A String Matching Interpretation for the Arimoto-Blahut Algorithm

Ram Zamir
Dept. of EE-Systems
Tel Aviv University
Israel

Kenneth Rose
Dept. of ECE
University of California at Santa Barbara
USA

Abstract — The Arimoto-Blahut (AB) algorithm is an iterative procedure for computing the rate-distortion function for a given source and distortion measure. It starts with an arbitrary (strictly positive) reproduction distribution, and iteratively produces a sequence of reproduction distributions (and corresponding rates) that converges to the point of prespecified slope on the rate-distortion curve, and to its associated reproduction distribution. We connect the AB algorithm with a mechanism of "natural type selection" in the random generation of codebooks for approximate (lossy) string matching. Consider a procedure that iterates two steps: (i) given a random codebook find the type of codewords most likely to match a source string at distortion D, and (ii) randomly regenerate the codebook using this type as reproduction distribution. We apply Csiszar and Tusnady's alternating minimization arguments to show that this procedure converges to the rate-distortion function and is, in fact, a string-matching relative of the Arimoto-Blahut algorithm.

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 [11] and, more recently, the Gold-Washing algorithm for lossy coding [10].

The lossless case was thoroughly investigated over the past two decades, and is now relatively well understood [11, 6]. LZ78 grows a tree, while observing the incoming source strings, in a way that ensures that, asymptotically, most codewords are typical sequences for the source. In Wyner and Ziv's "database" interpretation of string matching [6] (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 codewords. The situation in lossy coding, however, is fundamentally different, as good codewords 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 codewords 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 codewords 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 ℓ-string $\mathbf{X}$, independently drawn from the source, is in the codebook goes to 1 as ℓ goes to infinity.

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 word $\mathbf{x}$ by codeword $\mathbf{y}$ is given by $\rho(\mathbf{x},\mathbf{y}) = \frac{1}{\ell}\sum_{i=1}^{\ell}\rho(x_i,y_i)$. (Throughout this paper $\mathcal{X}$, $\mathcal{Y}$ and ρ are assumed finite.) We say that $\mathbf{y}$ "d-matches" $\mathbf{x}$, or that a "d-match" event occurred if $\rho(\mathbf{x},\mathbf{y}) \le d$. The rate-distortion function of the source is defined as

$$R(P,d) = \min_{W:\ \rho(P,W)\le d} I(P,W) \tag{1}$$

where $W = \{W(y|x)\}$ denotes a transition distribution from $\mathcal{X}$ to $\mathcal{Y}$, $I(P,W)$ 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 $P \circ W \triangleq \{P(x)W(y|x)\}$. Let W^* be a transition distribution that achieves the minimum in (1), i.e., $R(P,d) = I(P,W^*)$. Further, by marginalizing the optimal joint distribution, $P \circ W^*$, we obtain the *optimal reproduction distribution*:

$$Q^* = Q^*(P,d) = \Big\{\sum_x P(x)W^*(y|x)\Big\}. \qquad (2)$$

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

Suppose, however, that a different distribution $Q \neq Q^*$ was used to generate the codebook. We denote the minimum coding rate in this case $R(P,Q,d)$. To guarantee a d-match with probability approaching 1, we need to draw $\exp[\ell(R(P,Q,d)+\epsilon)]$ codewords, where $R(P,Q,d) > R(P,Q^*,d) \equiv R(P,d)$, [5, 4, 8, 7].

It is thus clear that in order to achieve optimum lossy compression, the codewords' type must approach Q^*. In the case of unknown source statistics, it is desirable to develop an adaptive mechanism which naturally selects codewords whose type approaches Q^*. The question of whether such a mechanism exists motivated this work.

Our main result shows the existence of a natural type selection mechanism for lossy string matching. More specifically, for large ℓ, almost all codewords in a codebook generated $\sim Q$ that d-match source strings $\sim P$ are typical with type $Q^*(P,Q,d)$. We define 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 $R(P,Q_{n+1},d) \leq R(P,Q_n,d)$. It is shown that, in fact,

$$Q_n \to Q^* \quad \text{and} \quad R(P,Q_n,d) \to R(P,d) \qquad (3)$$

as $n \to \infty$. Interestingly, this procedure simulates the iterations of a fixed-distortion variant of the Blahut algorithm, or Arimoto-Blahut algorithm (BA) [2], as it iteratively computes the rate-distortion function starting with Q_0 as reproduction distribution. The convergence of Q_n and $R(P,Q_n,d)$ in (3) follows from the work of Csiszar and Tusnady [3] on alternating minimization in the divergence sense. We also provide "fixed-slope" versions of this behavior, corresponding to coding with minimum weighted rate-distortion.

The extended summary below brings the main results and Theorems; a complete discussion and proofs appear in [9].

II Approximate String Matching with a Random Codebook

This section formulates the problem of lossy source coding with a random codebook of arbitrary (typically "mismatched") generating distribution, and reviews results concerning the expected encoding performance [7, 4, 10].

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. Let $\mathbf{x}$ denote a source word of ℓ letters from $\mathcal{X}$ having type (empirical distribution) $P_{\mathbf{x}}$.

Consider first the conditional probability that codeword $\mathbf{Y}_i$ d-matches string $\mathbf{x}$, given that $\mathbf{Y}_i$ is of type $Q' \in \mathcal{Q}_\ell$. We follow the convention 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'. Define the *"minimum mutual information with constrained output Q"* as

$$I_m(P\|Q,d) = \min_{W \in \mathcal{W}_{P,Q,d}} I(P,W), \qquad (4)$$

if $\mathcal{W}_{P,Q,d}$ is not empty, and ∞ otherwise, 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: $\sum_x P(x)W(y|x) = Q(y)$, and $\rho(P,W) \leq d$. It follows from the definition (4) that, given P and Q, there is a minimum distortion $\tilde{d}_{min}(P,Q)$ below which $I_m(P\|Q,d)$ is infinite.

Theorem 1 (Conditional d-Match Exponent) *As $\ell \to \infty$, for any type $Q' \in \mathcal{Q}_\ell$, and $d > \tilde{d}_{min}(P_{\mathbf{x}}, Q')$,*

$$-\frac{1}{\ell}\log\Big(\Pr\Big\{\rho(\mathbf{x},\mathbf{Y}_i) \leq d \;\Big|\; \mathbf{Y}_i \in T(Q')\Big\}\Big) = I_m(P_{\mathbf{x}}\|Q',d) + o(1) \qquad (5)$$

where $o(1) \to 0$.

Let $N_\ell = N_\ell(P,Q,d)$ denote the index of the first codeword in the codebook that d-matches an independent random source string $\mathbf{X}$ generated $\sim P$, i.e., $\rho(\mathbf{X},\mathbf{Y}_i) > d$ for $1 \leq i \leq N_\ell - 1$ and $\rho(\mathbf{X},\mathbf{Y}_{N_\ell}) \leq d$.

Theorem 2 (Mismatched Coding Rate) *Assume $Q(y) > 0$ $\forall y \in \mathcal{Y}$. As $\ell \to \infty$,*

- $\log(N_l)/l \to R(P,Q,d)$ *in probability,*

where

$$\begin{aligned} R(P,Q,d) &= \min_{Q'}\Big\{I_m(P\|Q',d) + \mathcal{D}(Q'\|Q)\Big\} && (6)\\ &= \min_{W:\, \rho(P,W)\leq d} \mathcal{D}(P\circ W\|P\times Q)\,. && (7)\end{aligned}$$

Thus, $R(P,Q,d)$ characterizes the asymptotic coding rate of this "mismatched" codebook.

II-1 Matching to Minimize the Rate-Distortion Lagrangian

The following "d-matching" variant corresponds to lossy source coding with a fixed slope. Fix the rate-distortion Lagrange parameter $s > 0$. Instead of looking 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 $N_{s,\ell} = N_{s,\ell}(P, Q, d)$ denote the index of this codeword.

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

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

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

$$d_s = d_s(P, Q) \stackrel{\Delta}{=} \arg\min_d \Big\{ R(P, Q, d) + sd \Big\}. \qquad (8)$$

If the above argmin is unique, then $\frac{1}{\ell}\log(N_{s,\ell}) \to R_s$ *and* $\rho(\mathbf{X}, \mathbf{Y}_{N_{s,\ell}}) \to d_s$ *in probability.*

III The Typical d-Matching Codeword

The form of $R(P, Q, d)$ in (6) gives rise to a useful interpretation of the d-matching mechanism. Assume $Q \ne Q^*$, where $Q^* = Q^*(P, d)$ is the optimum reproduction distribution (2), and let

$$Q^*(P, Q, d) \stackrel{\Delta}{=} \arg\min_{Q'} \Big\{ I_m(P \| Q', d) + \mathcal{D}(Q' \| Q) \Big\} \qquad (9)$$

denote a reproduction distribution which achieves $R(P, Q, d)$ (6).

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" in $\mathcal{X}^\ell$ covered by a code word 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 $Q' \ne Q^*$. But these codewords are very rare. Their frequency in the codebook is only $\exp(-\ell\mathcal{D}(Q^*\|Q))$. The types near $Q^*(P, Q, d)$ minimize the exponent sum in (9), 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 this optimal type.

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 $Q^*(P, Q, d)$, i.e.,

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

provided that $Q^*(P, Q, d)$ is unique, where

$$Q^*_\ell(P, Q, d) = \arg\max_{Q'} \Pr\Big(\mathbf{Y}_{N_\ell} \in T(Q')\Big)$$

denotes the most likely type of the d-matching codeword $\mathbf{Y}_{N_\ell}$.

In fact, we obtain a stronger result. Let Q_{N_ℓ} denote the (random) type of Y_{N_ℓ}. Theorem 4 below states that the distribution of Q_{N_ℓ} collapses asymptotically on $Q^*(P, Q, d)$.

Theorem 4 (The Selected Type) *As* $\ell \to \infty$*,*

$$I_m(P\|Q_{N_\ell}) + \mathcal{D}(Q_{N_\ell}\|Q) \to R(P, Q, d) \qquad \textit{in prob.}$$

Moreover, if $Q^*(P, Q, d)$ *is unique, then*

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

A similar result is obtained for string matching to minimize the rate-distortion Lagrangian.

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

As explained in the previous section, the typical d-matching type in the codebook, $Q^*_\ell(P, Q, d)$, 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 initial generating distribution Q_0, at each iteration the "favorite" type in the codebook is identified and generates a new codebook. This recursion results in a sequence of types in $\mathcal{Q}_\ell$

$$Q^{(\ell)}_n = Q^*_\ell(P, Q^{(\ell)}_{n-1}, d) \qquad n = 1, 2, \ldots. \qquad (10)$$

If $Q^*_\ell(P, Q^{(\ell)}_{n-1}, d)$ is not unique, then arbitrarily pick any member of the optimum set as $Q^{(\ell)}_n$.

As the word length ℓ goes to infinity, the sequence $Q^{(\ell)}_n$ approaches the sequence of distributions defined by the recursion

$$Q_n = Q^*(P, Q_{n-1}, d) \qquad n = 1, 2, \ldots \qquad (11)$$

where $Q^*(P, Q, d)$ is defined in (9). Note that Q_n achieves $R(P, Q_{n-1}, d)$ in the minimization of (6).

Recursion (11) corresponds to matching with fixed distortion. For the fixed slope variant of the coding scheme we have, as $\ell \to \infty$, the recursion

$$Q_n = Q^*_s(P, Q_{n-1}) \qquad n = 1, 2, \ldots \qquad (12)$$

where Q_s^* is defined as $Q_s^*(P,Q) = Q^*(P,Q,d_s)$. In this recursion both the rate and the distortion vary with n according to

$$R_n = R(P, Q_n, d_n) \quad \text{and} \quad d_n = d_s(P, Q_n) \qquad (13)$$

where d_s is defined in (8).

As shown in the full paper, 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 (12), (13), coincide with the sequence of tentative reproduction distributions and associated rates and distortions, respectively, in the iterations of the Blahut-Arimoto (BA) algorithm [1, 2] as it computes the rate-distortion function of the source starting from an initial distribution Q_0. Recursion (11), which corresponds to iterations with *fixed distortion* d, can be viewed as a stochastic simulation of the alternating minimization procedure of Csiszar and Tusnady [3] 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. These dualities imply the next theorem.

Theorem 5 (Natural Type Selection) *For any initial distribution Q_0 which is positive everywhere in $\mathcal{Y}$, the fixed-distortion recursion (11) satisfies as $n \to \infty$*

$$\begin{aligned} R(P, Q_n, d) &\to R(P, d) \\ Q_n &\to Q^* \end{aligned} \qquad (14)$$

where $Q^* = Q^*(P,d)$ is a reproduction distribution that achieves the rate-distortion bound $R(P,d)$. The sequence $R(P, Q_n, d)$ is monotonically non-increasing.

The fixed-slope recursion (12), (13) converges to the point of slope $-s$ on the rate-distortion function and to a corresponding optimal reproduction distribution, i.e., $R_n \to R(P, d_s^*)$, $d_n \to d_s^*$, and $Q_n \to Q^*$, where $d_s^* \triangleq \arg\min_d\{R(P,d) + sd\}$ and $Q^* \triangleq Q^*(P, d_s^*)$.

An interesting consequence of Theorem 5 is that the optimum reproduction distribution Q^* is the only "stable" codebook distribution.

The convergence in the first part of Theorem 5 follows from the Csiszar-Tusnady Theorem [3] for general alternating minimization procedures. By (7) and (1), we can write the rate-distortion function as the double minimization

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

It is easy to verify that the sets of joint distributions $\{P \times Q : \text{ any } Q\}$ and $\{P \circ W : \rho(P,W) \le d\}$ are convex. Furthermore, 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 (11) defines a sequence, $(P \times Q_{n-1}) \to (P \circ W_n) \to (P \times Q_n) \to \ldots$, of alternating minimizations between the above convex sets, where distance is measured by the divergence. By [3, thm. 3] the sequences of divergences and distributions converge to the minimum divergence, i.e., to $R(P,d)$, and to a reproduction distribution Q^* that achieves it, respectively.

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

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

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

[5] 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.

[6] 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.

[7] En-Hui Yang and John Kieffer. On the performance of data compression algorithms based upon string matching. *IEEE Trans. Information Theory*, IT-44:47–65, January 1998.

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

[9] R. Zamir and K. Rose. Natural type selection in approximate string matching. *IEEE Trans. Information Theory*, submitted Nov. 1998.

[10] 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.

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