Performance Analysis of Fano Coding

Stanislav Krajči*, Chin-Fu Liu†, Ladislav Mikeš*, and Stefan M. Moser†‡
*Institute of Computer Science, Pavol Jozef Šafárik University, Košice, Slovak Republic
†Department of Electrical and Computer Engineering, National Chiao-Tung University (NCTU), Hsinchu, Taiwan
‡Signal and Information Processing Lab, ETH Zurich, Switzerland

***Abstract*—A rigorous performance analysis of Fano coding is presented, providing an upper bound on the average codeword length of binary and ternary Fano codes for an arbitrary discrete memoryless source. The performance bound is slightly better than Shannon's well-known bound for Shannon coding.**

As a by-product a novel general lower bound on Shannon entropy is derived that might be of interest also in a different context. This bound is expressed with the help of variational distance and provides a special case of a reverse Pinsker inequality.

***Index Terms*—Entropy lower bound, Fano coding, reverse Pinsker inequality, Shannon entropy, variational distance.**

I. Introduction

Around 1948, both Claude E. Shannon [1] and Robert M. Fano [2] independently proposed two different source coding algorithms for an efficient description of a discrete memoryless source. Unfortunately, in spite of being different, both schemes became known under the same name *Shannon–Fano coding*.

There are several reasons for this mixup. For one thing, in the discussion of his coding scheme, Shannon mentions Fano's scheme and calls it "substantially the same" [1, p. 17]. For another, both Shannon's and Fano's coding schemes are similar in the sense that they both are efficient, but *suboptimal*[1] prefix-free coding schemes[2] with a similar performance.

In [4, Section 5.9] the scheme by Shannon is even called *Shannon–Fano–Elias coding*, also including Peter Elias in the name. The reason here lies in the many contributions of Elias towards *arithmetic coding*, whose roots lie directly in Shannon's algorithm of 1948. Note, however, that Elias denied having invented arithmetic coding [5, Section 1.2] and he obviously was not involved in the invention of the Shannon code. In this paper, we will stick to the names *Shannon coding* and *Fano coding*.

It is worth noting that even though Fano coding stems from the very dawn of information theory, to the best of our knowledge, there does not exist any rigorous analysis of its performance. Indeed, we believe that Shannon came up with his own algorithm exactly because Fano coding is rather difficult to analyze. This paper tries to fill this gap: it provides a performance analysis and presents an upper bound on the average codeword length for binary and ternary Fano codes. We conjecture that the presented bound also holds for general D-ary Fano coding. The bound is slightly better than Shannon's well-known bound on Shannon coding. This confirms that Fano coding — while still suboptimal — usually performs slightly better than Shannon coding.

As the gap to the best possible lossless compression, the difference between the average codeword length of a D-ary code and the source entropy normalized by $\log \mathrm{D}$ is also known as *average redundancy*. So this work presents an upper bound on the average redundancy for Fano codes. One can find quite a large body of research about redundancy in the literature (see [6], [7], [8] and references therein). These works assume certain structured sources (like a binary memoryless source or a Markov source) in combination with a block parser of fixed message size n. These length-n source sequences are then fed to a source coding scheme (usually Huffman or Shannon coding), and the behavior of the redundancy as a function of n is investigated, in particular for large n and asymptotically as n tends to infinity. In this paper, we do not make any assumption on the source or the source parser, but simply consider an arbitrary (finite-alphabet) distribution at the input of the encoder and then analyze the performance of the code.

As a by-product of our analysis, we also succeeded in finding a new lower bound on the Shannon entropy of an arbitrary probability distribution $\mathbf{p}$ in terms of the variational distance between $\mathbf{p}$ and the uniform distribution $\mathbf{q}_{\mathrm{u}}$. This new bound can also be used to find a novel *reverse Pinsker inequality* (see [9], [10] and references therein) on the relative entropy $\mathscr{D}(\mathbf{p}\|\mathbf{q}_{\mathrm{u}})$ between $\mathbf{p}$ and the uniform distribution $\mathbf{q}_{\mathrm{u}}$.

The remainder of this paper is structured as follows. Next, in Section II, we will recall the definition of Fano coding and present the new bound on its performance, and Section III reviews some properties and relations of Shannon entropy and variational distance that leads to a new lower bound on entropy in terms of variational distance.

II. Fano Coding and Main Result

We are given an r-ary random message where the ith message symbol appears with probability p_i, $i = 1, \ldots, r$ ($r \in \mathbb{N}$). Usually, we write the *probability mass function (PMF)* of this random message as a probability vector $\mathbf{p} = (p_1, \ldots, p_r)$.

Definition 1: Let $\mathrm{D} \geq 2$ be an integer. A D*-ary prefix-free code* for an r-ary random message is a mapping that assigns to the ith message symbol a D-ary codeword $\mathbf{c}_i$ of (variable) length $l(p_i)$, $i = 1, \ldots, r$.

[1] The optimal coding scheme was later found by Fano's student David A. Huffman [3].

[2] Note that in literature, prefix-free codes are sometimes also given the somewhat paradoxical name *prefix code*.

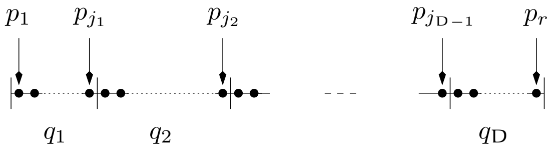

Fig. 1. A *split* $(j_0, \ldots, j_{\mathrm{D}})$ separating the r probabilities of a PMF $\mathbf{p}$ into D groups.

The performance of a code is measured by its average codeword length L:

$$\mathrm{L} \triangleq \sum_{i=1}^{r} p_i \, l(p_i). \tag{1}$$

Definition 2: We use $\mathrm{D}+1$ integers $(j_0, \ldots, j_{\mathrm{D}})$,

$$0 \triangleq j_0 < j_1 < j_2 < \cdots < j_{\mathrm{D}-1} < j_{\mathrm{D}} \triangleq r, \tag{2}$$

to describe a *split* of the vector of probabilities $(p_1, \ldots, p_r)$ into D *groups*

$$(p_{j_{\ell-1}+1}, \ldots, p_{j_\ell}), \quad \ell = 1, \ldots, \mathrm{D}, \tag{3}$$

with the total probability of each group denoted by

$$q_\ell \triangleq p_{j_{\ell-1}+1} + \cdots + p_{j_\ell}, \quad \ell = 1, \ldots, \mathrm{D} \tag{4}$$

(see Fig. 1). Note that here we have assumed that $r \geq \mathrm{D}$. If $r < \mathrm{D}$, then a split assigns to the first r groups exactly one probability value $q_\ell = p_\ell$, $\ell = 1, \ldots, r$, while the remaining $\mathrm{D} - r$ groups are assigned a value $q_\ell = 0$, $\ell = r+1, \ldots, \mathrm{D}$. To ease our notation, we introduce the following notational agreement: if $q_\ell = 0$, we set

$$\frac{p_{j_\ell}}{q_\ell} = \frac{0}{0} \triangleq 1. \tag{5}$$

Definition 3: A D-ary *Fano code* for a random message with PMF $\mathbf{p}$ is generated according to the following algorithm:

Step 1: Arrange the message symbols in order of decreasing probability: $p_1 \geq p_2 \geq \cdots \geq p_r$.

Step 2: Split the ordered probability vector into D groups in such a way that the total probabilities q_ℓ ($\ell = 1, \ldots, \mathrm{D}$) of each group are *as similar as possible*, i.e., such that

$$\frac{1}{\mathrm{D}} \sum_{\ell=1}^{\mathrm{D}} \sum_{\ell'=1}^{\mathrm{D}} |q_\ell - q_{\ell'}| \tag{6}$$

is minimized. We call such a split a *Fano split*.

Step 3: Assign the digit 0 to the first group, the digit 1 to the second group, ..., and the digit $\mathrm{D}-1$ to the last group. This means that the codewords for the symbols in the first group will all start with 0, and the codewords for the symbols in the second group will all start with 1, etc.

Step 4: Recursively apply Step 2 and Step 3 to each of the D groups, subdividing each group into further D groups and adding bits to the codewords until each symbol is the single member of a group.

p_1	p_2	p_3	p_4	p_5
0.35	0.25	0.15	0.15	0.1
0.6		0.4		
0		1		
0.35	0.25	0.15	0.15	0.1
		0.15	0.25	
0	1	0	1	
			0.15	0.1
			0	1
00	01	10	110	111

Fig. 2. Construction of a binary Fano code according to Example 4.

p_1	p_2	p_3	p_4	p_5
0.35	0.25	0.15	0.15	0.1
0.35	0.25	0.4		
0	1	2		
		0.15	0.15	0.1
		0	1	2
0	1	20	21	22

Fig. 3. Construction of a ternary Fano code according to Example 4.

Note that effectively this algorithm constructs a tree and that therefore the Fano code is prefix-free.

Example 4: Let us generate a binary ($\mathrm{D} = 2$) Fano code for a random message with five symbols having probabilities

$$\begin{aligned} &p_1 = 0.35, \quad p_2 = 0.25, \quad p_3 = 0.15, \\ &p_4 = 0.15, \quad p_5 = 0.1. \end{aligned} \tag{7}$$

Since the symbols are already ordered in decreasing order of probability, Step 1 can be omitted. We hence want to split the list into two groups, both having as similar total probability as possible. If we split in $\{1\}$ and $\{2,3,4,5\}$, we have a total probability 0.35 on the left and 0.65 on the right; the split $\{1,2\}$ and $\{3,4,5\}$ yields 0.6 and 0.4; and $\{1,2,3\}$ and $\{4,5\}$ gives 0.75 and 0.25. We see that the second split is best. So we assign 0 as a first digit to $\{1,2\}$ and 1 to $\{3,4,5\}$.

Now we repeat the procedure with both subgroups. Firstly, we split $\{1,2\}$ into $\{1\}$ and $\{2\}$. This is trivial. Secondly, we split $\{3,4,5\}$ into $\{3\}$ and $\{4,5\}$ because 0.15 and 0.25 is closer to each other than 0.3 and 0.1 that we would have gotten by splitting into $\{3,4\}$ and $\{5\}$. Again we assign the corresponding codeword digits.

Finally, we split the last group $\{4,5\}$ into $\{4\}$ and $\{5\}$. We end up with the five codewords $\{00, 01, 10, 110, 111\}$. This whole procedure is shown in Fig. 2.

In Fig. 3 a ternary Fano code is constructed for the same random message.

Remark 5: We would like to point out that there are cases where the algorithm given in Definition 3 does not lead to a unique design: There might be two different ways of dividing the list into D groups such that the total probabilities are

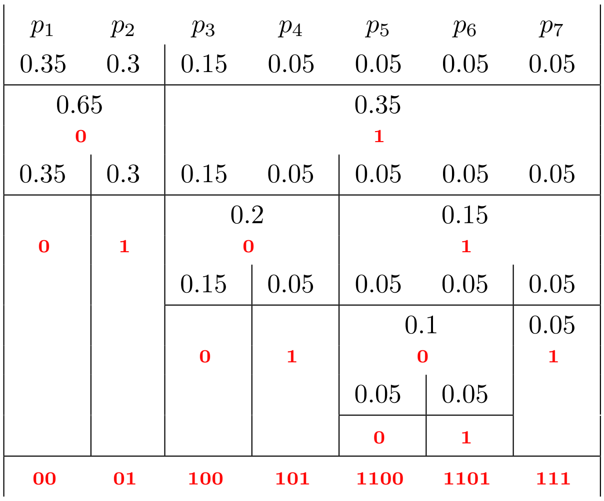

Fig. 4. One possible Fano code for the random message given in (8).

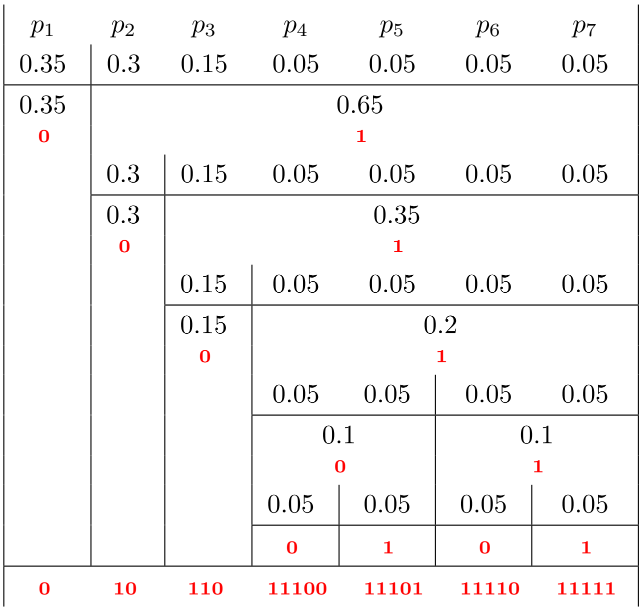

Fig. 5. A second possible Fano code for the random message given in (8).

as similar as possible. Since the algorithm does not specify what to do in such a case, we are free to choose any possible way. Unfortunately, however, these different choices can lead to codes with different performance.

As an example consider the following PMF:

$$\begin{aligned} &p_1 = 0.35, \quad p_2 = 0.3, \quad p_3 = 0.15, \quad p_4 = 0.05, \\ &p_5 = 0.05, \quad p_6 = 0.05, \quad p_7 = 0.05. \end{aligned} \tag{8}$$

One finds that in total there exist six different possible designs of a Fano code (two of which are shown in Figs. 4 and 5), some with an average codeword length of $\mathsf{L} = 2.45$ and some with $\mathsf{L} = 2.4$.

The main result of this paper is the following bound on the performance of Fano codes.

Theorem 6: For $\mathsf{D} \in \{2, 3\}$, the average codeword length L of a D-ary Fano code for an r-ary random message with PMF $\mathbf{p}$ satisfies

$$\mathsf{L}(\mathbf{p}) \leq \frac{\mathsf{H}(\mathbf{p})}{\log \mathsf{D}} + 1 - p_{\min}, \tag{9}$$

where $p_{\min} = \min_i p_i$ denotes the probability of the least likely symbol.

We conjecture that this bound also holds true for $\mathsf{D} \geq 4$.

Equivalently, we could write that the redundancy of Fano codig is upper-bounded by $1 - p_{\min}$. Note that this bound is slightly better than the performance bound of Shannon codes [1]:

$$\mathsf{L}_{\text{Shannon}}(\mathbf{p}) < \frac{\mathsf{H}(\mathbf{p})}{\log \mathsf{D}} + 1. \tag{10}$$

III. Entropy and Variational Distance

In this section, we compare *Shannon entropy* and *variational distance*. Based on results from [11], we succeed in finding a new result that we believe is of interest by itself and that might be useful in other areas unrelated to Fano coding, too.

Definition 7: The *Shannon entropy* (or short *entropy*) of an r-ary random variable of PMF $\mathbf{p}$ is defined as

$$\mathsf{H}(\mathbf{p}) = \mathsf{H}(p_1, \ldots, p_r) \triangleq -\sum_{i=1}^{r} p_i \log p_i. \tag{11}$$

For notational convenience we always understand $0 \log 0$ to be equal to 0. It is well-known and straightforward to prove that

$$0 \leq \mathsf{H}(\mathbf{p}) \leq \log r. \tag{12}$$

Definition 8: Let $\mathbf{p}$ and $\mathbf{q}$ be two PMFs over the same finite alphabet $\mathcal{X}$ of size r. The *variational distance* between $\mathbf{p}$ and $\mathbf{q}$ is defined as

$$\mathscr{V}(\mathbf{p}, \mathbf{q}) \triangleq \sum_{i=1}^{r} |p_i - q_i|. \tag{13}$$

It is obvious from the definition that $\mathscr{V}(\mathbf{p}, \mathbf{q}) \geq 0$ with equality if, and only if, $\mathbf{p} = \mathbf{q}$. While maybe slightly less obvious, it follows from the triangle inequality that $\mathscr{V}(\mathbf{p}, \mathbf{q}) \leq 2$.

A. Relation between $\mathsf{H}(\cdot)$ and $\mathscr{V}(\cdot, \cdot)$

Since $\mathscr{V}(\cdot, \cdot)$ satisfies all required conditions of a norm, it is correct to think of the variational distance $\mathscr{V}(\mathbf{p}, \mathbf{q})$ as a distance between $\mathbf{p}$ and $\mathbf{q}$. It describes how similar (or different) two random experiments are. Note, however, that for the situation when there are no constraints on the (finite) alphabet size, a small variational distance does not guarantee similar entropy values: even if $\mathscr{V}(\mathbf{p}, \mathbf{q}) < \epsilon$, it is still possible that $\mathsf{H}(\mathbf{p}) - \mathsf{H}(\mathbf{q}) > \delta$, for any choices of $\epsilon, \delta > 0$ [11, Theorem 1]. Once we fix the alphabet size $|\mathcal{X}|$, this cannot happen anymore.

In the remainder of this section and without loss of generality, we assume that the given PMF $\mathbf{p} = (p_1, \ldots, p_{|\mathcal{X}|})$ is ordered such that

$$p_1 \geq p_2 \geq \cdots \geq p_r > 0 = p_{r+1} = \cdots = p_{|\mathcal{X}|}, \tag{14}$$

where r denotes the support size of $\mathbf{p}$. Moreover, we use

$$(\cdot)^+ \triangleq \max\{\cdot, 0\}. \tag{15}$$

The following results have been derived in [11]. They answer the question of how to maximize (or minimize) entropy $\mathsf{H}(\mathbf{q})$ under a given similarity constraint, i.e., $\mathbf{q}$ should be similar to a given $\mathbf{p}$: $\mathscr{V}(\mathbf{p}, \mathbf{q}) \leq \epsilon$.

Proposition 9 ([11, Theorem 2]): For some given $0 \leq \epsilon \leq 2$ and some $\mathbf{p}$ satisfying (14), choose $\mu, \nu \in \mathbb{R}$ such that

$$\sum_{i=1}^{|\mathcal{X}|} (p_i - \mu)^+ = \frac{\epsilon}{2} \tag{16}$$

and

$$\sum_{i=1}^{|\mathcal{X}|} (\nu - p_i)^+ = \frac{\epsilon}{2}. \tag{17}$$

If $\nu \geq \mu$, define $\mathbf{q}_{\text{max}}$ to be the uniform distribution on $\mathcal{X}$,

$$q_{\text{max},i} \triangleq \frac{1}{|\mathcal{X}|}, \quad i = 1, \ldots, |\mathcal{X}|, \tag{18}$$

and if $\nu < \mu$, define $\mathbf{q}_{\text{max}}$ as

$$q_{\text{max},i} \triangleq \begin{cases} \mu & \text{if } p_i > \mu, \\ p_i & \text{if } \nu \leq p_i \leq \mu, \\ \nu & \text{if } p_i < \nu, \end{cases} \quad i = 1, \ldots, |\mathcal{X}|. \tag{19}$$

Then

$$\max_{\mathbf{q}\colon\, \mathscr{V}(\mathbf{p},\mathbf{q}) \leq \epsilon} \mathsf{H}(\mathbf{q}) = \mathsf{H}(\mathbf{q}_{\text{max}}). \tag{20}$$

Note the structure of the maximizing distribution: we cut the largest values of $\mathbf{p}$ to a constant level μ and add this probability to the smallest values to make them all constant equal to ν. The middle range of the probabilities is not touched. So, under the constraint that we cannot twiddle $\mathbf{p}$ too much, we should try to approach a uniform distribution by equalizing the extremes. See Fig. 6 for an illustration of this.

It is quite obvious that $\mathsf{H}(\mathbf{q}_{\text{max}})$ depends on the given ϵ. Therefore for a given $\mathbf{p}$ and for $0 \leq \epsilon \leq 2$, we define

$$\psi_{\mathbf{p}}(\epsilon) \triangleq \mathsf{H}(\mathbf{q}_{\text{max}}) \tag{21}$$

with $\mathbf{q}_{\text{max}}$ given in (18) and (19). One can show that $\psi_{\mathbf{p}}(\epsilon)$ is a concave (and therefore continuous) and strictly increasing function in ϵ.

Proposition 10 ([11, Theorem 3]): Let $0 \leq \epsilon \leq 2$ and $\mathbf{p}$ be given and assume that $\mathbf{p}$ satisfies (14). If $1 - p_1 \leq \frac{\epsilon}{2}$, define

$$\mathbf{q}_{\text{min}} \triangleq (1, 0). \tag{22}$$

Otherwise, let k be the largest integer such that

$$\sum_{i=k}^{r} p_i \geq \frac{\epsilon}{2} \tag{23}$$

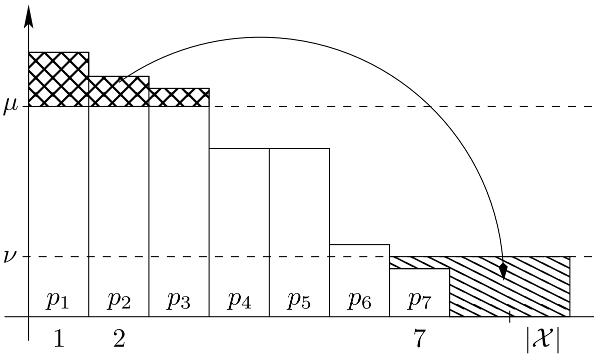

Fig. 6. Example demonstrating how a PMF with seven nonzero probabilities is changed to maximize entropy under a variational distance constraint ($|\mathcal{X}| = 9$, $r = 7$). The maximizing distribution is $\mathbf{q}_{\text{max}} = (\mu, \mu, \mu, p_4, p_5, p_6, \nu, \nu, \nu)$.

and define $\mathbf{q}_{\text{min}}$ as

$$q_{\text{min},i} \triangleq \begin{cases} p_1 + \frac{\epsilon}{2} & \text{if } i = 1, \\ p_i & \text{if } i = 2, \ldots, k-1, \\ \sum_{j=k}^{r} p_j - \frac{\epsilon}{2} & \text{if } i = k, \\ 0 & \text{if } i = k+1, \ldots, |\mathcal{X}|, \end{cases} \\ i = 1, \ldots, |\mathcal{X}|. \tag{24}$$

Then

$$\min_{\mathbf{q}\colon\, \mathscr{V}(\mathbf{p},\mathbf{q}) \leq \epsilon} \mathsf{H}(\mathbf{q}) = \mathsf{H}(\mathbf{q}_{\text{min}}). \tag{25}$$

Note that to minimize entropy, we need to change the PMF to make it more concentrated. To that goal the few smallest probability values are set to zero and the corresponding amount is added to the single largest probability. The middle range of the probabilities is not touched. So, under the constraint that we cannot twiddle $\mathbf{p}$ too much, we should try to approach the $(1, 0, \ldots, 0)$-distribution by removing the tail and enlarge the largest peak. See Fig. 7 for an illustration of this.

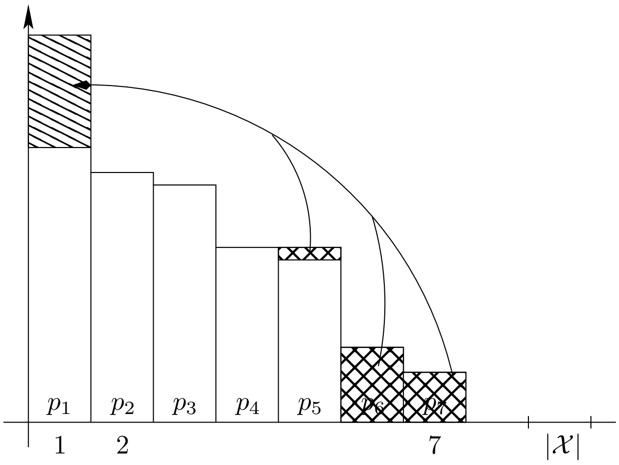

Fig. 7. Example demonstrating how a PMF with seven nonzero probabilities is changed to minimize entropy under a variational distance constraint ($r = 7$). The minimizing distribution is $\mathbf{q}_{\text{min}} = (p_1 + \epsilon/2, p_2, p_3, p_4, p_5 + p_6 + p_7 - \epsilon/2, 0, 0, 0, 0)$.

Also here, $\mathsf{H}(\mathbf{q}_{\text{min}})$ depends on the given ϵ. For a given $\mathbf{p}$ and for $0 \leq \epsilon \leq 2$, we define

$$\varphi_{\mathbf{p}}(\epsilon) \triangleq \mathsf{H}(\mathbf{q}_{\text{min}}) \tag{26}$$

with $\mathbf{q}_{\text{min}}$ defined in (22)–(24). One can show that $\varphi_{\mathbf{p}}(\epsilon)$ is a continuous and strictly decreasing function in ϵ.

We may, of course, also ask the question the other way around: For a given PMF $\mathbf{p}$ and a given entropy value H, what is the choice of a PMF $\mathbf{q}$ such that $\mathsf{H}(\mathbf{q}) = \mathsf{H}$ is achieved and such that $\mathbf{q}$ is most similar to $\mathbf{p}$ with respect to variational distance?

Proposition 11 ([11, Theorem 4]): Let $0 \leq t \leq \log|\mathcal{X}|$ and $\mathbf{p}$ satisfying (14) be given. Then

$$\begin{aligned} &\min_{\mathbf{q}\colon \mathsf{H}(\mathbf{q})=t} \mathscr{V}(\mathbf{p},\mathbf{q}) \\ &= \begin{cases} 2(1-p_1) & \text{if } t = 0, \\ \varphi_{\mathbf{p}}^{-1}(t) & \text{if } 0 < t \leq \mathsf{H}(\mathbf{p}), \\ \psi_{\mathbf{p}}^{-1}(t) & \text{if } \mathsf{H}(\mathbf{p}) < t < \log|\mathcal{X}|, \\ \sum_{i=1}^{r} \left|p_i - \frac{1}{|\mathcal{X}|}\right| + \frac{|\mathcal{X}|-r}{|\mathcal{X}|} & \text{if } t = \log|\mathcal{X}|, \end{cases} \end{aligned} \tag{27}$$

with $\psi_{\mathbf{p}}^{-1}(\cdot)$ and $\varphi_{\mathbf{p}}^{-1}(\cdot)$ being the inverse of the functions defined in (21) and (26), respectively.

Note that this result actually is a direct consequence of Proposition 9 and 10 and the fact that $\psi_{\mathbf{p}}(\epsilon)$ and $\varphi_{\mathbf{p}}(\epsilon)$ both are continuous and monotonic functions that have a unique inverse.

B. Lower Bound on Entropy in Terms of Variational Distance

We will now use the results summarized in Section III-A to derive a new lower bound on entropy. Note that without any assumptions or constraints, the only possible lower bound on entropy is the trivial bound

$$\mathsf{H}(\mathbf{p}) \geq 0. \tag{28}$$

Using the results from the previous section, we can now improve on this lower bound by taking into account the PMF $\mathbf{p}$.

Theorem 12: For a given $r \in \{2,3,\ldots\}$, consider a random variable with support size r and PMF $\mathbf{p} = (p_1, p_2, \ldots, p_r)$. Then the entropy $\mathsf{H}(\mathbf{p})$ can be lower-bounded as follows:

$$\mathsf{H}(\mathbf{p}) \geq \log r - \frac{r \log r}{2(r-1)} \sum_{i=1}^{r} \left| p_i - \frac{1}{r} \right|. \tag{29}$$

This lower bound has a beautiful interpretation: Let $\mathbf{q}_{\text{u}} = \left(\frac{1}{r}, \ldots, \frac{1}{r}\right)$ be the uniform PMF. Then (29) can be rewritten as follows:

$$\mathsf{H}(\mathbf{q}_{\text{u}}) - \mathsf{H}(\mathbf{p}) \leq \mathscr{V}(\mathbf{q}_{\text{u}}, \mathbf{p}) \cdot \frac{r \log r}{2(r-1)}. \tag{30}$$

Now recall that the entropy of a uniformly distributed random variable is equal to the logarithm of the alphabet size, and if the distribution is not uniform, then the entropy is smaller. So, Theorem 12 gives an upper bound on this reduction in terms of the variational distance between the PMF and the uniform PMF.

An immediate consequence of this lower bound on the entropy is as follows.

Corollary 13: For a given $r \in \{2,3,\ldots\}$, consider a random variable with support size r and PMF $\mathbf{p} = (p_1, p_2, \ldots, p_r)$. Then the *relative entropy* (or *Kullback–Leibler divergence*) between $\mathbf{p}$ and the uniform distribution $\mathbf{q}_{\text{u}}$ can be upper-bounded as follows:

$$\mathscr{D}(\mathbf{p} \| \mathbf{q}_{\text{u}}) \leq \frac{r \log r}{2(r-1)} \mathscr{V}(\mathbf{p}, \mathbf{q}_{\text{u}}). \tag{31}$$

Acknowledgment

This work was supported in part by grant VEGA 1/0131/09 and by the Agency of the Slovak Ministry of Education for the Structural Funds of the EU, under project ITMS:26220120007.

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423 and 623–656, Jul. and Oct. 1948.

[2] R. M. Fano, "The transmission of information," Research Laboratory of Electronics, Mass. Inst. of Techn. (MIT), Technical Report No. 65, Mar. 17, 1949.

[3] D. A. Huffman, "A method for the construction of minimum-redundancy codes," *Proc. of IRE*, vol. 40, no. 9, pp. 1098–1101, Sept. 1952.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley, 2006.

[5] J. Sayir, "On coding by probability transformation," Ph.D. dissertation, ETH Zürich, 1999, Diss. ETH No. 13099. [Online]. Available: http://e-collection.ethbib.ethz.ch/view/eth:23000

[6] P. R. Stubley, "On the redundancy of optimum fixed-to-variable length codes," in *Proc. Data Compression Conf.*, Snowbird, UT, USA, Mar. 29–31, 1994, pp. 90–97.

[7] W. Szpankowski, "Asymptotic average redundancy of Huffman (and other) block codes," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2434–2443, Nov. 2000.

[8] N. Merhav and W. Szpankowski, "Average redundancy of the Shannon code for Markov sources," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7186–7193, Nov. 2013.

[9] D. Berend, P. Harremoës, and A. Kontorovich, "Minimum KL-divergence on complements of l_1 balls," *IEEE Trans. Inf. Theory*, vol. 60, no. 6, pp. 3172–3177, Jun. 2014.

[10] I. Sason, "On reverse Pinsker inequalities," Irwin and Joan Jacobs Center for Communication and Information Technologies, Technion, Haifa, Israel, Tech. Rep. CCIT Report #882, Mar. 2015.

[11] S.-W. Ho and R. W. Yeung, "The interplay between entropy and variational distance," *IEEE Trans. Inf. Theory*, vol. 56, no. 12, pp. 5906–5929, Dec. 2010.