NONASYMPTOTIC LOWER BOUND ON THE MAXIMAL CUMULATIVE REDUNDANCY OF UNIVERSAL CODING

Vladimir B. Balakirsky and Frans M. J. Willems

Eindhoven University of Technology, Electrical Engineering Department
P.O.Box 513, 5600 MB Eindhoven, the Netherlands
v.b.balakirsky,f.m.j.willems@ele.tue.nl

We address the problem of estimating the maximal cumulative redundancy of universal coding and present a nonasymptotic lower bound for the Bernoulli sources with the unknown parameter that can be used for different assignments of the length distributions of both the messages and codewords; this includes the variable-to-variable length coding. The approach can be directly generalized to the other source models.

INTRODUCTION, NOTATION, AND STATEMENT OF THE PROBLEM

Let us consider *a Bernoulli source* with the parameter $\theta \in (0,1)$, i.e., the probability that the source generates a binary sequence u of length $a(u)+b(u)$ consisting of $a(u)$ zeroes and $b(u)$ ones is equal to

$$P_\theta(u) = (1-\theta)^{a(u)}\theta^{b(u)}. \tag{1}$$

Suppose also that $\mathcal{S} \subset \{0,1\}^*$ is a finite set of binary sequences chosen in such a way that $\{P_\theta(u), u \in \mathcal{S}\}$ is the probability distribution on $\mathcal{S}$ for any $\theta \in (0,1)$,

$$\sum_{u\in\mathcal{S}} P_\theta(u) = 1. \tag{2}$$

The set $\mathcal{S}$ will be referred to as *the source segment set*, and we represent it as a union of pairwise disjoint subsets

$$\mathcal{S}(a,b) = \{\ u \in \mathcal{S} :\ a(u) = a,\ b(u) = b\ \},\ \ a,b \geq 0. \tag{3}$$

Then (1)–(3) imply

$$\sum_{a,b} |\mathcal{S}(a,b)|(1-\theta)^a\theta^b = 1. \tag{4}$$

For any *encoding* $f : \mathcal{S} \to \{0,1\}^*$ such that

$$\sum_{u \in \mathcal{S}} 2^{-|f(u)|} \leq 1 \tag{5}$$

where $|f(u)|$ denotes the length of the binary sequence $f(u)$, let us introduce the function

$$\overline{\varrho}_\theta(f, \mathcal{S}) = \sum_{u \in \mathcal{S}} P_\theta(u) \cdot |f(u)| - \sum_{u \in \mathcal{S}} P_\theta(u) \log \frac{1}{P_\theta(u)} \tag{6}$$

which is known as *the cumulative redundancy* of coding the sequences belonging to the source segment set ("coding" is understood as a pair (encoding, decoding) where the decoding is assumed to be noiseless). The function $\overline{\varrho}_\theta(f, \mathcal{S})$ is non–negative, as it follows from the Shannon's coding theorem, and can serve as a measure of the coding performance. There are several explicit constructions of codes for the case when the parameter θ is known. In particular, one can assign the Huffman or arithmetic code and get the redundancy which is not greater to 1 and 2 bits, respectively. The case when f should be assigned without knowing the θ is regarded as *universal coding.* Then a proper measure of the performance is the maximal value of the function $\overline{\varrho}_\theta(f, \mathcal{S})$ for all $\theta \in (0,1)$ distinguished as *the maximal cumulative redundancy of universal coding*,

$$\overline{\varrho}(f, \mathcal{S}) = \max_{0<\theta<1} \overline{\varrho}_\theta(f, \mathcal{S}). \tag{7}$$

One of the explicit constructions of universal codes is based on the context tree weighting method [5] which guarantees the cumulative redundancy not greater to the function depending on the average source segment length,

$$L(\mathcal{S}) = \sum_{u \in \mathcal{S}} P_\theta(u) \cdot |u|$$

as $\log \sqrt{L(\mathcal{S})}$. This means that $\overline{\varrho}(f, \mathcal{S})$ is not bounded from above by a constant like for non–universal codes. It is also known [2] that the cumulative redundancy of any universal code of a fixed length (the fixed-to-variable length coding) is not less than a function proportional to $\log \sqrt{L(\mathcal{S})}$. However, as it follows from the arguments used in the proof, this lower bound can be applied only in the asymptotics when $L(\mathcal{S})$ is very big and, for the practical cases, we could only say that the maximal redundancy is non–negative. Another important class of source coding problems concerns different assignments of the length distributions of the sequences belonging to the source segment set and the code sequences, and one usually distinguishes between the fixed-to-variable, variable-to-fixed, and variable-to-variable length coding. Therefore we address the problem : *derive a nonasymptotic lower bound on the maximal cumulative redundancy of universal coding that can be used for any source segment set satisfying (2) and any encoding function satisfying (5).*

THE KRICHEVSKY–TROFIMOV ESTIMATOR AND ITS PROPERTIES

The universal coding technique developed in [3], [4], [5] and other papers essentially uses the properties of the function introduced in [1] and known as *the Krichevsky–Trofimov estimator*,

$$P_e(a,b) = \int_0^1 \frac{(1-\theta)^a \theta^b}{\pi\sqrt{(1-\theta)\theta}}\, d\theta, \quad \text{for all } \ a,b \geq 0. \tag{8}$$

The function $P_e(a,b), a,b \geq 0$, has the following remarkable properties :

(**P0**) $P_e(0,0) = 1$;

(**P1**) $P_e(a,b) = P_e(a+1,b) + P_e(a,b+1)$, *for all* $a,b \geq 0$;

(**P2**) *for all* $a,b \geq 0$ *such that* $a+b > 0$, *the value of* $P_e(a,b)$ *can be recurrently computed based on* (**P0**) *and equations*

$$P_e(a+1,b) = \frac{a+1/2}{a+b+1} P_e(a,b), \quad P_e(a,b+1) = \frac{b+1/2}{a+b+1} P_e(a,b);$$

(**P3**) *for all* $a,b \geq 0$ *such that* $a+b > 0$,

$$P_e(a,b) = \frac{\prod_{i=1}^a (i-1/2) \prod_{j=1}^b (j-1/2)}{(a+b)!}$$

where the product is understood as 1 when the lower limit is greater to the upper one and

$$\frac{1}{2} \cdot \frac{1}{\sqrt{a+b}} \leq \frac{P_e(a,b)}{\left(\frac{a}{a+b}\right)^a \left(\frac{b}{a+b}\right)^b} \leq \sqrt{\frac{2}{\pi}} \cdot \frac{1}{\sqrt{a+b}}. \tag{9}$$

Furthermore, using (4), (8), and (**P0**) we have

$$\sum_{u \in \mathcal{S}} P_e(a(u), b(u)) = \sum_{a,b} |\mathcal{S}(a,b)| P_e(a,b)$$

$$= \int_0^1 \frac{1}{\pi\sqrt{(1-\theta)\theta}} \left[\sum_{a,b} |\mathcal{S}(a,b)|(1-\theta)^a\theta^b \right] d\theta = 1 \qquad (10)$$

i.e., $P_e = \{\, P_e(a(u), b(u)),\ u \in \mathcal{S} \,\}$ *is the probability distribution on the source segment set.*

The upper bound on the cumulative redundancy proved in [5] can be presented as

$$\begin{aligned} \exists f^*:\ \ \overline{\varrho}_\theta(f^*, \mathcal{S}) &\leq 1 + \sum_{a,b} |\mathcal{S}(a,b)|(1-\theta)^a\theta^b \log \frac{(1-\theta)^a\theta^b}{P_e(a,b)} & (11) \\ &\leq 2 + \sum_{a,b} |\mathcal{S}(a,b)|(1-\theta)^a\theta^b \log \sqrt{a+b}. & (12) \end{aligned}$$

The term 1 at the right-hand side of (11) comes from the assumption that we use the Huffman code for the sequences included in the source segment set and P_e is the probability distribution on this set, and the sum on a, b in (11) can be obtained using the approach developed in the proof of Theorem 2 in [5] since the coincidence of the lengths of the code sequences is not used in the analysis. The inequality (12) follows from (9).

We try to derive a converse statement to the method of proving the inequality (12) rather than to the final result and conclude with the bound

$$\overline{\varrho}(f, \mathcal{S}) \geq \int_0^1 \frac{1}{\pi\sqrt{1-\theta)\theta}} \left[\sum_{a,b} |\mathcal{S}(a,b)|(1-\theta)^a\theta^b \log \frac{(1-\theta)^a\theta^b}{P_e(a,b)} \right] d\theta$$

for any source segment set $\mathcal{S}$ and encoding f satisfying (2) and (5). This inequality is proved in the next section.

LOWER BOUND ON THE MAXIMAL CUMULATIVE REDUNDANCY OF UNIVERSAL CODING

Theorem : *If (2) and (5) holds, then the maximal cumulative redundancy of universal coding defined by (6), (7) satisfies the inequality*

$$\overline{\varrho}(f, \mathcal{S}) \geq H(P_e(\mathcal{S})) - H_e(\mathcal{S}) \qquad (13)$$

where

$$H(P_e(\mathcal{S})) = \sum_{a,b} |\mathcal{S}(a,b)| P_e(a,b) \log \frac{1}{P_e(a,b)} \qquad (14)$$

and

$$H_e(\mathcal{S}) = \sum_{a,b} |\mathcal{S}(a,b)| \int_0^1 \frac{(1-\theta)^a\theta^b}{\pi\sqrt{(1-\theta)\theta}} \left[a\log\frac{1}{1-\theta} + b\log\frac{1}{\theta} \right] d\theta. \tag{15}$$

Proof : Let us upper-bound the maximal cumulative redundancy by the expected value of $\overline{\varrho}_\theta(f,\mathcal{S})$ over the ensemble specified by the probability density function $\pi^{-1}(1-\theta)^{-1/2}\theta^{-1/2}, \theta \in (0,1)$, i.e.,

$$\overline{\varrho}(f,\mathcal{S}) \geq \int_0^1 \frac{1}{\pi\sqrt{(1-\theta)\theta}} \overline{\varrho}_\theta(f,\mathcal{S})\, d\theta. \tag{16}$$

Let us denote

$$P(u) = \frac{2^{-|f(u)|}}{\sum_{u'\in\mathcal{S}} 2^{-|f(u')|}}\ , \quad \text{for all } \ u \in \mathcal{S}.$$

Then

$$\overline{\varrho}_\theta(f,\mathcal{S}) = \sum_{u\in\mathcal{S}} P_\theta(u)\log\frac{P_\theta(u)}{P(u)} - \log\sum_{u\in\mathcal{S}} 2^{-|f(u)|} \geq \sum_{u\in\mathcal{S}} P_\theta(u)\log\frac{P_\theta(u)}{P(u)}$$

where the inequality follows from (5). Therefore we can continue (16) as

$$\begin{aligned}
\overline{\varrho}(f,\mathcal{S}) &\geq \int_0^1 \frac{1}{\pi\sqrt{(1-\theta)\theta}} \left[\sum_{u\in\mathcal{S}} P_\theta(u)\log\frac{P_\theta(u)}{P(u)} \right] d\theta \\
&= \sum_{u\in\mathcal{S}} \int_0^1 \frac{P_\theta(u)}{\pi\sqrt{(1-\theta)\theta}} \left[\log\frac{P_\theta(u)}{P(u)} + \log\frac{P_e(a(u),b(u))}{P_e(a(u),b(u))} \right] d\theta \\
&= \sum_{u\in\mathcal{S}} \int_0^1 \frac{P_\theta(u)}{\pi\sqrt{(1-\theta)\theta}} \log\frac{P_\theta(u)}{P_e(a(u),b(u))}\, d\theta \\
&\quad + \sum_{u\in\mathcal{S}} \int_0^1 \frac{P_\theta(u)}{\pi\sqrt{(1-\theta)\theta}} \log\frac{P_e(a(u),b(u))}{P(u)}\, d\theta \\
&= \sum_{a,b} |\mathcal{S}(a,b)| \int_0^1 \frac{(1-\theta)^a\theta^b}{\pi\sqrt{(1-\theta)\theta}} \log\frac{(1-\theta)^a\theta^b}{P_e(a,b)}\, d\theta \\
&\quad + \sum_{u\in\mathcal{S}} P_e(a(u),b(u))\log\frac{P_e(a(u),b(u))}{P(u)} \\
&\geq \sum_{a,b} |\mathcal{S}(a,b)| \int_0^1 \frac{(1-\theta)^a\theta^b}{\pi\sqrt{(1-\theta)\theta}} \log\frac{(1-\theta)^a\theta^b}{P_e(a,b)}\, d\theta
\end{aligned}$$

where the inequality follows from (10) and the non–negativity of the divergence, and obtain (13) using notations (14), (15). Q.E.D.

Note that the inequality (13) can be also presented in such a way that the maximal cumulative redundancy of universal coding is not less than the mutual information between the continuous random variable θ distributed over the interval $(0,1)$ in accordance with the probability density function $\pi^{-1}(1-\theta)^{-1/2}\theta^{-1/2}, \theta \in (0,1)$, and the discrete random variable $u \in \mathcal{S}$. The transition probabilities are defined as $(1-\theta)^{a(u)}\theta^{b(u)}$ for all $\theta \in (0,1)$ and $u \in \mathcal{S}$.

NUMERICAL RESULTS FOR FIXED-TO-VARIABLE CODING

Suppose that the source segment set is specified as the set $\mathcal{S}^{(n)}$ consisting of all binary sequences of length n. Then

$$|\mathcal{S}^{(n)}(a, n-a)| = \binom{n}{a}, \quad \text{for all} \quad a = 0, \ldots, n$$

and $\mathcal{S}^{(n)}(a,b) = \emptyset$ for all a, b such that $a + b \neq n$. In this case, the upper bound (12) can be expressed as

$$\exists f^* : \; \overline{\varrho}_\theta(f^*, \mathcal{S}^{(n)}) \leq 2 + \log\sqrt{n}. \tag{17}$$

Since

$$\sum_{a,b} |\mathcal{S}^{(n)}(a,b)|(1-\theta)^a\theta^b \left[a \log \frac{1}{1-\theta} + b \log \frac{1}{\theta} \right] = nh(\theta)$$

where $h(\theta) = -(1-\theta)\log(1-\theta) - \theta\log\theta$ is the binary entropy function, we also have

$$H_e(\mathcal{S}^{(n)}) = nh_e$$

where

$$h_e = \int_0^1 \frac{h(\theta)}{\pi\sqrt{(1-\theta)\theta}}\, d\theta.$$

Thus, the bound (13) can be presented as

$$\begin{aligned} \overline{\varrho}(f, \mathcal{S}^{(n)}) \;\; &\geq \;\; \sum_{a=0}^{n} \int_0^1 \left[\omega_\theta^{(n)}(a) \cdot D_\theta^{(n)}(a) \right] d\theta \\ &= \;\; -\sum_{a=0}^{n} \binom{n}{a} P_e(a, n-a) \log P_e(a, n-a) - nh_e \end{aligned} \tag{18}$$

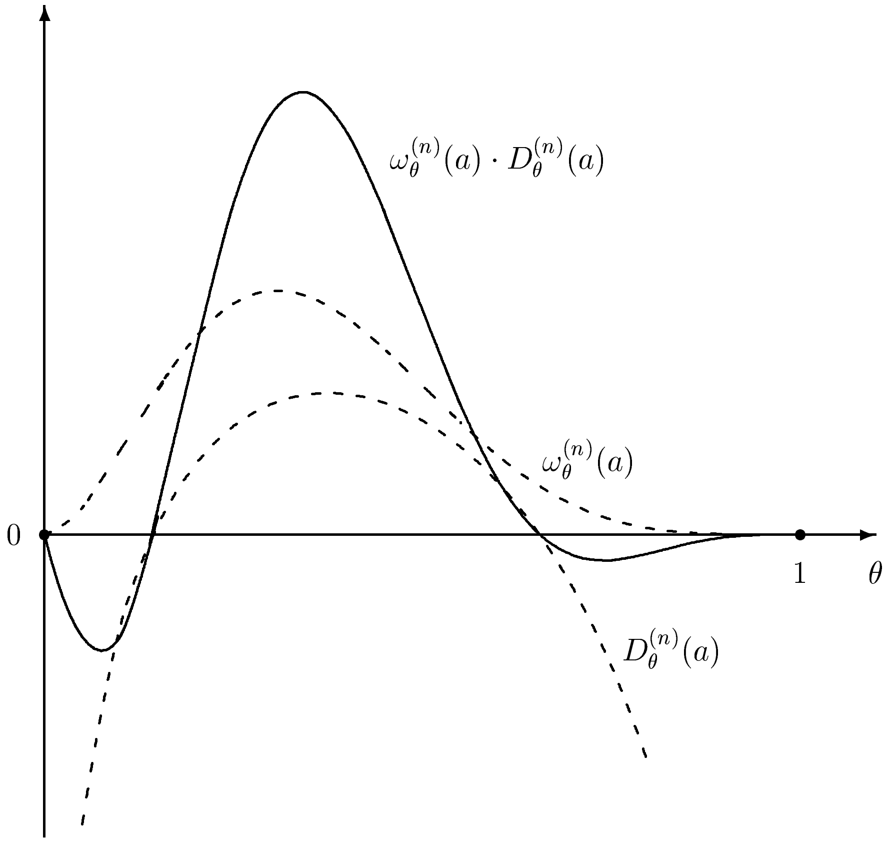

Figure 1: The functions $\omega_\theta^{(n)}(a), D_\theta^{(n)}(a)$, and their product contributing to the lower bound (18) for $a = 5$ and $n = 8$.

where

$$\omega_\theta^{(n)}(a) = \binom{n}{a} \cdot \frac{(1-\theta)^a \theta^{n-a}}{\pi\sqrt{(1-\theta)\theta}}\,, \quad D_\theta^{(n)}(a) = \log \frac{(1-\theta)^a \theta^{n-a}}{P_e(a, n-a)}\,.$$

The functions $\omega_\theta^{(n)}(a), D_\theta^{(n)}(a)$, and their product contributing to the lower bound (18) for $a = 5$ and $n = 8$ are shown in Figure 1 and some numerical results obtained from the bounds (17), (18) are given in Table 1.

Table 1: The values of the upper and lower bounds on the maximal cumulative redundancy of universal coding for the Bernoulli sources generating binary sequences of length $n = 2^k$ for $k = 1, \ldots, 9$.

n	$\log \sqrt{n}$	Upper bound	Lower bound
2	0.500	2.500	0.695
4	1.000	3.000	1.014
8	1.500	3.500	1.385
16	2.000	4.000	1.791
32	2.500	4.500	2.218
64	3.000	5.000	2.651
128	3.500	5.500	3.072
256	4.000	6.000	3.452
512	4.500	6.500	3.737

REFERENCES

[1] R.E. Krichevsky and V.K. Trofimov, "The Performance of Universal Encoding", IEEE Trans. Inform. Theory, IT–27(2), pp. 199–207, 1981.

[2] J. Rissanen, "Universal Coding, Information, Prediction, and Estimation", IEEE Trans. Inform. Theory, IT–30(4), pp. 629–636, 1984.

[3] Yu.M. Shtarkov, "Universal Sequential Coding of Single Messages", Probl. Inform. Transm., 23(3), pp. 3–17, 1987.

[4] Tj.J. Tjalkens and F.M.J. Willems, "A Universal Variable-to-Fixed Length Source Code Based on Lawrence's Algorithm", IEEE Trans. Inform. Theory, IT–38(2), pp. 247–253, 1992.

[5] F.M.J. Willems, Yu.M. Shtarkov, and Tj.J. Tjalkens, "The Context-Tree Weighting Method : Basic Properties", IEEE Trans. Inform. Theory, IT–41(3), pp. 653–664, 1995.