On the Excess Entropy of a Mixture of Sources and the Average Redundancy Rate of Adaptive Block Codes

Yuriy A. Reznik *
RealNetworks, Inc.
2601 Elliott Avenue, Suite 1000
Seattle, WA 98121
E-mail: yreznik@ieee.org

Anatoly V. Anisimov
Faculty of Cybernetics, Kiev University
2 Academician Glushkov Avenue
03680, Kiev, Ukraine
E-mail: ava@mi.unicyb.kiev.ua

Abstract

It is well known that the average redundancy rate of sample-based universal block codes is monotonically decreasing with the length of samples used for their construction. We show, that this is no longer true if samples and blocks of symbols to be compressed are produced by two different sources. In this case, the formula for the average redundancy receives an additional term which changes the overall dependency on the sample length ℓ, such that there exists a point $\ell = \ell^*$, where the redundancy is minimal. This optimal sample size is asymptotically: $\ell^* = \frac{m-1}{2\,D(T\,||\,S)} + O\left(\frac{1}{n}\right)$, where m is a cardinality of the alphabet, n is a block size, and $D\,(T\,||\,S)$ is the relative entropy between the sources T, S producing samples and input data correspondingly. We use this finding to show how to design optimal sequential and random-access-capable compression systems based on adaptive block 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 [11].

Sample-based (or *adaptive*) block codes represent a transitional case: while the exact parameters of the source are not known, one can obtain finite length sequences of symbols (or *samples*) produced by this source in the past. It turns out, that using such samples it is possible to construct codes that are more efficient in encoding this particular source than pure universal codes [8, 9].

In this paper we will show that under certain conditions *sample-based block codes can outperform universal codes even in a case when samples are produced by a different source* from one we are about to encode. We establish this fact by deriving an asymptotic expression for the average redundancy rate of adaptive block codes under mixed sources, which generalizes the previous well-known result of Krichevsky (cf. [10, Theorem 1], [11, Theorem 3.4.1]).

The most interesting new property of such codes is that under different sources their average redundancy can be minimized by properly selecting the length of samples used for their construction. We will obtain an asymptotic formula for the optimal sample length, and will show how it can be used to guide the design of practical data compression schemes based on the adaptive codes.

2 Parameters of Pure and Mixed Memoryless Sources

Let Ω be a class of memoryless sources over some finite alphabet A, $|A| = m$, $2 \leqslant m < \infty$. By $P_X(\alpha)$ we denote a probability with which a source $X \in \Omega$ produces a symbol $\alpha \in A$. We assume that $P_X(\alpha)$ has the standard properties: $0 \leqslant P_X(\alpha) \leqslant 1$ ($\forall\, \alpha \in A$), and $\sum_{\alpha\in A} P_X(\alpha) = 1$. By $H(X)$ we denote the *entropy* of the source X:

$$H(X) = -\sum_{\alpha\in A} P_X(\alpha)\log P_X(\alpha)\,, \qquad (1)$$

and by $D\,(X\,||\,Y)$ we denote the *relative entropy* (or *Kullback-Leibler distance*) between two sources $X, Y \in \Omega$:

$$D\,(X\,||\,Y) = \sum_{\alpha\in A} P_X(\alpha)\log\frac{P_X(\alpha)}{P_Y(\alpha)}\,. \qquad (2)$$

The above parameters are standard in the information theory and readers are referred to textbooks

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

[2] or [16] for discussions of their properties and means.

Consider now two sources $X, Y \in \Omega$, and a real number γ, such that $0 \leqslant \gamma \leqslant 1$. By $X \overset{\gamma}{\bowtie} Y$ we denote a stochastic source, which probabilities represent weighted (with parameter γ) average of probabilities of sources Y and Y ($\forall \alpha \in A$):

$$P_{X \overset{\gamma}{\bowtie} Y}(\alpha) = \gamma\, P_X(\alpha) + (1-\gamma)\, P_Y(\alpha)\,. \qquad (3)$$

Indeed, $X \overset{\gamma}{\bowtie} Y \in \Omega$, and we call it a *mixed source* based on sources X and Y.

From (3) and concavity of the entropy (1), it follows that:

$$H\left(X \overset{\gamma}{\bowtie} Y\right) \geqslant \gamma\, H(X) + (1-\gamma)\, H(Y)\,,$$

which illustrates a well-known effect of increase of the entropy due to the mixture of sources [2].

Our main attention, however, will be focused on the residual quantity:

$$\Delta\left(X \overset{\gamma}{\bowtie} Y\right) = H\left(X \overset{\gamma}{\bowtie} Y\right) - \gamma\, H(X) - (1-\gamma)\, H(Y)\,, \qquad (4)$$

which will call the *excess entropy* of a mixed source $X \overset{\gamma}{\bowtie} Y$.

The following lemma establishes several simple relations between $\Delta\left(X \overset{\gamma}{\bowtie} Y\right)$ and relative entropies between the original and mixed sources.

Lemma 1. *The excess entropy* $\Delta\left(X \overset{\gamma}{\bowtie} Y\right)$ *has the following properties:*

$$\begin{aligned} \Delta\left(X \overset{\gamma}{\bowtie} Y\right) &= \gamma\, D\left(X \,||\, X \overset{\gamma}{\bowtie} Y\right) \qquad (5) \\ &\quad + (1-\gamma)\, D\left(Y \,||\, X \overset{\gamma}{\bowtie} Y\right)\,, \\ \lim_{\gamma \to 0} \frac{\Delta\left(X \overset{\gamma}{\bowtie} Y\right)}{\gamma} &= D\left(X \,||\, Y\right)\,, \qquad (6) \\ \lim_{\gamma \to 1} \frac{\Delta\left(X \overset{\gamma}{\bowtie} Y\right)}{1-\gamma} &= D\left(Y \,||\, X\right)\,. \qquad (7) \end{aligned}$$

In our subsequent discussion we will show that the excess entropy $\Delta(\,.\,)$ arises in the analysis of *sample-based universal block codes* [10]. Limit properties (6) and (7) will be crucial for understanding asymptotic behavior of these codes.

3 Block Codes for Known, Unknown, and Partially Known Sources

Consider a word w of length $|w| = n$ produced by a source $S \in \Omega$. Since S is memoryless, the probability of w is a product of probabilities of its letters, i.e.:

$$P_S\left(w\right) = \prod_{\alpha \in A} P_S(\alpha)^{r_\alpha(w)}\,, \qquad (8)$$

where $r_\alpha(w)$ denotes the number of letters α in w. Indeed, $\sum_{\alpha\,\in A} r_\alpha(w) = |w|$.

A *block code* ϕ is an injective mapping between words $w \in A^n$ and binary sequences (or *codewords*) $\phi\left(w\right)$:

$$\phi : A^n \to \{0,1\}^*\,, \qquad (9)$$

where the codewords have a property that $\phi\left(w_{i_1}\right) \phi\left(w_{i_2}\right) \ldots \phi\left(w_{i_s}\right) = \phi\left(w_{j_1}\right) \phi\left(w_{j_2}\right) \ldots \phi\left(w_{i_t}\right)$ always implies that $s = t$ and $\phi\left(w_{i_k}\right) = \phi\left(w_{j_k}\right)$ $(k = 1, \ldots, s)$, i.e. the code ϕ is *decipherable*.

By Φ we will denote a set of all possible decipherable codes (9).

The *average cost* of an encoding ϕ of a source S is:

$$C\left(\phi, n, S\right) = \frac{1}{n} \sum_{w\,\in A^n} P_S(w)\, |\phi(w)|\,, \qquad (10)$$

where $|\phi(w)|$ denotes the length of a codeword $\phi(w)$. The *average redundancy rate* of such an encoding is:

$$R\left(\phi, n, S\right) = C\left(\phi, n, S\right) - H(S)\,. \qquad (11)$$

The classic problem of source coding is to find a code ϕ_S that minimizes the redundancy $R\left(\phi_S, n, S\right)$ of an encoding of a known source S:

$$R\left(\phi_S, n, S\right) = \inf_{\phi\,\in\,\Phi} R\left(\phi, n, S\right)\,.$$

Huffman, Shannon, and Gilbert-Moore algorithms are very well known examples of possible (exact or approximate) solutions of this problem [7, 2, 11]. These codes have a similar (with a difference in a constant factor) redundancy rate of $R\left(\phi_S, n, S\right) = O\left(\frac{1}{n}\right)$ [11].

A more interesting problem is to find a code ϕ_Ω that minimizes the worst case redundancy of encoding of any source $S \in \Omega$:

$$R\left(\phi_\Omega, n, \Omega\right) = \inf_{\phi\,\in\,\Phi} \sup_{S\,\in\,\Omega} R\left(\phi, n, S\right)\,.$$

Such codes are called *universal codes* [5, 3, 11] and they are particularly useful in situations when the parameters of a source to be encoded are not known.

In order to construct a universal block code ϕ_Ω one can *estimate probabilities* $P_e(w)$ of words $w \in A^n$ first, and then apply any convenient coding technique, such as Huffman, Shannon, or Gilbert-Moore code to a source with probabilities $P_e(w)$ [11]. The appropriate (and, in fact, nearly-optimal [12]) estimates of words' probabilities can be obtained by using a well-known Krichevsky-Trofimov (KT) formula [10]:

$$P_e\left(w\right) = \Gamma\left(\frac{m}{2}\right) \frac{\prod_{\alpha\in A} \Gamma\left(r_\alpha(w) + \frac{1}{2}\right)}{\pi^{\frac{m}{2}}\, \Gamma\left(|w| + \frac{m}{2}\right)}\,, \qquad (12)$$

where m is the cardinality of the alphabet A, and $\Gamma(x)$ is a Γ-function.

It has been shown (cf. [10, 11]), that the average redundancy rate of universal block codes ϕ_Ω decreases with the block size n as:

$$\begin{aligned} R\left(\phi_\Omega, n, \Omega\right) &= \frac{1}{n}\left[\frac{m-1}{2}\log\frac{n}{m} + O\left(m\right)\right] \quad (13) \\ &= O\left(\frac{\log n}{n}\right), \end{aligned}$$

which is somewhat slower than the $O\left(\frac{1}{n}\right)$ convergence rate of codes for a known source.

Yet another problem in coding theory is to construct a set of codes $\phi_{\ell,\Omega} = \left\{\phi_u \middle| \, u \in A^\ell\right\}$ based on observed sequences of symbols (or *samples*) u of length ℓ produced by a source $S \in \Omega$, such that their worst case average redundancy (under the same source) is minimal:

$$R\left(\phi_{\ell,\Omega}, n, \Omega\right) = \inf_{\{\phi_u\}\subset\Phi} \sup_{S\in\Omega} \sum_{u\in A^\ell} P_S(u) R\left(\phi_u, n, S\right).$$

Such codes are called *sample-based* or *adaptive block codes*, and they are useful in a (typical in practice) situation when one has to encode a source known to have produced a given sample sequence in the past.

The idea of sample-based codes belongs to R. E. Krichevsky [8], who has shown that the average redundancy rate of such codes is (cf. [9, 10]):

$$R(\phi_{\ell,\Omega}, n, \Omega) \sim \frac{m-1}{2\,n}\log\frac{\ell+n}{\ell}, \qquad (14)$$

where ℓ is the sample length. Based on (14) it is clear that the use of sufficiently long ($\ell = O(n)$) samples will make $R(\phi_{\ell,\Omega}, n, \Omega) = O\left(\frac{1}{n}\right)$, which is the order of the redundancy rate of codes built for a known source. It is also clear that further increase ($\ell \gg n$) of the sample length ℓ will only reduce the constant factor in (but not the order of) this expression.

While this result fully explains the effectiveness of adaptive codes when the sample and input blocks are produced by the same source, in practice, however, we much more often have a situation when we know that *the sample is produced by a similar, but not necessarily the same source as one we are about to encode*. Realization of this fact was our main motivation for this research, and in the next two sections we will provide a formal setting of this problem and present the results of our analysis.

4 Average Redundancy Rate of Adaptive Block Codes under Different Sources

Assume that at the time of code construction we have access to a *sample* sequence u of length $|u| = \ell$ produced by some source $T \in \Omega$, and we hope that the actual source that will be compressed $S \in \Omega$ is similar to T:

$$D\left(T \,||\, S\right) \leqslant \delta, \qquad (15)$$

where $\delta < \infty$ is some positive constant.

Following Krichevsky [11], we construct adaptive block codes $\phi_{\ell,T} = \left\{\phi_u \middle| \, u \in A^\ell\right\}$ using the estimates

$$P_e\left(w|u\right) = \frac{P_e\left(u\,w\right)}{P_e\left(u\right)}, \qquad (16)$$

of *conditional probabilities*[1] of words $w \in A^n$. The average (with respect to both word w and the sample u) redundancy rate of such codes[2] is:

$$\begin{aligned} &R\left(\phi_{\ell,T}, n, S\right) \\ &\quad= \frac{1}{n}\sum_{u\in A^\ell}\sum_{w\in A^n} P_T(u)P_S(w)\left|\phi_u(w)\right| - H(S) \\ &\quad= \frac{1}{n}\sum_{u\in A^\ell}\sum_{w\in A^n} P_T(u)P_S(w)\left\lceil -\log P_e(w|u)\right\rceil \\ &\qquad - H(S), \qquad (17) \end{aligned}$$

and our main goal will be to evaluate the asymptotic behavior of this quantity in a case when $D\left(T \,||\, S\right) \neq 0$.

5 Main Results

We are now ready to present our results regarding the performance of adaptive block codes under different sources T and S.

Theorem 1. *The average redundancy rate of adaptive block codes $\phi_{\ell,T}$ is asymptotically (for large n and ℓ):*

$$\begin{aligned} R\left(\phi_{\ell,T}, n, S\right) = \frac{1}{n}\Big[&(\ell+n)\,\Delta\left(T \overset{\ell/(\ell+n)}{\bowtie} S\right) \\ &+ \frac{m-1}{2}\log\frac{\ell+n}{\ell} + O\left(1\right)\Big], \quad (18) \end{aligned}$$

where m is the cardinality of the alphabet, n is a block size, ℓ is the length of sample sequences used to construct these codes, and $\Delta\left(T \overset{\ell/(\ell+n)}{\bowtie} S\right)$ is the excess entropy of a $\frac{\ell}{\ell+n}$-mixed source based on sources T and S producing samples and input blocks correspondingly.

[1] Note that since both sources S and T are memoryless, the true probability of a word w is indeed the same $P_{T,S}(u|w) = \frac{P_{T,S}(u\,w)}{P_T(u)} = \frac{P_T(u)\,P_S(w)}{P_T(u)} = P_S(w)$. However, the estimated conditional probability $P_e(w|u)$ may be different from $P_e(w)$. Most importantly, the estimate $P_e(w|u)$ is based on a longer overall sequence $|u\,w| = \ell + n$, which means that it is more accurate than $P_e(w)$.

[2] For simplicity, we assume that the final encoding is done by using the Shannon code.

The first immediate consequence of Theorem 1 and a property (7) is that the use of $\ell = O(n)$ and longer samples from a different source $T \not\equiv S$ will actually make the redundancy of such codes $O(1)$-large. This means that such codes are no longer universal (no redundancy rate convergence with the block size n) and consecutively, they are of little practical interest:

Corollary 1. *When* $\frac{\ell}{\ell+n} = \gamma = O(1)$*:*

$$R\left(\phi_{\ell,T}, n, S\right) = \frac{1}{1-\gamma}\,\Delta\Big(T \overset{\gamma}{\bowtie} S\Big) + O\left(\frac{1}{n}\right) = O(1)\,. \tag{19}$$

Corollary 2. *When* $\ell \gg n$*:*

$$R\left(\phi_{\ell,T}, n, S\right) = D\left(S \,||\, T\right) + O\left(\frac{1}{n}\right) = O(1)\,. \tag{20}$$

On the other hand, if we let input blocks to be much longer than the sample ($n \gg \ell$), then due to the second limit property of the excess entropy (6) we arrive at the following expression.

Corollary 3. *When* $n \gg \ell$*:*

$$\begin{aligned} &R\left(\phi_{\ell,T}, n, S\right) \\ &\quad = \frac{1}{n}\left[\frac{m-1}{2}\log\frac{n}{\ell} + \ell\, D\left(T \,||\, S\right) + O\left(1\right)\right]. \end{aligned} \tag{21}$$

Observe, that the main terms depending on ℓ: $-\frac{m-1}{2}\log\ell$ and $\ell\, D\left(T \,||\, S\right)$ in (21) have the opposite directions (and different speeds) of growth, which means, that their sum must have a point of minima. I.e., we can prove the following.

Theorem 2. *If* $D\left(T \,||\, S\right) \neq 0$*, then there exists a sample length* ℓ^* *such that*

$$R\left(\phi_{\ell^*,T}, n, S\right) = \min_{\ell} R\left(\phi_{\ell,T}, n, S\right)\,. \tag{22}$$

Corollary 4. *The optimal length of samples* ℓ^* *is asymptotically (for large* n*):*

$$\ell^* = \frac{m-1}{2\, D\left(T \,||\, S\right)} + O\left(\frac{1}{n}\right)\,. \tag{23}$$

Corollary 5. *The minimum average redundancy rate* $R\left(\phi_{\ell^*,T}, n, S\right)$ *is asymptotically (for large* n*):*

$$\begin{aligned} R\left(\phi_{\ell^*,T}, n, S\right) = \frac{1}{n}\Bigg[&\frac{m-1}{2}\log n \\ &+ \frac{m-1}{2}\log\frac{2\,e\, D\left(T \,||\, S\right)}{m-1} + O\left(1\right)\Bigg]. \end{aligned} \tag{24}$$

Now, based on the formula (24) we can find the maximum distance between sources $D\left(T \,||\, S\right)$ when the adaptive block codes $\phi_{\ell^*,T}$ have equal or better efficiency than pure universal codes (13).

Theorem 3. *Adaptive block codes constructed using samples from a source* T *and applied to a source* S *can achieve a lower average redundancy than universal codes if:*

$$D\left(T \parallel S\right) < \delta_1 = \frac{1}{2} + O\left(\frac{1}{m}\right)\,. \tag{25}$$

In other words, we can conclude that adaptive block codes are useful even in situations when samples are produced by a different (within $D\left(T \parallel S\right) \leqslant \delta_1$) source. However, in order to achieve their maximum efficiency such codes shall be based on samples of length $\ell = \ell^*$ (23). Using longer or shorter samples will only increase their redundancy. This constitutes a principal difference in the behavior of these codes compared to a case when samples are produced by the same source.

6 Applications

In this section we will show how our results can be used to guide the design of two data compression systems based on adaptive block codes. The first system provides random access to compressed data, and for this reason all samples are generated by an additional (embedded in both encoder and decoder) source. The second system is an example of a sequential data compression scheme, where samples are being taken from previously compressed (or decompressed) blocks.

6.1 Adaptive Coding System Based on an Additional Source

Consider a situation when we need to encode a set of words $w_i \in A^n$ produced by different memoryless sources S_i $(i = 1, \ldots, N)$. These words will have to be encoded independently from one another, but both encoder and decoder will have access to samples u_i produced by another source T, such that:

$$\max_i D\left(T \,||\, S_i\right) \leqslant \delta_T\,, \tag{26}$$

where $\delta_T < \frac{1}{2}$ is a known constant.

To implement such coding system we will use adaptive block codes $\phi_{\ell,T}$, and we will want to find a sample length ℓ^*, such that:

$$\begin{aligned} &R\left(\phi_{\ell^*,T}, n, S_1, \ldots, S_N\right) \\ &\quad = \sum_{i=1}^{N} R\left(\phi_{\ell^*,T}, n, S_i\right) = \min_{\ell}\left\{\sum_{i=1}^{N} R\left(\phi_{\ell,T}, n, S_i\right)\right\}. \end{aligned}$$

Using (21), we can show that:

$$\begin{aligned} &R\left(\phi_{\ell,T} n, S_1, \ldots, S_N\right) \\ &\quad = \frac{N}{n}\left[\frac{m-1}{2}\log\frac{n}{\ell} + \frac{\ell}{N}\sum_{i=1}^{N} D\left(T \,||\, S_i\right) + O\left(1\right)\right], \end{aligned}$$

and due to Theorem 2 and formula (23):

$$\ell^* = \frac{m-1}{2} \frac{N}{\sum_{i=1}^{N} D\left(T \,||\, S_i\right)} + O\left(\frac{1}{n}\right) . \quad (27)$$

Using the upper bound for $D\left(T \,||\, S_i\right)$ (26), we can show that

$$\ell^* \geqslant \frac{m-1}{2\,\delta_T} , \quad (28)$$

which gives us a very simple estimate for the sample length ℓ that can be used to design such codes.

6.2 Adaptive Coding System Using Preceding Blocks as Samples

As in the previous section, we have to encode a set of words $w_i \in A^n$ produced by different memoryless sources S_i $(i = 1, \ldots, N)$, but now we are allowed to take samples from the previous words (except for the first one):

$$u_i = \left[\begin{array}{ll} \emptyset\,, & \text{if} \quad i = 1\,, \\ \text{prefix}_{1\ldots\ell}\left(w_{i-1}\right)\,, & \text{if} \quad i = 2, \ldots, N\,. \end{array} \right.$$

Assume also, that:

$$\max_{i>1} D\left(S_{i-1} \,||\, S_i\right) \leqslant \delta_\Delta\,, \quad (29)$$

where $\delta_\Delta < \frac{1}{2}$ is a known constant.

Following the same arguments as in the previous section we can show that the optimal length of samples ℓ^* in this coding system is

$$\ell^* = \frac{m-1}{2} \frac{N-1}{\sum_{i=2}^{N} D\left(S_{i-1} \,||\, S_i\right)} + O\left(\frac{1}{n}\right) . \quad (30)$$

Using (29) this estimate can be further simplified to

$$\ell^* \geqslant \frac{m-1}{2\,\delta_\Delta}\,. \quad (31)$$

Based on these estimates, we can conclude that in order to design an optimal adaptive block coding system it is not sufficient to simply have a "good" reference source. What is also important is to know how distant (in the $D(.\,||\,.)$ metric) it can be from the actual source(s) we are about to encode.

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. 2002 Data Compression Conference* (Snowbird, UT, 2002) 222-231.

[5] 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.)

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

[7] E. N. Gilbert and E. F. Moore, Variable-Length Binary Encodings, The Bell System Tech. Journal, 7 (1959) 932–967.

[8] R. E. Krichevsky, The Connection Between the Redundancy and Reliability of Information about the Source, *Probl. Inform. Trans.*, 4 (3) (1968) 48–57 (in Russian).

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

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

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

[12] R. E. Krichevskiy, Laplace's law of succession and universal encoding, *IEEE Trans. Information Theory*, **44** (1998) 296–303.

[13] Yu. A. Reznik and W. Szpankowski, Asymptotic Average Redundancy of Adaptive Block Codes, *2003 IEEE International Symposium on Informaion Theory* (Yokohama, Japan, 2003) – accepted.

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

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

[16] R. W. Yeung, *A First Course in Information Theory* (Kluwer Academic/Plenum Publishers, New York, NY, 2002).

A Sketches of Proofs

Using formula (17) and an inequality $\lceil x \rceil \leqslant 1 + x$ (x is a positive real number), we can upper bound the average redundancy rate of an adaptive block code (17) as:

$$\begin{aligned} & R\left(\phi_{\ell,T}, n, S\right) \\ & \leqslant \frac{1}{n}\left[1 - \sum_{u\in A^\ell}\sum_{w\in A^n} P_T(u)P_S(w)\log P_e\left(w|u\right)\right] \\ & \quad -H(S)\,, \end{aligned} \tag{32}$$

where where $P_T(u)$ and $P_S(w)$ are probabilities of words u and w produced by memoryless sources T and S correspondingly, and $P_e\left(w|u\right)$ is the KT-estimated conditional probability (16) of the word w.

Based on (16), we can split the central sum in (32) into:

$$\begin{aligned} & -\sum_{u\in A^\ell}\sum_{w\in A^n} P_T(u)P_S(w)\log P_e\left(w|u\right) \\ & = -\sum_{u\in A^\ell}\sum_{w\in A^n} P_T(u)P_S(w)\log P_e\left(u\,w\right) \\ & \qquad + \sum_{u\in A^\ell} P_S(u)\log P_e\left(u\right) \\ & = (\ell+n)\,C_e(\ell,T,n,S) - \ell\,C_e(\ell,T)\,, \end{aligned} \tag{33}$$

where

$$C_e(\ell,T) = -\frac{1}{\ell}\sum_{u\in A^\ell} P_T(u)\log P_e\left(u\right)\,, \tag{34}$$

is the average rate of the KT-estimator processing ℓ-symbols words produced by the source T, and

$$\begin{aligned} C_e(\ell,T,n,S) \;=\; & \frac{1}{\ell+n}\sum_{u\in A^\ell}\sum_{w\in A^n} P_T(u)P_S(w)\times \\ & \times\log P_e\left(u\,w\right)\,, \end{aligned} \tag{35}$$

is the average rate of the KT-estimator processing sequences with ℓ first symbols produced by the source T and the remaining n symbols produced by the source S.

Using Krichevsky's technique [11, 3.4.4–3.4.10] (Stirling's approximation for $\log P_e(u)$, and Jensen inequalities for sums over $-\log(x)$ and $x\log(x)$) we can show that:

$$C_e(\ell,T) = H(T) + \frac{1}{\ell}\left[\frac{m-1}{2}\log\frac{\ell}{m} + r_1\right], \tag{36}$$

where the remaining term r_1 is such that:

$$-\frac{m-2}{2}\log e \leqslant \lim_{\ell\to\infty} r_1 + \frac{1}{2}\log 2 - \delta_m \leqslant \frac{m}{2}\log e\,,$$

where

$$\begin{aligned} \delta_m \;=\; & \frac{m-1}{2}\log m - \frac{m}{2}\log 2e + \frac{1}{2}\log 4\pi \\ & -\log\Gamma\left(\frac{m}{2}\right) = O\left(\frac{1}{m}\right). \end{aligned}$$

Similarly, we can show that the average rate of the KT-estimator under a mixed source (35) satisfies:

$$\begin{aligned} C_e(\ell,T,n,S) = {} & H\left(T^{\,\ell/(\ell+n)}\!\bowtie S\right) \\ & + \frac{1}{\ell+n}\left[\frac{m-1}{2}\log\frac{\ell+n}{m} + r_2\right], \end{aligned} \tag{37}$$

where the remaining term r_2 is such that:

$$-\frac{m}{2}\log e \leqslant \lim_{\ell,n\to\infty} r_2 + \frac{1}{2}\log 2 - \delta_m \leqslant \frac{m}{2}\log e\,.$$

By applying estimates (36) and (37) in (32,33) we can show that the average redundancy rate of an adaptive code is

$$\begin{aligned} R\left(\phi_{\ell,T}, n, S\right) \leqslant \frac{1}{n}\Big[& (\ell+n)\,\Delta\left(T^{\,\ell/(\ell+n)}\!\bowtie S\right) \\ & + \frac{m-1}{2}\log\frac{\ell+n}{\ell} + r_2 - r_1 + 1\Big], \end{aligned}$$

which, leads to a formula (18) claimed in the Theorem 1.

The subsequent claims are simple and natural consequences of the Theorem 1, and they can be easily repeated following the order in which they appear in the main text.