\{V^\ell : \left[V^\ell\right]_x = P^\ell, \ \mathbb{E}_{V^\ell}(\overline{\rho}_d\left(\mathbf{X}_1^\ell, \mathbf{Y}_1^\ell\right)) \le d'\}$ will induce $\overline{Q}^*_{P^\ell,Q^\ell,d'}$. Hence, by Theorem 1, the recursion in (15), achieves a sequence of alternating minimization across convex sets.

$$\begin{aligned} &\overline{V}^*_{P^\ell,Q_{0,\ell},d'} \to (P^\ell \times \overline{Q}^*_{P^\ell,Q_{0,\ell},d'}) \to \\ &\quad \overline{V}^*_{P^\ell,Q_{1,\ell},d'} \to (P^\ell \times \overline{Q}^*_{P^\ell,Q_{1,\ell},d'}) \dots, \end{aligned} \tag{32}$$

where,

$$\overline{V}^*_{P^\ell,Q_{n,\ell},d'} \triangleq \arg \min_{\substack{V^\ell:\left[V^\ell\right]_x=P^\ell,\\ \mathbb{E}_{V^\ell}\left(\overline{\rho}_d(\mathbf{X}_1^\ell,\mathbf{Y}_1^\ell)\right)\le d'}} \mathcal{D}\left(V^\ell || P^\ell \times Q_{n,\ell}\right). \tag{33}$$

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., $R(P^\ell, 0)$, and the corresponding optimum reproduction distribution $\overline{Q}^*_{P^\ell,0}$ on $\mathcal{Y}^\ell$ asymptotically in K and n. $\square$

Theorem 3: The marginal probability measure of $Q_{n,\ell}$ on $\mathcal{Y}$, denoted by $Q_{n,\ell}^{(1)}$, converges in the weak convergence sense to the optimal probability measure $Q^*_{P,d}$ that achieves the rate-distortion function $R(P,d)$ as ℓ and n go to infinity.

Proof: Let $\mathbf{x}_1^\ell(i)$ and $\mathbf{y}_1^\ell(j(i))$, $i = 1, 2, \dots, K$, be a sequence of d-matching ℓ-length words that are generated with the product probability measures P^ℓ and $Q_{n-1,\ell}$ over the alphabets $\mathcal{X}^\ell$ and $\mathcal{Y}^\ell$, respectively. In other words, $\rho(\mathbf{x}_1^\ell(i), \mathbf{y}_1^\ell(j(i))) \le d, \forall i \in \{1, 2, \dots, K\}$. Theorem 3 of [7] showed that the marginal probability measure of the d-matching codewords converges in the weak convergence sense to $Q^*_{P,Q^{(1)}_{n-1,\ell},d}$, defined in (7), as ℓ goes to infinity, i.e.,

$$\mathbf{z}_1^\ell \triangleq \mathbf{y}_1^\ell(j(i)), \quad Q^{(1)}_{\mathbf{y}_1^\ell(j(i))} = \frac{1}{\ell}\sum_{m=1}^{\ell} \delta_{z_m}, \quad \forall i, \tag{34}$$

$$Q^{(1)}_{\mathbf{y}_1^\ell(j(i))} \Longrightarrow Q^*_{P,Q^{(1)}_{n-1,\ell},d}, \quad \text{as } \ell \to \infty, \tag{35}$$

where z_m is the m-th sample in the vector $\mathbf{z}_1^\ell$, and $Q^{(1)}_{\mathbf{y}_1^\ell(j(i))}$ is the marginal empirical probability measure of $\mathbf{y}_1^\ell(j(i))$ on the alphabet $\mathcal{Y}$. Hence, by the definition of $Q_{n,\ell}$ in (15), the marginal probability measure $Q^{(1)}_{n,\ell}$ converges weakly to $Q^*_{P,Q^{(1)}_{n-1,\ell},d}$ as $\ell \to \infty$, as well. The rate-distortion function in (3) can be rewritten as [4], [7]

$$R(P,d) = \inf_{Q} \inf_{\substack{V:[V]_x=P,\\ \mathbb{E}_V(\rho(X,Y))\le d}} \mathcal{D}\left(V||P \times Q\right), \tag{36}$$

here the inner infimum is taken over all joint distributions V of the random variables (X, Y) such that the x-marginal of V is P, and the expected distortion $\mathbb{E}_V(\rho(X,Y)) \le d$. Finally, similar to Theorem 2, the marginal distributions obtained by the recursion in (15), as $\ell \to \infty$, result in a sequence of alternating minimization across convex sets, i.e.,

$$\begin{aligned} &V^*_{P,Q^{(1)}_0,d} \to (P \times Q^*_{P,Q^{(1)}_0,d}) \to \\ &\quad V^*_{P,Q^{(1)}_1,d} \to (P \times Q^*_{P,Q^{(1)}_1,d}) \dots, \end{aligned} \tag{37}$$

where, $Q^{(1)}_n$ is the marginal probability measure of Q_n, and,

$$V^*_{P,Q^{(1)}_n,d} \triangleq \arg \min_{\substack{V:[V]_x=P,\\ \mathbb{E}_V(\rho(X,Y))\le d}} \mathcal{D}\left(V||P \times Q^{(1)}_n\right). \tag{38}$$

The sequence of divergences will converge to the minimum divergence, i.e., $R(P,d)$, and the marginal probability measure $Q^{(1)}_{n,\ell}$ will converge to the corresponding optimum reproduction distribution $Q^*_{P,d}$ asymptotically in n and ℓ. $\square$

V. Conclusion

This paper generalizes the practically-effective and asymptotically-optimal iterative NTS lossy codebook generating algorithm in [5] to sources with more general alphabet spaces. We assume that the source and reproduction alphabets are complete separable metric spaces (often called Polish spaces). Similar to [5], an ML framework is leveraged to identify the next iteration codebook reproduction distribution after observing sequence of K d-matching events. We show that for finite code-word length ℓ, the codebook reproduction distribution on $\mathcal{Y}^\ell$ converges to the optimal distribution $\overline{Q}^*_{P^\ell,d'}$, in the weak convergence sense, as $K \to \infty$ and $n \to \infty$, that achieves the rate distortion function $R(P^\ell, d')$ albeit for an auxiliary distortion measure $\overline{\rho}_d$, and distortion constraint $d' = 0$. Additionally, we show that asymptotically in ℓ, the marginal codebook reproduction distribution on $\mathcal{Y}$ converges to the optimal distribution $Q^*_{P,d}$ that achieves the rate-distortion $R(P,d)$ function for the original distortion measure ρ.

VI. Acknowledgement

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

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

[5] 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.

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

[7] 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.

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

[9] 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.

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

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

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

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