Asymptotic Average Redundancy of Adaptive Block Codes

Yuriy A. Reznik *
RealNetworks, Inc.
2601 Elliott Avenue, Seattle, WA 98121
E-mail: yreznik@real.com

Wojciech Szpankowski†
Department of Computer Science, Purdue University
W. Lafayette, IN 47907
E-mail: spa@cs.purdue.edu

Abstract

Krichevsky has shown that the average redundancy rate of an adaptive block code for memoryless sources is $R = \frac{m-1}{2\,n} \log \frac{n+t}{t+1} + O\left(\frac{1}{n}\right)$, where m is a cardinality of source's alphabet, n is a block size, and t is a size of a sample used to construct this code. This led him to a conclusion that using samples with $t = n$ is sufficient to make $R = O\left(\frac{1}{n}\right)$, which is the order of the redundancy rate of block codes for a known source.

In this paper we obtain an exact expression for an $O\left(\frac{1}{n}\right)$ term in Krichevsky's formula, which makes it possible to perform a much more accurate comparison between adaptive block codes and codes for known sources, such as block-based Shannon, Gilbert-Moore, or Huffman codes.

1 Introduction

In order to construct a *minimum-redundancy code* for a memoryless source one needs to know its exact parameters (probabilities of symbols). When such parameters are not known, the best option available is to use *universal codes* – ones that minimize the worst case redundancy for a class of sources [6, 3, 14].

Adaptive codes represent a transitional case: while the exact parameters of a source are not known, one still can obtain a finite length sequence of symbols (or a *sample*) produced by the source. This sequence can be used by both encoder and decoder to construct codes that are more efficient in processing this particular source than pure universal codes.

The idea and the original analysis of such *sample-based universal block codes* (or *adaptive block codes*) belong to R. E. Krichevsky [12]. In particular, he has shown

*On leave from the Institute of Mathematical Machines and Systems, Kiev, Ukraine.

†This research was supported in part by NSF Grants NCR-9415491 and C-CR-9804760, and contract 1419991431A from sponsors of CERIAS at Purdue.

(cf. [13, Theorem 1], [14, Theorem 3.4.1]), that the average redundancy rate of an adaptive block code is asymptotically

$$R = \frac{m-1}{2\,n} \log \frac{n+t}{t+1} + O\left(\frac{1}{n}\right) , \tag{1}$$

where m is a cardinality of the source's alphabet, n is a block size, and t is a size of a sample used to construct this code.

From (1) it is clear, that by using samples of length $t = n$ it is possible to lower the total redundancy rate of such a code to $R = O\left(\frac{1}{n}\right)$, which is the order of the redundancy rate of block codes for a known source (such as block-based Shannon, Gilbert-Moore, or Huffman codes [2, 14, 16]).

However, in order to be able to compare adaptive codes and codes for a known source more precisely one clearly needs to know the exact expression of an $O\left(\frac{1}{n}\right)$ term in (1). Obtaining the exact expression for this and subsequent terms describing the average behavior of adaptive block codes will be the main goal of this paper.

2 Definitions and Main Result

Consider a memoryless (or Bernoulli) source S producing symbols from a binary alphabet $A = \{0, 1\}$ with probabilities $P = \{p, q\}$, $0 < p, q < 1$, $p + q = 1$. By $H(S)$ we denote the (Shannon's) entropy of this source [2]:

$$H(S) = -p \log p - (1-p) \log (1-p) \, . \tag{2}$$

If a is a word of length $|a|$ produced by the source S, then its probability:

$$P_S\,(a) = p^{\,r_0(a)}\,q^{\,r_1(a)} \, . \tag{3}$$

where $r_0(a)$ and $r_1(a)$ denote the numbers of symbols 0 and 1 correspondingly in a word a. Clearly, $r_0(a) + r_1(a) = |a|$.

In a case when the parameters of the source are not known, one still can estimate the probability of a given word a by using the Krichevsky-Trofimov (or KT) formula [13]:

$$P_{KT}\,(a) = \frac{\Gamma\left(r_0(a) + 1/2\right)\,\Gamma\left(r_1(a) + 1/2\right)}{\pi \Gamma\left(|a| + 1\right)} , \tag{4}$$

where $\Gamma(x)$ is a Γ-function. In fact, this formula follows from sequential use of so-called "add-1/2" rule for estimating individual symbols' probabilities ([15]), and it is well known that it guarantees a uniform (in p) convergence to the actual probability (3) when the length of a word $|w| \to \infty$.

Consider now a situation when in a addition to a word a to be encoded, we also have access to a *sample* sequence w of length $|w| = t$ produced by the same source S. The idea of *adaptive coding* is to construct (unique for each sample w) codes ϕ_w, which map blocks of symbols $a \in A^n$ of length $|a| = n$ from S into *uniquely decipherable* codewords $\phi_w(a) \in \{0,1\}^*$.

The average rate (or per-symbol cost) of such a code for a source S is:

$$C\,(n, t, S) = \frac{1}{n} \sum_{|w|=t} \sum_{|a|=n} P_S(w) P_S(a)\,|\phi_w(a)| \ , \tag{5}$$

and the main goal is to find codes ϕ_w such that their *average redundancy rate*:

$$R\left(n,t,S\right) = C\left(n,t,S\right) - H(S)\,, \tag{6}$$

is minimal.

Krichevsky has shown that the optimal (within a $+1$ difference in rate) adaptive block code ϕ_w produces codewords with lengths:

$$\left|\phi_w(a)\right| = \left\lceil -\log P_{KT}\left(a|w\right)\right\rceil\,, \tag{7}$$

where

$$P_{KT}\left(a|w\right) = \frac{P_{KT}\left(w\,a\right)}{P_K T\left(w\right)}\,, \tag{8}$$

is a KT-estimated *conditional probability* of a word a:

$$P_S\left(a|w\right) = \frac{P_S\left(w\,a\right)}{P_S\left(w\right)}\,. \tag{9}$$

Our main result regarding the asymptotic average redundancy rate of adaptive block codes is formulated below.

Theorem 1 *The average redundancy rate of an adaptive block code for a binary memoryless source has the following asymptotic expansion:*

$$\begin{aligned} R\left(n,t,S\right) &= \frac{1}{n}\left\{\frac{1}{2}\log\frac{t+n}{t} + 1 - \Delta(n,t,S) + \frac{1-4\,p\,q}{24\,pq}\,\frac{n}{t\,(t+n)}\right. \\ &\left. - \frac{1-3\,pq}{24\,p^2q^2}\,\frac{(n+2\,t)\,n}{t^2\,(t+n)^2} + O\left(\frac{1}{t^3}+\frac{1}{(t+n)^3}\right)\right\}\,, \end{aligned} \tag{10}$$

where n is a block size, and t is a size of a sample used to construct this code, $p, q = 1-p$ are probabilities of symbols 0 and 1 correspondingly, and $\Delta(n,t,S)$ is an oscillating function of bounded magnitude $|\Delta(n,t,S) - 1/2| \leqslant 1/2$, and which exact formula is given by:

$$\begin{aligned} \Delta(n,t,S) &= \sum_{s=0}^{t}\sum_{k=0}^{n}\binom{n}{k}p^k\,q^{n-k}\binom{t}{s}p^s\,q^{t-s}\left\langle\log\frac{\Gamma\left(s+1/2\right)\,\Gamma\left(t-s+1/2\right)}{\Gamma\left(t+1\right)}\right. \\ &\left. -\log\frac{\Gamma\left(s+k+1/2\right)\,\Gamma\left(t-s+n-k+1/2\right)}{\Gamma\left(t+n+1\right)}\right\rangle\,. \end{aligned} \tag{11}$$

3 Analysis

We need to evaluate asymptotic behavior of the following expression:

$$C\left(n,t,S\right) = \frac{1}{n}\sum_{|w|=t}\sum_{|a|=n} P_S(w)P_S(a)\left\lceil -\log P_{KT}\left(a|w\right)\right\rceil\,. \tag{12}$$

Using an identity $\lceil x\rceil = 1 + \lfloor x\rfloor = 1 + x - \langle x\rangle$, where x is a positive real (non-integer) number, and $\langle x\rangle$ is its fractional part, we can rewrite (12) as:

$$C\left(n,t,S\right) = \frac{1}{n}\left[1 - \sum_{|w|=t}\sum_{|a|=n} P_S(w)P_S(a)\log P_{KT}\left(a|w\right) - \Delta(n,t,S)\right]\,, \tag{13}$$

where

$$\Delta(n,t,S) \quad = \quad \sum_{|w|=t} \sum_{|a|=n} P_S(w) P_S(a) \left\langle -\log P_{KT}\left(a|w\right) \right\rangle \,. \tag{14}$$

Based on (8) and the fact that $P_S(w)P_S(a) = P_S(w\,a)$ (S is memoryless), we can split the central sum in (13) into:

$$\begin{aligned}
&-\sum_{|w|=t} \sum_{|a|=n} P_S(w) P_S(a) \log P_{KT}\left(a|w\right) = \\
&\quad -\sum_{|w|=t} \sum_{|a|=n} P_S(w) P_S(a) \left(\log P_{KT}\left(w\,a\right) - \log P_{KT}\left(w\right)\right) \\
&\quad = -\sum_{|wa|=t+n} P_S(wa) \log P_{KT}\left(w\,a\right) + \sum_{|w|=t} P_S(w) \log P_{KT}\left(w\right) \\
&\quad = (t+n)\, C_{KT}(t+n,S) - t\, C_{KT}(t,S)\,,
\end{aligned} \tag{15}$$

where

$$C_{KT}(n,S) = -\frac{1}{n} \sum_{|a|=n} P_S(a) \log P_{KT}\left(a\right)\,, \tag{16}$$

is the average rate of the KT-estimator processing n-symbols words produced by S.

3.1 Asymptotic average rates of empirical entropy and KT-estimator

Consider a word a, produced by the source S, such that $|a| = n$, and $r_0(a) = k$, for some integers $0 \leqslant k \leqslant n$. Then, KT-estimated probability of a

$$P_{KT}\left(a\right) = \frac{\Gamma\left(k+1/2\right)\,\Gamma\left(n-k+1/2\right)}{\pi \Gamma\left(n+1\right)}\,. \tag{17}$$

Using Stirling's approximation (and excluding cases when $k = 0, n$), we can show that:

$$\begin{aligned}
-\log P_{KT}\left(a\right) \quad &= \quad n\,F\left(a\right) + \frac{1}{2}\log n + \frac{1}{2}\log\frac{\pi}{2} + \frac{1}{12\,n} \\
&\quad + \frac{1}{24\,k} + \frac{1}{24\left(n-k\right)} + O\left(\frac{1}{k^3} + \frac{1}{(n-k)^3}\right)\,,
\end{aligned} \tag{18}$$

where:

$$F\left(a\right) = -\frac{k}{n}\log\left(\frac{k}{n}\right) - \frac{n-k}{n}\log\left(\frac{n-k}{n}\right)\,. \tag{19}$$

is an empirical entropy (or Fitingof entropy) [**?**, 13] of a word a.

The average rate of the empirical entropy $F(a)$ under source S is:

$$\begin{aligned}
&\sum_{|a|=n} P_S(a)F(a) \\
&= -\sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k}\left[\frac{k}{n}\log\left(\frac{k}{n}\right)+\frac{n-k}{n}\log\left(\frac{n-k}{n}\right)\right] \\
&= \log n - \sum_{k=1}^{n}\binom{n}{k}p^k q^{n-k}\frac{k}{n}\log(k) - \sum_{k=0}^{n-1}\binom{n}{k}p^k q^{n-k}\frac{n-k}{n}\log(n-k) \\
&= \log n - p\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1}q^{n-k}\log(k) - q\sum_{k=0}^{n-1}\binom{n-1}{k}p^k q^{n-1-k}\log(n-k) \\
&= \log n - p\, f(n-1,k,p) - q\, f(n-1,k,q)\,. \qquad (20)
\end{aligned}$$

where:

$$f(n,k,\theta) = \sum_{k=0}^{n}\binom{n}{k}\theta^k(1-\theta)^{n-k}\log(k+1)\ . \qquad (21)$$

We immediately notice that (21) belongs to a class of binomial sums that arise in probability theory (see, e.g. [1, p. 92]), and that for large n we must have (a uniform in θ) convergence $f(n,k,\theta)\to\log(1+\theta\, n)$). However, in order to obtain a more detailed asymptotic of $f(n,k,\theta)$ one must use analytic techniques, such as analytic depoissonization [9, 10], or singularity analysis of generating functions [8].

In fact, the last approach in application to a class of polylogarithmic Bernoulli sums has been recently used by P. Flajolet [7], and in particular, he has shown that

$$\sum_{k=1}^{n}\binom{n}{k}\theta^k(1-\theta)^{n-k}\log k = \log(\theta\, n) + \frac{\theta-1}{2\,\theta\, n} - \frac{\theta^2-6\,\theta+5}{12\,\theta^2 n^2} + O\left(\frac{1}{n^3}\right),$$

which is a very similar sum to one that we need to evaluate (21). To take advantage of this existing result, we simply replace $\log(1+k)$ in (21) with:

$$\log(k+1) = \log(k) + \frac{1}{k+1} + \frac{1}{2}\frac{1}{(k+1)(k+2)} + \frac{5}{6}\frac{1}{(k+1)(k+2)(k+3)} + \ldots \qquad (22)$$

The evaluation of the sums containing factorial powers of k yields:

$$\begin{aligned}
S_1(n,\theta) &= \sum_{k=1}^{n}\binom{n}{k}\theta^k(1-\theta)^{n-k}\frac{1}{k+1} \\
\binom{n+1}{k+1}\theta^{k+1}(1-\theta)^{n-k} &= \frac{1}{\theta\,(n+1)}\left[1-(1-\theta)^n-\theta\, n\,(1-\theta)^n\right] \\
&= \frac{1}{\theta\, n} - \frac{1}{\theta\, n^2} + O\left(\frac{1}{n^3}\right), \qquad (23) \\
S_2(n,\theta) &= \sum_{k=1}^{n}\binom{n}{k}\theta^k(1-\theta)^{n-k}\frac{1}{(k+1)(k+2)} \\
&= \frac{1}{\theta^2\,(n+1)\,(n+2)}\left[1-(1-\theta)^n-\theta\, n\,(1-\theta)^n-\frac{\theta^2\, n\,(n+1)}{2}(1-\theta)^n\right] \\
&= \frac{1}{\theta^2\, n^2} + O\left(\frac{1}{n^3}\right), \qquad (24)
\end{aligned}$$

and it is clear that the contribution of the subsequent terms in (22) to the sum (21) is within $O\left(\frac{1}{n^3}\right)$.

Combining the above formulae, we obtain

$$f(n,k,\theta) = \log(\theta\, n) + \frac{1+\theta}{2\,\theta\, n} - \frac{\theta^2 + 6\,\theta - 1}{12\,\theta^2 n^2} + O\left(\frac{1}{n^3}\right), \tag{25}$$

and subsequently (after plugging (25) in (20) and some simple algebra):

$$\sum_{|a|=n} P_S(a) F(a) = H(S) - \frac{1}{2\,n} + \frac{pq-1}{12\,p\,q\,n^2} + O\left(\frac{1}{n^3}\right), \tag{26}$$

which is an (up to $O\left(\frac{1}{n^3}\right)$-term) accurate asymptotic expression for the average rate of empirical entropy.

Remark 1 The average rate of empirical entropy has already been studied by Krichevsky (cf. [13, Lemma 1] and [14, Lemma 3.2.1]). His conclusion was that:

$$-\frac{1}{n} \leqslant \sum_{|a|=n} P_S(a) F(a) - H(S) \leqslant 0\,.$$

Our formula (26) confirms and refines this statement.

We now focus our attention on the average rate of the KT-estimator (16). Using our asymptotic expression (18) and replacing $\frac{1}{k}$ and $\frac{1}{n-k}$ with the appropriate factorial powers we can show that:

$$\begin{aligned} C_{KT}(n,S) &= -\frac{1}{n} \sum_{|a|=n} P_S(a) \log P_{KT}(a) \\ &= \sum_{|a|=n} P_S(a) F(a) + \frac{1}{n} \left\{ \frac{1}{2} \log n + \frac{1}{2} \log \frac{\pi}{2} + \frac{1}{12\,n} \right. \\ &\qquad \left. + \frac{1}{24} \left[S_1(n,p) + S_2(n,p) \right] + \frac{1}{24} \left[S_1(n,q) + S_2(n,q) \right] \right\} \\ &\qquad + O\left(\frac{1}{n^4}\right), \end{aligned} \tag{27}$$

where $S_1(n,\theta)$ and $S_2(n,\theta)$ are already familiar sums (23) and (24).

Now by using (26) and expanding all the expressions in (27) we finally obtain:

$$\begin{aligned} C_{KT}(n,S) &= H(S) + \frac{1}{2\,n} \left\{ \log n + \log \frac{\pi}{2} - 1 - \frac{1 - 4\,p\,q}{12\,pq\,n} + \frac{1 - 3\,pq}{12\,p^2 q^2\, n^2} \right\} \\ &\quad + O\left(\frac{1}{n^4}\right). \end{aligned} \tag{28}$$

3.2 Asymptotic average rate of the adaptive block code

Recall, that according to (13-16):

$$C\,(n,t,S) = \frac{1}{n} \left[1 + (t+n)\, C_{KT}(t+n,S) - t\, C_{KT}(t,S) - \Delta(n,t,S) \right], \tag{28}$$

where $C_{KT}(n,S)$ is the average rate of the KT-estimator (16) and $\Delta(n,t,S)$ is a fluctuating term studied in the previous section ().

By applying our asymptotic result (28) for $C_{KT}(n,S)$ and combining the remaining (after some cancellations) terms we arrive at

$$\begin{aligned} C\,(n,t,S) &= H(S) + \frac{1}{n}\left\{\frac{1}{2}\log\frac{t+n}{t} + 1 - \Delta(n,t,S) + \frac{1-4\,p\,q}{24\,pq}\,\frac{n}{t\,(t+n)}\right. \\ &\qquad \left. - \frac{1-3\,pq}{24\,p^2q^2}\,\frac{(n+2\,t)\,n}{t^2\,(t+n)^2} + O\left(\frac{1}{t^3} + \frac{1}{(t+n)^3}\right)\right\}, \end{aligned}$$

which is exactly the formula (10) claimed by our theorem.

References

[1] A. A. Borovkov, *Probability Theory* (Gordon & Breach, 1998)

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

[3] L. D. Davisson, Universal Noiseless Coding, *IEEE Trans. Inform. Theory*, 19 (6) (1973) 783–795.

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

[5] M. Drmota and R. Tichy, *Sequences, Discrepancies, and Applications* (Springer-Verlag, Berlin, 1997).

[6] B. M. Fitingof, Optimal Coding in the Case of Unknown and Changing Message Statistics, *Probl. Inform. Transm.*, 2, (2) (1965) 3–11 (in Russian) 1–7 (English Transl.)

[7] P. Flajolet, Singularity analysis and asymptotics of Bernoulli sums, *Theoretical Computer Science*, 215 (1999) 371–381.

[8] P. Flajolet and A. Odlyzko, Singularity analysis of generating functions, *SIAM J. Discrete Math.*, 3 (2) (1990) 216–240.

[9] P. Jacquet, and W. Szpankowski, Entropy computations via analytic depoissonization, Tech. Report CSD-TR-96-085, Purdue University (1996)

[10] P. Jacquet, and W. Szpankowski, Analytical depoissonization and its applications, *Theoretical Computer Science*, 201 (1-2) (1998) 1–62.

[11] D. Knuth, *The Art of Computer Programming. Sorting and Searching. Vol. 3* (Addison-Wesley, Reading MA, 1973).

[12] R. E. Krichevsky, Optimal Source Coding Based on Observation, *Probl. Inform. Trans.*, 11 (1) (1975) 37–48 (in Russian).

[13] R. E. Krichevsky and V. K. Trofimov, The Performance of Universal Encoding, *IEEE Trans. Information Theory*, **27** (1981) 199–207.

[14] R. E. Krichevsky, Universal Data Compression and Retrieval. (Kluwer, Norwell, MA, 1993).

[15] R. E. Krichevskiy, Laplace's Law of Succession and Universal Encoding, *IEEE Trans. Information Theory*, **44** (1998) 296–303.

[16] W. Szpankowski, Asymptotic Average Redundancy of Huffman (and Other) Block Codes, *IEEE Trans. Information Theory*, **46** (7) (2000) 2434–2443.

[17] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, (John Wiley & Sons, New York, 2001).