On Effective Stochastic Mechanisms for On-The-Fly Codebook Regeneration

Ahmed Elshafiy, Mahmoud Namazi, 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

***Abstract*—This paper proposes an effective mechanism for stochastic codebook generation for lossy coding, using source examples. Earlier work has shown that the rate-distortion bound can be asymptotically achieved by a "natural type selection" (NTS) mechanism which iteratively considers asymptotically long source strings (from given distribution P) and regenerates the codebook according to the type of the first codeword to "d-match" the source string (i.e., satisfy the distortion constraint), where the sequence of codebook generating types converges to the optimal reproduction distribution. While ensuring optimality, earlier results had a significant practical flaw, due to the order of limits at which the convergence is achieved. More specifically, NTS iterations indexed by n presume asymptotically large ℓ, but the codebook size grows exponentially with ℓ. The reversed order of limits is practically preferred, wherein most codebook regeneration iterations involve manageable string lengths. This work describes a dramatically more efficient mechanism to achieve the optimum within a practical framework. It is specifically shown that it is sufficient to individually encode many source strings of short fixed length ℓ, then find the maximum likelihood estimate for the distribution Q_{n+1} that would have generated the observed sequence of d-matching codeword strings, then use Q_{n+1} to generate a new codebook for the next iteration. The sequence of distributions $Q_1, Q_2, \dots$ converges to the optimal reproduction distribution $Q^*_\ell(P,d)$, achievable at finite length ℓ. It is further shown that $Q^*_\ell(P,d)$ converges to the optimal reproduction distribution $Q^*(P,d)$ that achieves the rate-distortion bound $R(P,d)$, asymptotically in string length ℓ.**

I. Introduction

Stochastic mechanisms for codebook generation and adaptation, based on source string matching, have appeared in both the lossless and the lossy coding literature. They had a major impact on lossless coding, where the seminal contributions of Lempel and Ziv [1]–[3] had considerable theoretical and practical implications. For example in LZ78 [3], the operating codebook is a tree that is grown, on-the-fly, based on observation of source strings, such that asymptotically its codewords consist (mostly) of typical source sequences. Asymptotic optimality is thus achieved for ergodic sources without prior knowledge of the source statistics.

Stochastic mechanisms for codebook generation have also been proposed for lossy coding, including for example gold-washing [4] and natural type selection [5], [6]. However, it is important to emphasize that the lossy coding case is fundamentally more challenging. Note that the optimal codebook generating distribution in lossless coding is simply the source distribution, so the stochastic mechanism's essential objective is to learn this distribution from observation of source strings. However, in lossy coding, the optimal codebook generating, or reproduction distribution Q^* differs from the source distribution P, as it depends on the distortion constraint d. This represents a non-trivial learning challenge, especially in the *non*-high resolution regime, where Q^* deviates significantly from P [5], [7]–[9]. For example, in the case of continuous alphabet sources with the squared error distortion measure, at small distortion (high resolution) $Q^* \approx P$, but as the distortion constraint is relaxed, i.e., d increases, Q^* increasingly differs from P, it shrinks, often becomes discrete, and eventually collapses on a single point when $d = d_{\max}$ [10].

Most relevant to this paper is the stochastic mechanism proposed in [5] for codebook generation in lossy coding of discrete sources. At each iteration indexed by n, and given a sufficiently large string length ℓ, the source string will favor codewords of type Q_{n+1} through a distortion match event. The favored codeword type, which is naturally selected by the source, is then used to regenerate the random codebook in the next iteration. It was shown that, asymptotically, the sequence of codebook generating types $Q_1, Q_2, \dots$ converges to the optimal reproduction distribution $Q^*(P,d)$ that achieves the minimum possible coding rate $R(P,d)$. The motivation for this work comes from the observation that this "natural type selection" (NTS) codebook generating algorithm suffers from a significant practical flaw due to the order of the limits required. To ensure convergence to the rate-distortion function we must first have asymptotically long strings ($\ell \to \infty$), and only then iteratively regenerate the codebook ($n \to \infty$). Performing such iterations with very long strings implies intractable d-match search complexity, as the codebook size grows exponentially with string length. Thus, as it stands, NTS is not easily amenable to practical implementations such as those that resulted in the phenomenal impact of the Lempel-Ziv algorithms for lossless coding.

This paper proposes an effective mechanism for on-the-fly stochastic generation of codebook. It builds on the principles of the original NTS approach, but delivers optimality within a practically implementable algorithm. In Theorem 1 and Theorem 2, the proposed codebook generating algorithm is shown to find the rate-distortion optimal reproduction distribution. It is specifically shown that for a fixed string length ℓ, the codebook reproduction distribution converges to the optimum achievable distribution $Q^*_\ell(P,d)$, from a set of distributions defined by ℓ. Additionally, $Q^*_\ell(P,d)$ can be made as close as needed to $Q^*(P,d)$ by the sending the string length to infinity. This implies that convergence to the optimal distribution is achieved by the reversed order of the limits, when compared

to the original NTS algorithm in [5]. Consequently, by implementing codebook regeneration iterations at a *manageable* string length, a dramatic decrease in d-match search complexity is accomplished, which is exponential in the string length and ultimately determines the complexity of the NTS iteration.

The remainder of this paper is organized as follows: Section II provides some relevant background, Section III summarizes the original NTS algorithm, Section IV introduces the proposed effective and practical mechanism of codebook generation that captures the benefits of NTS at dramatically reduced complexity, by leveraging a maximum likelihood estimation framework. Asymptotic convergence to the rate-distortion bound achieving distribution is established in Section IV, and conclusions are drawn in Section V.

II. Relevant Background

Recall the structure of a random codebook for lossless or lossy coding. Let $\mathbf{x}_1$, $\mathbf{x}_2$, ... be a sequence of source strings/vectors of length ℓ, where the source is discrete, memoryless and drawn from distribution $P = \{P(x), x \in \mathcal{X}\}$ over the input alphabet $\mathcal{X}$. By Shannon's lossless coding theorem, if we generate an independent and identically distributed (i.i.d.) codebook of $\exp\left(\ell(H(P)+\epsilon)\right)$ codeword strings from the source distribution P, then the probability of finding in the codebook another, independently generated source string goes to one as ℓ goes to ∞, where $H(P)$ is the source entropy:

$$H(P) = -\sum_{x\in\mathcal{X}} P(x)\log\left(P(x)\right). \tag{1}$$

Throughout the paper, the logarithm function is taken to the base e. In the lossy coding case, we define a distortion function $\rho : \mathcal{X}\times\mathcal{Y}\to[0,\infty)$, where $\mathcal{Y}$ is the output or reproduction discrete alphabet. The distortion seen between vectors $\mathbf{x}$ and $\mathbf{y}$ is given by,

$$\rho(\mathbf{x},\mathbf{y}) = \frac{1}{\ell}\sum_{i=1}^{\ell}\rho(x_i,y_i). \tag{2}$$

Next, define a "d-match" event as the event that $\rho(\mathbf{x},\mathbf{y}) \le d$ is satisfied. Shannon's lossy coding theorem specifies that if we generate an i.i.d. codebook of length $\exp\left(\ell(R(P,d)+\epsilon)\right)$ from optimal distribution $Q^*(P,d)$, then the probability of finding a d-match to an independently generated source string of length ℓ goes to one as ℓ goes to ∞ [11], wherein $R(P,d)$ is the rate-distortion function and $Q^*(P,d)$ is the optimal reproduction distribution, i.e.,

$$R(P,d) = \min_{W:\rho(P,W)\le d} I(P,W), \qquad d\ge 0, \tag{3}$$

$$\rho(P,W) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x)W(y|x)\rho(x,y), \tag{4}$$

$$I(P,W) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x)W(y|x)\log\frac{W(y|x)}{\sum_{x'\in\mathcal{X}}P(x')W(y|x')}. \tag{5}$$

Here, $I(P,W)$ denotes the mutual information in terms of source distribution P and the transition (conditional) distribution $W = \{W(y|x) : y\in\mathcal{Y}, x\in\mathcal{X}\}$. Let $W^* = W^*(P,d)$ be the minimizing transition distribution in (3), i.e. $R(P,d) = I(P,W^*)$. The optimal reproduction distribution $Q^*(P,d)$ is obtained by marginalizing the joint distribution $\left\{P(x)W^*(y|x)\right\}$, i.e.,

$$Q^*(P,d) = [P\circ W^*]_y = \left\{\sum_{x\in\mathcal{X}} P(x)W^*(y|x)\right\}, \tag{6}$$

where $P\circ W^*$ denotes the joint distribution, and $[P\circ W^*]_y$ is the y-marginal of the joint distribution, while $[P\circ W^*]_x$ is the x-marginal of the joint distribution. Next, suppose that a different reproduction distribution $Q \ne Q^*(P,d)$ is used to generate the random codebook. In that case, the minimum coding rate, denoted as $R(P,Q,d)$, which is required to guarantee a d-match event as ℓ goes to ∞ is given by [12],

$$R(P,Q,d) = \min_{Q'}\{I_m(P||Q',d) + \mathcal{D}(Q'||Q)\} > R(P,d), \tag{7}$$

where $\mathcal{D}(\cdot||\cdot)$ denotes the Kullback Leibler (KL) divergence function, and $I_m(P||Q,d)$ is the minimum mutual information with constrained reproduction distribution Q, i.e.,

$$I_m(P||Q,d) = \begin{cases} \min\limits_{W\in\mathcal{W}(P,Q,d)} I(P,W) & \text{if } \mathcal{W}(P,Q,d) \text{ is non empty} \\ \infty & \text{otherwise} \end{cases}, \tag{8}$$

$$\mathcal{W}(P,Q,d) = \Big\{W : [P\circ W]_y = Q, \rho(P,W)\le d\Big\}. \tag{9}$$

III. Natural Type Selection

Let us define $Q^*(P,Q,d)$ as the reproduction distribution that achieves $R(P,Q,d)$, i.e.,

$$Q^*(P,Q,d) = \arg\min_{Q'}\left\{I_m(P||Q',d) + \mathcal{D}(Q'||Q)\right\}. \tag{10}$$

Furthermore, let N_ℓ denote the index of the first codeword in the codebook that d-matches a source vector $\mathbf{x}$:

$$\rho(\mathbf{x},\mathbf{y}_i) > d, \text{ for } 1\le i\le N_\ell - 1, \text{ and } \rho(\mathbf{x},\mathbf{y}_{N_\ell})\le d, \tag{11}$$

where $\mathbf{y}_i$ is the i-th codeword in the codebook generated from distribution $Q = \{Q(y) : y\in\mathcal{Y}\}$. Theorem 4 of [5] shows that the empirical type of the first d-matching codeword $Q_{N_\ell}(P,Q,d)$ converges in probability to $Q^*(P,Q,d)$ asymptotically as $\ell\to\infty$. Note that $Q^*(P,Q,d)$ is more efficient in coding the source than Q, the codebook generating distribution, however $Q^*(P,Q,d)$, is not as efficient as $Q^*(P,d)$. This intuitively gives rise to the recursive algorithm in [5]. Starting with an arbitrary strictly positive initial codebook generating distribution $Q_{0,\ell}$, the type of the first codeword to d-match a source string is used to generate a new codebook. 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" algorithm. This recursion results in a sequence of reproduction distributions,

$$Q_{n,\ell} = Q_{N_\ell}(P, Q_{n-1,\ell}, d) \tag{12}$$

$$Q_n = \lim_{\ell\to\infty} Q_{n,\ell} = Q^*(P,Q_{n-1},d), \quad n = 1,2,\ldots \tag{13}$$

Moreover, Theorem 5 of [5], states that the recursion in (13) asymptotically achieves the optimal reproduction distribution $Q^*(P,d)$, and the corresponding rate achieves the rate-distortion bound, i.e.,

$$Q^*(P,d) = \lim_{n\to\infty}\lim_{\ell\to\infty} Q_{n,\ell}, \tag{14}$$

$$R(P,d) = \lim_{n\to\infty}\lim_{\ell\to\infty} R(P,Q_{n,\ell},d). \tag{15}$$

The above result, while ensuring optimality, still has a significant practical flaw due to its order of limits. The convergence as $n\to\infty$ presumes that ℓ is already very large. In other words, the limit in ℓ is taken before the limit in n. In practice, however, it is the reversed order of limits that "makes more sense". One would like to implement codebook regeneration iterations at manageable string length. Ideally, one would like to derive the asymptotic behavior for finite string lengths. More significantly, it is noteworthy that the codebook size grows exponentially with the string length ℓ. Equivalently, a d-match event occurs with probability that decays exponentially with ℓ. This means that the d-match search complexity, which is at the heart of the NTS iteration, explodes when ℓ increases. This provides strong motivation for the proposed approach we describe in the next section.

IV. Practically Effective And Asymptotically Optimal Natural Type Selection

An interesting question to be answered is: can a more effective algorithm be devised such that convergence to the optimal reproduction distribution $Q^*(P,d)$ is achieved but through a reversed order of limits? I.e., can we achieve $Q^*(P,d)$ by first sending n to infinity for a finite ℓ, and then sending ℓ to infinity? Obviously, for finite ℓ, the type of the d-matching codeword is restricted in resolution to $1/\ell$, i.e., the frequency of letters in the codeword is multiple of $1/\ell$. Such a low resolution of types may cause difficulties for an iterative algorithm that advances by potentially small adjustments to the distribution. We circumvent this shortcoming by estimating the *general* reproduction distribution that would have generated a sequence of d-matching codeword strings. In other words, we consider the maximum likelihood estimate of the distribution given a set of K observed d-matching codeword strings.

Lemma 1: The Maximum Likelihood (ML) estimate of the reproduction distribution, given a set of K d-matching codewords, is the *average* of the d-matching codeword types, i.e.,

$$\hat{Q}^{\mathrm{ML}}_{n+1,\ell,K} = \frac{1}{K}\sum_{i=1}^{K} Q_{n,\ell}(\mathbf{y}_{j(i)}), \tag{16}$$

where $\mathbf{y}_{j(i)}$ is the ℓ-length codeword, of index $j(i)$, that achieves a d-match event to the i-th source string, and $Q_{n,\ell}(\mathbf{y}_{j(i)})$ is its corresponding type.

The proof of Lemma 1 is given in Appendix A. This result immediately suggests a modified and substantially more effective variant of the NTS recursive algorithm. Starting with an arbitrary and strictly positive initial reproduction distribution $Q_{0,\ell}$, the *average* type of the set of K d-matching codewords is used to generate a new codebook. This recursion yields a sequence of reproduction distributions, i.e.,

$$Q_{n,\ell,K} = \frac{1}{K}\sum_{j=1}^{K} Q_{n-1,\ell}(\mathbf{y}_j), \tag{17}$$

$$Q_{n,\ell} = \lim_{K\to\infty} Q_{n,\ell,K}, \quad n = 1,2,\ldots \tag{18}$$

In the following analysis, we establish that the sequence of reproduction distributions of the modified NTS algorithm, despite the fact that it maintains a fixed and finite string length ℓ, converges asymptotically, in probability, as $n\to\infty$ and $K\to\infty$ to the optimal achievable reproduction distribution $Q^*_\ell(P,d)$, to be defined in Theorem 2. First, we start by defining the set of all d-matching joint types for a fixed sourceword or codeword string length ℓ as,

$$\begin{aligned}\mathcal{V}_\ell(d) \triangleq \Big\{ & V : V = P'\circ W', P'\in\mathcal{P}_\ell, \\ & Q' = [V]_y, Q'\in\mathcal{Q}_\ell, \rho(P',W')\le d\Big\},\end{aligned} \tag{19}$$

where $\mathcal{P}_\ell$ and $\mathcal{Q}_\ell$ are the sets of all possible ℓ-length string types over the input alphabet $\mathcal{X}$, and the output alphabet $\mathcal{Y}$, respectively. Next, define the set $E_\ell(P,d)$ as,

$$E_\ell(P,d) \triangleq \Big\{ V : V\in \mathrm{Conv}(\mathcal{V}_\ell(d)),\ [V]_x = P\Big\}, \tag{20}$$

where $\mathrm{Conv}(\mathcal{V}_\ell(d))$ is the convex hull of all the joint types in the set $\mathcal{V}_\ell(d)$. We can now state our first main result.

Theorem 1: The reproduction distribution of the modified recursive NTS algorithm in (17) converges asymptotically as $K\to\infty$, and for a fixed string length ℓ to the distribution $Q^*_\ell(P,Q,d)$ in probability, i.e.,

$$\begin{aligned} Q_{n,\ell} &= \lim_{K\to\infty} Q_{n,\ell,K} = Q^*_\ell(P,Q,d), \quad Q = Q_{n-1,\ell} \\ V^*_\ell &= V^*_\ell(P,Q,d) \triangleq \arg\min_{V\in E_\ell(P,d)} \mathcal{D}\left(V||P\times Q\right), \\ Q^*_\ell(P,Q,d) &= \left[V^*_\ell(P,Q,d)\right]_y. \end{aligned} \tag{21}$$

Proof: Let $\mathbf{x}_i$ and $\mathbf{y}_{j(i)}$, $i=1,2,\ldots,K$, be a sequence of d-matching ℓ-length memory-less vectors that are generated according to P and Q over discrete alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $P_{i,\ell}$ be the i-th instantaneous type of $\mathbf{x}_i$, $Q_{j,\ell}$ be the j-th instantaneous type of $\mathbf{y}_{j(i)}$, and $W_{i,\ell}$ be the i-th instantaneous channel, i.e.,

$$\begin{aligned} &\left[P_{i,\ell}\circ W_{i,\ell}\right]_y = Q_{j,\ell}, \\ &\rho(P_{i,\ell},W_{i,\ell}) \le d, \quad i = 1,2,\ldots,K. \end{aligned} \tag{22}$$

Let $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ be the $K\ell$-length vectors constructed by concatenating $\{\mathbf{x}_1,\ldots,\mathbf{x}_K\}$ vectors of length ℓ, and $\{\mathbf{y}_{j(1)},\ldots,\mathbf{y}_{j(K)}\}$ vectors of length ℓ, respectively. Additionally, define $E_{K,\ell}(P,d)$ as,

$$\begin{aligned} E_{K,\ell}(P,d) \triangleq \Big\{ & V : V = \frac{1}{K}\sum_{i=1}^{K} P_{i,\ell}\circ W_{i,\ell}, P_{i,\ell}\in\mathcal{P}_\ell, \\ & Q_{j,\ell}\in\mathcal{Q}_\ell, \rho(P_{i,\ell},W_{i,\ell})\le d, P_{i,\ell}(x) = \frac{N(x|\mathbf{x}_i)}{\ell}, \mathbf{X}\sim P\Big\}, \end{aligned} \tag{23}$$

where $N(x|\mathbf{x}_i)$ is the number of occurrence of letter $x\in\mathcal{X}$ in the vector $\mathbf{x}_i$. Additionally, $\mathbf{X}\sim P$ indicates that the elements

of the vectors $\mathbf{x}_i$ are generated i.i.d. according to P. Note that the type of the concatenated vectors $\bar{\mathbf{x}}$, and $\bar{\mathbf{y}}$ is equal to the average of the set of types $\{P_{i,\ell}\}$, and $\{Q_{j,\ell}\}$, respectively, i.e.,

$$P_{\bar{\mathbf{x}}} = \frac{1}{K}\sum_{i=1}^{K} P_{i,\ell},\ Q_{\bar{\mathbf{y}}} = \frac{1}{K}\sum_{j=1}^{K} Q_{j,\ell}. \tag{24}$$

By strong law of large numbers, $P_{\bar{\mathbf{x}}}$ almost surely converges to the generating distribution P as $K \to \infty$. We will show that for any $\delta > 0$, and sufficiently large K,

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

Thus, conditioning on the event that the joint type of $(\bar{\mathbf{x}},\bar{\mathbf{y}})$ belongs to $E_{K,\ell}(P,d)$, the type of $\bar{\mathbf{y}}$ is with high probability close in the divergence-sense to $Q^*_\ell(P,Q,d)$. Since closeness in divergence implies closeness in $\mathcal{L}_1$ sense [11], this establishes Theorem 1. We start by verifying that, as $K \to \infty$, and by (20) and (23), $E_{K,\ell}(P,d)$ approaches $E_\ell(P,d)$, hence define,

$$D^* = \min_{V \in E_\ell(P,d)} \mathcal{D}(V||P\times Q). \tag{26}$$

Then following [11],

$$\mathbb{P}(\mathcal{D}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}}||P\times Q) > D^*+3\delta, V_{\bar{\mathbf{x}},\bar{\mathbf{y}}} \in E_{K,\ell}(P,d)) = \sum_{\substack{V' \in E_{K,\ell}(P,d)\cap \mathcal{P}_{K\ell}\times\mathcal{Q}_{K\ell}:\\ \mathcal{D}(V'||P\times Q) > D^*+3\delta}} \mathbb{P}(T(V')), \tag{27}$$

where the probability of type class of V' is denoted by $\mathbb{P}(T(V'))$. Then, by [11],

$$\mathbb{P}(\mathcal{D}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}}||P\times Q) > D^*+3\delta, V_{\bar{\mathbf{x}},\bar{\mathbf{y}}} \in E_{K,\ell}(P,d)) \leq \sum_{\substack{V' \in E_{K,\ell}(P,d)\cap \mathcal{P}_{K\ell}\times\mathcal{Q}_{K\ell}:\\ \mathcal{D}(V'||P\times Q) > D^*+3\delta}} \exp\left(-K\ell\mathcal{D}(V'||P\times Q)\right), \tag{28}$$

$$\mathbb{P}(\mathcal{D}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}}||P\times Q) > D^*+3\delta, V_{\bar{\mathbf{x}},\bar{\mathbf{y}}} \in E_{K,\ell}(P,d)) \leq \sum_{\substack{V' \in E_{K,\ell}(P,d)\cap \mathcal{P}_{K\ell}\times\mathcal{Q}_{K\ell}:\\ \mathcal{D}(V'||P\times Q) > D^*+3\delta}} \exp\left(-K\ell\left(D^*+3\delta\right)\right), \tag{29}$$

$$\mathbb{P}(\mathcal{D}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}}||P\times Q) > D^*+3\delta, V_{\bar{\mathbf{x}},\bar{\mathbf{y}}} \in E_{K,\ell}(P,d)) \leq (K\ell+1)^{|\mathcal{X}||\mathcal{Y}|}\exp\left(-K\ell\left(D^*+3\delta\right)\right), \tag{30}$$

since there are only a polynomial number of joint types. Next, we observe that,

$$\mathbb{P}(\mathcal{D}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}}||P\times Q) \leq D^*+2\delta, V_{\bar{\mathbf{x}},\bar{\mathbf{y}}} \in E_{K,\ell}(P,d)) = \sum_{\substack{V' \in E_{K,\ell}(P,d)\cap \mathcal{P}_{K\ell}\times\mathcal{Q}_{K\ell}:\\ \mathcal{D}(V'||P\times Q) \leq D^*+2\delta}} \mathbb{P}(T(V')), \tag{31}$$

$$\mathbb{P}(\mathcal{D}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}}||P\times Q) \leq D^*+2\delta, V_{\bar{\mathbf{x}},\bar{\mathbf{y}}} \in E_{K,\ell}(P,d)) \geq \sum_{\substack{V' \in E_{K,\ell}(P,d)\cap \mathcal{P}_{K\ell}\times\mathcal{Q}_{K\ell}:\\ \mathcal{D}(V'||P\times Q) \leq D^*+2\delta}} \frac{\exp\left(-K\ell\mathcal{D}(V'||P\times Q)\right)}{(K\ell+1)^{|\mathcal{X}||\mathcal{Y}|}}, \tag{32}$$

$$\mathbb{P}(\mathcal{D}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}}||P\times Q) \leq D^*+2\delta, V_{\bar{\mathbf{x}},\bar{\mathbf{y}}} \in E_{K,\ell}(P,d)) \geq \frac{\exp\left(-K\ell\left(D^*+2\delta\right)\right)}{(K\ell+1)^{|\mathcal{X}||\mathcal{Y}|}}, \tag{33}$$

since for sufficiently large K, there exists at least one term in the summation, i.e., there exists one $K\ell$ type V' in $E_{K,\ell}(P,d)$ such that,

$$\mathcal{D}(V'||P\times Q) \leq D^*+2\delta. \tag{34}$$

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

$$\mathbb{P}(\mathcal{D}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}} \in E_{K,\ell}(P,d))) \geq \frac{\exp\left(-K\ell\left(D^*+2\delta\right)\right)}{(K\ell+1)^{|\mathcal{X}||\mathcal{Y}|}}. \tag{35}$$

By Bayes' law we get,

$$\mathbb{P}(\mathcal{D}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}}||P\times Q) > D^*+3\delta | V_{\bar{\mathbf{x}},\bar{\mathbf{y}}} \in E_{K,\ell}(P,d)) \leq (K\ell+1)^{2|\mathcal{X}||\mathcal{Y}|}\exp\left(-K\ell\delta\right). \tag{36}$$

By the "Pythagorean" theorem [11], we have,

$$\mathcal{D}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}}||V^*_\ell) + \mathcal{D}(V^*_\ell||P\times Q) \leq \mathcal{D}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}}||P\times Q), \tag{37}$$

Hence, $\mathcal{D}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}}||P\times Q) \leq D^*+3\delta$ implies that,

$$\mathcal{D}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}}||V^*_\ell) \leq 3\delta. \tag{38}$$

Finally, by the data processing inequality, we have,

$$\mathcal{D}(Q_{\bar{\mathbf{y}}}||Q^*(P,Q,d)) \leq \mathcal{D}(V_{\bar{\mathbf{x}},\bar{\mathbf{y}}}||V^*_\ell) \tag{39}$$

since, both are the respective y-marginals of the joint types. Hence, Theorem 1 follows from (36) as desired.

Theorem 2: For a strictly positive initial distribution $Q_{0,\ell}$, the recursion in (18) achieves,

$$Q_{n,\ell} \to Q^*_\ell(P,d), \quad \text{as } n \to \infty\ , \tag{40}$$

$$\begin{matrix} R_\ell(P,d) \to R(P,d) \\ Q^*_\ell(P,d) \to Q^*(P,d) \end{matrix}, \quad \text{as } \ell \to \infty\ , \tag{41}$$

where $Q^*(P,d)$ is the optimum reproduction distribution that achieves the rate distortion function $R(P,d)$, and $Q^*_\ell(P,d)$ is the optimum achievable reproduction distribution for finite length ℓ, that achieves $R_\ell(P,d)$, i.e.,

$$\begin{aligned} &R_\ell(P,Q,d) \triangleq \min_{V\in E_\ell(P,d)} \mathcal{D}\left(V||P\times Q\right), \\ &\mathcal{W}_\ell(P,d) \triangleq \{W : P\circ W = V, V \in E_\ell(P,d)\}, \\ &W^*_\ell(P,d) = W^*_\ell \triangleq \arg\min_{W\in\mathcal{W}_\ell(P,d)} I(P,W), \\ &R_\ell(P,d) \triangleq \min_Q R_\ell(P,Q,d) = I(P,W^*_\ell), \\ &Q^*_\ell(P,d) \triangleq \left[P\circ W^*_\ell(P,d)\right]_y. \end{aligned} \tag{42}$$

Proof: It can be shown using Csiszár–Tusnády Theorem 3 of [13] for general alternating minimization procedures across convex sets, that the convergences stated in (40) is guaranteed. From (42), we have,

$$R_\ell(P,d) = \min_Q \min_{V\in E_\ell(P,d)} \mathcal{D}\left(V||P\times Q\right). \tag{43}$$

It is straight forward to verify that the sets of joint distributions $\{P\times Q : \text{any } \mathrm{Q}\}$, and $E_\ell(P,d)$ are convex sets. Furthermore,

it should be noted that for a fixed V, the reproduction distribution which minimizes $\mathcal{D}(V||P\times Q)$ is the y-marginal of V. On the other hand, for a fixed Q and distortion constraint d, the joint distribution which minimizes $\mathcal{D}(V||P\times Q)$ over $E_\ell(P,d)$ will induce $Q^*_\ell(P,Q,d)$. Hence, by the result of Theorem 1, the recursion in (18), achieves a sequence of alternating minimization across convex sets, i.e.,

$$\begin{aligned} V^*_\ell(P,Q_{0,\ell},d)\to(P\times Q^*_\ell(P,Q_{0,\ell},d))\to\\ V^*_\ell(P,Q_{1,\ell},d)\to(P\times Q^*_\ell(P,Q_{1,\ell},d))\dots \end{aligned} \tag{44}$$

It should be noted that the distance in the alternating minimization of (44) is measured by divergence. Hence, by [13], the sequences of divergences and distributions will converge to the minimum divergence, i.e., $R_\ell(P,d)$, and the corresponding optimum reproduction distribution $Q^*_\ell(P,d)$. Next, to show the second part of Theorem 2 stated in (41), first verify that the minimum coding rate with constrained reproduction distribution, $R(P,Q,d)$, can be rewritten as [5],

$$R(P,Q,d)=\min_{W:\rho(P,W)\le d} I(P,W)+\mathcal{D}([P\circ W]_y\,||Q), \tag{45}$$

$$R(P,Q,d)=\min_{W:\rho(P,W)\le d}\mathcal{D}\left(P\circ W||P\times Q\right), \tag{46}$$

Hence, $R(P,d)$ follows from (46) as,

$$R(P,d)=\min_{Q}\min_{W:\rho(P,W)\le d}\mathcal{D}(P\circ W||P\times Q). \tag{47}$$

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

$$E_\ell(P,d)\to\{V: V=P'\circ W', P'=P, \rho(P',W')\le d\}. \tag{48}$$

Consequently, as $\ell\to\infty$, and from (42), (43), and (47),

$$R_\ell(P,Q,d)\to\min_{W:\rho(P,W)\le d}\mathcal{D}\left(P\circ W||P\times Q\right)=R(P,Q,d), \tag{49}$$

$$R_\ell(P,d)\to R(P,d). \tag{50}$$

Thus by the definition of $Q^*_\ell(P,d)$ in (42), the second part of Theorem 2 follows.

V. Conclusion

This paper proposes a modified and more effective NTS approach for a stochastic generation of random codebook in the lossy coding settings. Unlike the original NTS approach in [5], the codebook generating distribution at each iteration is not restricted in resolution to the type of the d-matching codeword. Instead, an ML estimation framework is leveraged to identify the most likely distribution that would have generated a set of d-matching codewords. It was further shown by Theorem 1 and Theorem 2, that the proposed codebook generating distribution, that emerges from the proposed stochastic algorithm, converges to the optimal codebook reproduction distribution asymptotically as $K\to\infty$, $n\to\infty$, and $\ell\to\infty$. A significant improvement is achieved in comparison with the original NTS algorithm by reversing the order of limits required to achieve convergence. This consequently reduces dramatically the complexity of finding a d-match in the codebook, which is the central operation of the NTS algorithm. Hence, the modified NTS approach is rendered significantly more appealing to practical applications.

Appendix A
Maximum Likelihood Estimation of Codebook Reproduction Distribution

Let $\mathcal{C}=\left\{\mathbf{y}_{j(1)},\mathbf{y}_{j(2)},\dots,\mathbf{y}_{j(K)}\right\}$ be the set of ℓ-length memoryless d-matching codewords to the input source examples $\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_K$, generated by $Q_{n,\ell,K}(y)$ over $y\in\mathcal{Y}$ (with n being the NTS iteration index), i.e.,

$$\rho(\mathbf{x}_i,\mathbf{y}_{j(i)})\le d,\qquad i\in\{1,2,\dots,K\}, \tag{51}$$

The ML estimator of the codebook generating distribution $Q_{n+1,\ell,K}$ would maximize the joint probability of generating the codewords in $\mathcal{C}$, and hence can be written as,

$$Q_{n+1,\ell,K}=\arg\max_{\hat{Q}\in\mathcal{Q}}\mathbb{P}(\mathbf{y}_{j(1)},\mathbf{y}_{j(2)},\dots,\mathbf{y}_{j(K)}|\hat{Q}), \tag{52}$$

where $\mathcal{Q}$ is the set of valid distributions, i.e.,

$$\mathcal{Q}=\left\{Q:\sum_{y\in\mathcal{Y}}Q(y)=1\right\}. \tag{53}$$

The likelihood function shown in (52) depends on the codewords $\mathbf{y}_{j(i)}, 1\le i\le K$ only through the codewords' types. Let $Q_{n,\ell}(\mathbf{y}_{j(i)})$ be the type of the d-matching codeword $\mathbf{y}_{j(i)}$. Then, the ML formulation in (52) can be written as,

$$Q_{n+1,\ell,K}=\arg\max_{\hat{Q}\in\mathcal{Q}}\mathbb{P}(Q_{n,\ell}(\mathbf{y}_{j(1)}),\dots Q_{n,\ell}(\mathbf{y}_{j(K)})|\hat{Q}). \tag{54}$$

Taking independence between codewords into consideration, we get [11],

$$Q_{n+1,\ell,K}=\arg\max_{\hat{Q}\in\mathcal{Q}}\prod_{i=1}^{K}\mathbb{P}(Q_{n,\ell}(\mathbf{y}_{j(i)})|\hat{Q}), \tag{55}$$

$$\begin{aligned} Q_{n+1,\ell,K}=&\arg\max_{\hat{Q}\in\mathcal{Q}}\prod_{i=1}^{K}\\ &\exp\left\{-\ell\left(H\left(Q_{n,\ell}(\mathbf{y}_{j(i)})\right)+\mathcal{D}\left(Q_{n,\ell}(\mathbf{y}_{j(i)})||\hat{Q}\right)\right)\right\}, \end{aligned} \tag{56}$$

$$Q_{n+1,\ell,K}=\arg\max_{\hat{Q}\in\mathcal{Q}}\exp\left\{-\ell\sum_{i=1}^{K}\left(H\left(Q_{n,\ell}(\mathbf{y}_{j(i)})\right)+\mathcal{D}\left(Q_{n,\ell}(\mathbf{y}_{j(i)})||\hat{Q}\right)\right)\right\}, \tag{57}$$

where $H(Q_{n,\ell}(\mathbf{y}_{j(i)}))$ denotes the entropy calculated over $Q_{n,\ell}(\mathbf{y}_{j(i)})$. The $\log_e(\cdot)$ function is monotonically increasing, and the entropy term $H(Q_{n,\ell}(\mathbf{y}_{j(i)}))$ doesn't depend on $\hat{Q}$, hence the expression in (57) simplifies to,

$$Q_{n+1,\ell,K}=\arg\min_{\hat{Q}\in\mathcal{Q}}\sum_{i=1}^{K}\left(\mathcal{D}\left(Q_{n,\ell}(\mathbf{y}_{j(i)})||\hat{Q}\right)\right) \tag{58}$$

In summary, the ML estimate of the codebook reproduction distribution is the one that minimizes the sum of KL divergences towards the types of the d-matching codewords. Then, it is straight forward to show from (58) that,

$$Q_{n+1,\ell,K}=\frac{1}{K}\sum_{i=1}^{K}Q_{n,\ell}(\mathbf{y}_{j(i)}). \tag{59}$$

Acknowledgment

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 transactions on Information Theory*, vol. IT-22, No. 1, pp. 75–81, 1976.

[2] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Transactions on information theory*, vol. 23, no. 3, pp. 337–343, 1977.

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

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

[5] R. Zamir and K. Rose, "Natural type selection in addaptive lossy compression," *IEEE Transactions on information theory*, vol. 47, pp. 99–111, Jan. 2001.

[6] Y. Kochman and R. Zamir, "Adaptive parametric vector quantization by natural type selection," in *Proceedings of the Data Compression Conference*, Mar 2002, pp. 392–401.

[7] Y. Steinberg and M. Gutman, "An algorithm for source coding subject to a fidelity criterion, based on string matching," *IEEE Transactions on information theory*, vol. IT-39, pp. 877–886, May 1993.

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

[9] ——, "A type generation model for adaptive lossy compression," in *Proc. of ISIT97*, Ulm, Germany, June 1997, p. 186.

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

[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 1991.

[12] E. H. Yang and J. Kieffer, "On the performance of data compression algorithms based upon string matching," *IEEE Transactions on information theory*, vol. 44, pp. 47–65, 1998.

[13] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," *Statistics and decisions*, vol. Supplement issue No. 1, pp. 205–237, 1984.