Discrete Applied Mathematics 244 (2018) 205–213

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

On the ratio of prefix codes to all uniquely decodable codes with a given length distribution

Adam Woryna

Silesian University of Technology, Institute of Mathematics, ul. Kaszubska 23, 44-100 Gliwice, Poland

ARTICLE INFO

Article history:
Received 20 August 2016
Received in revised form 1 January 2018
Accepted 19 February 2018
Available online 9 April 2018

Keywords:
Uniquely decodable code
Prefix code
Length distribution
Kraft's inequality
Sardinas–Patterson algorithm

ABSTRACT

We investigate the ratio $\rho_{n,L}$ of prefix codes to all uniquely decodable codes over an n-letter alphabet and with length distribution L. For any integers $n \geq 2$ and $m \geq 1$, we construct a lower bound and an upper bound for $\inf_L \rho_{n,L}$, the infimum taken over all sequences L of length m for which the set of uniquely decodable codes with length distribution L is non-empty. As a result, we obtain that this infimum is always greater than zero. Moreover, for every $m \geq 1$ it tends to 1 when $n \to \infty$, and for every $n \geq 2$ it tends to 0 when $m \to \infty$. In the case $m = 2$, we also obtain the exact value for this infimum.

1. Introduction and the results

In this paper, we study variable-length codes with a given *length distribution* $L = (a_1, \ldots, a_m)$ ($a_i \geq 0$), that is finite sequences $(v_1, \ldots, v_m)$ of words $v_i \in X^*$ (so-called *code words*) over a given finite alphabet X such that for every $1 \leq i \leq m$ the length $|v_i|$ of the word v_i is equal to a_i. An important and the most-studied class of variable-length codes are prefix codes. Therefore, given a particular class of codes, it is natural to ask about the contribution of prefix codes in this class. This contribution may be though of as the ratio to all codes in this class. Recall that a *prefix code* is an injective sequence $(v_1, \ldots, v_m)$ of non-empty code words v_i such that no code word is a prefix (initial segment) of another code word. It is known that every prefix code $(v_1, \ldots, v_m)$ is *uniquely decodable*, which means that the following condition holds: if $v_{i_1}v_{i_2}\ldots v_{i_t} = v_{j_1}v_{j_2}\ldots v_{j_{t'}}$ for some $t, t' \geq 1$, $1 \leq i_s, j_{s'} \leq m$, $1 \leq s \leq t$, $1 \leq s' \leq t'$, then $t = t'$ and $i_s = j_s$ for every $1 \leq s \leq t$. Obviously, not every uniquely decodable code is a prefix code. For example, the code $(0, 01)$ over the binary alphabet is uniquely decodable, but it is not a prefix code. Also, not every injective code is uniquely decodable (as an example may serve the injective code (v_1, v_2, v_3) with the code words $v_1 = 0$, $v_2 = 01$, $v_3 = 10$, which satisfy $v_2v_1 = v_1v_3$).

Given an integer $n \geq 2$ and a finite sequence $L = (a_1, \ldots, a_m)$ of positive integers, let $UD_n(L)$ denote the set of all uniquely decodable codes over an n-letter alphabet and with length distribution L, and let $PR_n(L) \subseteq UD_n(L)$ denote the subset of prefix codes. According to the Kraft–McMillan theorem [6], we have: $UD_n(L) \neq \emptyset$ if and only if $PR_n(L) \neq \emptyset$ if and only if $\sum_{i=1}^{m} n^{-a_i} \leq 1$. Thus, for every $n \geq 2$ and $m \geq 1$ the set

$$\mathcal{L}_{n,m} := \{L : |L| = m,\ UD_n(L) \neq \emptyset\} = \{L : |L| = m,\ PR_n(L) \neq \emptyset\}$$

is infinite. If we also denote

$$\mathcal{L}_n := \{L : UD_n(L) \neq \emptyset\} = \{L : PR_n(L) \neq \emptyset\},$$

then we have $\mathcal{L}_n = \bigcup_{m\geq 1} \mathcal{L}_{n,m}$. In particular, the sets $\mathcal{L}_n$ ($n \geq 2$) form an increasing sequence: $\mathcal{L}_2 \subseteq \mathcal{L}_3 \subseteq \ldots$.

E-mail address: adam.woryna@polsl.pl.

https://doi.org/10.1016/j.dam.2018.02.026
0166-218X/

In the present paper, we study the asymptotic behaviour of the quotients

$$\rho_{n,L} := \frac{|PR_n(L)|}{|UD_n(L)|}, \quad n \geq 2, \quad L \in \mathcal{L}_n.$$

Since every prefix code is uniquely decodable, we have $0 \leq \rho_{n,L} \leq 1$. In [8], we have shown that $\rho_{n,L} = 1$ if and only if L is constant. We derived that result from the following estimation:

Theorem 1 *([8], Theorem 1).* *If $L \in \mathcal{L}_n$ is non-constant, then*

$$\frac{|UD_n(L)|}{|PR_n(L)|} \geq 1 + \frac{r_a \cdot r_b}{|PR_n((a, b))|} = 1 + \frac{r_a \cdot r_b}{n^{a+b} - n^{\max\{a,b\}}},$$

where a and b are arbitrary two different values of L and r_a (resp. r_b) is the number of those elements in L which are equal to a (resp. to b).

For every $n \geq 2$ and $m \geq 1$, let us define the infimum

$$\xi_{n,m} := \inf_{L\in\mathcal{L}_{n,m}} \frac{|PR_n(L)|}{|UD_n(L)|} = \inf_{L\in\mathcal{L}_{n,m}} \rho_{n,L}.$$

In particular $\xi_{n,1} = 1$ and $0 \leq \xi_{n,m} < 1$ for all $n, m \geq 2$. Since the set $\mathcal{L}_{n,m}$ is infinite, one may ask if there exist $n \geq 2$, $m \geq 1$ such that $\xi_{n,m} = 0$. For the first result of the present paper, we construct in Section 2 a positive lower bound for the quotients $\rho_{n,L}$, which negatively answers this question. Namely, if we define

$$\varsigma_{n,m} := \frac{n - (m)_{n-1}}{n^{\left\lfloor \frac{m}{n-1} \right\rfloor + 1}}, \tag{1}$$

where $(m)_{n-1}$ is the remainder from the division of m by $n - 1$, then we obtain the following result:

Theorem 2. *Let $n \geq 2$, $m \geq 1$ and $L \in \mathcal{L}_{n,m}$. Then the quotient $\rho_{n,L} = |PR_n(L)|/|UD_n(L)|$ is not smaller than $q_{n,m} \cdot \varsigma_{n,m}^{m-1}$, where*

$$q_{n,m} := \begin{cases} 1, & n \geq m, \\ \dfrac{(m-1)!}{(m-1)^{m-1}}, & n < m. \end{cases}$$

Moreover, if the sequence L is injective (i.e. all values in L are distinct), then $\rho_{n,L}$ is not smaller than the product

$$\varpi_{n,m} := \left(1 - \frac{1 - n^{-1}}{n-1}\right)\left(1 - \frac{1 - n^{-2}}{n-1}\right)\ldots\left(1 - \frac{1 - n^{-m+1}}{n-1}\right). \tag{2}$$

As a direct consequence of the above theorem, we obtain:

Corollary 1. *For all $n \geq 2$ and $m \geq 1$ the infimum $\xi_{n,m} = \inf_{L\in\mathcal{L}_{n,m}} \rho_{n,L}$ is not smaller than $q_{n,m} \cdot \varsigma_{n,m}^{m-1}$. Moreover, for every $m \geq 1$, we have $\lim_{n\to\infty} \xi_{n,m} = 1$.*

To derive Theorem 2, we consider the set $I_n(L)$ of all injective codes over an n-letter alphabet and with length distribution L. Since $UD_n(L) \subseteq I_n(L)$, the following inequality holds: $\rho_{n,L} \geq |PR_n(L)|/|I_n(L)|$. For the required bound, we apply the general formulae for the cardinalities of the sets $PR_n(L)$ and $I_n(L)$ to estimate the quotient on the right-hand side of the above inequality. As for the formula for $|PR_n(L)|$, we derived it in [8] by using a well-known combinatorial construction (a so-called Kraft's construction) of an arbitrary prefix code from $PR_n(L)$. Namely, if $\widetilde{L} = (\nu_1, \ldots, \nu_t)$ is the sequence of the values of L ordered from the smallest to the largest (i.e. $\nu_1 < \nu_2 < \cdots < \nu_t$) and if r_i ($1 \leq i \leq t$) is the number of those elements in L which are equal to ν_i, then we obtained (see Section 2 in [8]):

$$|PR_n(L)| = \prod_{i=1}^{t} \binom{N_i}{r_i} r_i!, \tag{3}$$

where $N_1 := n^{\nu_1}$ and $N_{i+1} := n^{\nu_{i+1}-\nu_i}(N_i - r_i)$ for $1 \leq i < t$.

For the second result of the present paper, we consider the numbers $\eta_{n,m}$ ($n \geq 2$, $m \geq 1$) defined as follows:

$$\eta_{n,m} := 1 + \sum_{i=1}^{m-1} \binom{m-1}{i} \frac{1}{n^i - 1}.$$

In particular, the following obvious inequality holds:

$$\eta_{n,m} \geq 1 + \sum_{i=1}^{m-1} \binom{m-1}{i} \frac{1}{n^i} = \left(1 + \frac{1}{n}\right)^{m-1}.$$

In Section 3, we use these numbers to find the following upper bound for the infimum $\xi_{n,m}$.