Abstract

Name of the Candidate	**Shivkumar K M**
SR Number	**04-02-00-17-12-11-1-08795**
Title of the Thesis	**On Some Questions Involving Prefix Codes**
Research Supervisor	**Prof. Navin Kashyap**
Degree Registered	**Doctor of Philosophy**
Department	**Electrical Communication Engineering**
Institute	**Indian Institute of Science, Bangalore**

Let $\mathcal{A}$ be a finite alphabet and $\mathcal{A}^*$ be the set of all finite sequences of the elements of $\mathcal{A}$. A *word* is any member of $\mathcal{A}^*$. A prefix code X is a set of words satisfying the prefix property, i.e., no word in the set is a prefix of any other word in the set. If X^* is defined as the collection of all concatenations of the words of X, then it can be seen that each of its elements can be expressed as a concatenation of the words of X in a unique manner. Any subset of $\mathcal{A}^*$ possessing this property is called a *uniquely decodable code* and the prefix codes constitute an important subclass of uniquely decodable codes. In our work, we look into the following questions involving prefix codes:

i) We first study a parameter associated with prefix codes. For a discrete source with source distribution P, the problem of constructing a prefix code over the alphabet $\mathcal{A}$ with the minimum expected length is one of the earliest problems addressed in information theory. Let $\mathcal{L}_D(P)$ ($D = |A|$) denote the minimum expected length of a prefix code for this source. This $\mathcal{L}_D(P)$ is the parameter of our interest and can be seen as a function—call it the minimum expected length function $\mathcal{L}_D$—over the set $\mathcal{P}_n$ of all probability mass functions (PMF) of the form $(p_1, p_2, \ldots, p_n)$. It is well known that $\mathcal{L}_D$ attains its maximum value at the uniform distribution $U_n = (1/n, 1/n, \ldots, 1/n)$. However, a characterization of all the PMFs at which this function attains a maximum value has not been carried out before, which we do in this work. This characterization also suggests the following problem: do the restrictions of $\mathcal{L}_D$ over certain subsets of $\mathcal{P}_n$ attain maximum values in their respective domains? If so, what are the PMFs at which these maximum values are attained? We give a partial solution to this problem for the binary case $D = 2$.

ii) We introduce the problem of finding a minimum expected length binary prefix code (hereafter known as an optimal code) among the prefix codes that satisfy the following constraint: all the possible concatenations of the codewords must satisfy the (d, k) runlength-limited (RLL) constraint, i.e., the number of zeros between any two successive ones in them is at least d and the length of any run of consecutive zeros is at most k. For certain (d, k) pairs, we show that this problem can be reduced to a well-studied problem of finding a prefix code with the minimum expected cost when each letter of the alphabet has a non-negative cost associated with it. Also, for these (d, k) pairs, we examine if the optimal codes satisfy a certain maximality property defined with respect to the prefix condition and the RLL constraint.

iii) We then study a property of prefix codes: of it being synchronous or not. A prefix code is said to be *synchronous* if there exists a word $x \in \mathcal{A}^*$ such that for all $w \in \mathcal{A}^*$, we have $wx \in X^*$. Capocelli et al. (1988) have given an algorithm to determine if a given prefix code is synchronous or not, which has subsequently been improved. In our work, we devise an algorithm based on the notion of dangling suffixes, similar to the classical Sardinas-Patterson test for determining whether or not a given code is uniquely decodable. We show that our algorithm has a much better worst-case performance when compared to that

of the improved version of the algorithm of Capocelli et al. In this process, we also slightly improve upon the known necessary and sufficient condition for a prefix code to be synchronous.

iv) Finally we look into a class of prefix codes called the bifix codes. A bifix code is a prefix code in which no codeword is a suffix of another codeword. For a finite sequence of non-decreasing natural numbers $\mathsf{L} = (l_1, l_2, \ldots, l_n)$, there is no known efficient algorithm to determine the existence of a bifix code whose sequence of codeword lengths is the same as L (henceforth referred to as a bifix code for L). For a finite sequence L taking only two distinct integer values (called a two-level sequence), we show that the problem of deciding the existence of a bifix code for L can be converted to a problem of finding a particular subset of vertices from certain graphs derived from de Bruijn graphs. We then conjecture an efficient way of finding these subsets. Ahlswede's conjecture (1996) is another problem which has led to a lot of work in the field of bifix codes. It states that if a sequence L has a Kraft sum ($\sum_i 2^{-l_i}$) less than or equal to $3/4$, then there exists a binary bifix code for L. This conjecture has been proved when L is a two-level sequence. We give an alternate proof of this by pointing out a new general way of constructing a bifix code for a two-level sequence L with Kraft sum less than or equal to $3/4$.