Natural Type Selection in Adaptive Lossy Compression*

Ram Zamir[1] and Kenneth Rose[2]

[1] Dpt. of EE-Systems, Tel Aviv University, Israel
zamir@eng.tau.ac.il
[2] Dpt. of ECE, University of California, Santa Barbara
rose@ece.ucsb.edu

Submitted to IEEE Trans. on Information Theory, October 1998
Revised November 1999 and July 2000

Abstract

Consider approximate (lossy) matching of a source string $\sim P$, with a random codebook generated from reproduction distribution Q, at a specified distortion d. Recent work determined the minimum coding rate, $R_1 = R(P, Q, d)$, for this setting. We observe that for large word length and with high probability, the matching codeword is typical with a distribution Q_1 which is different from Q. If a new random codebook is generated $\sim Q_1$, then the source string will favor code words which are typical with a new distribution Q_2, resulting in minimum coding rate $R_2 = R(P, Q_1, d)$, and so on. We show that the sequences of distributions $Q_1, Q_2 \ldots$ and rates $R_1, R_2, \ldots$, generated by this procedure, converge to an optimum reproduction distribution Q^*, and the rate-distortion function $R(P, d)$, respectively. We also derive a fixed rate-distortion *slope* version of this *natural type selection* process. In the latter case, an iteration of the process stochastically simulates an iteration of the Blahut-Arimoto algorithm for rate-distortion function computation (without recourse to prior knowledge of the underlying source distribution). To strengthen these limit statements, we also characterize the steady state error of these procedures when iterating at a finite string length. Implications of the main results provide fresh insights into the workings of lossy variants of the Lempel-Ziv algorithm for adaptive compression.

Key words: approximate string matching, universal/adaptive compression, rate-distortion, types, Blahut-Arimoto algorithm, alternating optimization, Lempel-Ziv coding, gold washing.

*This work was presented in part at the Information Theory Workshop, Haifa, Israel, June 1996, at the International Symposium on Information Theory, Ulm, Germany, June 1997, and at the Canadian Workshop on Information Theory, Kingston, Canada, June 1999. This work was supported in part by the National Science Foundation under Grant no. NCR-9314335, the Binational Science Foundation no. 9800309, the University of California MICRO Program, Cisco Systems, Inc., Conexant Systems, Inc., Dialogic Corp., Fujitsu Laboratories of America, Inc., General Electric Co., Hughes Network Systems, Lernout & Hauspie Speech Products, Lockheed Martin, Lucent Technologies, Inc., Qualcomm, Inc., and Texas Instruments, Inc.

I Introduction

Codebook adaptation, based on the frequency of codeword use, plays a central role in many universal source coding algorithms. Prominent examples in the information theoretic literature include the Lempel-Ziv 78 (LZ78) algorithm for lossless coding [26] and, more recently, the Gold-Washing algorithm for lossy coding [24].

The lossless case was thoroughly investigated over the past two decades, and is now relatively well understood [26, 17]. LZ78 grows a tree, while observing the incoming source strings, in a way that ensures that, asymptotically, most code words are typical sequences for the source. In Wyner and Ziv's "database" interpretation of string matching [17] (derived to explain the workings of the LZ77 variant), long source strings are encoded by specifying a pointer to a matching string in the database. This provides another angle on how the asymptotic optimality hinges on the same fundamental phenomenon, as the database, assuming it is generated by the same source (or by an independent copy of it), contains mostly typical strings.

We see that in lossless coding, although the source statistics may be unknown, the source history provides us with good "examples" that can serve as useful code words. The situation in lossy coding, however, is fundamentally different, as good code words are not necessarily statistically equivalent, or even similar to the source. Thus, the expected role of codebook adaptation in the lossy coding case, is not that of matching the source statistics but, more generally, that of "type selection". In other words, instead of emulating the source statistics, we seek the optimal reproduction type for this source. In light of the spontaneous way in which good code words are selected in the LZ algorithm, we ask ourselves whether a mechanism of *natural* type selection also exists in the framework of lossy source coding.

Recall the structure of the random code that achieves the entropy and the rate-distortion function in lossless and lossy coding, respectively. Let $X_1, X_2, \ldots$ be a discrete memoryless source of letters from $\mathcal{X}$, generated by a distribution $P = \{P(x),\ x \in \mathcal{X}\}$. By Shannon's first coding theorem, if we generate a codebook of $\exp(\ell(H(P)+\epsilon))$ independent code words of length ℓ from the source, where

$$H(P) = -\sum_{x\in\mathcal{X}} P(x)\log P(x)$$

is the source entropy in natural units, then the probability that some independent ℓ-string $\mathbf{X} \sim P$ is in the codebook goes to 1 as ℓ goes to infinity. (Note that logarithms are base e throughout this paper.)

For lossy coding we assume a reproduction alphabet $\mathcal{Y}$ and a distortion measure $\rho : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$, such that the distortion incurred by reproducing source string $\mathbf{x}$ by codeword $\mathbf{y}$ is given by $\rho(\mathbf{x}, \mathbf{y}) = \frac{1}{\ell} \sum_{i=1}^{\ell} \rho(x_i, y_i)$. We assume that $\mathcal{X}$, $\mathcal{Y}$ and ρ are finite and, for technical convenience, that

$$\text{for every } x \in \mathcal{X} \text{ there exists some } y \in \mathcal{Y} \text{ such that } \rho(x, y) = 0 \tag{1}$$

(see [2]). We say that $\mathbf{y}$ "d-matches" $\mathbf{x}$, or that a "d-match" event occurred if

$$\rho(\mathbf{x}, \mathbf{y}) \leq d.$$

The rate-distortion function of the source $X \sim P$ is defined as

$$R(P, d) = \min_{W:\, \rho(P,W) \leq d} I(P, W), \qquad d \geq 0, \tag{2}$$

where $W = \{W(y|x)\}$ denotes a transition distribution from $\mathcal{X}$ to $\mathcal{Y}$, and where

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

denotes the mutual information, and

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

denotes the average distortion, induced by the joint distribution $P \circ W \triangleq \{P(x)W(y|x)\}$ [2]. The minimum in (2) exists and is finite (although not necessarily unique) by the continuity and finiteness of I over the compact, non-empty set $\{W : \rho(P, W) \leq d\}$. Let $W^* = W^*(P, d)$ be a minimizing transition distribution, i.e., $R(P, d) = I(P, W^*)$. By marginalizing the joint distribution $P \circ W^*$, we obtain an *optimal reproduction distribution*:

$$Q^* = Q^*(P, d) = \Big\{ \sum_x P(x) W^*(y|x) \Big\}. \tag{3}$$

Shannon's lossy coding theorem states that if we draw a codebook of $\exp(\ell(R(P, d) + \epsilon))$ independent code words of length ℓ from $\mathcal{Y}$ with i.i.d. symbol distribution Q^*, then the probability that an independent source ℓ-string, $\mathbf{X}$, finds a d-matching codeword in the codebook approaches 1 as $\ell \to \infty$ [5].

Suppose, however, that a different distribution $Q \neq Q^*$ was used to generate the codebook. We denote the minimum coding rate in this case $R(P, Q, d)$. In other words, to guarantee a d-match with probability approaching 1, we need to draw $\exp[\ell(R(P, Q, d) + \epsilon)]$ code words. Clearly, $R(P, Q, d) > R(P, Q^*, d) \equiv R(P, d)$. The function $R(P, Q, d)$ is given in Theorem 2 of the next section and, in fact, appeared in several recent papers in the area of lossy string matching [16, 9, 12, 22, 18, 8]. From the above it is clear that, in order to achieve optimum lossy compression, the code words' type must approach Q^* (see [15] for a similar condition in channel coding). In the case of unknown source statistics, it is desirable to develop an adaptive mechanism which naturally selects code words whose type approaches Q^*. The question of whether such a mechanism exists motivated this work.

The main results of this paper (Theorems 5 and 6) show the existence of a natural type selection mechanism for lossy string matching. More specifically, we show that for large ℓ and with high probability, the first codeword in a codebook generated $\sim Q$, that d-match a source string $\sim P$, is typical with the "favorite" type $Q^*(P, Q, d)$ which emerges from the optimization operation that determines $R(P, Q, d)$. We consider the sequence of types Q_n, $n = 1, 2, \ldots$, where $Q_{n+1} = Q^*(P, Q_n, d)$, which is effectively generated by a procedure that repeatedly alternates between d-matching, and codebook re-generation using the most likely d-matching type. The type Q_{n+1} "delivers better rate" than Q_n in the sense that $R(P, Q_{n+1}, d) \leq R(P, Q_n, d)$. In Theorem 5 it is shown that, in fact,

$$Q_n \to Q^* \quad \text{and} \quad R(P, Q_n, d) \to R(P, d) \tag{4}$$

as $n \to \infty$. Interestingly, this procedure simulates the iterations of a fixed-distortion variant of the Blahut algorithm, or Blahut-Arimoto algorithm (BA) [1, 3], as it iteratively computes the rate-distortion function starting with Q_0 as reproduction distribution. The proof of convergence of Q_n and $R(P, Q_n, d)$ in (4) builds on the alternating minimization approach of Csiszár and Tusnády [6, 7].

The above result still has a significant weakness due to its order of limits. The convergence as $n \to \infty$ in (4) presumes that ℓ is large. In other words, the limit in ℓ is taken before the limit in n. In practice, however, one is interested in the reversed order of limits and in the asymptotic behavior for finite word length. This shortcoming is eliminated by Theorem 6 that shows that convergence is ensured when the order of limits is reversed (if Q^* is unique), and quantifies the "steady state error" for finite word length.

The basic tools for our analysis, namely, the rate function $R(P, Q, d)$ and the favorite type $Q^*(P, Q, d)$, are derived in Section II and Section III, respectively, in Theorems 1 - 4. (Further discussion of their properties and usefulness is relegated to an appendix.) These sections also provide "fixed-slope" versions of these quantities, corresponding to coding for minimum rate-distortion Lagrangian (Theorem 3). The evolution of types by alternating d-matching and code re-generation is presented in Section IV, where the convergence results (Theorems 5 and 6) are stated. In Section V, the relation to the Blahut algorithm is shown, and proofs are provided for Theorems 5 and 6.

A preliminary version of our results appeared in [22, 23]. Theorem 1, on the conditional d-match event and the lower mutual information, was proved by Zhang, Yang and Wei in [25] in the context of redundancy analysis. Yang and Kieffer [18] proved a version of Theorem 2, regarding the rate function $R(P, Q, d)$, in the more general database string matching setting, possibly with memory in the database. The main results in this work build on, and is inspired by the intuition obtained from the simpler derivation in [22, 23]. Other related work in the literature, in particular [9, 11, 12, 10], will be discussed in later sections in the context of the corresponding results. Relations with the Gold-Washing algorithm of Zhang and Wei [24], as well as possible generalizations and implications of our result on adaptive/universal lossy compression will be discussed in further detail in Section VI.

II Preliminaries: Lossy Compression with a Mismatched Random Codebook

This section formulates the problem of lossy source coding with a random codebook of an arbitrary (typically "mismatched") generating distribution, and reviews results concerning the expected encoding performance. Both fixed-distortion and "fixed-slope" settings will be considered.

Let distribution $Q = \{Q(y) : \ y \in \mathcal{Y}\}$ be used to randomly generate infinitely many codewords of length ℓ: $\mathbf{Y}_1, \mathbf{Y}_2, \ldots$. Specifically, each symbol in each codeword is drawn independently from $\mathcal{Y}$ according to distribution Q. To avoid resolvable but unnecessary complications, we will usually assume that Q is strictly positive, i.e., $Q(y) > 0 \ \ \forall y \in \mathcal{Y}$. Let $\mathbf{x}$ denote a given source string of ℓ letters from $\mathcal{X}$ having type (empirical distribution) $P_{\mathbf{x}}$ (see e.g., [5]). For some fixed distortion level $d \geq 0$, we consider the "d-match" event $\{\rho(\mathbf{x}, \mathbf{Y}) \leq d\}$ where codeword $\mathbf{Y}$ d-matches source string $\mathbf{x}$, and where ρ satisfies (1). Since all symbols in all codewords are generated i.i.d., the probability of d-match depends on $\mathbf{x}$ only through its type $P_{\mathbf{x}}$.

Let random variable N_ℓ denote the index of the first code word in the codebook that d-matches $\mathbf{x}$, i.e.,

$$\rho(\mathbf{x}, \mathbf{Y}_i) > d \ \text{ for } \ 1 \leq i \leq N_\ell - 1 \ \text{ and } \ \rho(\mathbf{x}, \mathbf{Y}_{N_\ell}) \leq d. \tag{5}$$

If an independent random source string $\mathbf{X}$ is generated i.i.d. according to $P = \{P(x) : \ x \in \mathcal{X}\}$, then it can be shown, using the technique of the conventional forward coding theorem (e.g., [5, sec. 13.6]), that the normalized logarithm of the random variable N_ℓ converges to a constant, $R(P, Q, d)$, in probability. Not surprisingly, the same constant appears as the limit of the "d-match exponent". See [18] and Theorem 2 below for the precise statement, and specification of $R(P, Q, d)$. Thus, for long words $R(P, Q, d)$ represents the exact rate in nats per source symbol required to encode a source string by specifying the index of a d-matching codeword in the random codebook.

II-1 Coding Rate at a Fixed Distortion

Let us consider first the conditional probability that codeword $\mathbf{Y}$ d-matches string $\mathbf{x}$, given that $\mathbf{Y}$ is of type $Q' \in \mathcal{Q}_\ell$. We follow the convention of [5, sec. 12.1] where $\mathcal{Q}_\ell$ denotes the set of all possible types (empirical distributions) of words in $\mathcal{Y}^\ell$, and $T(Q')$, the "type class" of Q', denotes the set of words in $\mathcal{Y}^\ell$ having type Q'. In order to characterize this probability, we use the quantity *"lower mutual information"* [25] (or *"minimum mutual* *information with constrained output Q"* [22, 23]), defined as

$$I_m(P\|Q,d) = \begin{cases} \min_{W\in\mathcal{W}_{P,Q,d}} I(P,W), & \text{if } \mathcal{W}_{P,Q,d} \text{ is not empty} \\ \infty, & \text{otherwise} \end{cases} \tag{6}$$

where $\mathcal{W}_{P,Q,d}$ denotes the set of all transition distributions W from $\mathcal{X}$ to $\mathcal{Y}$ such that input distribution P induces distribution Q on the output, and an expected distortion which is at most d:

$$\mathcal{W}_{P,Q,d} = \{W : \sum_x P(x)W(y|x) = Q(y),\ \rho(P,W) \le d\}.$$

By the convexity of the closed set $\mathcal{W}_{P,Q,d}$ and the finiteness and continuity of I, the minimum in (6) exists whenever $\mathcal{W}_{P,Q,d}$ is not empty. The following theorem, [25], gives an asymptotic characterization of the conditional d-match probability.

Theorem 1 (Conditional d-Match Exponent, Zhang, Yang and Wei [25]) *Let source string $\mathbf{x}$ be of type $P_{\mathbf{x}}$, and let code word $\mathbf{Y}$ be drawn i.i.d. $\sim Q$. For any type $Q' \in \mathcal{Q}_\ell$ such that $I_m(P_{\mathbf{x}}\|Q',d) < \infty$,*

$$-\frac{1}{\ell}\log\Big(\Pr\Big\{\rho(\mathbf{x},\mathbf{Y}) \le d \,\Big|\, \mathbf{Y} \in T(Q')\Big\}\Big) = I_m(P_{\mathbf{x}}\|Q',d) + o(1) \tag{7}$$

where $o(1) \to 0$ as $\ell \to \infty$ uniformly in $P_{\mathbf{x}}$, Q' and d.

We note in passing that we have another proof of (7) that builds on Sanov's theorem and is closer in spirit to the main results of this paper. However, it is not reproduced here for conciseness and we rely instead on the proof in [25]. In [22, 23] we introduced I_m as the rate of a "pure type" (or a fixed decomposition) codebook. From (2), for fixed P and d, I_m is minimized by the optimum reproduction distribution $Q^* = Q^*(P,d)$ that realizes the rate-distortion function, i.e., $\min_Q I_m(P\|Q,d) = I_m(P\|Q^*,d) = R(P,d)$.

The probability that a codeword generated according to distribution Q will be of type $Q' \in \mathcal{Q}_\ell$ is determined by the divergence $\mathcal{D}$ via

$$-\frac{1}{\ell}\log\Big(\Pr\{\mathbf{Y} \in T(Q')\}\Big) = \mathcal{D}(Q'\|Q) + o(1), \tag{8}$$

where $o(1) \to 0$ as $\ell \to \infty$ uniformly in Q' (e.g., [5]). By taking the expectation of the conditional d-match probability (7) over the type probabilities (8), and using the fact that the cardinality (number of types) $|\mathcal{Q}_\ell|$ grows polynomially with ℓ, it is easy to obtain an asymptotic characterization of the *un*conditional d-match probability for any source string $\mathbf{x}$:

$$\begin{aligned}\Pr\Big\{\rho(\mathbf{x},\mathbf{Y})\le d\Big\} &= \sum_{Q'\in\mathcal{Q}_\ell} \exp\Big(-\ell[I_m(P_{\mathbf{x}}\|Q',d)+o(1)]\Big)\cdot\exp\Big(-\ell[\mathcal{D}(Q'\|Q)+o(1)]\Big)\\ &= \max_{Q'}\ \exp\Big(-\ell[I_m(P_{\mathbf{x}}\|Q',d)+\mathcal{D}(Q'\|Q)+o(1)]\Big).\end{aligned}\tag{9}$$

The next theorem, [18], states this result formally, and relates it to the coding rate, $\lim_{\ell\to\infty}\log(N_\ell)/\ell$, for a random source string $\mathbf{X}$, where N_ℓ was defined in (5). Note that as the d-match probability of (9) depends on the source string $\mathbf{x}$ (in fact on $P_{\mathbf{x}}$), the d-match probability of a *given random string* $\mathbf{X}$ is a random variable whose asymptotic behavior is given by the theorem.

Theorem 2 (Mismatched Coding Rate, Yang and Kieffer [18]) *If a source string* $\mathbf{X}$ *is drawn i.i.d.* $\sim P$ *and code words* $\{\mathbf{Y}\}$ *are drawn i.i.d.* $\sim Q$*, where* Q *is strictly positive, i.e.,* $Q(y)>0\ \forall y\in\mathcal{Y}$*, then*

(a) $-\frac{1}{\ell}\log\Big(\Pr\Big\{\rho(\mathbf{X},\mathbf{Y})\le d\ \Big|\ \mathbf{X}\Big\}\Big)\to R(P,Q,d)$ *, and*

(b) $\log(N_l)/l\to R(P,Q,d)$ *,*

in probability as the word length $\ell\to\infty$*, where*

$$R(P,Q,d)=\min_{Q'}\Big\{I_m(P\|Q',d)+\mathcal{D}(Q'\|Q)\Big\}.\tag{10}$$

Note that for strictly positive Q, $R(P,Q,d)$ is finite, since $I_m(P\|Q',d)$ is finite for some Q' (e.g., it is equal to $R(P,d)$ for $Q'=Q^*$), and since $\mathcal{D}(Q'\|Q)$ is finite for all Q'. The minimum in (10) exists since the minimization set (the simplex) is compact, and we show below that it is unique.

Note that the formula for $R(P,Q,d)$ in [18] is slightly different from (10), as discussed below. In [22, 23] (10) was derived as a special case of a general formula for the coding rate of a random code with "type spectrum" $w(Q)$(see Appendix B). Other characterizations of the performance of mismatched codes and waiting times in string matching appeared in [9, 11, 12, 8].

The expression (10) contains a double minimization, over W (implicit in the definition of I_m) and over Q'. It is possible to convert it into the single minimization:

$$R(P,Q,d) \quad=\quad \min_{W:\,\rho(P,W)\le d}\Big\{I(P,W)+\mathcal{D}([P\circ W]_y\|Q)\Big\}\tag{11}$$

$$= \min_{W:\,\rho(P,W)\le d} \mathcal{D}(P\circ W\|P\times Q) \tag{12}$$

where $[P\circ W]_y = \{\sum_x P(x)W(y|x)\}$ denotes the output (y-marginal) distribution of $P\circ W$. The formulation of (11) was used in [18] to express the waiting time exponent in approximate string matching. The concise form of (12) follows since for $Q' = [P\circ W]_y$

$$\begin{aligned}
\mathcal{D}(P\circ W\|P\times Q) &= \sum_{x,y} P(x)W(y|x)\log\left(\frac{W(y|x)}{Q(y)}\right)\\
&= \sum_{x,y} P(x)W(y|x)\log\left(\frac{W(y|x)}{Q'(y)}\right) + \sum_y Q'(y)\log\left(\frac{Q'(y)}{Q(y)}\right)\\
&= I(P,W) + \mathcal{D}(Q'\|Q). \qquad (13)
\end{aligned}$$

Other useful formulations of $R(P,Q,d)$ exist.[1] Since (12) is a minimization of a strictly convex function (divergence) over a convex set, the minimizing joint distribution $P\circ W$ is unique whenever $R(P,Q,d) < \infty$ (see [5, eq. (12.152)]). Hence, the y-marginal of this joint distribution is the *unique* minimizer of (10). As was the case for $I_m(P\|Q,d)$, and not surprisingly, the minimization of $R(P,Q,d)$ over Q results in the rate-distortion function, i.e.,

$$\min_Q R(P,Q,d) = R(P,Q^*,d) = R(P,d). \tag{14}$$

The form of $R(P,Q,d)$ in (10) provides useful insight into the d-matching mechanism, as discussed in Section III and in [21]. It also leads to upper bounds on $R(P,Q,d)$ via the relation $R(P,Q,d) \le I_m(P\|Q,d)$ (which is obtained by substituting the possibly sub-optimal $Q' = Q$ in (10)). Appendix A reviews properties of the function $R(P,Q,d)$, and demonstrates its applicability to bounds by giving a simple proof for the Steinberg-Gutman theorem [16] which states that $R(P,P,d) \le R(P,d/2)$.

II-2 Fixed R-D Slope Coding (Minimum R-D Lagrangian)

The following "d-matching" variant corresponds to lossy source coding with a fixed slope [5, sec. 13.8] and [20]. Fix the rate-distortion Lagrange parameter $s > 0$. Instead of looking

[1] An interesting alternative form of $R(P,Q,d)$ appeared in [18]. It states $R(P,Q,d) = \min \mathcal{D}(U;Y|X)$, where the minimization is over all U jointly distributed with $X \sim P$ such that $E\rho(X,U) \le d$, and where $\mathcal{D}(U;Y|X)$ denotes the conditional divergence between the random variables U and Y given X, where $Y \sim Q$ is independent of X.

for the first codeword that satisfies $\rho(\mathbf{x}, \mathbf{Y}_i) \le d$, we search for the index i that minimizes the weighted sum

$$\frac{1}{\ell}\log(i) + s \cdot \rho(\mathbf{x}, \mathbf{Y}_i).$$

As Theorem 3 below shows, the term $\frac{1}{\ell}\log(i)$ plays the role of "rate", that is, the number of nats per source symbol required to specify the minimizing codeword. Let the random variable $N_{s,\ell}$ denote the index of this codeword. By setting $i = 1$ we see that the minimum weighted sum is upperbounded by $s\rho_{max}$, where $\rho_{max} = \max_{x,y}\rho(x,y) < \infty$, and consequently $N_{s,\ell}$ never exceeds $\exp\{\ell s \rho_{max}\}$.

Theorem 3 (Fixed-Slope Matching) *As $\ell \to \infty$,*

$$\frac{1}{\ell}\log(N_{s,\ell}) \to R_s$$

and

$$\rho(\mathbf{X}, \mathbf{Y}_{N_{s,\ell}}) \to d_s$$

in probability, where $R_s = R(P, Q, d_s)$, and

$$d_s = d_s(P,Q) \triangleq \arg\min_d \Big\{ R(P,Q,d) + sd \Big\}. \tag{15}$$

The proof is given in Appendix C. The minimum in (15) exists and is unique since $R(P,Q,d)$ is a strictly convex $\cup$ function of d (see Appendix A). It follows that (R_s, d_s) is a point on the $R(P,Q,d)$ curve whose slope is $-s$ (if the first derivative is discontinuous at this point than $-s$ lies in the discontinuity interval). Clearly, by varying s we obtain the entire curve.

III The Typical d-Matching Codeword

The form of $R(P,Q,d)$ in (10) gives rise to a useful interpretation of the d-matching mechanism. Assume that $Q \ne Q^*$, i.e., Q is not an optimum reproduction distribution achieving the rate-distortion function (3), and let

$$Q^*(P,Q,d) \triangleq \arg\min_{Q'} \Big\{ I_m(P\|Q', d) + \mathcal{D}(Q'\|Q) \Big\} \tag{16}$$

denote a reproduction distribution that achieves $R(P,Q,d)$ in (10). As shown in the previous section, $Q^*(P,Q,d)$ exists and is unique whenever $R(P,Q,d) < \infty$, and it can be expressed as the output (y-marginal) distribution of $(P \circ W^*_{P,Q,d})$, where

$$P \circ W^*_{P,Q,d} \triangleq \arg \min_{V:\ V=P\circ W,\ \rho(P,W)\leq d} \mathcal{D}(V \| P \times Q) \tag{17}$$

Most of the codewords in the codebook are of type $\approx Q$. However, a type Q codeword will only d-match very unlikely source strings. More specifically, the probability that a source word will fall into the "d-ball"

$$\Big\{\mathbf{x}:\ \rho(\mathbf{x},\mathbf{y}) \leq d\Big\} \ \subset \mathcal{X}^\ell$$

covered by a code word $\mathbf{y}$ of type Q is $\approx \exp(-\ell I_m(P\|Q,d))$, which is very low except when Q is nearly optimal. The most efficient codewords in this respect are those having type $\approx Q^*$, as their d-ball probability is $\approx \exp(-\ell R(P,d)) \gg \exp(-\ell I_m(P\|Q',d))$ for any $Q' \neq Q^*$. But these codewords are very rare. Their frequency in the codebook is only $\approx \exp(-\ell \mathcal{D}(Q^*\|Q))$. The types near $Q^*(P,Q,d)$ minimize the exponent sum in (16), and thus strike the optimal balance between the type's coding efficiency and its frequency of occurrence in the codebook. Consequently, the d-matching code word typically belongs to one of these "good" types.

To put this intuitive interpretation on more concrete grounds, we will show that, indeed, for large ℓ the d-matching mechanism favors codewords whose type is close to the optimizing distribution $Q^*(P,Q,d)$. Let Q_{N_ℓ} denote the (random) type of the first d-matching codeword $\mathbf{Y}_{N_\ell}$. (For completeness, if no match occurred define Q_{N_ℓ} arbitrarily; we will see shortly that a match occurs with probability one.) Theorem 4 below states that the distribution of Q_{N_ℓ} collapses asymptotically on $Q^*(P,Q,d)$.

Theorem 4 ("Favorite Type") *If Q is strictly positive, then as $\ell \to \infty$,*

$$Q_{N_\ell} \to Q^*(P,Q,d) \qquad \textit{in prob.}$$

As shown above, if Q is strictly positive (i.e., $Q(y) > 0,\ \forall y$), then $R(P,Q,d)$ is finite, implying that $Q^*(P,Q,d)$ exists and is unique. With a slight modification of the proof it can be shown that the theorem holds under the weaker condition that $R(P,Q,d)$ is finite.

Proof: Let $\mathbf{x}$ be an arbitrary source ℓ-string. Given that Q is strictly positive, and from (1), it follows that the probability of the d-match event $\rho(\mathbf{x}, \mathbf{Y}_i) \leq d$ is positive. Further, as the code words are independently generated, we have a sequence of independent matching trials with positive matching probability and hence the d-match event occurs with probability one:

$$\Pr\Big(N_\ell < \infty \;\Big|\; \mathbf{X} = \mathbf{x}\Big) = 1. \tag{18}$$

The independent generation of code words also implies that the conditional distribution of the first d-matching code word $\mathbf{Y}_{N_\ell}$, given the fixed source string $\mathbf{X} = \mathbf{x}$, is equal to the conditional distribution of any code word $\mathbf{Y} \sim Q$, given that $\rho(\mathbf{x}, \mathbf{Y}) \leq d$:

$$\Pr\Big\{\mathbf{Y}_{N_\ell} = \mathbf{y} \;\Big|\; \mathbf{X} = \mathbf{x}\Big\} = \Pr\Big\{\mathbf{Y} = \mathbf{y} \;\Big|\; \rho(\mathbf{x}, \mathbf{Y}) \leq d\Big\}. \tag{19}$$

It is worthwhile to note that the above is only valid given a fixed known source string. A random source string $\mathbf{X}$ (which is drawn once) introduces dependence between the matching trials.

Since the symbols in each codeword are i.i.d. and the distortion measure is additive, the distributions in (19) depend only on the types. Specifically, let A and B be sets of distributions on $\mathcal{X}$ and $\mathcal{Y}$, respectively, and define

$$f(B, A, \ell) \triangleq \Pr\Big\{Q_{\mathbf{Y}} \in B \;\Big|\; P_{\mathbf{X}} \in A,\; \rho(P_{\mathbf{X}}, W_{\mathbf{Y}|\mathbf{X}}) \leq d\Big\} \tag{20}$$

where $Q_{\mathbf{Y}}$ and $P_{\mathbf{X}}$ denote the types of $\mathbf{X}$ and $\mathbf{Y}$, and $W_{\mathbf{Y}|\mathbf{X}}$ denotes the conditional type induced by $(\mathbf{X}, \mathbf{Y})$. Then, since $\rho(\mathbf{x}, \mathbf{Y}) \leq d$ is equivalent to $\rho(P_{\mathbf{x}}, W_{\mathbf{Y}|\mathbf{x}}) \leq d$ (note that since the latter's arguments are the induced types, it is not an expectation but the exact empirical matching distortion of the strings), identity (19) implies that

$$\Pr\Big\{Q_{N_\ell} \in B \;\Big|\; P_{\mathbf{X}} = P'\Big\} = f(B, P', \ell) \tag{21}$$

where P' denotes the type of $\mathbf{x}$. We may therefore compute the probability that Q_{N_ℓ} falls inside the set B as

$$\begin{aligned}
\Pr\Big\{Q_{N_\ell} \in B\Big\} &= \sum_{P' \in \mathcal{P}_\ell} \Pr\{P_{\mathbf{X}} = P'\} \;\; \Pr\Big\{Q_{N_\ell} \in B \;\Big|\; P_{\mathbf{X}} = P'\Big\} \\
&= E\Big\{f(B, P_{\mathbf{X}}, \ell)\Big\}
\end{aligned} \tag{22}$$

where $\mathcal{P}_\ell$ denotes the set of all types of ℓ-strings, and $E\{\cdot\}$ denotes expectation.

The key idea behind Theorem 4 is that (20) satisfies a "conditional limit theorem" [5, sec. 12.6]. Roughly, the conditional limit theorem implies that, conditioned on the event that the joint type of $(\mathbf{X}, \mathbf{Y})$ belongs to a convex set of distributions E, the joint type of $(\mathbf{X}, \mathbf{Y})$ converges in probability to the distribution $V^* = \arg\min_{V \in E} \mathcal{D}(V \| P \times Q)$ as $\ell \to \infty$. See [5, eq. (12.170)]. We cannot apply the conditional limit theorem directly, since we want to condition on the event that the type of $\mathbf{X}$ is close to P, which is, quite surprisingly, an *atypical* event when conditioned on $\rho(\mathbf{X}, \mathbf{Y}) \leq d$. To overcome this obstacle, we use a modified version of the conditional limit theorem, proved in Appendix E, in which the set E changes with ℓ. Let the sequence $P_1, P_2, \ldots$ be such that $P_\ell \in \mathcal{P}_\ell$ and $P_\ell \to \widetilde{P}$. Then, conditioned on the event $\left\{P_{\mathbf{X}} = P_\ell,\ \rho(P_{\mathbf{X}}, W_{\mathbf{Y}|\mathbf{X}}) \leq d\right\}$, the type of $\mathbf{Y}$ converges to $Q^*_{\widetilde{P},Q,d}$ in probability as $\ell \to \infty$. Using definition (20), this implies that for any open set B

$$f(B, P_\ell, \ell) \to \begin{cases} 1, & \text{if } Q^*(\widetilde{P}, Q, d) \in B \\ 0, & \text{if } Q^*(\widetilde{P}, Q, d) \notin \bar{B} \end{cases} \tag{23}$$

as $\ell \to \infty$, where $\bar{B}$ denotes the closure of B.

Equipped with (23) we now return to the distribution of Q_{N_ℓ}. Since $\mathbf{X}$ is i.i.d.$\sim P$, we have by the strong law of large numbers

$$P_{\mathbf{X}} \to P \quad \text{almost surely, as } \ell \to \infty. \tag{24}$$

Applying (23) with $\widetilde{P} = P$, and combining with (24), we have

$$f(B, P_{\mathbf{X}}, \ell) \to \begin{cases} 1, & \text{if } Q^*(P, Q, d) \in B \\ 0, & \text{if } Q^*(P, Q, d) \notin \bar{B} \end{cases} \quad \text{almost surely} \tag{25}$$

as $\ell \to \infty$. Since $f(\cdot)$ is bounded between zero and one, almost sure convergence implies convergence in expectation, which by (22) implies

$$\Pr\left\{Q_{N_\ell} \in B\right\} = E\left\{f(B, P_{\mathbf{X}}, \ell)\right\} \to \begin{cases} 1, & \text{if } Q^*(P, Q, d) \in B \\ 0, & \text{if } Q^*(P, Q, d) \notin \bar{B} \end{cases} \quad \text{as } \ell \to \infty. \tag{26}$$

The theorem now follows since we may take the set B as small as we want. $\square$

Remark: It is worth noting that the source word and the code word have different roles in the above analysis: with high probability the d-matching event occurs with a *typical* source word (of type $\sim P$) but with an *atypical* code word (of type $\sim Q^*(P,Q,d)$). This is due to the asymmetry in the probabilistic setting itself: *many* code words $\mathbf{Y} \sim Q$ are drawn to d-match a *single* source word $\mathbf{X} \sim P$.

Theorem 4 implies that by repeated d-matching of long enough source strings, one can identify the "favorite types". For future use we need a convergence result for a non-random representative type. To this end, define the *average favorite type* as the expected value of Q_{N_ℓ}, i.e.,

$$\overline{Q}_\ell(P,Q,d) = \sum_{Q'} Q' \ \Pr\{Q_{N_\ell} = Q'\}. \tag{27}$$

Since $\mathbf{X}$ and $\mathbf{Y}$ are memoryless and the distortion measure is additive, the d-matching code word $\mathbf{Y}_{N_\ell}$ is uniformly distributed over the type class of Q_{N_ℓ}. Thus, $\overline{Q}_\ell(P,Q,d)$ is also the marginal distribution of each component of $\mathbf{Y}_{N_\ell}$, e.g.,

$$\overline{Q}_\ell(P,Q,d)[y] = \Pr\{Y_{N_\ell,1} = y\} \tag{28}$$

where $Y_{N_\ell,1}$ denotes the first component of $\mathbf{Y}_{N_\ell}$.

Corollary 1

$$\overline{Q}_\ell(P,Q,d) \to Q^*(P,Q,d) \quad \textit{as } \ell \to \infty.$$

Proof: In view of (27), the corollary amounts to convergence *in expectation* of Q_{N_ℓ} to $Q^*(P,Q,d)$. Since $Q_{N_\ell}(y)$ is bounded between zero and one, this follows from convergence *in probability*, as given by Theorem 4. □

Appendix F further shows that the convergence in Corollary 1 is uniform over all Q's which are bounded away from zero.

Similar results are obtained for string matching to minimize the rate-distortion Lagrangian (see Section II.2): Let $-s$ be the desired slope, and define

$$Q_s^*(P,Q) = Q^*(P,Q,d_s) \tag{29}$$

where $d_s = d_s(P,Q)$ is as defined in (15). Then, it follows from Theorems 3 and 4 that as $\ell \to \infty$

$$Q_{N_{s,\ell}} \to Q_s^*(P,Q) \quad \text{in prob.}$$

where $Q_{N_{s,\ell}}$ denotes the random type of $\mathbf{Y}_{N_{s,\ell}}$.

IV Type Evolution by Alternating d-Matching and Code Re-Generation

As explained in the previous section, for large ℓ the average favorite type in the codebook, $\overline{Q}_\ell(P,Q,d)$, is close to a distribution $Q^*(P,Q,d)$ which is more efficient for coding $\mathbf{X}$ than Q, the codebook generating distribution - although it is not as efficient as Q^*, the rate-distortion function achieving distribution. This motivates the following iterative procedure: Starting with an arbitrary initial generating distribution Q_0, at each iteration the average favorite type in the codebook is identified and generates a new codebook. This recursion results in a sequence of codebook distributions

$$Q_{\ell,n} = \overline{Q}_\ell(P, Q_{\ell,n-1}, d) \qquad n = 1, 2, \ldots \tag{30}$$

where $\overline{Q}_\ell(P,Q,d)$ is given in (28) or (27).

As the word length ℓ goes to infinity, the sequence $Q_{\ell,n}$ approaches the sequence of distributions defined by the recursion

$$Q_n = Q^*(P, Q_{n-1}, d) \qquad n = 1, 2, \ldots \tag{31}$$

where $Q^*(P,Q,d)$ is defined in (16). Note that Q_n achieves $R(P,Q_{n-1},d)$ in the minimization of (10).

Recursion (31) corresponds to matching with fixed distortion. For the fixed slope variant of the coding scheme we have, as $\ell \to \infty$, the recursion

$$Q_n = Q_s^*(P, Q_{n-1}) \qquad n = 1, 2, \ldots \tag{32}$$

where Q_s^* is defined in (29). In this recursion both the rate and the distortion vary with n according to

$$R_n = R(P, Q_n, d_n) \quad \text{and} \quad d_n = d_s(P, Q_n) \tag{33}$$

where the function $d_s(\cdot)$ is defined in (15).

As shown in the next section, the sequences of distributions $Q_0, Q_1, \ldots$ and associated coding rates $R_1, R_2 \ldots$ and distortions $d_0, d_1, \ldots$ of the fixed-slope recursion (32), (33), coincide with the sequence of tentative reproduction distributions and associated rates and distortions, respectively, in the iterations of the Blahut-Arimoto (BA) algorithm [1, 3, 6, 4] as it computes the rate-distortion function of the source starting from an initial distribution Q_0. Recursion (31), which corresponds to iterations with *fixed distortion* d, can be viewed as a stochastic simulation of the alternating minimization procedure of Csiszár and Tusnády [7] which computes the constrained double minimization of $\mathcal{D}(P \circ W \| P \times Q)$ (over W and Q), and hence leads to the rate-distortion function at distortion level d. This observation leads to the next theorem.

Theorem 5 (Natural Type Selection for $\ell = \infty$) *For any strictly positive initial distribution Q_0, the fixed-distortion recursion (31) satisfies as $n \to \infty$*

$$\begin{aligned} R(P, Q_n, d) &\to R(P, d) \\ Q_n &\to Q^* \end{aligned} \tag{34}$$

where Q^* is a reproduction distribution that achieves the rate-distortion bound $R(P, d)$. (If $Q^*(P, d)$ of (3) is not unique then Q^* depends on Q_0.) Moreover, the sequence $R(P, Q_n, d)$ is monotonic non-increasing.

The fixed-slope recursion (32), (33) follows the steps of the BA algorithm and converges to a point of slope $-s$ on the rate-distortion function as well as to a corresponding optimal reproduction distribution, i.e., $R_n \to R(P, d^*)$, $d_n \to d^*$, and $Q_n \to Q^*$, where d^* minimizes $R(P, d) + sd$ and Q^* achieves $R(P, d^*)$ (if these minimizers are not unique then d^* and Q^* depend on Q_0).

The proof of Theorem 5 is given in the next section. An interesting consequence of Theorem 5 is that only an optimum reproduction distribution Q^* is a "stable" codebook distribution.

Theorem 5 may be viewed as a double limit, where the word length goes to infinity first, and then the number of iterations of codebook re-generation is taken to infinity, i.e.,

$$\lim_{n \to \infty} \lim_{\ell \to \infty} R(P, Q_{\ell, n}, d) = R(P, d). \tag{35}$$

However, the reverse order of limits is desirable for practical considerations, as it implies that we can approximate the optimum recursion by matching with a *finite* word length. The next theorem states this stronger claim under a slightly more restrictive condition.

Theorem 6 (Natural Type Selection for finite ℓ) *Suppose that $Q^* = Q^*(P,d)$ of (3) is unique, and that the initial distribution Q_0 is strictly positive. Furthermore, assume that the sequence Q_n of (31) (corresponding to $\ell = \infty$) which starts from Q_0 is bounded away from zero, i.e., $Q_n(y) \geq q_{min}$ for all y and n for some positive q_{min}. (In particular, Q_0 and $Q^*(P,d) = \lim_n Q_n$ are lower bounded by q_{min}.) Then the sequence of codebooks in recursion (30) arbitrarily well approximates, for large enough ℓ, the optimum codebook for compression with distortion d, i.e.,*

$$\lim_{\ell\to\infty} \limsup_{n\to\infty} \mathcal{D}(Q^* \| Q_{\ell,n}) = 0 \tag{36}$$

and hence $\lim_{\ell\to\infty} \limsup_{n\to\infty} R(P, Q_{\ell,n}, d) = R(P,d)$.

The proof is given in the next section. We believe that the theorem holds for some optimum Q^* even if $Q^*(P,d)$ is not unique, and even without boundedness of Q_n. Note that even though for some fixed q and poorly chosen initial distribution Q_0 the condition $Q_n(y) \geq q$ may be violated for a few n's, the problem can be resolved by enforcing the condition on the procedure update steps. It can be shown that the convergence property of the BA is unaffected by such modification of the update rule.

A similar statement holds with respect to matching with a fixed slope. We leave for future work the investigation of a stochastic version of the type selection recursion, where the generating distribution for the next codebook is determined by the (random) type of the d-matching codeword, i.e.,

$$Q_{\ell,n+1} = Q^{(n)}_{N_\ell},$$

where $Q^{(n)}_{N_\ell}$ obtained by matching with a codebook $\sim Q_{\ell,n}$.

V A String-Matching Interpretation of the Blahut-Arimoto Algorithm

Theorem 5 above follows from the analogy between the iterations of string matching / codebook re-generation and the BA algorithm, as the length of the string goes to infinity. We first recall a parametric solution for $R(P,Q,d)$, [19], paralleling the parametric representation of the rate-distortion function [2, 3, 6].

Proposition 1 (Parametric Form - Yang and Zhang [19]) *For fixed P and Q, $R(P,Q,d)$ as a function of d is given parametrically in terms of the parameter $s \geq 0$, by*

$$\begin{aligned} R_s &= R_s(P,Q) = \mathcal{D}(P \circ W_s \| P \times Q) \\ d_s &= d_s(P,Q) = \rho(P, W_s) \end{aligned} \tag{37}$$

where the transition distribution W_s is given by

$$W_s(y|x) = \frac{Q(y)\exp(-s\rho(x,y))}{\sum_{y'} Q(y')\exp(-s\rho(x,y'))}. \tag{38}$$

The pair (d_s, R_s) is the point of slope $-s$ on the $R(P,Q,d)$-versus-d curve, as given in (15).

It follows that at distortion level d_s, W_s above is the minimizing transition distribution in (11), (12) and (17), while the "favorite" reproduction distribution in the codebook for large ℓ (16), (29) is given by

$$Q^*(P,Q,d_s)[y] = \sum_x P(x) W_s(y|x) = \sum_x P(x) \frac{Q(y)\exp(-s\rho(x,y))}{\sum_{y'} Q(y')\exp(-s\rho(x,y'))}. \tag{39}$$

Substituting in (10), we can rewrite R_s above as $R_s = I(P, W_s) + \mathcal{D}(Q^*(P,Q,d_s)\|Q)$.

Let us compare the above with the BA algorithm [3, thm. 6]. It is easy to see that if $d = d_s(P,Q)$, then $Q^*(P,Q,d)$ is equal to the output distribution obtained after one iteration of BA for slope parameter s and initial distribution Q. Thus, in our terms BA can be written recursively as

$$\begin{aligned} Q_n &= Q^*(P, Q_{n-1}, d_{n-1}) \\ d_n &= d_s(P, Q_n) \qquad n = 1, 2, \ldots \end{aligned} \tag{40}$$

coinciding with the recursion defining Q_n and d_n in (32) and (33). We are now ready to prove Theorem 5.

Proof of Theorem 5: We start with the second part of the theorem, i.e., with the fixed slope recursion. This part follows directly from the convergence of the BA algorithm. Specifically, as shown by Csiszár [6, eq. (12) and thm. 1], recursion (40) satisfies $Q_n \to Q^*$ and $d_n \to d^*$, where d^* minimizes $R(P,d) + sd$, and Q^* achieves $R(P,d^*)$. (To compare with [6] make the substitutions $W \to \mathbf{Q}$ and $Q \to \mathbf{q}$, and note that we also have $W_n \to W^*$, where W_n is the W_s of (38) associated with $Q = Q_{n-1}$, and where $Q_n = [P \circ W_n]_y$, and $W^* = W^*(P, d)$ is the optimum transition distribution (3). See also [4].)

We now turn to the fixed distortion recursion. The convergence in the first part of Theorem 5 follows from the Csiszár-Tusnády Theorem [7] for general alternating minimization procedures. By (12) and (14), we can write the rate-distortion function as the double minimization

$$R(P, d) = \min_{Q} \min_{W:\rho(P,W)\leq d} \mathcal{D}(P \circ W \| P \times Q). \tag{41}$$

It is easy to verify that the sets of joint distributions $\{P \times Q : \text{ any } Q\}$ and $\{P \circ W : \rho(P, W) \leq d\}$ are convex. Moreover, for fixed W, the output distribution Q which minimizes $\mathcal{D}(P \circ W \| P \times Q)$ is the y-marginal of $P \circ W$; while for fixed Q, the conditional distribution W which minimizes $\mathcal{D}(P \circ W \| P \times Q)$ under the distortion constraint d induces an output $Q^*(P, Q, d)$. Thus, the recursion of (31) defines a sequence,

$$(P \times Q_{n-1}) \to (P \circ W_n) \to (P \times Q_n) \to \ldots,$$

of alternating minimization between the above convex sets, where distance is measured by the divergence. See Figure 1a. By [7, thm. 3] the sequences of divergences and distributions converge to the minimum divergence, i.e., to $R(P, d)$, and to an optimal reproduction distribution Q^*, respectively. $\square$

Consider Figure 1b, which shows what happens to the geometric picture of Figure 1a when ℓ is finite. We see that each projection $Q_{\ell,n+1} = \overline{Q}_\ell(P, Q_{\ell,n}, d)$ is now shifted within a radius $\epsilon = \epsilon(\ell)$ with respect to the "ideal projection" $Q^*(P, Q_{\ell,n}, d)$, where $\epsilon(\ell) \to 0$ as $\ell \to \infty$. The desired result follows from the fact that for a finite "projection error" ϵ the sequence of projections enters eventually an ϵ' neighborhood of $Q^*(P, d)$, such that $\epsilon \to 0$ implies $\epsilon' \to 0$. This process is very similar to the behavior of the Least Mean Square algorithm for adaptive filtering, which is a "noisy version" of the gradient descent method, or more generally, the behavior of a noisy tracking loop which passes from an "acquisition phase" to a "locked phase".

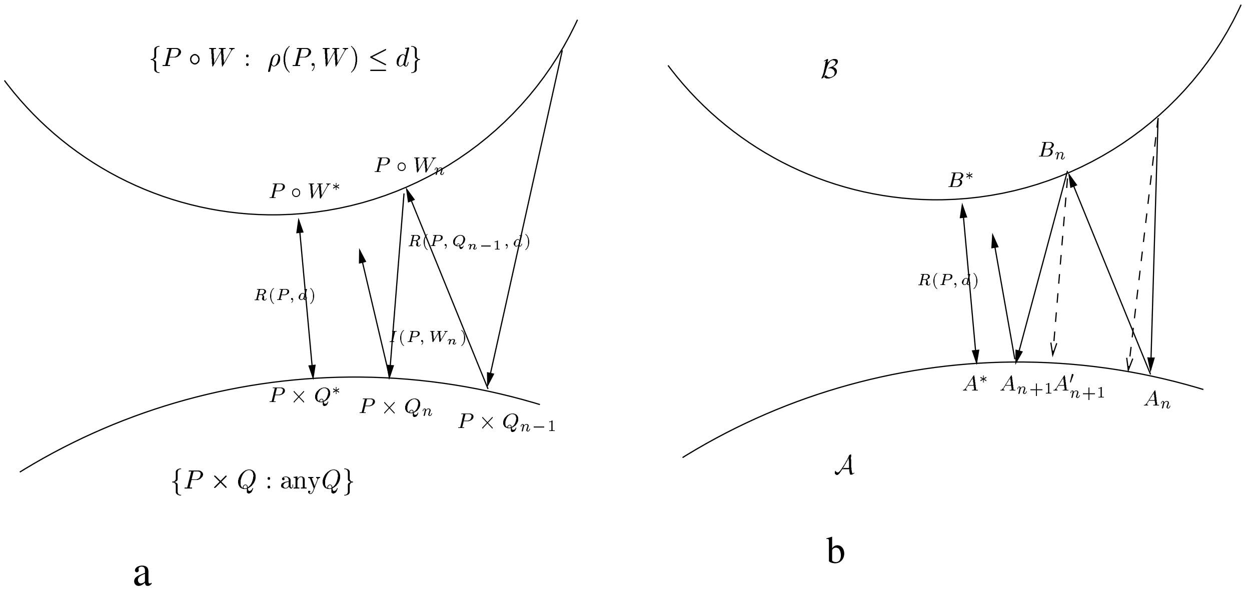


Figure 1: Alternating projections between convex sets: (a) ideal case corresponding to $\ell = \infty$; (b) "noisy projections" corresponding to finite ℓ.

Proof of Theorem 6: The proof modifies the derivation in [6, 4, 7] to "noisy iterations". Let $Q^* = Q^*(P,d)$ denote the optimum reproduction distribution (3), assumed here to be unique, and let $W^* = W^*(P,d)$ denote a corresponding optimum transition distribution. Define L_ϵ, for $\epsilon \geq 0$, as the maximum of $\mathcal{D}(Q^* \| Q')$ over all Q' such that $R(P,Q',d) - R(P,d) \leq \epsilon$. We show below that

$$\limsup_{n\to\infty} \mathcal{D}(Q^* \| Q_{\ell,n}) \leq L_{\epsilon(\ell)} + \epsilon(\ell) \tag{42}$$

where $\epsilon(\ell) \to 0$ as $\ell \to \infty$. By the uniqueness of Q^*, it follows that $R(P,Q',d)$ is strictly convex in Q' near Q^*, thus $L_\epsilon \to 0$ as $\epsilon \to 0$, and the theorem follows.

We establish (42) by a sequence of lemmas. Let $Q'_{\ell,n}$ denote the "ideal projection", $Q^*(P, Q_{\ell,n-1}, d)$, at the n-th step. The first lemma corresponds to the basic ingredient of Csiszar and Tusnady's proof [7] regarding a single "noiseless" iteration:

Lemma 1

$$R(P, Q_{\ell,n}, d) - R(P,d) \leq \mathcal{D}(Q^* \| Q_{\ell,n}) - \mathcal{D}(Q^* \| Q'_{\ell,n+1}).$$

Proof: To simplify notations, let $A_n = P \times Q_{\ell,n}$, $A^* = P \times Q^*$, $B_n = P \circ W^*(P, Q_{\ell,n}, d)$, $B^* = P \circ W^*$ and $A'_n = P \times Q'_{\ell,n}$, where $W^*(P,Q,d)$ is defined in (17). See Figure 1b. Thus $R(P,d) = \mathcal{D}(B^* \| A^*)$, $R(P, Q_{\ell,n}, d) = \mathcal{D}(B_n \| A_n)$, and $I(P, W_{n+1}) = \mathcal{D}(B_n \| A'_{n+1})$. Since B_n minimizes $\mathcal{D}(B \| A_n)$ over $\mathcal{B}$, and A'_{n+1} minimizes $\mathcal{D}(B_n \| A)$ over $\mathcal{A}$, it follows that

$$A_n \to B_n \to A'_{n+1}$$

are two steps of alternating projections between $\mathcal{A}$ and $\mathcal{B}$, where $\mathcal{A}$ and $\mathcal{B}$ denote the minimization sets in (41). The step $A_n \to B_n$ and the "Pythagorean" theorem for divergence [5, thm. 12.6.1] (called also "the three points property" [7, lem. 3]) imply

$$\mathcal{D}(B^* \| A_n) \geq \mathcal{D}(B^* \| B_n) + \mathcal{D}(B_n \| A_n).$$

The step $B_n \to A'_{n+1}$ and the "four points property" [7, lem. 3] imply

$$\mathcal{D}(B^* \| A'_{n+1}) \leq \mathcal{D}(B^* \| B_n) + \mathcal{D}(B^* \| A^*) \ .$$

Combining these two we have

$$R(P, Q_{\ell,n}, d) - R(P,d) \leq \mathcal{D}(B^* \| A_n) - \mathcal{D}(B^* \| A'_{n+1}) \ . \tag{43}$$

The lemma now follows by observing that the right hand side of (43) can be written explicitly as

$$\sum_{x,y} P(x) W^*(y|x) \log \frac{Q'_{\ell,n+1}(y)}{Q_{\ell,n}(y)} = \mathcal{D}(Q^* || Q_{\ell,n}) - \mathcal{D}(Q^* || Q'_{\ell,n+1})$$

since $\sum_x P(x) W^*(y|x) = Q^*(y)$. $\square$

Since ℓ is finite but large, the average favorite type $Q_{\ell,n+1}$ is roughly as close to Q^* in divergence sense as $Q'_{\ell,n+1}$:

Lemma 2 *If $Q_{\ell,n}$ is bounded away from zero, i.e., $Q_{\ell,n}(y) \geq q \ \ \forall y$, then*

$$\mathcal{D}(Q^* \| Q_{\ell,n+1}) \leq \mathcal{D}(Q^* \| Q'_{\ell,n+1}) + \epsilon(q, \ell)$$

where $\epsilon(q,\ell) \to 0$ as $\ell \to \infty$ for any $q > 0$.

Proof: Corollary 1 and Appendix F imply that $Q_{\ell,n+1} \to Q'_{\ell,n+1}$ as $\ell \to \infty$ uniformly whenever $Q_{\ell,n}$ is bounded away from zero. Lemma 6 in Appendix D then implies that $\mathcal{D}(Q^*\|Q_{\ell,n}) \to \mathcal{D}(Q^*\|Q'_{\ell,n})$ as $\ell \to \infty$ uniformly, as desired. □

To justify a lower bound on $Q_{\ell,n}$ we use the closeness of $Q_{\ell,n}$ to Q^* in the divergence sense, and the strict positiveness of Q^* assumed in the theorem:

Lemma 3 *If $\mathcal{D}(Q^*\|Q_{\ell,n}) \leq q_{min}^2/8$, then $Q_{\ell,n}(y) \geq q_{min}/2 \ \ \forall y$, where q_{min} is the bound on the sequence Q_n in Theorem 6.*

Proof: By the divergence bound on $\mathcal{L}_1$ distance [5, lem. 16.3.1] $\|Q^* - Q_{\ell,n}\|_1 \leq \sqrt{2\mathcal{D}(Q^*\|Q_{\ell,n})} \leq q_{min}/2$. The lemma now follows since $Q^*(y) \geq q_{min} \ \ \forall y$ by the assumption of the theorem. □

To further simplify notations, let $\Delta_n = R(P, Q_{\ell,n}, d) - R(P, d)$, $L_n = \mathcal{D}(Q^*||Q_{\ell,n})$ and $L'_n = \mathcal{D}(Q^*||Q'_{\ell,n})$. Note that Δ_n, L_n and L'_n are non negative quantities. Using this notation Lemma 1 becomes $\Delta_n \leq L_n - L'_{n+1}$, Lemma 2 becomes $Q_{\ell,n} \geq q \implies L_{n+1} \leq L'_{n+1} + \epsilon(q,\ell)$, and Lemma 3 becomes $L_n \leq q_{min}^2/8 \implies Q_{\ell,n} \geq q_{min}/2$. Combining the three lemmas we thus have

$$L_n \leq q_{min}^2/8 \implies \Delta_n \leq L_n - L_{n+1} + \epsilon(q_{min}/2, \ell). \tag{44}$$

Recall the definition of the function L_ϵ just before (42), and recall that $L_\epsilon \to 0$ as $\epsilon \to 0$. Let ℓ be large enough so that

$$\delta(\ell) \triangleq L_{\epsilon(q_{min}/2,\ell)} + \epsilon(q_{min}/2, \ell) \leq q_{min}^2/8, \tag{45}$$

and let $\epsilon = \epsilon(q_{min}/2, \ell)$.

Lemma 4 *If $L_n \leq L_\epsilon + \epsilon$, then $L_{n+1} \leq L_\epsilon + \epsilon$.*

Proof: Observe first that for ℓ and ϵ as above, $L_n \leq L_\epsilon + \epsilon$ implies $L_n \leq q_{min}^2/8$, which by (44) implies

$$\Delta_n \leq L_n - L_{n+1} + \epsilon. \tag{46}$$

Consider now two cases, $L_\epsilon \le L_n \le L_\epsilon + \epsilon$, and $L_n \le L_\epsilon$. In the first case, by the definition of L_ϵ, $L_n \ge L_\epsilon$ implies $\Delta_n \ge \epsilon$. By (46) this implies $L_{n+1} \le L_n$, and the lemma holds since $L_n \le L_\epsilon + \epsilon$ by assumption. In the second case, i.e., $L_n \le L_\epsilon$, the lemma holds since (46) and the non negativeness of Δ_n imply that $L_{n+1} \le L_n + \epsilon$. $\square$

Corollary 2 ("black hole") *By induction, if $L_{n'} \le L_\epsilon + \epsilon$ for some n', then $L_n \le L_\epsilon + \epsilon$ for all $n \ge n'$.*

We conclude that for δ as in (45), if the sequence $Q_{\ell,n}$ enters a δ-divergence neighborhood of Q^* (i.e., $\{Q' : \mathcal{D}(Q^*\|Q') \le \delta\}$), it remains there. Furthermore, δ can be made arbitrarily small by choosing ℓ sufficiently large.

To establish (42), it remains to show that for any valid initialization Q_0 and $\delta > 0$, the sequence $Q_{\ell,n}$ enters this δ-neighborhood for sufficiently large ℓ. This follows from two facts: (i) the sequence Q_n of (31) (corresponding to ideal BA) converges to Q^* by Theorem 5, so it must enter any $\delta/2$-neighborhood, $\delta > 0$, within a finite number of, say, $N = N(\delta)$ steps; (ii) any finite segment of the sequence $Q_{n,\ell}$ arbitrarily approaches the corresponding segment of Q_n as $\ell \to \infty$, as shown in Lemma 5 below. These two facts imply that $Q_{n,\ell}$ enters the δ-neighborhood after N iterations. Thus, (42) and Theorem 6 follow by combining Lemma 5 and Corollary 2.

Lemma 5 *$Q_{\ell,n} \to Q_n$ as $\ell \to \infty$ for any n, where Q_n is defined in (31).*

Proof: The proof is by induction. By Corollary 1, as $\ell \to \infty$

$$Q_{1,\ell} = \overline{Q}_\ell(P, Q_0, d) \to Q^*(P, Q_0, d) = Q_1,$$

which shows the induction basis. As for the induction step, suppose the statement holds for some n. In particular, since $Q_n(y) \ge q_{min}$ by the assumption of the theorem, we have $Q_{\ell,n}(y) \ge q_{min}/2$ for all sufficiently large ℓ. Then, by the uniform convergence in Corollary 1 shown in Appendix F, $Q_{\ell,n+1} = \overline{Q}_\ell(P, Q_{\ell,n}, d)$ is arbitrarily close to $Q^*(P, Q_{\ell,n}, d) = Q'_{\ell,n+1}$ for <u>all</u> sufficiently large ℓ. On the other hand, since $Q_{\ell,n} \to Q_n$ by the induction assumption, we have by the continuity in Q of $Q^*(P, Q, d)$ at $Q = Q_n$ (see Appendix A), that $Q'_{\ell,n+1} =$

$Q^*(P, Q_{\ell,n}, d)$ goes to $Q^*(P, Q_n, d) = Q_{n+1}$ as $\ell \to \infty$. Combining the two facts we obtain that $Q_{\ell,n+1} \to Q_{n+1}$ as $\ell \to \infty$, which proves the induction step. $\square$

VI Discussion: Iterations with Increasing Word Length

This paper originated from an attempt to prove that a *lossy* version of the LZ78 algorithm [26] converges to the rate-distortion bound, following prior investigation by others [14]. More specifically, we wanted to show that most of the leaves of a tree code, grown over alphabet $\mathcal{Y}$ by successive d-matches of strings from $X_1, X_2, \ldots$, will asymptotically be of type $\to Q^*$. See [22, 23] for a more detailed description of the algorithm. This convergence would imply that the coding rate goes to $R(P, d)$. So far we have not found such a proof.

As an alternative line of attack, we proposed a type selection model, where at each iteration the length of the words was incremented [22, 23]. Restating this model in terms of the parameters of the recursion in Section IV: the word length is $\ell = n$, where $n = 1, 2, \ldots$ is the iteration index, and at $n = 1$ the initial types (i.e., all symbols of $\mathcal{Y}$) have a uniform distribution, i.e., $Q_1(y) = 1/|\mathcal{Y}|$. Since ℓ increases at each iteration, the model defines a mechanism of inducing probabilities from types in $\mathcal{Q}_\ell$ to types in $\mathcal{Q}_{\ell+1}$; types of different orders are connected in a tree-like structure, and a type $Q \in \mathcal{Q}_{\ell+1}$ inherits its frequency in the codebook, $w_{\ell+1}(Q)$, from the d-matching probabilities of its parent types in the tree, i.e.,

$$w_{\ell+1}(Q) = \frac{1}{|\mathcal{Y}|} \sum_{Q' \in \mathcal{A}_\ell(Q)} \Pr\Big(Q_{N_\ell} \in T(Q')\Big)$$

where $\mathcal{A}_\ell(Q)$, the "parent types" of Q, denotes the set (of size $|\mathcal{Y}|$) of types of order ℓ from which Q descends by adding one symbol, and the probability above is measured with respect to d-matching a source $\sim P$ with a database having type spectrum $w_\ell(\cdot)$ (as defined in Appendix B).

Our proof of convergence of the type spectrum $w_\ell(Q) \to Q^*$ as $\ell \to \infty$ in the above model is cumbersome, and limited to $|\mathcal{Y}| = 2$. The convergence result in this paper, although requiring taking a double limit (on ℓ and n), is more insightful and elegant.

The phenomena of natural type selection is a "Darwinian" relative of the mechanism of "gold-washing", proposed and analyzed by Zhang and Wei [24]. The gold-washing algorithm, a *fixed rate* universal lossy source coding scheme, adapts to the source by promoting frequently used codewords to higher position in a codebook list, while randomly generating new codewords. This algorithm is shown to converge to the distortion-rate function (the inverse of $R(P, d)$), by first taking the time (i.e., the encoding/adaptation process) to infinity, and then taking the word length to infinity - paralleling the formulation of Theorem 6. The proof requires some technical conditions upon the distortion measure, and that Q^* be unique, although it is suspected that these conditions are unnecessary for convergence.

It would be interesting to further explore the relations between the (fixed distortion / fixed slope) natural type selection model and the gold-washing algorithm, as well as other iterative universal lossy compression schemes [13, 10]. Extensions to more general alphabets and distributions, and to 'sliding window" string matching with a database, are left for future work. For example, replacing the marginal distribution in (28) by the joint distribution of some sub-vector of the d-matching codeword $\mathbf{Y}_{N_\ell}$ may allow to generalize the "selection" step of the procedure above to sources and codebooks with memory.

Appendix

A. Properties and Application of $I_m(P\|Q, d)$ and $R(P, Q, d)$

The quantity $I_m(P\|Q, d)$ is a convex $\cup$ function of P, Q and d, while $R(P, Q, d)$ is a convex $\cup$ function of Q and d. This follows from the convexity properties of the mutual information and the divergence, the duality of P and Q in the definition of I_m, and by the convexity of the minimization sets $\mathcal{W}_{P,Q,d}$ and $\{W : \rho(P, W) \leq d\}$ in (6) and (12), respectively. Moreover, from the "slope" dependent form of $R(P, Q, d)$ in Proposition 1 (in Section V) it follows directly that $R(P, Q, d)$ is a *strictly* convex function of d. Similar to the rate-distortion function both quantities are non-increasing functions of d, and due to convexity they are continuous in the range of P, Q and d for which they are finite. For fixed P and Q, both quantities are zero for $d \geq d_{max} = \sum_{x,y} P(x)Q(y)\rho(x, y)$, and are finite for some interval of d's below d_{max}. By definition (10), $R(P, Q, d) \leq I_m(P\|Q, d)$, so both quantities are bounded by $\min\{\log|\mathcal{X}|, \log|\mathcal{Y}|\}$ whenever I_m is finite (a bound on mutual information). For strictly positive Q, and assuming ρ satisfies (1) (i.e., $\rho(x, y) = 0$ for at least one y for each x), $R(P,Q,d)$ is finite for all P and $d \geq 0$. The continuity of the divergence implies also the continuity of the (unique) minimizers in (10), $W^*_{P,Q,d}$ and $Q^*(P,Q,d)$, with respect to P,Q and d whenever they exist. See [18, 19] for further details.

Substituting $Q' = Q^*$ in the minimization defining $R(P,Q,d)$, we obtain the Koga-Arimoto bound $R(P,Q,d) \leq R(P,d) + \mathcal{D}(Q^*\|Q)$ [9].

For $\mathcal{X} = \mathcal{Y}$ and the Hamming distortion, $I_m(P\|Q,0)$ is infinite for all $Q \neq P$, and $I_m(P\|P,0) = H(P)$. Thus, in this case $R(P,Q,0) = H(P) + \mathcal{D}(P\|Q)$. For example, if Q is uniform over $\mathcal{X}$, $\mathcal{D}(P\|Q) = \log|\mathcal{X}| - H(P)$, and consequently $R(P,Q,0) = \log|\mathcal{X}|$, corresponding to zero compression! See [21].

We next show an implication of the simple bound $R(P,Q,d) \leq I_m(P\|Q,d)$. Recall an early result by Steinberg and Gutman [16] considering approximate string matching of a source $\sim P$ with a database also drawn from P. They show that for distortion measures $\rho : \mathcal{X}\times\mathcal{X} \to \mathcal{R}^1$ satisfying the triangle inequality, the rate of such a coding scheme is at most $R(P,d/2)$. Since the Steinberg Gutman algorithm performs non-overlapped matching, it coincides for the memoryless case with the setting of random codebook matching discussed here, so its coding rate is given by $R(P,P,d)$. Let $W^*_{d/2}$ denote a transition distribution achieving $R(P,d/2)$, and let $W^{-*}_{d/2}$ denote the backward distribution from $\mathcal{Y}$ to $\mathcal{X}$ induced by $P \circ W^*_{d/2}$. Clearly the output distribution of the channel $P \circ W^*_{d/2} \circ W^{-*}_{d/2}$ is P. Furthermore, since the distortion measure satisfies the triangle inequality, the average distortion of this channel is at most d. Thus by the definition of the minimum mutual information $I(P, W^*_{d/2} \circ W^{-*}_{d/2}) \geq I_m(P\|P,d) \geq R(P,P,d)$. On the other hand, by the data-processing inequality $I(P, W^*_{d/2} \circ W^{-*}_{d/2}) \leq I(P, W^*_{d/2}) = R(P,d/2)$, proving the Steinberg-Gutman Theorem $R(P,P,d) \leq R(P,d/2)$.

B. Codebook with Type Spectrum $w(Q)$

In [22, 23] we showed a slightly more general form of Theorem 2 regarding a codebook with "type spectrum" $w(Q)$. Assume that the codewords are drawn according to a weighting function $w(Q)$ over the types $Q \in \mathcal{Q}_\ell$, and a uniform distribution within each type, i.e., $\Pr(\mathbf{Y}_i \in T(Q)) = w(Q)/|T(Q)|$. Then, it follows from Theorem 1 that as $\ell \to \infty$ the coding rate is given by

$$R(P, w(\cdot), d) = \min_{Q'} \Big\{ I_m(P\|Q', d) - \frac{1}{\ell} \log w(Q') \Big\}.$$

This formula specializes to Theorem 1 if the codebook contains only one type (i.e., $w(Q') = 1$ if $Q' = Q$ and zero otherwise), and specializes to Theorem 2 if the codebook is generated i.i.d. $\sim Q$, i.e., $w(Q') = \exp(-\ell[\mathcal{D}(Q'\|Q)+o(1)])$. Interestingly, for a *uniform* type spectrum $w(Q') = 1/|\mathcal{Q}_\ell|$, $\forall Q'$, the coding rate coincides with the rate-distortion function $R(P, d)$, since $|\mathcal{Q}_\ell|$ is polynomial in ℓ so the term $\frac{1}{\ell} \log w(Q')$ vanishes. This fact is utilized in the universal coding scheme of [10].

C. Proof of Theorem 3

We will show that as $\ell \to \infty$

$$\frac{1}{\ell} \log(N_{s,\ell}) + s\rho(\mathbf{X}, \mathbf{Y}_{N_{s,\ell}}) \to R_s + sd_s \quad \text{ in prob.},$$

which by the strict convexity in d of $R(P, Q, d)$ implies the term-wise convergence stated in the theorem. On the one hand,

$$\limsup_{\ell\to\infty} \frac{1}{\ell} \log(N_{s,\ell}) + s\rho(\mathbf{X}, \mathbf{Y}_{N_{s,\ell}}) \le R_s + sd_s \quad \text{ in prob.},$$

follows from the result for the *fixed*-distortion version (Theorem 2), since choosing the specified distortion to be $d = d_s$ guarantees $\frac{1}{\ell} \log(N_\ell) \to R_s$. The reverse direction follows from the fact that all the codewords prior to $\mathbf{Y}_{N_{s,\ell}}$ must have higher distortion, implying the identity

$$N_{s,\ell} = N_\ell(D_{s,\ell}) \tag{47}$$

where $N_\ell(d)$ denotes the index of the first word that satisfies distortion d, and $D_{s,\ell}$ denotes the (random) distortion of the word selected by the s-slope algorithm (i.e., $D_{s,\ell} = \rho(\mathbf{X}, \mathbf{Y}_{N_{s,\ell}})$). This identity holds for any source string and codebook realization. The desired reverse inequality amounts to

$$\liminf_{\ell\to\infty} \log(N_{s,\ell})/\ell + s\rho(\mathbf{X}, \mathbf{Y}_{N_{s,\ell}}) \ge R_s + sd_s \quad \text{in probability.}$$

Now, the identity (47) implies

$$\log(N_{s,\ell})/\ell + s\rho(\mathbf{X}, \mathbf{Y}_{N_{s,\ell}}) = \log(N_\ell(D_{s,\ell}))/\ell + sD_{s,\ell}$$

which, by Theorem 2, is with probability going to one greater or equal to

$$R(P, Q, D_{s,\ell}) - \epsilon + sD_{s,\ell}$$

which, by the definition of (d_s, R_s) in (15) is greater or equal to

$$R_s + sd_s - \epsilon.$$

Since $\epsilon > 0$ is arbitrary, this proves the converse.

Note that we assumed above that Theorem 2 holds uniformly in the distortion d. Indeed, by combining (9) with the simple bound $\Pr\{N_\ell \leq n | \mathbf{X} = \mathbf{x}\} = 1 - (1-p)^n \leq np$, where $p = \Pr\{\rho(\mathbf{x}, \mathbf{Y}) \leq d\}$, we obtain as $\ell \to \infty$

$$\Pr\Big\{\frac{1}{\ell}\log(N_\ell) \leq R(P_{\mathbf{x}}, Q, d) - \epsilon \;\Big|\; \mathbf{X} = \mathbf{x}\Big\} \to 0 \tag{48}$$

for any $\epsilon > 0$, uniformly in $\mathbf{x}$, Q and d.

D. A Uniform Bound on Divergence Difference

Lemma 6 *Let P be a distribution over a finite alphabet of size α with smallest non-zero letter probability p_m. Let $B > 0$, and Q be a distribution such that $D(P\|Q) < \log(B)$. For any ϵ, $0 \leq \epsilon < c = (B\alpha)^{-1/p_m}$, if $\|Q' - Q\|_1 < \epsilon$, then*

$$\Big|D(P\|Q') - D(P\|Q)\Big| < \frac{\epsilon}{c - \epsilon}\log(e).$$

Proof: Let $\mathcal{S} = \{i : P(i) > 0\}$, so $p_m = \min_{i\in\mathcal{S}} P(i)$. Since $-\sum_i P(i)\log P(i) = H(P) \leq \log(\alpha)$, the bound $D(P\|Q) < \log(B)$ implies that $P(i)\log(1/Q(i)) < \log(B) + \log(\alpha)$, and therefore

$$Q(i) > c > 0$$

for all $i \in \mathcal{S}$. Combining with $|Q'(i) - Q(i)| < \epsilon$ we have

$$\frac{c-\epsilon}{c} \leq \frac{Q'(i)}{Q(i)} \leq \frac{c+\epsilon}{c},$$

which, by applying the inequality $1 - 1/x \leq \ln(x) \leq x - 1$, implies

$$\Big|\ln\Big(\frac{Q(i)}{Q'(i)}\Big)\Big| \leq \frac{\epsilon}{c-\epsilon},$$

for all $i \in \mathcal{S}$. The lemma follows by substituting this bound in

$$D(P\|Q') - D(P\|Q) = \sum_{i\in\mathcal{S}} P(i)\log\Big(\frac{Q(i)}{Q'(i)}\Big).$$

□

E. Conditional Limit Theorem with Varying Condition

Assume $\mathbf{X}$ and $\mathbf{Y}$ are independent i.i.d. vectors with generating distribution P and Q, respectively. Let the sequence of types $P_1, P_2, \ldots$, where $P_i \in \mathcal{P}_i$, converge to some $\widetilde{P}$, where $P_i(x) = 0$ whenever $\widetilde{P}(x) = 0$, and $\mathcal{D}(\widetilde{P}\|P) + R(\widetilde{P}, Q, d) < \infty$. For $\ell = 1, 2, \ldots$ let

$$E_\ell = \left\{ V = P' \circ W' : \; P' = P_\ell, \; \rho(P', W') \leq d \right\}$$

denote a set of joint distributions on $\mathcal{X} \times \mathcal{Y}$. We will show that for any $\delta > 0$ and sufficiently large ℓ

$$\Pr\Big(\mathcal{D}(Q_{\mathbf{Y}} \| Q^*(\widetilde{P}, Q, d)) > 3\delta \;\Big|\; V_{\mathbf{X},\mathbf{Y}} \in E_\ell \Big) \leq (\ell+1)^{2|\mathcal{X}||\mathcal{Y}|} e^{-\ell\delta}. \tag{49}$$

Thus, conditioning on the event that the joint type of $(\mathbf{X}, \mathbf{Y})$ belongs to E_ℓ, the type of $\mathbf{Y}$ is with high probability close in the divergence sense to $Q^*(\widetilde{P}, Q, d)$ as $\ell \to \infty$. Since closeness in divergence implies closeness in $\mathcal{L}_1$ sense, this establishes (23) in the proof of Theorem 4. In the next Appendix we further show that (49) holds uniformly over all Q's which are bounded away from zero.

To establish (49), first verify that

$$\mathcal{D}(P_\ell \circ W \| P \times Q) = \mathcal{D}(P_\ell \| P) + \mathcal{D}(P_\ell \circ W \| P_\ell \times Q),$$

implying by (12) that

$$\begin{aligned} \min_{V \in E_\ell} \mathcal{D}(V \| P \times Q) &= \mathcal{D}(P_\ell \| P) + R(P_\ell, Q, d) &(50)\\ &\to \mathcal{D}(\widetilde{P} \| P) + R(\widetilde{P}, Q, d) &(51) \end{aligned}$$

as $\ell \to \infty$ by the continuity of the divergence. Define

$$\mathcal{D}^* = \min_{V \in E} \mathcal{D}(V \| P \times Q) \;\; = \;\; \mathcal{D}(\widetilde{P} \| P) + R(\widetilde{P}, Q, d),$$

where $E = \{V = P' \circ W' : \; P' = \widetilde{P}, \; \rho(P', W') \leq d\}$, and the second equality follows from (12). Recall that $\mathcal{D}^*$ is finite by assumption. Following [5, eq. (12.158)]

$$\Pr\Big(\mathcal{D}(V_{\mathbf{X},\mathbf{Y}} \| P \times Q) > \mathcal{D}^* + 3\delta \Big) \leq (\ell+1)^{|\mathcal{X}||\mathcal{Y}|} e^{-\ell[\mathcal{D}^* + 3\delta]}. \tag{52}$$

On the other hand, following [5, eq. (12.162)], for sufficiently large ℓ

$$\Pr\Big(\mathcal{D}(V_{\mathbf{X},\mathbf{Y}} \| P \times Q) \leq \mathcal{D}^* + 2\delta, \;\; V_{\mathbf{X},\mathbf{Y}} \in E_\ell \Big) \geq \frac{1}{(\ell+1)^{|\mathcal{X}||\mathcal{Y}|}} e^{-\ell[\mathcal{D}^* + 2\delta]}, \tag{53}$$

since for sufficiently large ℓ: (i) by (50) there is at least one ℓ-type V' in E_ℓ such that

$$\mathcal{D}(V'\|P \times Q) \leq \mathcal{D}(P_\ell\|P) + R(P_\ell, Q, d) + \delta \ ; \tag{54}$$

and (ii) by (51)

$$\mathcal{D}(P_\ell\|P) + R(P_\ell, Q, d) \leq \mathcal{D}^* + \delta. \tag{55}$$

It further follows from (53) that

$$\Pr\Big(V_{\mathbf{X},\mathbf{Y}} \in E_\ell\Big) \geq \frac{1}{(\ell+1)^{|\mathcal{X}||\mathcal{Y}|}} e^{-\ell[\mathcal{D}^*+2\delta]},$$

since the probability of the intersection is less than or equal to the probability of each event, implying by Bayes' law and (52) that

$$\Pr\Big(\mathcal{D}(V_{\mathbf{X},\mathbf{Y}}\|P \times Q) > \mathcal{D}^* + 3\delta \ \Big|\ V_{\mathbf{X},\mathbf{Y}} \in E_\ell\Big) \leq (\ell+1)^{2|\mathcal{X}||\mathcal{Y}|} e^{-\ell\delta}. \tag{56}$$

Now, define V^* as the (unique) minimizer $V^* = \arg\min_{V \in E} \mathcal{D}(V\|P \times Q)$. By the "Pythagorean" theorem $\mathcal{D}(V_{\mathbf{X},\mathbf{Y}}\|P \times Q) \leq \mathcal{D}^* + 3\delta$ implies $\mathcal{D}(V_{\mathbf{X},\mathbf{Y}}\|V^*) \leq 3\delta$ (see [5, eq. (12.168)]). Moreover, by the divergence data processing inequality (which follows from chain rule for divergence [5, thm. 2.5.3]), $\mathcal{D}(Q_{\mathbf{Y}}\|Q^*_{\widetilde{P},Q,d}) \leq \mathcal{D}(V_{\mathbf{X},\mathbf{Y}}\|V^*)$, since $Q_{\mathbf{Y}}$ and $Q^*_{\widetilde{P},Q,d}$ are the y-marginals of $V_{\mathbf{X},\mathbf{Y}}$ and V^*, respectively. Hence (49) follows from (56) as desired.

F. Uniform Convergence in Theorem 4 and Corollary 1

We now examine the conditions for uniform convergence in (49), which implies uniform convergence in Theorem 4 and in Corollary 1. We will show uniform convergence over all codebook generation distributions which are bounded away from zero, i.e.,

$$Q(y) \geq q_{min} > 0 \ \ \forall y.$$

We have two inequalities, (52) and (53), that together establish (49). The former is clearly uniform in P, Q and d. The latter requires ℓ to be "sufficiently large" for (54) and (55) to hold, and this may depend on Q. We first examine (54). Observe that it is possible to approximate a joint distribution $V = P' \circ W$ with an ℓ-type $V' = P' \circ W'$ such that $|V'(x,y) - V(x,y)| < 1/\ell$ for all (x,y), and the average distortion does not increase, i.e., $\rho(P', W') \leq \rho(P', W)$. For $d > 0$ we can obtain V' by rounding *down* all components of V

to the nearest multiple of $1/\ell$, and then, for each x, successively rounding *up* components $\{V(x,y),\ y \in \mathcal{Y}\}$ associated with the *lowest* values of $\rho(x,y)$, $y \in \mathcal{Y}$, until the desired constraint $\sum_{y\in\mathcal{Y}} V'(x,y) = P'(x)$ is satisfied. For $d = 0$ V must be a type and rounding is unnecessary, so we will restrict attention to positive d's. Now suppose that the following partial derivatives are uniformly bounded:

$$\left\| \frac{\partial \mathcal{D}(V\|P \times Q)}{\partial V} \right\| < B_1. \tag{57}$$

Then, the ℓ-type V' above satisfies (54) for all $\ell > B_1/\delta$. As for (55), we can write $R(P_\ell, Q, d) - R(\widetilde{P}, Q, d) \leq \|P_\ell - \widetilde{P}\| B_3$ provided that the partial derivatives of $R(P,Q,d)$ with respect to nonzero components of P are uniformly bounded for all Q

$$\left\| \frac{\partial R(P,Q,d)}{\partial P} \right\| < B_2. \tag{58}$$

(Note that $\widetilde{P}(x) = 0$ implies $P_\ell(x) = 0$.) Hence (54) and (55), and consequently (53) hold uniformly whenever (57) and (58) hold.

The derivative in (57), taken with respect to some component $V(x,y)$, is given by

$$\frac{\partial}{\partial V(x,y)} \sum_{x,y} V(x,y) \log\Big(\frac{V(x,y)}{P(x)Q(y)}\Big) = 1 + \log\Big(\frac{V(x,y)}{P(x)Q(y)}\Big) \tag{59}$$

For all V of interest the x-marginal is zero if $P(x) = 0$. Thus the argument of the logarithm above is upper bounded by $1/(p_{min} q_{min})$, where p_{min} is the lowest nonzero component of P. To lower bound the logarithm argument, we note from (38) that for an optimal transition distribution $W^*_{\widetilde{P},Q,d}$ the ratio $W(y|x)/Q(y)$ is lower bounded by $\exp(-s\rho_{max})$ where s is the slope $\partial R(\widetilde{P}, Q, d)/\partial d$. It is easy to verify that under condition (1),

$$R(\widetilde{P}, Q, 0) \leq H(\widetilde{P}) + \mathcal{D}(\widetilde{P}\|Q) \leq \log|\mathcal{X}| + \log(1/q_{min})$$

(set $Q' = \widetilde{P}$ in (10)). Hence by the convexity of $R(P,Q,d)$ in d (see Appendix A) the slope s is upper bounded by $s_{max} = \log(|\mathcal{X}|/q_{min})/d$. Thus, the ratio $V(x,y)/P(x)Q(y)$ at $V(x,y) = \widetilde{P}(x) W^*_{\widetilde{P},Q,d}(y|x)$ is lower bounded by $\widetilde{P}(x)\exp(-s_{max}\rho_{max})$. Since we are only interested in the derivative at the $1/\ell$-neighborhood of optimal points $V(x,y) = \widetilde{P}(x) W^*_{\widetilde{P},Q,d}(y|x)$, and only for non-zero $\widetilde{P}(x)$ (since if $\widetilde{P}(x) = 0$ rounding of $V(x,y)$ is not needed), we see that

$$\log\Big(\frac{1}{p_{min} q_{min}}\Big) \geq \frac{\partial \mathcal{D}(V\|P \times Q)}{\partial V(x,y)} \geq \log\Big(\tilde{p}_{min} \exp(-s_{max}\rho_{max}) - 1/(\ell p_{min} q_{min})\Big)$$

where the righthand side is valid for $\ell > \exp(s_{max}\rho_{max})/(\tilde{p}_{min}p_{min}q_{min})$, and $\tilde{p}_{min}$ is the smallest nonzero component of $\widetilde{P}$. Thus the derivative (57) is uniformly absolutely upper bounded for all Q's which are bounded away from zero.

We turn to the derivative in (58). In view of (12), we want to show that $\frac{d}{dP}\min_{W:\rho(P,W)\leq d}\mathcal{D}(P\circ W\|P\times Q)$ is absolutely bounded for all Q which are bounded away from zero. This follows easily from the fact, shown below, that the partial derivative $\frac{\partial}{\partial P}\mathcal{D}(P\circ W\|P\times Q)$ is absolutely bounded for all P and W, if $Q(y)\geq q_{min}\ \forall y$. Specifically, the partial derivative with respect to some component $P(x)$, is given by

$$\frac{\partial}{\partial P(x)}\sum_y P(x)W(y|x)\log\Big(\frac{W(y|x)}{Q(y)}\Big) = \sum_y W(y|x)\log W(y|x) + \sum_y W(y|x)\log(1/Q(y)) \tag{60}$$

which is absolutely upper bounded by $\log|\mathcal{Y}| + \log(1/q_{min})$.

Hence we have shown that (57) and (58) hold uniformly for all Q's which are bounded away from zero, and the assertion follows.

Acknowledgment

The authors are grateful to Meir Feder, Yiannis Kontoyiannis, Tamas Linder and En-Hui Yang for enlightening discussions which contributed much to this work. We also thank Phil Chou and the anonymous reviewers for useful comments and suggestions.

References

[1] S. Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Trans. Information Theory*, 18:14–20, Jan. 1972.

[2] T. Berger. *Rate Distortion Theory: A Mathematical Basis for Data Compression.* Prentice-Hall, Englewood Cliffs, NJ, 1971.

[3] R. E. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Trans. Information Theory*, IT-18:460–473, 1972.

[4] P. Boukris. An upper bound on the speed of convergence of the Blahut algorithm for computing rate-distortion functions. *IEEE Trans. Information Theory*, pages 708–709, Sept. 1973.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, New York, 1991.

[6] I. Csiszár. On the computation of rate-distortion functions. *IEEE Trans. Information Theory*, pages 122–124, 1974.

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

[8] A. Dembo and Y. Kontoyiannis. The asymptotics of waiting times between stationary processes, allowing distortion. *Ann. Appl. Prob.*, 9:413–429, 1999.

[9] H. Koga and S. Arimoto. Asymptotic properties of algorithms for data compression with fidelity criterion based upon string matching. In *Proc. of ISIT, Trondheim, Norway*, page 264, June 1994.

[10] Y. Kontoyiannis. An implementable lossy version of the Lempel-Ziv algorithm - part I: Optimality for memoryless sources. *IEEE Trans. Information Theory*, 45:2293–2305, 1999.

[11] A. Lapidoth. On the role of missmatch in rate-distortion theory. *IEEE Trans. Information Theory*, 43:38–47, Jan. 1997.

[12] T. Luczak and W. Szpankowski. A suboptimal lossy data compression based on approximate pattern matching. *IEEE Trans. Information Theory*, 43:1439–1451, Sept. 1997.

[13] D. Manor and M. Feder. An iterative technique for universal lossy compression of individual sequences. In *Proc. of the Data Comp. Conf.*, Snowbird, Utah, March 1997.

[14] Morita and Kobayashi. An extenssion of LZW coding algorithm to source coding subject to a fideleity criterion. In *In Proc. of the 4th Joint Swedish-Soviet Int. Conf. on Inf. Th.*, pages 105–109, June 1992.

[15] S. Shamai and S. Verdú. The empirical distribution of good codes. *IEEE Trans. Information Theory*, pages 836–846, May 1997.

[16] Y. Steinberg and M. Gutman. An algorithm for source coding subject to a fidelity criterion, based on string matching. *IEEE Trans. Information Theory*, IT-39:877–886, May 1993.

[17] A.D. Wyner and J. Ziv. Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Trans. Information Theory*, IT-35:1250–1258, Nov. 1989.

[18] E.H. Yang and J. Kieffer. On the performance of data compression algorithms based upon string matching. *IEEE Trans. Information Theory*, IT-44:47–65, Jan. 1998.

[19] E.H. Yang and Z. Zhang. On the redundancy of lossy source coding with abstracts alphabets. *IEEE Trans. Information Theory*, IT-45:1092–1110, May 99.

[20] E.H. Yang, Z. Zhang, and T. Berger. Fixed slope universal lossy data compression. *IEEE Trans. Information Theory*, pages 1465–1476, Sept. 1997.

[21] R. Zamir. The index entropy in mismatched lossy source coding. In *Proc. of ISIT 2000, Sorrento, Italy*, page 124, June 2000.

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

[23] R. Zamir and K. Rose. A type generation model for adaptive lossy compression. In *Proc. of ISIT97*, page 186, Ulm, Germany, June 1997.

[24] Z. Zhang and V. Wei. An on-line universal lossy data compression algorithm via continuous codebook refinement-part I: Basic results. *IEEE Trans. Information Theory*, IT-42:803–821, May 1996.

[25] Z. Zhang, E.H. Yang, and V. Wei. The redundancy of source coding with a fidelity criterion - Part one: Known statistics. *IEEE Trans. Information Theory*, IT-43:71–91, Jan. 1997.

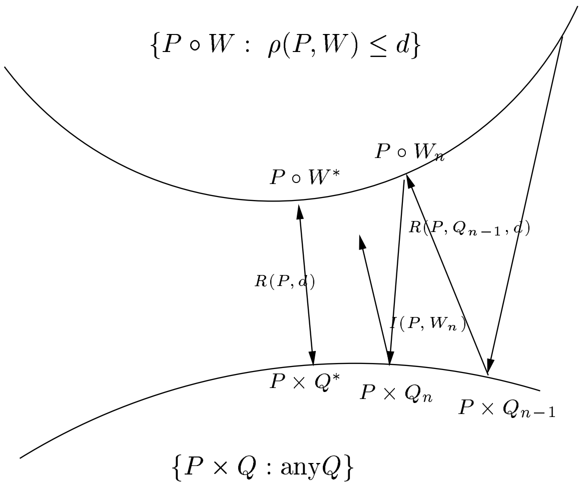
[26] J. Ziv and A. Lempel. Compression of individual sequences via variable rate coding. *IEEE Trans. Information Theory*, IT-24:530–536, Sept. 1978.

List of Figures

List of Footnotes

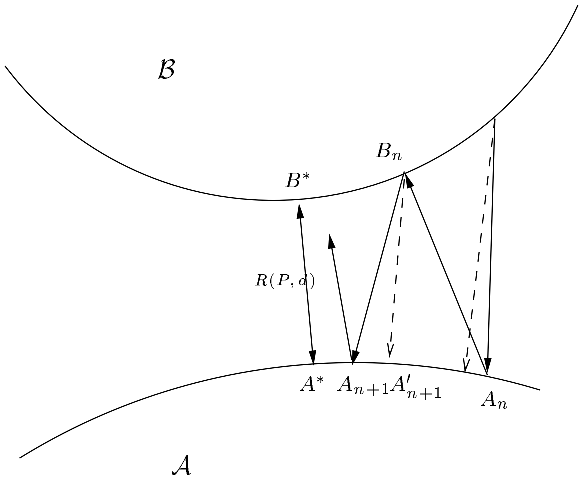
Footnote 1: An interesting alternative form of $R(P, Q, d)$ appeared in [18]. It states $R(P, Q, d) = \min \mathcal{D}(U; Y|X)$, where the minimization is over all U jointly distributed with $X \sim P$ such that $E\rho(X, U) \leq d$, and where $\mathcal{D}(U; Y|X)$ denotes the conditional divergence between the random variables U and Y given X, where $Y \sim Q$ is independent of X.

FIGURE 1



$\{P \circ W : \ \rho(P,W) \le d\}$
$P \circ W_n$
$P \circ W^*$
$R(P, Q_{n-1}, d)$
$R(P, d)$
$I(P, W_n)$
$P \times Q^*$
$P \times Q_n$
$P \times Q_{n-1}$
$\{P \times Q : \text{any} Q\}$

a



$\mathcal{B}$
B_n
B^*
$R(P, d)$
A^*
A_{n+1}
A'_{n+1}
A_n
$\mathcal{A}$

b

Ram Zamir: a Biography

Ram Zamir was born in Ramat-Gan, Israel in 1961. He received the B.Sc., M.Sc. (summa cum laude) and D.Sc. (with distinction) degrees from Tel-Aviv University, Israel, in 1983, 1991, and 1994, respectively, all in electrical engineering. In the years 1994 - 1996 he spent a post-doctoral period at Cornell University, Ithaca, NY, and at the University of California, Santa Barbara. Since 1996 he has been with the department of Elect. Eng. - Systems at Tel Aviv University. Dr. Zamir has been consulting in the areas of radar and communications, mainly in developing algorithms and in the design of signals, and has been teaching information theory, communications systems and communications circuits at Tel Aviv University.

Dr. Zamir received the Israel Ministry of Communications Award in 1993, and the Wolfson Post-Doctoral Research Award in 1994, and visited the Technical University of Budapest under a joint program of the Israeli and Hungarian academies of science in summer 1995. His research interests include information theory, communication and remote sensing systems, and signal processing.

Ken Rose: a Biography

Kenneth Rose received the B.Sc. (summa cum laude) and M.Sc. (magna cum laude) degrees in electrical engineering from Tel-Aviv University, Israel, in 1983 and 1987, respectively, and the Ph.D. degree in electrical engineering from the California Institute of Technology, in 1990.

From July 1983 to July 1988 he was employed by Tadiran Ltd, Israel, where he carried out research in the areas of image coding, image transmission through noisy channels, and general image processing. From September 1988 to December 1990 he was a graduate student at Caltech. In January 1991 he joined the Department of Electrical and Computer Engineering, University of California at Santa Barbara, where he is currently an associate professor. His research interests are in information theory, source and channel coding, image coding and processing, speech and general pattern recognition, and nonconvex optimization in general.

Dr. Rose is currently Editor for Source/Channel Coding for the IEEE Transactions on Communications. He was co-recipient of the William R. Bennett Prize Paper Award of the IEEE Communications Society (1990).