The Index Entropy of a Mismatched Codebook*

Ram Zamir

Dept. of Electrical Engineering - System
Tel-Aviv University, Tel-Aviv 69978, Israel
zamir@eng.tau.ac.il

Accepted for publication in the IEEE tr. on Info. Theory

August 27, 2001

Abstract

Entropy coding is a well known technique to reduce the rate of a quantizer. It plays a particularly important role in universal quantization, where the quantizer codebook is usually not matched to the source statistics. We investigate the gain due to entropy coding by considering the entropy of the index of the first codeword, in a mismatched random codebook, that D-matches the source word. We show that the index entropy is strictly lower than the "un-coded" rate of the code, provided that the entropy is conditioned on the codebook. The number of bits saved by conditional entropy coding is equal to the divergence between the "favorite type" (the limiting empirical distribution of the first D-matching codeword) and the codebook generating distribution. Specific examples are provided.

Key Words: mismatched source coding, favorite type, entropy coded quantization, universal quantization.

*This work was presented in part in the International Symposium on Information Theory, June 2000, Sorrento, Italy. This work was supported in part by the United States - Israel Binational Science Foundation grant 1998-309.

I. Introduction

Entropy coding is an efficient method for enhancing quantizer performance [6, 2]. This paper investigates the role of entropy coding when the quantizer codebook is *mismatched* with respect to the source distribution.

The setting of optimum entropy coding of a mismatched codebook is typical of universal quantization [16]. Although our main motivation originates from structured (e.g., lattice) codes and real valued processes, we confine our discussion to random codes and to discrete memoryless sources. Our results are based on recent work by Yang and Kieffer [11] and by Zamir and Rose [13] in the area of mismatched codes.

Specifically, we investigate the entropy rate of the index of the first codeword in $\{\mathbf{Y}_1, \mathbf{Y}_2, \ldots\}$ which satisfies $d(\mathbf{X}, \mathbf{Y}_n) \leq D$ ("D-match"), where $\mathbf{X}$ is a source l-word generated i.i.d $\sim P$, and $\{\mathbf{Y}_1, \mathbf{Y}_2, \ldots\}$, the codebook, is an infinite list of random code l-words drawn i.i.d. $\sim Q$. Let N_l denote the index n of the first D-matching code word. Our main result characterizes the entropy rate of N_l in terms of single letter information quantities. It implies, for example, that the difference between the index entropy and the conditional index entropy (conditional on the codebook) is given by

$$\lim_{l\to\infty} \frac{1}{l} H(N_l) - \lim_{l\to\infty} \frac{1}{l} H(N_l | \mathbf{Y}_1, \mathbf{Y}_2, \ldots) = D(Q^*_{P,Q,D} \| Q)$$

where $D(\cdot)$ denotes divergence, and the distribution $Q^*_{P,Q,D}$, called the "favorite type", is the limiting empirical distribution of the first D-matching code word [13].

Section II defines the setting, introduces some useful information quantities, and gives the main result. Section III illustrates this result by an explicit example, and Section IV gives the proof. In a future work with I. Kontoyiannis [8] we extend these concepts to general alphabets and source distributions, and demonstrate their tight relation to dithered lattice (universal) quantization [16, 12].

II. Preliminaries and Main Result

Consider coding a source string $\mathbf{X} = X_1...X_l$ from a finite alphabet $\mathcal{X}$, into a code word $\mathbf{y} = y_1...y_l$ from a finite alphabet $\mathcal{Y}$, under the distortion constraint

$$d(\mathbf{X}, \mathbf{y}) \triangleq 1/l \sum_{i=1}^{l} d(X_i, y_i) \leq D \tag{1}$$

where $d : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ is a finite distortion measure. If (1) holds we say that "$\mathbf{y}$ D-matches $\mathbf{X}$". We shall assume that $\mathbf{X}$ is generated by a memoryless source having a distribution $P = \{P(x),\ x \in \mathcal{X}\}$.

Suppose a random codebook $\mathbf{Y}_1, \mathbf{Y}_2, ...$ of words in $\mathcal{Y}^l$ is generated such that each letter in each code word is independent and identically distributed as $Q = \{Q(y),\ y \in \mathcal{Y}\}$. Let N_ℓ denote the index of the first codeword that satisfies (1), i.e.,

$$d(\mathbf{X}, \mathbf{Y}_i) > D \quad i = 1, ..., N_l - 1, \qquad d(\mathbf{X}, \mathbf{Y}_{N_l}) \leq D.$$

To avoid technical subtleties we assume that the distortion measure is such that every source letter has a perfect reconstruction letter, i.e., for each x

$$d(x, y) = 0 \qquad \text{for some } y. \tag{2}$$

We also assume that $Q(y) > 0$ for all y in $\mathcal{Y}$. It follows that for any source string $\mathbf{x}$ and $D \geq 0$, there is a positive probability $p_{match} > 0$ that each codeword $\mathbf{Y}_i$ will D-match $\mathbf{x}$. As a consequence

$$\Pr\{N_l < \infty \mid \mathbf{X} = \mathbf{x}\} = 1,$$

i.e., a D-match is found in the codebook with probability one. See, e.g., [10, 15, 9, 11, 5] and the references therein for various settings of lossy source coding and the related topic of approximate string matching.

In [11], Yang and Kieffer show that $1/l \log(N_l)$ converges to a constant,

$$\frac{1}{l} \log(N_l) \to R(P, Q, D) \quad \text{as } l \to \infty \text{ in probability}, \tag{3}$$

and they characterize $R(P,Q,D)$ in terms of information theoretic quantities. A modification of their formula for $R(P,Q,D)$, which appears in [13], has the form

$$R(P,Q,D) = \min_{Q'} \Big\{ I_m(P\|Q',D) + D(Q'\|Q) \Big\} \tag{4}$$

where $D(\cdot)$ denotes divergence (or relative entropy), and I_m denotes "lower mutual information" [15]:

$$I_m(P\|Q,D) = \min_{W \in \mathcal{W}_{P,Q,D}} I(P,W), \tag{5}$$

where $I(P,W)$ denotes the mutual information associated with input distribution P and transition distribution W from $\mathcal{X}$ to $\mathcal{Y}$, and the minimization is over the set $\mathcal{W}_{P,Q,D}$ of all W's that induce output distribution Q and average distortion less than or equal to D, i.e., $\sum_x P(x)W(y|x) = Q(y)$ and $\sum_{x,y} d(x,y)P(x)W(y|x) \leq D$. If no such W exists, i.e., the set $\mathcal{W}_{P,Q,D}$ is empty, then $I_m(P\|Q,D)$ is equal to infinity. See [13, sec. II.A].

An implication of (3) is that for large word length l, approximately $2^{lR(P,Q,D)}$ code words suffice to guarantee a D-match with high probability. This result may give some insight into the workings of adaptive lossy source coding schemes, where during the adaptation phase the codebook is sub-optimum with respect to the source. See, e.g., [14] in the context of lossy Lempel-Ziv algorithms.

In [13], Zamir and Rose address the random "type" T_{N_l} of the D-matching codeword $\mathbf{Y}_{N_l}$. Here, the type of a vector means its empirical distribution [4, 3], i.e., for any $y \in \mathcal{Y}$, $T_{\mathbf{Y}}(y)$ denotes the number of times y occurs in $\mathbf{Y}$ divided by l. They show that for strictly positive Q and large word length, T_{N_l} concentrates around a limiting distribution [13, thm. 4]:

$$T_{N_l}(y) \to Q^*_{P,Q,D}(y), \quad \text{as } l \to \infty \text{ in probability}, \quad \text{for all } y \in \mathcal{Y} \tag{6}$$

where $Q^*_{P,Q,D}$ is the unique distribution Q' which achieves the minimum in (4):

$$Q^*_{P,Q,D} = \arg\min_{Q'} \Big\{ I_m(P\|Q',D) + D(Q'\|Q) \Big\}. \tag{7}$$

We call this concentration point "the favorite type" (although $Q^*_{P,Q,D}$ is in general not an l-type).

The intuition underlying (6) comes from the roles of the two terms in the minimum in (4): The lower mutual information $I_m(P\|Q',D)$ characterizes the "covering efficiency" of a type Q' - it is the D-match probability exponent given that the codeword's type is Q', while the divergence $D(Q'\|Q)$ characterizes the frequency of a type Q' in the codebook - the frequency is $\approx 2^{-lD(Q'\|Q)}$. The common types in the codebook are close to Q, but their covering efficiency is small; the most covering-efficient types in the codebook are close to an *optimum reproduction distribution*,

$$Q^*_{P,D} = \arg\min_Q R(P,Q,D),$$

that achieves the rate-distortion function $R(P,D)$ [1], but these types are too rare in the codebook. The "favorite type" $Q^*_{P,Q,D}$ strikes the optimum balance between covering efficiency and frequency in the codebook.

Another interesting interpretation for $Q^*_{P,Q,D}$, shown in [13], is that it amounts to one iteration of a fixed-distortion version of the Blahut algorithm as it computes $Q^*_{P,D}$ starting from Q.

It follows from (6) that most of the $2^{lR(P,Q,D)}$ codewords in the codebook are asymptotically useless; only those having a type close to $Q^*_{P,Q,D}$ - whose fraction in the codebook is only $\approx 2^{-lD(Q^*_{P,Q,D}\|Q)}$ - have a good chance to be the first to D-match the source word. In a sense, we are paying extra $D(Q^*_{P,Q,D}\|Q)$ bits in coding rate for the random appearance of types in the codebook. Our main result shows that this redundancy can be removed by entropy coding conditioned on the codebook.

Let

$$H(N_l) = -\sum_{n=1}^{\infty} \Pr\{N_l = n\} \log\Big(\Pr\{N_l = n\}\Big)$$

denote the entropy of the index of the first D-matching code word, and let

$$H(N_l|\mathbf{Y}_1,\mathbf{Y}_2,...) = \lim_{M\to\infty} \sum_{\mathbf{y}_1,...,\mathbf{y}_M} \Pr\{\mathbf{Y}_1 = \mathbf{y}_1,...,\mathbf{Y}_M = \mathbf{y}_M\} H(N_l|\mathbf{y}_1,...,\mathbf{y}_M)$$

denote its conditional entropy given the codebook, where the limit holds since the sequence of conditional entropies is non-increasing with M. Conditioning on the codebook means that we design the entropy code only *after* generating the random codebook.

Theorem 1 *If the codebook generating distribution Q is positive everywhere, then the un-conditional index entropy satisfies*

$$\lim_{l\to\infty} \frac{1}{l} H(N_l) = R(P,Q,D), \tag{8}$$

while the conditional index entropy satisfies

$$\begin{aligned} \lim_{l\to\infty} \frac{1}{l} H(N_l|\mathbf{Y}_1,\mathbf{Y}_2,...) &= I_m(P\|Q^*_{P,Q,D}, D) && (9)\\ &= R(P,Q,D) - D(Q^*_{P,Q,D}\|Q). && (10)\end{aligned}$$

Thus conditioning the entropy coding on the codebook saves $D(Q^*_{P,Q,D}\|Q)$ bits in coding rate.

The proof is given in Section III. Roughly speaking, for a random codebook the index N_l is uniformly distributed over the entire range $(1...2^{lR(P,Q,D)})$, since the "favorite type" can appear anywhere in that range, and therefore the *un*conditional entropy in (8) does not improve the coding rate in (3). As suggested by I. Kontoyiannis [7], conditioning on the codebook amounts to re-ordering the codewords according to their probability to be the first to D-match the source, which effectively constructs a sub-codebook of size

$$\frac{2^{lR(P,Q,D)}}{2^{lD(Q^*_{P,Q,D}\|Q)}}.$$

The number of bits saved by conditioning, $D(Q^*_{P,Q,D}\|Q)$, is strictly positive, unless the generating distribution Q is an optimal reproduction distribution, i.e., $Q = Q^*_{P,D}$. On the other hand, since in general $Q^*_{P,Q,D} \neq Q^*_{P,D}$, the conditional index entropy does not achieve the rate-distortion function of the source, so

$$R(P,Q,D) \geq \lim_{l\to\infty} \frac{1}{l} H(N_l|\mathbf{Y}_1^\infty) \geq R(P,D)$$

where $\mathbf{Y}_1^\infty$ denotes $\mathbf{Y}_1, \mathbf{Y}_2, \ldots$. Equality holds in both inequalities above if and only if $Q = Q^*_{P,D}$; see [13]. The next section gives an explicit example for the quantities in this inequality.

III. An Example

In the following special case we have explicit formulas for $R(P, Q, D)$ and $Q^*_{P,Q,D}$, and hence for the conditional index entropy.

Let $\mathcal{X} = \mathcal{Y} = \{0, \ldots, |\mathcal{X}| - 1\}$. Consider symmetric distortion measure of the form

$$d(x, y) = d(y - x)$$

where the subtraction is modulo-$|\mathcal{X}|$. Define $H_{max}(D)$ and V_D as the maximum-entropy under a D-constraint and the maximum-entropy achieving distribution, respectively [3]:

$$H_{max}(D) = H(V_D) = \max_{V:\ \sum_x V(x)d(x) \leq D} H(V) \ .$$

Lemma 1 (Information rates for a uniform codebook) *For symmetric distortion measure, uniform codebook generating distribution $Q(y) = 1/|\mathcal{Y}|\ \forall y$, and any source P,*

$$R(P, Q, D) = \log |\mathcal{X}| - H_{max}(D)$$

and

$$Q^*_{P,Q,D} = P * V_D$$

where the $*$ sign denotes a circular convolution (i.e., $P * V_D$ is the distribution of the independent sum, modulo-$|\mathcal{X}|$, of a random variable $\sim P$ and a random variable $\sim V_D$). Hence the conditional index entropy (9) is given by

$$I_m(P \| Q^*_{P,Q,D}, D) = H(P * V_D) - H_{max}(D). \tag{11}$$

Proof: If $Ed(Y - X) \leq D$, where the subtraction is modulo-$|\mathcal{X}|$, then by the properties of the entropy function

$$H(Y|X) = H(Y - X|X) \leq H(Y - X) \leq H_{max}(D)$$

with equality if and only if Y can be written as the independent sum of X and "noise" $\sim V_D$. Thus, supposing Y' achieves $I_m(P\|Q', D)$, we obtain the "reversed Shannon lower bound"

$$I_m(P\|Q', D) = H(Y') - H(Y'|X) \geq H(Q') - H_{max}(D) ,$$

with equality iff $Q' = P * V_D$. On the other hand, since Q is uniform, the divergence in (4) can be written for any Q' as $D(Q'\|Q) = H(Q) - H(Q') = \log|\mathcal{X}| - H(Q')$. Substituting in (4) we obtain the lower bound $R(P, Q, D) \geq \log|\mathcal{X}| - H_{max}(D)$. The lemma now follows since substituting $Q' = P * V_D$ in (4) achieves this lower bound.

QED

In particular, in the symmetric binary-Hamming case, i.e., $\mathcal{X} = \mathcal{Y} = \{0, 1\}$, $Q(0) = Q(1) = 1/2$, $d(x, y) = x \oplus y$, and for a Bernoulli(p) source, we have by Lemma 1

$$\begin{aligned} R(P, Q, D) &= 1 - H_B(D) && (12) \\ Q^*_{P,Q,D}(1) &= p * D && (13) \\ I_m(P\|Q^*_{P,Q,D}, D) &= H_B(p * D) - H_B(D), && (14) \end{aligned}$$

where $H_B(p) = -p\log(p) - (1 - p)\log(1 - p)$ denotes the binary entropy.

Note that $R(P, Q, D)$ is highly redundant and is independent of the source distribution. In contrast, the conditional index entropy $I_m(P\|Q^*_{P,Q,D}, D)$ is significantly lower, and for low distortions it is close to the Shannon lower bound on the rate-distortion function, $R_{SLB}(D) = H(P) - H_{max}(D)$ [1].

It is interesting to observe that we can rewrite (11) as

$$I_m(P\|Q^*_{P,Q,D}, D) = H(X + N_D) - H(N_D) = I(X; X + N_D),$$

where N_D is independent of X and is distributed as $\sim V_D$; i.e., the conditional index entropy is equal to the mutual information obtained when passing the source X through a modulo-additive noise channel with noise N_D. This relation allows to use the technique of [12] to obtain a universal (source independent) bound on the rate loss with respect to the rate-distortion function. This result will be discussed in more detail in a forthcoming paper with I. Kontoyiannis [8]. There we also extend the "favorite type theorem" (6) and the index entropy result above to continuous sources and reproduction alphabets, and to sources with memory.

IV. Proof of Theorem

In the proof we need the following three propositions, the proofs of which appear in the appendix.

Proposition 1 ("almost flat" geometric distribution entropy) *Suppose the random variable G is geometrically distributed, with* $\Pr(G = n) = p[1-p]^{n-1}$, $n = 1, 2, \ldots$. *Then, its entropy is given by*

$$H(G) = \frac{H_B(p)}{p}. \tag{15}$$

If $p = 2^{-lE}$, then

$$\lim_{l\to\infty} \frac{1}{l} H(G) = E.$$

Proposition 2 (D-match exponents) *Suppose* $\mathbf{Y}$ *is generated i.i.d.* $\sim Q$, *with* $Q[y] > 0\ \forall y \in \mathcal{Y}$. *Let $T_{\mathbf{x}}$ and $T_{\mathbf{Y}}$ denote the types of the l-strings* $\mathbf{x}$ *and* $\mathbf{Y}$, *respectively. Then*

(a) $\Pr\Big\{d(\mathbf{x}, \mathbf{Y}) \le D\Big\} = 2^{-l[R(T_{\mathbf{x}}, Q, D) + o(1)]}$

where $o(1) \to 0$ as $l \to \infty$ uniformly in $\mathbf{x}$*; and*

(b) $\Pr\Big\{d(\mathbf{x}, \mathbf{Y}) \le D \;\Big|\, T_{\mathbf{Y}} = Q'\Big\} = 2^{-l[I_m(T_{\mathbf{x}}\|Q', D) + o(1)]}$

where $o(1) \to 0$ as $l \to \infty$ uniformly for all $\mathbf{x}$ *and Q' such that $I_m(T_{\mathbf{x}}\|Q', D)$ is finite. Otherwise (i.e., if $I_m(T_{\mathbf{x}}\|Q', D) = \infty$) the probability for D-match is zero.*

Proposition 3 (Properties of rate functions) *Assume $Q[y] > 0 \forall y \in \mathcal{Y}$. Then $R(P, Q, D)$ as a function of P is continuous and bounded by* $\log|\mathcal{X}| + \log(1/q_{min})$, *where* $q_{min} = \min_y Q[y]$. *Furthermore, $I_m(P\|Q, D)$ is continuous and bounded by* $\log|\mathcal{Y}|$ *over some closed set of pairs (P, Q), and is equal to infinity outside this set.*

A. The un-conditional case (8)

We start with showing that $\lim_{l\to\infty} \frac{1}{l} H(N_l) = R(P,Q,D)$. Consider the conditional entropy of N_l given a specific source string $\mathbf{x}$

$$H(N_l|\mathbf{X}=\mathbf{x}) = -\sum_{n=1}^{\infty} p_n(\mathbf{x}) \log(p_n(\mathbf{x})) \tag{16}$$

where $p_n(\mathbf{x})$ is the conditional probability that $N_l = n$ given $\mathbf{X} = \mathbf{x}$. Since the code words $\{\mathbf{Y}_i\}$ are independent and identically distributed, we have

$$p_n(\mathbf{x}) = [1 - p_{match}(\mathbf{x})]^{n-1} p_{match}(\mathbf{x}) \tag{17}$$

where $p_{match}(\mathbf{x}) = \Pr\Big(d(\mathbf{x},\mathbf{Y}) \leq D\Big)$. Combining Proposition 1 and part (a) of Proposition 2, we get

$$\frac{1}{l} H(N_l|\mathbf{X}=\mathbf{x}) = R(T_{\mathbf{x}}, Q, D) + o(1) \tag{18}$$

where $o(1) \to 0$ as $\ell \to \infty$ uniformly in $\mathbf{x}$. Since $p_n(\mathbf{x})$, the conditional distribution of N_l, depends on $\mathbf{x}$ only through its type, we also have

$$\frac{1}{l} H(N_l|T_{\mathbf{X}} = P') = R(P', Q, D) + o(1). \tag{19}$$

Using the fact that the $o(1)$ term converges uniformly, we take an expectation over the source string type $T_{\mathbf{X}}$, and get

$$\frac{1}{l} H(N_l|T_{\mathbf{X}}) = \sum_{P' \in \mathcal{P}_l} \Pr(T_{\mathbf{X}} = P')\ R(P', Q, D) + o(1) \tag{20}$$

where $\mathcal{P}_l$ denotes the set of types of strings in $\mathcal{X}^l$, and $o(1) \to 0$ as $\ell \to \infty$. But the source string is memoryless, so by the law of large numbers $T_{\mathbf{X}}$ converges to P in probability. Furthermore, by Proposition 3 $R(P', Q, D)$ as a function of P' is continuous and bounded, so convergence in probability implies convergence in the mean. We conclude that

$$\lim_{l\to\infty} \frac{1}{l} H(N_l|T_{\mathbf{X}}) = R(P,Q,D).$$

Now, by the properties of the mutual information function we have

$$
\begin{aligned}
0 \leq H(N_l) - H(N_l|T_{\mathbf{X}}) &= I(T_{\mathbf{X}}; N_l) && (21)\\
&\leq H(T_{\mathbf{X}}) && (22)\\
&\leq \log |\mathcal{P}_l| && (23)\\
&= O(\log(l)) && (24)
\end{aligned}
$$

where (23) follows since $\mathcal{P}_l$ is the support of $T_{\mathbf{X}}$, and (24) since the size of $\mathcal{P}_l$ (the number of l-types) grows polynomialy with l [4, 3]. Thus $\frac{1}{l}[H(N_l) - H(N_l|T_{\mathbf{X}})] \to 0$, and the first part of the theorem follows.

B. The conditional case (9): direct half

We turn to the conditional index entropy (9). Consider first the "direct" half of the statement, i.e.,

$$\limsup_{l\to\infty} \frac{1}{l} H(N_l|\mathbf{Y}_1^\infty) \leq I_m(P\|Q^*_{P,Q,D}, D) \tag{25}$$

where we use the concise notation $\mathbf{Y}_1^\infty$ to denote $\mathbf{Y}_1, \mathbf{Y}_2, \ldots.$ Let $\mathcal{P}$ and $\mathcal{Q}$ denote the sets of all distributions over $\mathcal{X}$ and $\mathcal{Y}$, respectively, and let B_δ denote a subset of $\mathcal{P} \times \mathcal{Q}$ to be specified later. Given $\mathbf{X}$ and $\mathbf{Y}_1^\infty$, define the sequence of "good type pair" indicators

$$J_n = \begin{cases} 1, & \text{if } (T_{\mathbf{X}}, T_{\mathbf{Y}_n}) \in B_\delta \\ 0, & \text{otherwise} \end{cases} \qquad \text{for } n = 1, 2, \ldots \tag{26}$$

where $T_{\mathbf{Y}_n}$ denotes the (random) type of the n-th codeword. Note that J_{N_l} indicates whether $(T_{\mathbf{X}}, T_{N_l})$ belongs to B_δ or not, i.e., if the source type and the type of the first D-matching codeword are a "good type pair". Partition the code $\{\mathbf{Y}_1^\infty\}$ into two sub-codes $\{\mathbf{Y}_n : J_n = 0\}$ and $\{\mathbf{Y}_n : J_n = 1\}$, called subcode-0 and subcode-1, respectively. Note that the partition depends on both the source type and the codebook. For $j = 0, 1$, let $N_l^{(j)}$ denote the index of the first codeword in subcode-j that D-matches the source $\mathbf{X}$, and let $M_l^{(j)}$ denote the number of codewords in subcode-j until the D-match was found in subcode-j, i.e.,

$$M_l^{(j)} = \left|\{\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_l^{(j)}}\} \bigcap \text{subcode-}j\right|. \tag{27}$$

If D-match is not found in subcode-j, then define $N_l^{(j)} = M_l^{(j)} = \infty$. Notice the identity

$$N_l^{(J_{N_l})} = N_l.$$

By (21) – (24)

$$H(N_l|\mathbf{Y}_1^\infty) \leq H(N_l|\mathbf{Y}_1^\infty, T_\mathbf{X}) + O(\log(l)). \tag{28}$$

Clearly, given the codebook $\mathbf{Y}_1^\infty$ and the source type $T_\mathbf{X}$, the index N_l has a one-to-one relation with the pair $(J_{N_l}, M_l^{(J_{N_l})})$. This pair has a one-to-one relation with the triplet $(J_{N_l}, J_{N_l}M_l^{(1)}, (1-J_{N_l})M_l^{(0)})$. Thus, by the properties of the entropy function we have

$$\begin{aligned}
H(N_l|\mathbf{Y}_1^\infty, T_\mathbf{X}) &= H(J_{N_l}, J_{N_l}M_l^{(1)}, (1-J_{N_l})M_l^{(0)}|\mathbf{Y}_1^\infty, T_\mathbf{X}) && (29)\\
&\leq H(J_{N_l}, M_l^{(1)}, (1-J_{N_l})M_l^{(0)}|\mathbf{Y}_1^\infty, T_\mathbf{X}) && (30)\\
&= H(J_{N_l}|\mathbf{Y}_1^\infty, T_\mathbf{X}) \\
&\quad + H(M_l^{(1)}|\mathbf{Y}_1^\infty, T_\mathbf{X}, J_{N_l}) \\
&\quad + H((1-J_{N_l})M_l^{(0)}|\mathbf{Y}_1^\infty, T_\mathbf{X}, J_{N_l}, M_l^{(1)}) && (31)\\
&\leq 1 + H(M_l^{(1)}|T_\mathbf{X}) + H((1-J_{N_l})M_l^{(0)}|\mathbf{Y}_1^\infty, J_{N_l}) && (32)
\end{aligned}$$

where (29) follows from the one-to-one relation $N_l \Leftrightarrow (J_{N_l}, J_{N_l}M_l^{(1)}, (1-J_{N_l})M_l^{(0)})$ when $(\mathbf{Y}_1^\infty, T_\mathbf{X})$ is given; (30) follows since $(J_{N_l}, J_{N_l}M_l^{(1)})$ is a function of $(J_{N_l}, M_l^{(1)})$; (31) follows from the chain rule; and (32) follows by bounding the entropy of the binary variable J_{N_l} by one bit, and since dropping conditions increases the average conditional entropy. The last term in (32) can be written as

$$\Pr(J_{N_l} = 0) \cdot H(M_l^{(0)}|\mathbf{Y}_1^\infty, J_{N_l} = 0) + \Pr(J_{N_l} = 1) \cdot 0. \tag{33}$$

Since the sub-index $M_l^{(0)}$ is a function of $\mathbf{X}$ and $\mathbf{Y}_1^\infty$, we have

$$H(M_l^{(0)}|\mathbf{Y}_1^\infty, J_{N_l} = 0) \leq H(\mathbf{X}|\mathbf{Y}_1^\infty, J_{N_l} = 0) \leq \log|\mathcal{X}|^l \tag{34}$$

where in the second inequality we bounded the entropy of $\mathbf{X}$ by the logarithm of its alphabet size. Substituting in (32) and (28), we get

$$\frac{1}{l}H(N_l|\mathbf{Y}_1^\infty) \leq \frac{1}{l}H(M_l^{(1)}|T_\mathbf{X}) + \Pr(J_{N_l}=0)\cdot\log|\mathcal{X}| + \frac{O(\log(l))}{l}. \tag{35}$$

Suppose we choose B_δ as some neighborhood of $(P, Q^*_{P,Q,D})$. Now, by the law of large numbers, and the favorite type theorem (6), we have as $l \to \infty$

$$T_\mathbf{X} \to P \quad \text{and } T_{N_l} \to Q^*_{P,Q,D} \tag{36}$$

in probability. Thus

$$\Pr\Big((T_\mathbf{X}, T_{N_l}) \notin B_\delta\Big) = \Pr(J_{N_l}=0) \to 0. \tag{37}$$

Hence, since $|\mathcal{X}| < \infty$, for such B_δ the second term in (35) vanishes asymptotically, and we are left to consider the first term of (35).

Observe that if $\Pr((T_\mathbf{x}, T_\mathbf{Y}) \in B_\delta) = 0$, then given $\mathbf{X} = \mathbf{x}$ codebook-1 is empty, so $M_l^{(1)} = \infty$ with probability one. Thus for such $\mathbf{x}$ $H(M_l^{(1)}|\mathbf{X}=\mathbf{x}) = 0$. Consider now the other case, i.e., $\mathbf{x}$ is such that $\Pr((T_\mathbf{x}, T_\mathbf{Y}) \in B_\delta) > 0$. We call such $\mathbf{x}$ "δ-typical". Observe that conditional on $\mathbf{X} = \mathbf{x}$, $M_l^{(1)}$ is geometrically distributed random variable with

$$\Pr(M_l^{(1)} = m \mid \mathbf{X} = \mathbf{x}) = [1 - p'_{match}(\mathbf{x})]^{m-1} p'_{match}(\mathbf{x}) \tag{38}$$

for $m = 1, 2, \ldots$, where

$$p'_{match}(\mathbf{x}) = \Pr\Big(d(\mathbf{x}, \mathbf{Y}) \leq D \;\Big|\; (T_\mathbf{x}, T_\mathbf{Y}) \in B_\delta\Big)$$

is the D-match probability in subcode-1. (This is analogous to the distribution of N_l (17).) Thus, by Proposition 1 the entropy of $M_l^{(1)}$ is asymptotically determined by the exponent of $p'_{match}(\mathbf{x})$. To characterize this exponent, we first characterize the set B_δ in the following lemma which is proved in the Appendix.

Lemma 2 ("Typical D-match Set") *For any $\delta > 0$ there exists a set $B_\delta \subset \mathcal{P} \times \mathcal{Q}$ such that* *(a)* $\Pr((T_{\mathbf{X}}, T_{N_l}) \in B_\delta) \to 1$ *as* $l \to \infty$*;*

(b) $\Big| I_m(P' \| Q', D) - I_m(P \| Q^*_{P,Q,D}, D) \Big| \le \delta,$ *for all pairs* $(P', Q') \in B_\delta$*.*

Lemma 2(a) guarantees (37). Lemma 2(b) and Proposition 2(b) imply that for a δ-typical $\mathbf{x}$,

$$\Pr\Big\{ d(\mathbf{x}, \mathbf{Y}) \le D \;\Big|\; (T_{\mathbf{x}}, T_{\mathbf{Y}}) \in B_\delta \Big\} \ge 2^{-l[I_m(P\|Q^*_{P,Q,D}, D) + \delta + o(1)]}, \tag{39}$$

where $o(1) \to 0$ as $l \to \infty$ uniformly over all δ-typical $\mathbf{x}$. Combining with Proposition 1 and (38), we get

$$\frac{1}{l} H(M_l^{(1)} | \mathbf{X} = \mathbf{x}) \le I_m(P \| Q^*_{P,Q,D}, D) + \delta + o(1), \tag{40}$$

where $o(1) \to 0$ as $l \to \infty$ uniformly over all δ-typical $\mathbf{x}$. Since, as discussed above, for a non δ-typical $\mathbf{x}$ we have $H(M_l^{(1)} | \mathbf{X} = \mathbf{x}) = 0$, (40) holds in fact for <u>all</u> $\mathbf{x}$. We now proceed similarly to the proof of the unconditional case, steps (18)-(20). Since the distribution of $M_l^{(1)}$ depends on $\mathbf{x}$ only through its type, we replace the conditioning on $\mathbf{X}$ by conditioning on $T_{\mathbf{X}}$. Then, noting that the $o(1)$ term converges to zero uniformly in $\mathbf{x}$, we average over $T_{\mathbf{X}}$ and take the limit in l to obtain

$$\limsup_{l \to \infty} \frac{1}{l} H(M_l^{(1)} | T_{\mathbf{X}}) \le I_m(P \| Q^*_{P,Q,D}, D) + \delta \; .$$

We now let $\delta \to 0$, which, in view of (35) and (37), establishes (25).

C. The conditional case (9): converse half

The "converse" half of (9) shows that the reversed inequality holds asymptotically, i.e.,

$$\liminf_{l \to \infty} \frac{1}{l} H(N_l | \mathbf{Y}_1^\infty) \ge I_m(P \| Q^*_{P,Q,D}, D). \tag{41}$$

The proof of the converse deviates from the subcode analysis of the index entropy in (32), because we cannot drop the conditioning on the codebook. Consequentially, we cannot consider the D-match trails as statistically independent, and we must take a different approach. The idea of the proof is based on the following simple lemma.

Lemma 3 (Small probabilities imply high entropy) *If the probability of each letter in some discrete alphabet is upper bounded by $1/M$ for some integer M (where M is less than or equal to the size of the alphabet), then the entropy is* <u>lower bounded</u> *by $\log(M)$. The minimum is achieved if and only if the distribution is uniform over some M letters.*

Proof: Given any probability distribution $p_1, p_2, \ldots$, such that $p_i \leq 1/M$ for all i,

$$H(p_1, p_2, \ldots) = \sum_i p_i \log \frac{1}{p_i} \geq \sum_i p_i \log(M) = \log(M)$$

with equality iff $p_i \in \{0, 1/M\}$.

QED

Starting similarly to the subcode analysis of the index entropy in (29)-(32), we have

$$\begin{aligned} H(N_l|\mathbf{Y}_1^\infty) &\geq H(N_l|\mathbf{Y}_1^\infty, T_\mathbf{X}) && (42)\\ &= H(J_{N_l}, J_{N_l}M_l^{(1)}, (1-J_{N_l})M_l^{(0)}|\mathbf{Y}_1^\infty, T_\mathbf{X}) && (43)\\ &\geq H(J_{N_l}M_l^{(1)}|\mathbf{Y}_1^\infty, T_\mathbf{X}, J_{N_l}) && (44)\\ &= \Pr(J_{N_l}=1) \cdot H(M_l^{(1)}|\mathbf{Y}_1^\infty, T_\mathbf{X}, J_{N_l}=1) && (45) \end{aligned}$$

where (44) follows from the chain rule and the nonnegativity of the entropy. We want to lower bound (45) in terms of $H(M_l^{(1)}|\mathbf{Y}_1^\infty, T_\mathbf{X})$. For that, consider the following chain of inequalities:

$$\begin{aligned} H(M_l^{(1)}|\mathbf{Y}_1^\infty, T_\mathbf{X}) &\leq H(M_l^{(1)}, J_{N_l}|\mathbf{Y}_1^\infty, T_\mathbf{X}) && (46)\\ &\leq H(M_l^{(1)}|\mathbf{Y}_1^\infty, T_\mathbf{X}, J_{N_l}) + 1 && (47)\\ &= \Pr(J_{N_l}=1) \cdot H(M_l^{(1)}|\mathbf{Y}_1^\infty, T_\mathbf{X}, J_{N_l}=1) \\ &+ \Pr(J_{N_l}=0) \cdot H(M_l^{(1)}|\mathbf{Y}_1^\infty, T_\mathbf{X}, J_{N_l}=0) + 1 && (48)\\ &\leq \Pr(J_{N_l}=1) \cdot H(M_l^{(1)}|\mathbf{Y}_1^\infty, T_\mathbf{X}, J_{N_l}=1) \\ &+ \Pr(J_{N_l}=0) \cdot l \log|\mathcal{X}| + 1 && (49) \end{aligned}$$

where (46) and (47) follow by the chain rule, the nonnegativity of the entropy, and by bounding the entropy of the binary variable J_{N_l} by one bit; and (49) follow similarly to (34). Now, suppose we take the set B_δ (which determines J_{N_l}) from Lemma 2. Then, since $|\mathcal{X}|$ is finite and $\Pr(J_{N_l}=0)\to 0$ by (37), the second term in (49) has the form $lo(1)$, where $o(1)\to 0$ as $l\to\infty$. Comparing with (45), we thus have the asymptotic lower bound

$$\frac{1}{l}H(N_l|\mathbf{Y}_1^\infty)\geq \frac{1}{l}H(M_l^{(1)}|\mathbf{Y}_1^\infty,T_{\mathbf{X}})-o(1) \tag{50}$$

where $o(1)\to 0$ as $l\to\infty$.

Our goal is therefore to lower bound $H(M_l^{(1)}|\mathbf{Y}_1^\infty,T_{\mathbf{X}})$. Consider

$$\Pr\Big(M_l^{(1)}=m \;\Big|\; \mathbf{Y}_1^\infty=\mathbf{y}_1^\infty, T_{\mathbf{X}}=P'\Big).$$

Let $n(m)$ denote the index of the m-th codeword in subcode-1. Since previous codewords can D-match the source before the $n(m)$-th codeword, the probability that $M_l^{(1)}=m$ cannot exceed the probability that $\mathbf{y}_{n(m)}$ D-matches the source string, i.e.,

$$\begin{aligned}
\Pr\Big(M_l^{(1)}=m \quad &\Big| \quad \mathbf{Y}_1^\infty=\mathbf{y}_1^\infty, T_{\mathbf{X}}=P'\Big)\\
&\leq \Pr\Big(d(\mathbf{X},\mathbf{y}_{n(m)})\leq D \;\Big|\; \mathbf{Y}_1^\infty=\mathbf{y}_1^\infty, T_{\mathbf{X}}=P'\Big) && (51)\\
&= \Pr\Big(d(\mathbf{X},\mathbf{y}_{n(m)})\leq D \;\Big|\; T_{\mathbf{X}}=P'\Big) && (52)\\
&= 2^{-l[I_m(P'\|Q',D)-o(1)]} && (53)\\
&\leq 2^{-l[I_m(P\|Q^*_{P,Q,D},D)-\delta-o(1)]} && (54)
\end{aligned}$$

where (52) follows since the source string is independent of the codebook; (53) follows from Proposition 2(b) assuming $T_{\mathbf{y}_{n(m)}}=Q'$; (54) follows because $\mathbf{y}_{n(m)}$ belongs to subcode-1, and by Lemma 2(b); and $o(1)\to 0$ as $l\to\infty$ uniformly in P' and Q'. Note that the upper bound in (54) is independent of m. Applying Lemma 3 we thus have

$$H\Big(M_l^{(1)} \;\Big|\; T_{\mathbf{X}}=P', \mathbf{Y}_1^\infty=\mathbf{y}_1^\infty\Big)\geq l[I_m(P\|Q^*_{P,Q,D},D)-\delta-o(1)] \tag{55}$$

where $o(1)\to 0$ as $l\to\infty$ uniformly over P' and $\mathbf{y}_1^\infty$. Since the lower bound is independent of $\mathbf{y}_1^\infty$ and P', it holds also with respect to the average conditional entropy $H(M_l^{(1)}|\mathbf{Y}_1^\infty, T_\mathbf{X})$. Combining with (50) we conclude that

$$\liminf_{l\to\infty} \frac{1}{l} H(N_l|\mathbf{Y}_1^\infty) \geq I_m(P\|Q^*_{P,Q,D}, D) - \delta.$$

Since $\delta > 0$ is arbitrary the converse part (41) follows.

QED

Appendix

A. Proof of Proposition 1

The proof of (15) follows directly from the series $\sum_{m=1}^\infty mq^m = q/(1-q)^2$. The case $p = 2^{-lE}$ follows straightforwardly from (15).

B. Proof of Proposition 2

Parts (a) and (b) are stated in [13, eq. 9] and [13, thm. 1], respectively, as derivatives of results due to Zhang, Yang and Wei. Specifically, part (b) follows from the formula, [15, sec. III.C],

$$p^n(B(r,t,d)) = 2^{o(n)-n(D(t\|p)+I_l(r,t,d))}$$

where n is the word length; $B(r,t,d) = \{\mathbf{x} : d(\mathbf{x},\mathbf{y}) \leq d, T_\mathbf{x} = t\}$ is the d-ball associated with some $\mathbf{y}$ of type r, projected on $\{\mathbf{x} : T_\mathbf{x} = t\}$; $p^n(\cdot)$ denotes probability with respect to $\mathbf{X}$; and $I_l(r,t,d)$ is our notation $I_m(t\|r,d)$. The $o(n)$ term comes from the multinom approximation (where $s(x,y)$ is the joint distribution whose marginals are $r(y)$ and $t(x)$)

$$\prod_{y\in\mathbf{Y}} \binom{nr(y)}{ns(1,y),\ldots,ns(|\mathcal{X}|,y)} = 2^{o(n)+nH(s|r)},$$

and it is bounded by (see, e.g., [3, thm. 12.1.3])

$$0 \leq o(n) \leq |\mathcal{X}||\mathcal{Y}|\log(n),$$

i.e., $o(n)/n \to 0$ as $n \to \infty$ uniformly in r and t. This proves part (b) since by Bayes law and [3, thm. 12.1.4]

$$\Pr\Big\{d(\mathbf{X},\mathbf{y}) \le D \;\Big|T_{\mathbf{X}} = t\Big\} = \frac{p^n(B(r,t,d))}{p^n(T_{\mathbf{X}} = t)} = 2^{o(n)-nI_l(r,t,d)}.$$

Part (a) follows again by the Bayes law,

$$\Pr\Big\{d(\mathbf{x},\mathbf{Y}) \le D\Big\} = \sum_{Q' \in \mathcal{Q}_n} \Pr\{T_{\mathbf{Y}} = Q'\} \cdot \Pr\Big\{d(\mathbf{x},\mathbf{Y}) \le D \;\Big|T_{\mathbf{Y}} = Q'\Big\},$$

which, using part (b), [3, thm. 12.1.4], and the fact that $|\mathcal{Q}_n| = O(\log(n))$, implies part (a). See [13, eq. 9].

C. Proof of Proposition 3:

The continuity properties of $R(P,Q,D)$ and $I_m(P,Q,D)$ are shown in [15, sec. III.C] and [13, app. A]. The bound on $R(P,Q,D)$ follows from the definition (4), since the divergence is bounded by $D(Q'\|Q) \le \log(1/q_{min})$ for all Q', while the lower mutual information I_m is bounded by $\log|\mathcal{Y}|$ whenever it is finite, e.g., for the optimum reproduction distribution Q^*, $I_m(P\|Q^*,D) = R(P,D) \le \log|\mathcal{Y}|$.

D. Proof of Lemma 2

By Proposition 3, $I_m(P'\|Q',D)$ is bounded (by $\log|\mathcal{Y}|$), and it is continuous over some closed set $S_D^* \subset \mathcal{P} \times \mathcal{Q}$ which contains $(P, Q^*_{P,Q,D})$. (Note that $\infty > R(P,Q,D) \ge I_m(P\|Q^*_{P,Q,D}, D)$ by Proposition 3 and (4).) Let S_δ be some small neighborhood of $(P, Q^*_{P,Q,D})$ (e.g., a ball or cube in the simplex $\mathcal{P} \times \mathcal{Q}$), and let

$$B_\delta = S_\delta \cap S_D^*.$$

Then, the continuity of $I_m(P'\|Q',D)$ implies part (b) for sufficiently small S_δ. Moreover, since $I_m(P'\|Q',D) = \infty$ outside S_D^*, we have by part (b) of Proposition 2,

$$\Pr(S_D^*) \triangleq \Pr((T_{\mathbf{X}}, T_{N_l}) \in S_D^*) = 1$$

implying that $\Pr(S_\delta \cap S_D^*) = \Pr(S_\delta)$. On the other hand, (36) implies that

$$\Pr(S_\delta) \to 1,$$

for any neighborhood S_δ. Thus also $\Pr(S_\delta \cap S_D^*) \to 1$, and part (a) of the lemma follows. QED

Acknowledgments

My joint work with K. Rose formed the basis for this work. I thank I. Kontoyiannis for the discussion that motivated this result. I thank T. Berger for introducing me to the work of Pinkston on entropy coded random codes. I thank T. Linder for helpful comments. I thank the anonymous reviewers who pointed out a critical error in an early version of the proof of Theorem 1, and showed ways to significantly simplify the proof.

References

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

[2] P. A. Chou, T. Lookabaugh, and R. M. Gray. Entropy constrained vector quantization. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-37:31–42, Jan. 1989.

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

[4] I. Csiszár and J. Korner. *Information Theory - Coding Theorems for Discrete Memoryless Systems.* Academic Press, New York, 1981.

[5] A. Dembo and I. Kontoyiannis. Source coding, large deviations and approximate pattern matching. *IEEE Trans. Information Theory*, preprint.

[6] H. Gish and N. J. Pierce. Asymptotically efficient quantization. *IEEE Trans. Information Theory*, IT-14:676–683, Sept. 1968.

[7] I. Kontoyiannis. Private communications. Aug. 1998.

[8] I. Kontoyiannis and R. Zamir. Mismatched codebooks in rate-distrotion theory. in preparation.

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

[10] 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.

[11] 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.

[12] R. Zamir and M. Feder. On universal quantization by randomized uniform / lattice quantizer. *IEEE Trans. Information Theory*, pages 428–436, March 1992.

[13] R. Zamir and K. Rose. Natural type selection in addaptive lossy compression. *IEEE Trans. Information Theory*, 47:99–111, Jan. 2001.

[14] 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.

[15] 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.

[16] J. Ziv. On universal quantization. *IEEE Trans. Information Theory*, IT-31:344–347, May 1985.