A One-to-One Code and Its Anti-redundancy

Wojciech Szpankowski*
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.

Abstract

One-to-one codes are "one shot" codes that assign a distinct codeword to source symbols and are not necessarily prefix codes (more generally, uniquely decodable). For example, such codes arise when there exists an "end of message" channel symbol. Interestingly, as Wyner proved in 1972, for such codes the average code length can be *smaller* than the source entropy. By how much? We call this difference the *anti-redundancy*. Various authors over the years have shown that the anti-redundancy can be as big as minus the logarithm of the source entropy. However, to the best of our knowledge precise estimates do not exist. In this note, we consider a block code of length n generated by a binary memoryless source, and prove that the average anti-redundancy is

$$-\frac{1}{2}\log_2 n + C + H(n) + o(1)$$

where C is a constant and either $H(n) = 0$ if $\log_2(1-p)/p$ is irrational (where p is the probability of generating a "0") or otherwise $H(n)$ is a fluctuating function as the code length increases. This relatively simple finding requires a combination of quite sophisticated analytic tools such as precise evaluation of Bernoulli sums, the saddle point method, and theory of distribution of sequences modulo 1.

Index Terms — Prefix codes, one-to-one codes, average redundancy, Bernoulli sums, saddle point method, distribution of sequences modulo 1.

*This research was supported by NSF Grant CCR-0208709, the NIH grant R01 GM068959-01, and AFOSR Grant FA8655-04-1-3074.

1 Introduction

Traditionally, source coding deals with prefix (or more generally, uniquely decodable) codes that are injection from an alphabet $\mathcal{A}$ into binary strings $\{0,1\}^*$. Already in 1948 Shannon observed that for such codes the average code length cannot be smaller than the entropy of the source. The next natural step is to ask by how much the average code length exceeds the entropy. This is called the average redundancy which is known to be nonnegative for prefix codes. Over the last twenty years a substantial literature was built to address this problem (e.g., [5] for some recent developments).

Occasionally, encodings are not necessarily prefix free. In *one-to-one codes* a distinct codeword is assigned to each source symbol and unique decodability is not required. Such codes are usually one shot codes and there is one designated an"end of message" channel symbol. Wyner [14] in 1972 proved that the average code length L is actually smaller than the source X entropy $H(X)$. A lower bound for the average code length of such codes was first established in [10] and then improved by Alon and Orlitsky [1] who proved that

$$L \geq H(X) - \log(H(X)+1) - \log e. \tag{1}$$

Some recent results on one-to-one codes are reported in [11].

As with prefix codes, one can study the difference between the average code length and the entropy which for one-to-one codes we shall call the *anti-redundancy*. Thus the anti-redundancy is defined as

$$\bar{R} = L - H(X)$$

and from [1] we conclude that $\bar{R} = \Omega(-\log H(X))$. A question arises whether this lower bound is a universal one for a class of sources. Alon and Orlitsky [1] showed that the lower bound is achievable for the geometric distribution. In this note we consider a block one-to-one code generated by a binary memoryless source over $\{0,1\}^n$ and analyze precisely the average anti-redundancy $\bar{R}_n$.

Let us briefly discuss our main results. We consider a source sequence $X_1^n = X_1 \dots X_n$ generated by a binary memoryless source with p being the probability of generating a "0". We assume $p \leq 1-p := q$ and order all probabilities $p^k(1-p)^{n-k}$ in a nondecreasing fashion assigning $\lfloor \log_2 j \rfloor$ to the jth message where $1 \leq j \leq 2^n$. Observe that for every $1 \leq k \leq n$ there are $\binom{n}{k}$ messages of the same probability that we order randomly. Our goal is to estimate the average code length

$$L_n = \sum_k p^k (1-p)^{n-k} \lfloor \log_2 j \rfloor$$

where $1 \leq j \leq 2^n$ and the summation is over all 2^n strings of length n (see next section for a precise definition). We shall prove that for $p < 1/2$

$$L_n = nH(p) - \frac{1}{2}\log_2 n + C + F(n) + o(1)$$

where $H(p) = -p\log_2 p - (1-p)\log_2(1-p)$, C is an explicitly computable constant, and $F(n) \equiv 0$ when $\log_2(1-p)/p$ is irrational and $F(n)$ is a fluctuating function of n when $\log_2(1-p)/p$ is rational. Thus the floor function appearing in the formula of L_n contributes fluctuation only to the third order asymptotic expansion.

To obtain our main result we need a battery of sophisticated analytic techniques. Namely, a formula to deal with sums of floor functions, asymptotics for the Bernoulli sums, the saddle point method, and the theory of distribution of sequences modulo 1. In the next section we present our main results that we shall prove in Section 3.

2 Main Results

We consider a binary memoryless source X over the binary alphabet $\mathcal{A} = \{0, 1\}$ generating a sequence $x_1^n = x_1, \ldots, x_n \in \mathcal{A}^n$. Let p and $q = 1 - p$ be the probabilities of generating a "0" and "1", respectively. Throughout this paper we assume that $p \le q$. Then $P(x_1^n) = p^k q^{n-k}$, where k is the number of 0s in x_1^n. We now list all 2^n probabilities in a nonincreasing order

$$q^n \left(\frac{p}{q}\right)^0 \ge q^n \left(\frac{p}{q}\right)^1 \ge \ldots \ge q^n \left(\frac{p}{q}\right)^n. \tag{2}$$

Let us assign consecutive natural numbers j $(1 \le j \le 2^n)$ to each probability on the list of $P(x_1^n)$. Clearly, there are $\binom{n}{k}$ equal probabilities $p^k q^{n-k}$. Define

$$A_k = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{k}, \qquad A_{-1} = 0.$$

Starting from the position A_{k-1} on the nonincreasing list of $P(x_1^n)$, the next $\binom{n}{k}$ probabilities are the same and equal to $p^k q^{n-k}$. We assign the code length

$$\lfloor \log_2(j) \rfloor = \lfloor \log_2(A_{k-1} + i) \rfloor$$

to the jth probability, where $1 \le i \le \binom{n}{k}$. Thus the average code length is

$$\begin{aligned} L_n &= \sum_{k=0}^{n} p^k q^{n-k} \sum_{j=A_{k-1}+1}^{A_k} \lfloor \log_2(j) \rfloor \qquad (3) \\ &= \sum_{k=0}^{n} p^k q^{n-k} \sum_{i=1}^{\binom{n}{k}} \lfloor \log_2(A_{k-1} + i) \rfloor. \end{aligned}$$

Our goal is to estimate L_n asymptotically for large n.

Let us first simplify the above formula for L_n. We need to handle the inner sum that contains the floor function. Define

$$S_n = \sum_{i=1}^{\binom{n}{k}} \lfloor \log_2(A_{k-1} + i) \rfloor.$$

To evaluate this sum we apply the following identity (cf. Knuth [9] Ex. 1.2.4-42)

$$\sum_{j=1}^{N} a_j = Na_n - \sum_{j=1}^{N-1} (a_{j+1} - a_j)$$

for any sequence a_j. After some tedious algebra, we finally reduce the formula for L_n to the following more explicit one

$$\begin{aligned}
L_n &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \lfloor \log_2 A_k \rfloor && (4)\\
&- \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} 2^{-\langle \log_2 A_k \rangle} && (5)\\
&+ \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \frac{1 + A_{k-1}}{\binom{n}{k}} \left(\log_2 \left(1 + \binom{n}{k} A_{k-1}^{-1} \right) + \langle \log_2 A_{k-1} \rangle - \langle \log_2 A_k \rangle \right) && (6)\\
&- 2 \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \frac{A_{k-1}}{\binom{n}{k}} \left(2^{-\langle \log_2 A_k \rangle} - 4 \cdot 2^{-\langle \log_2 A_{k-1} \rangle} \right) && (7)
\end{aligned}$$

where $\langle x \rangle = x - \lfloor x \rfloor$ is the fractional part of x.

In the next section, we evaluate asymptotically sums (4)– (7) leading to our main result of this paper.

Theorem 1 *Consider a binary memoryless source and the one-to-one block code described above. Then for* $p < \frac{1}{2}$

$$\begin{aligned}
L_n &= nH(p) - \frac{1}{2} \log_2 n - 1 - \frac{1}{2 \ln 2} + \log_2 \frac{1-p}{(1-2p)\sqrt{pq\pi}} + \frac{1-p}{1-2p} \log_2 \frac{2-3p}{1-p} && (8)\\
&+ \frac{5-4p}{1-2p} \left(\frac{1}{2 \ln 2} + G(n) \right) + F(n) + o(1)
\end{aligned}$$

where $H(p) = -p \log_2 p - (1-p) \log_2 (1-p)$, *and* $G(n) = F(n) = 0$ *if* $\log_2 \frac{1-p}{p}$ *is irrational. If* $\log_2 \frac{1-p}{p} = N/M$ *for some integers* M, N *such that* $\gcd(N, M) = 1$, *then* $G(n)$ *and* $F(n)$ *are oscillating functions of complicated nature. For example,*

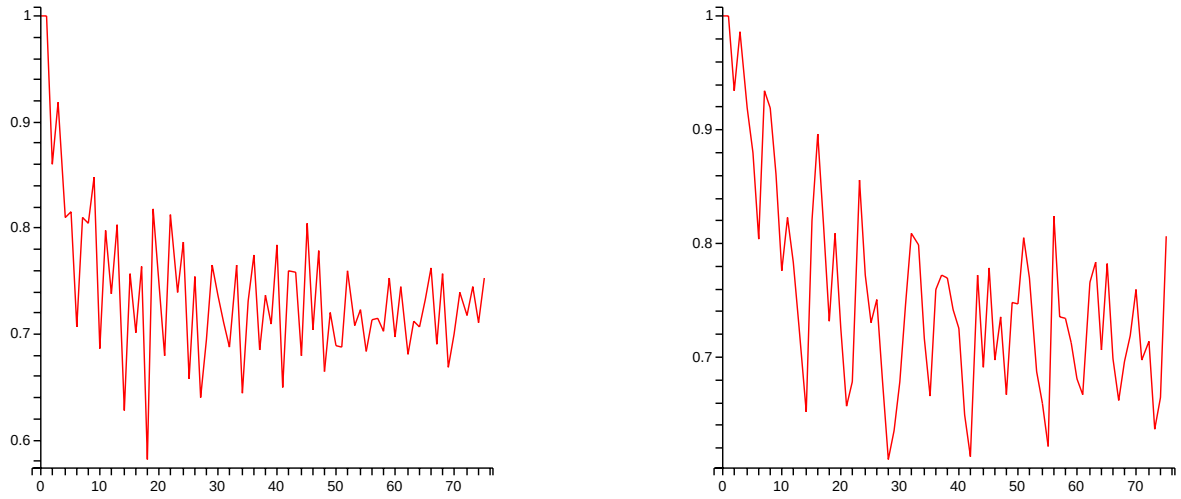
$$F(n) = \frac{1}{M\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} \left(\left\langle M \left(n\beta - \log \left(\frac{1-2p}{1-p} \sqrt{2\pi pqn} \right) - \frac{x^2}{2 \ln 2} \right) \right\rangle - \frac{1}{2} \right) dx$$

where $\beta = -\log_2 (1-p)$.
For $p = \frac{1}{2}$, *then*

$$L_n = nH(1/2) - 1 + 2^{-n}(n-2)$$

for every $n \geq 1$.

In view of the above result, we again see that asymptotic behavior of the redundancy or anti-redundancy depends on the rationality/irrationality of $\log_2 (1-p)/p$. In Figure 2 we plotted



(a) (b)

Figure 1: The "constant" part of the average anti-redundancy versus n for: (a) irrational $\alpha = \log_2(1-p)/p$ with $p = 1/\pi$; (b) rational $\alpha = \log_2(1-p)/p$ with $p = 1/9$.

a "constant" part of the anti-redundancy. We observe change of "mode" when switching from $\alpha = \log_2(1-p)/p$ irrational (cf. Fig. 2(a)) to rational (cf. Fig. 2(b). This phenomenon was already observed in [5, 12] for Huffman and Shannon codes. This indicates that the bounds derived in [1, 14] cannot be improved.

Finally, observe that the lower bound (which is tight as proved in [1]) suggests the leading term of $\bar{R} - n$ to be $-\log n$ (e.g., for geometric distribution). In this paper we prove that for the binomial distribution the average anti-redundancy is asymptotically equal to $-\frac{1}{2}\log n$.

3 Analysis

In this section we analyze asymptotically the four terms of L_n as presented in (4)–(7). We start with (4) that we split as follows

$$\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \lfloor \log_2 A_k \rfloor = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \log_2 A_k - \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \langle \log_2 A_k \rangle,$$

and define

$$a_n \quad = \quad \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \log_2 A_k, \tag{9}$$

$$b_n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \langle \log_2 A_k \rangle. \tag{10}$$

Both sums, and most discussed in this section, fall under the so called *Bernoulli sum* analyzed in [7, 8] as well as in [3, 4]. In order to evaluate these sums we first need to estimate asymptotically A_k around $k = np$ which is presented in the next lemma. Throughout the paper we shall use small positive constants $\delta > 0$ and $\varepsilon > 0$ that can change from line to line.

Lemma 1 *For large n and $p < 1/2$*

$$A_{np} = \frac{1-p}{1-2p} \frac{1}{\sqrt{2\pi np(1-p)}} 2^{nH(p)} \left(1 + O(n^{-1/2})\right). \tag{11}$$

More precisely, for an $\varepsilon > 0$ and $k = np + \Theta(n^{1/2+\varepsilon})$ we have

$$A_k = \frac{1-p}{1-2p} \frac{1}{\sqrt{2\pi np(1-p)}} \left(\frac{1-p}{p}\right)^k \frac{1}{(1-p)^n} \exp\left(-\frac{(k-np)^2}{2p(1-p)n}\right) \left(1 + O(n^{-\delta})\right) \tag{12}$$

for some $\delta > 0$.

Proof. We use the saddle point method [13]. Let's first define the generating function of A_k, that is,

$$A_n(z) = \sum_{k=0}^{n} A_k z^k = \frac{(1+z)^n - 2^n z^{n+1}}{1-z}.$$

Thus by Cauchy's formula [13]

$$\begin{aligned} A_k &= \frac{1}{2\pi i} \oint \frac{(1+z)^n - 2^n z^{n+1}}{1-z} \frac{dz}{z^{k+1}} \\ &= \frac{1}{2\pi i} \oint \frac{1}{1-z} 2^{n \log(1+z) - (k+1)\log z} dz. \end{aligned}$$

Define $h(z) = n \log(1+z) - (k+1) \log z$. Then the saddle point z_0 is such that $h'(z_0) = 0$ and one finds $z_0 = (k+1)/(n-k+1) = p/(1-p)$ while $h''(z_0) = q^3/p$. Thus by the saddle point method

$$A_k = \frac{1}{1-z_0} \frac{1}{\sqrt{2\pi h''(z_0)}} 2^{nh(z_0)} (1 + O(n^{-1/2})).$$

After some algebra we prove (11). In a similar manner, as shown explicitly in [3], we prove (12). ∎

We also need to approximate the binomial distribution around the mean. We shall use the following well known lemma that is a simple consequence of Stirling's approximation.

Lemma 2 *Let $p_n(k) = \binom{n}{k} p^k q^{n-k}$ where $q = 1 - p$ be the binomial distribution. Then for $|k - pn| \le n^{1/2+\varepsilon}$ we have*

$$p_n(k) = \frac{1}{\sqrt{2\pi p(1-p)n}} \exp\left(-\frac{(k-pn)^2}{2p(1-p)n}\right) + O(n^{-\delta}) \tag{13}$$

uniformly as $n \to \infty$. *Furthermore*

$$\sum_{|k-np|>\sqrt{np}n^{1/2+\varepsilon}} p_n(k) < 2n^{-\varepsilon}e^{-n^{2\varepsilon}/2} \tag{14}$$

for large n.

Now we are in a position to estimate a_n and b_n. Observe that based on (14) of Lemma 2 we can restrict the sum to $|k-pn| \le n^{1/2+\varepsilon}$. In fact, by Lemma 1 we have

$$\log A_k = \log A_{np} + \alpha(k-np) - \frac{(k-np)^2}{2pqn\ln 2} + O(n^{-\delta}). \tag{15}$$

Using Lemma 2 we arrive at

$$a_n = \log A_{np} - \frac{1}{2\ln 2} + O(n^{-\delta})$$

which proves the desire result after applying (11).

The above formula could be also concluded from [8] where it was proved that for a large class of functions f (not growing too fast) the following is true

$$S_n = \sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k} f(k) = f(np) + \sum_{i=1}^{\infty}\sum_{j\ge 2i} c_{ij} n^i f^{(j)}(np) \tag{16}$$

where $f^{(j)}(np)$ is the jth derivative of f at np and c_{ij} are constant coefficients.

Now, we deal with the second sum (10), namely

$$b_n = \sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k}\langle \log_2 A_k\rangle.$$

To evaluate it we need a completely different technique already used in [4, 12] to estimate the redundancy of the Huffman code and arithmetic codes. Observe first that from (12) we find that for $|k-pn| \le n^{1/2+\varepsilon}$

$$\log A_k = \alpha k + n\beta - \log_2 \omega\sqrt{n} - \frac{(k-np)^2}{2pqn\ln 2} + O(n^{-\delta})$$

where $\omega = (1-2p)\sqrt{2\pi pq}/(1-p)$. In order to estimate b_n we need to understand asymptotics of the following sum

$$\sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k}\left\langle \alpha k + n\beta - \log_2\omega\sqrt{n} - \frac{(k-np)^2}{2pqn\ln 2}\right\rangle.$$

Asymptotics of the above sum depend upon rationality or irrationality of α as proved in [4] (cf. also [6, 12]). In fact, the next lemma follows directly from the analysis of [4].

Lemma 3 *Let $0 < p < 1$ be a fixed real number and $f : [0,1] \to \mathbf{R}$ be a Riemann integrable function.*

(i) *If α is irrational, then*

$$\lim_{n\to\infty} \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f\left(\left\langle k\alpha + y - (k-np)^2/(2pqn\ln 2)\right\rangle\right) = \int_0^1 f(t)\,dt, \qquad (17)$$

where the convergence is uniform for all shifts $y \in \mathbf{R}$.

(ii) *Suppose that $\alpha = \frac{N}{M}$ is a rational number with integers N, M such that $\gcd(N,M) = 1$. Then uniformly for all $y \in \mathbf{R}$*

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} f\left(\left\langle k\alpha + y - (k-np)^2/(2pqn\ln 2)\right\rangle\right) = \int_0^1 f(t)\,dt + H_M(y) \qquad (18)$$

where

$$H_M(y) := \frac{1}{M}\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} \left(\left\langle M\left(y - \frac{x^2}{2\ln 2}\right)\right\rangle - \int_0^1 f(t)\,dt\right) dx$$

is a periodic function with period $\frac{1}{M}$.

Using this lemma we immediately show that for α irrational

$$b_n = \frac{1}{2} + o(1),$$

while for $\alpha = N/M$ we have

$$b_n = \frac{1}{2} + H_M\left(\beta n - \log_2\left(\frac{1-2p}{1-p}\sqrt{2\pi pqn}\right)\right)$$

as shown in Theorem 1.

Now we consider term (7) which we split into two terms

$$\begin{aligned} c_n &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \frac{1+A_{k-1}}{\binom{n}{k}} \log_2\left(1 + \binom{n}{k} A_{k-1}^{-1}\right), \\ d_n &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \frac{1+A_{k-1}}{\binom{n}{k}} \left(\langle \log_2 A_{k-1}\rangle - \langle \log_2 A_k\rangle\right). \end{aligned}$$

By Lemma 1 for $|k-np| \le n^{1/2+\varepsilon}$ we have

$$\frac{A_k}{\binom{n}{k}} = \frac{1-p}{1-2p} + O(n^{-\delta})$$

for some $\delta > 0$, thus directly

$$d_n = \frac{1-p}{1-2p}\log_2\frac{2-3p}{1-p} + o(1).$$

By Lemma 3 we conclude that $D_n = o(1)$.

Thus, to complete the proof of Theorem 1 we need to evaluate (8) which we recall below

$$e_n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \frac{A_{k-1}}{\binom{n}{k}} \left(2^{-\langle \log_2 A_k \rangle} - 4 \cdot 2^{-\langle \log_2 A_{k-1} \rangle} \right).$$

which can be estimated using Lemma 3. In particular, for α irrational one finds

$$e_n = \frac{1-p}{1-2p} \frac{1}{\ln 2} + o(1).$$

The rational case can be treated in a similar manner, however, the formula is quite complicated. This completes the proof of Theorem 1 since the case $p = 0.5$ is trivial.

Acknowledgment

The author thanks Prof. M. Drmota (TU Wien) for many useful discussions.

References

[1] N. Alon and A. Orlitsky, A Lower Bound on the Expected Length of One-to-One Codes, *IEEE Trans. Information Theory*, 40, 1670-1672, 1994.

[2] T. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York 1991.

[3] M. Drmota, A Bivariate Asymptotic Expansion of Coefficients of Powers of Generating Functions, *Europ. J. Combinatorics*, 15, 139-152, 1994.

[4] M. Drmota, H-K. Hwang, and W. Szpankowski, Precise Average Redundancy of an Idealized Arithmetic Coding, *Proc. Data Compression Conference*, 222-231, Snowbird, 2002.

[5] M. Drmota and W. Szpankowski, Precise Minimax Redundancy and Regret, *IEEE Trans. Information Theory*, 50, No. 11, 2004.

[6] M. Drmota and R. Tichy, *Sequences, Discrepancies, and Applications*, Springer Verlag, Berlin, Heidelberg, 1997.

[7] P. Flajolet, Singularity Analysis and Asymptotics of Bernoulli Sums, *Theoretical Computer Science*, 215, 371–381, 1999.

[8] P. Jacquet and W. Szpankowski, Entropy Computations via Analytic Depoissonization, *IEEE Trans. Information Theory*, 45, 1072-1081, 1999.

[9] D. Knuth, *The Art of Computer Programming: Fundamental Algorithms,* Vol. 1. Addison-Wesley, 1997.

[10] S. K. Leung-Yan-Cheong, T. Cover, Some Equivalences between Shannon Entropy and Kolmogorov Complexity, *IEEE Trans. Information Theory*, 24, 331-338, 1978.

[11] S. Savari and A. Naheta, Bounds on the Expected Cost of One-to-One Codes, *2004 ISIT*, p. 92, Chicago, 2004.

[12] W. Szpankowski, Asymptotic Average Redundancy of Huffman (and Other) Block Codes, *IEEE Trans. Information Theory*, 46, 2434-2443, 2000.

[13] W. Szpankowski, *Average Case Analysis of Algorithms in Sequences*, John Wiley & Sons, New York, 2001.

[14] A. D. Wyner, An Upper Bound on the Entropy Series, *Inform. Control*, 20, 176-181, 1972.