entropy

Article

Entropic Bounds on the Average Length of Codes with a Space

Roberto Bruno † and **Ugo Vaccaro** *,†

Department of Computer Science, University of Salerno, 84084 Fisciano, Italy; rbruno@unisa.it
* Correspondence: uvaccaro@unisa.it
† These authors contributed equally to this work.

Abstract: We consider the problem of constructing prefix-free codes in which a designated symbol, a *space*, can only appear at the end of codewords. We provide a linear-time algorithm to construct *almost*-optimal codes with this property, meaning that their average length differs from the *minimum possible* by at most one. We obtain our results by uncovering a relation between our class of codes and the class of one-to-one codes. Additionally, we derive upper and lower bounds to the average length of optimal prefix-free codes with a space in terms of the source entropy.

Keywords: codes; entropy; average length; prefix-free codes; one-to-one codes

Citation: Bruno, R.; Vaccaro, U. Entropic Bounds on the Average Length of Codes with a Space. *Entropy* **2024**, *26*, 283. https://doi.org/10.3390/e26040283

Academic Editors: Igal Sason and Boris Ryabko

Received: 13 November 2023
Revised: 15 March 2024
Accepted: 23 March 2024
Published: 26 March 2024

1. Introduction

Modern natural languages achieve the unique parsability of written texts by inserting a special character (i.e., a *space*) between words [1] (See [2] for a few exceptions to this rule). Classical Information Theory, instead, studies codes that achieve the unique parsability of texts by imposing diverse combinatorial properties on the codeword set: e.g., the prefix property, unique decipherability, etc. [3]. With respect to the *efficiency* of such codes (usually measured via the average number of code symbols per source symbol), it is well known that the Shannon entropy of the information source constitutes a fundamental lower bound for it. On the other hand, if one drops the property of the unique parsability of code messages into individual codewords, and simply requires that different source symbols be encoded with different codewords, one can obtain codes (known as *one-to-one codes*) with efficiency below the source Shannon entropy (although not too much below; see, e.g., [4,5]).

Jaynes [6] took the approach of directly studying source codes in which a designated character of the code alphabet is *exclusively* used as a word delimiter. More precisely, Jaynes studied the possible decrease of the noiseless channel capacity (see [7], p. 8) associated with any code that uses a designated symbol as an end-of-codeword mark, as compared with the noiseless channel capacity of an unconstrained code. Quite interestingly, Jaynes proved that the decrease of the noiseless channel capacity of codes with an end-of-codeword mark becomes negligible, as the maximum codeword length increases.

In this paper, we study the problem of constructing prefix-free codes where a specific symbol (referred to as a 'space') can only be positioned at the end of codewords. We refer to this kind of prefix code as *prefix codes ending with a space*. We develop a linear-time algorithm that constructs 'almost'-optimal codes with this characteristic, in the sense that the average length of the constructed codes is at most one unit longer than the *shortest possible* average length of any prefix-free code in which the space can appear only at the end of codewords. We prove this result by highlighting a connection between our type of codes and the well-known class of one-to-one codes. We also provide upper and lower limits of the average length of optimal prefix codes ending with a space, expressed in terms of the source entropy and the cardinality of the code alphabet.

The paper is structured as follows. In Section 2, we illustrate the relationships between prefix codes ending with a space and one-to-one codes. Specifically, we prove that, from one-to-one codes, one can easily construct prefix codes ending with a space, and we give an upper bound on the average length of the constructed codes. Successively, we show that, if we remove all the spaces from the codewords of prefix codes ending with a space, one obtains a one-to-one code. This result allows us to prove that the average length of our prefix codes ending with a space differs from the minimum possible by at most one. In Sections 3 and 4, we derive upper and lower bounds on the average length of optimal prefix codes ending with a space in terms of the source entropy and the cardinality of the code alphabet.

2. Relations between One-to-One Codes and Prefix Codes Ending with a Space

Let $S = \{s_1, \ldots, s_n\}$ be the set of source symbols, $\mathbf{p} = (p_1, \ldots, p_n)$ be a probability distribution on the set S (that is, p_i is the probability of source symbol s_i), and $\{0, \ldots, k-1\}$ be the code alphabet. We denote by $\{0, \ldots, k-1\}^+$ the set of all non-empty sequences on the code alphabet $\{0, \ldots, k-1\}$, $k \geq 2$, and by $\{0, \ldots, k-1\}^+\sqcup$ the set of all non-empty k-ary sequences that ends with the special symbol $\sqcup$, i.e., the *space* symbol.

A *prefix-free code* ending with a space is a one-to-one mapping:

$$C : S \longmapsto \{0, \ldots, k-1\}^+ \cup \{0, \ldots, k-1\}^+\sqcup$$

in which no codeword $C(s)$ is a prefix of another codeword $C(s')$, for any $s, s' \in S$, $s \neq s'$.

A k-ary *one-to-one code* (see [4,5,8–11] and the references therein quoted) is a bijective mapping $D : S \longmapsto \{0, \ldots, k-1\}^+$ from S to the set of all non-empty sequences over the alphabet $\{0, \ldots, k-1\}$, $k \geq 2$.

The average length of an arbitrary code for the set of source symbols $S = \{s_1, \ldots, s_n\}$, with probabilities $\mathbf{p} = (p_1, \ldots, p_n)$, is $\sum_{i=1}^{n} p_i \ell_i$, where ℓ_i is the number of alphabet symbols in the codeword associated with the source symbol s_i.

Without loss of generality, we assume that probability distribution $\mathbf{p} = (p_1, \ldots, p_n)$ is ordered, that is $p_1 \geq \ldots \geq p_n$. Under this assumption, it is apparent that the *best* one-to-one code proceeds by assigning the shortest codeword (e.g., in the binary case, codeword 0) to the highest probability source symbol s_1, the next shortest codeword 1 to the source symbol s_2, the codeword 00 to s_3, the codeword 01 to s_4, and so on.

An equivalent approach for constructing an optimal one-to-one code, which we will use later, proceeds as follows: Let us consider the first n non-empty k-ary strings according to the *radix* order [12] (that is, the k-ary strings are ordered by length and, for equal lengths, ordered according to the lexicographic order). We assign the strings to the symbols $s_1, \ldots, s_n$ in S by increasing the string length and, for equal lengths, by inverse order according to the lexicographic order. For example, in the binary case, we assign the codeword 1 to the highest probability source symbol s_1, the codeword 0 to the source symbol s_2, the codeword 11 to s_3, the codeword 10 to s_4, and so on. Therefore, one can see that, in the general case of a k-ary code alphabet, $k \geq 2$, an optimal one-to-one code of minimal average length assigns a codeword of length ℓ_i to the i-th symbol $s_i \in S$, where ℓ_i is given by:

$$\ell_i = \lfloor \log_k((k-1)\, i + 1) \rfloor. \tag{1}$$

Moreover, the codewords of an optimal k-ary one-to-one code can be represented as the nodes of a k-ary tree of maximum depth $h = \lceil \log_k(n - \lceil n/k \rceil) \rceil$, where, for each node v, the k-ary string (codeword) associated with v is obtained by concatenating all the labels in the path from the root of the tree to v.

It is evident that, if we apply the above encoding to a *sequence* of source symbols, the obtained binary sequence is *not* uniquely parsable in terms of individual codewords. Let us see how one can recover unique parsability by appending a space $\sqcup$ to judiciously chosen codewords of an optimal one-to-one code. To gain insight, let us consider the following example. Let $S = \{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9, s_{10}\}$ be the set of source symbols, and and let us assume that the code alphabet is $\{0, 1\}$. Under the standing hypothesis that $p_1 \geq \ldots \geq p_{10}$, one has that the best prefix-free code C one can obtain by the procedure of appending a space $\sqcup$ to codewords of the optimal one-to-one code for S is the following:

$$\begin{aligned}
C(s_1) &= 1\sqcup \\
C(s_2) &= 0\sqcup \\
C(s_3) &= 11 \\
C(s_4) &= 10 \\
C(s_5) &= 01\sqcup \\
C(s_6) &= 00\sqcup \\
C(s_7) &= 011 \\
C(s_8) &= 010 \\
C(s_9) &= 001 \\
C(s_{10}) &= 000.
\end{aligned}$$

Observe that we started from the codewords of the optimal one-to-one code constructed according to the *second* procedure previously described. Moreover, note that the codewords associated with symbols s_1, s_2, s_5, and s_6 necessitate the space character $\sqcup$ at their end; otherwise, the unique parsability of some encoded sequences of source symbols would not be guaranteed. On the other hand, the codewords associated with symbols s_3, s_4, s_7, s_8, s_9, and s_{10} do not necessitate the space character $\sqcup$. Indeed, the codeword set

$$\{1\sqcup, 0\sqcup, 11, 10, 01\sqcup, 00\sqcup, 011, 010, 001, 000\}$$

satisfies the prefix-free condition (i.e., no codeword is a prefix of any other); therefore, it guarantees the unique parsability of any coded message in terms of individual codewords.

The idea of the above example can be generalized, as shown in the following lemma.

Lemma 1. *Let $S = \{s_1, \ldots, s_n\}$ be the set of source symbols and $\mathbf{p} = (p_1, \ldots, p_n)$, $p_1 \geq \ldots \geq p_n > 0$, be a probability distribution on S. Let $\{0, \ldots, k-1\}$ be the $k \geq 2$-ary code alphabet. We can construct a prefix-free code $C : S \longmapsto \{0, \ldots, k-1\}^+ \cup \{0, \ldots, k-1\}^+\sqcup$, in $O(n)$, such that its average length $L(C)$ satisfies*

$$L(C) = L_+ + \sum_{i=1}^{\frac{k^{h-1}-1}{k-1}-1} p_i + \sum_{i=\frac{k^h+k^{h-1}-2}{k-1}-\lceil n/k\rceil}^{\frac{k^h-1}{k-1}-1} p_i \tag{2}$$

$$\leq L_+ + \sum_{i=1}^{\lceil n/k\rceil-1} p_i, \tag{3}$$

where L_+ is the average length of an optimal one-to-one code $D : S \longmapsto \{0, \ldots, k-1\}^+$ and $h = \lceil \log_k(n - \lceil n/k\rceil)\rceil$.

Proof. Under the standing hypothesis that the probabilities of the source symbols are ordered from the largest to the smallest, we show how to construct a prefix-free code—by appending the special character $\sqcup$ to the end of (some) codewords of an optimal one-to-one code for S—having the average length upper bounded by (2).

Among the class of all the prefix-free codes that one can obtain by appending the character $\sqcup$ to the end of (some) codewords of an optimal one-to-one code for S, we aim to construct the one with the minimum average length. Therefore, we have to ensure that, in the k-ary tree representation of the code, the following basic condition holds: For any pair of nodes v_i and v_j, $i < j$, associated with the symbols s_i and s_j, the depth of the node v_j is not smaller than the depth of the node v_i. In fact, if it were otherwise, the average length of the code could be improved.

Therefore, by recalling that $h = \lceil \log_k(n - \lceil n/k\rceil)\rceil$ is the height of the k-ary tree associated with an optimal one-to-one code, we have that the prefix-free code of the minimum average length that one can obtain by appending the special character $\sqcup$ to the end of (some) codewords of an optimal one-to-one code for S assigns a codeword of length ℓ_i to the i-th symbol $s_i \in S$, where ℓ_i is given by:

$$\ell_i = \begin{cases} \lfloor \log_k((k-1)\,i+1) \rfloor + 1, & \text{if } i \leq \frac{k^{h-1}-1}{k-1} - 1, \\ \lfloor \log_k((k-1)\,i+1) \rfloor + 1, & \text{if } i \geq \frac{k^h+k^{h-1}-2}{k-1} - \lceil n/k \rceil \\ & \text{and } i \leq \frac{k^h-1}{k-1} - 1, \\ \lfloor \log_k((k-1)\,i+1) \rfloor, & \text{otherwise.} \end{cases} \tag{4}$$

We stress that the obtained prefix-free code is not necessarily a prefix-free code $C : S \longmapsto \{0,\ldots,k-1\}^+ \cup \{0,\ldots,k-1\}^+ \sqcup$ of minimum average length. Now, we justify the expression (4). First, since the probabilities $p_1,\ldots,p_n$ are ordered in non-increasing fashion, the codeword lengths ℓ_i of the code are ordered in non-decreasing fashion, that is $\ell_1 \leq \ldots \leq \ell_n$. Therefore, in the k-ary tree representation of the code, it holds the desired basic condition: For any pair of nodes v_i and v_j, $i < j$, associated with the symbols s_i and s_j, the depth of the node v_i is smaller than or equal to the depth of the node v_j.

Furthermore, we need to append the space character only to the k-ary strings that are the prefix of some others. Therefore, let us consider the first n non-empty k-ary strings according to the *radix* order [12], in which, we recall, the k-ary strings are ordered by length and, for equal lengths, ordered according to the lexicographic order. We have that the number of strings that are a prefix of some others is exactly $\lceil \frac{n}{k} \rceil - 1$. One obtains this number by seeing the strings as corresponding to nodes in a k-ary tree with labels $0,\ldots,k-1$ on the edges. The number of strings that are a prefix of some others (among the n strings) is *exactly* equal to the number of internal nodes (except the root) in such a tree. This number of internal nodes is equal to $\lceil \frac{N-1}{k} \rceil - 1$, where N is the total number of nodes that, in our case, is equal to $N = n+1$ (i.e., N counts also the root of the tree).

Moreover, starting from the optimal one-to-one code constructed according to our second method, that is by assigning k-ary strings to the symbols by increasing length and, for equal lengths, by inverse order according to the lexicographic order, one can verify that the $\lceil \frac{n}{k} \rceil - 1$ internal nodes are associated with the codewords of the symbols s_i, for i that goes from 1 to $\frac{k^{h-1}-1}{k-1} - 1$, and from $\frac{k^h+k^{h-1}-2}{k-1} - \lceil n/k \rceil$ to $\frac{k^h}{k-1} - 2$.

In fact, since the height of the k-ary tree is $h = \lceil \log_k(n - \lceil n/k \rceil) \rceil$ and since all the levels of the tree, except the last two, are full, we need to append the space to all symbols from 1 to $\frac{k^{h-1}-1}{k-1} - 1$. While on the second-to-last level, we have to append the space only to the remaining internal nodes associated with the symbols s_i, where i goes from $\frac{k^h+k^{h-1}-2}{k-1} - \lceil n/k \rceil$ to $\frac{k^h-1}{k-1} - 1$. Those remaining nodes are exactly, among all the nodes in the second-to-last level, the ones associated with the symbols that have smaller probabilities. Thus, we obtain (4).

Summarizing, we can construct a prefix-free code $C : S \longmapsto \{0,\ldots,k-1\}^+ \cup \{0,\ldots,k-1\}^+ \sqcup$, in $O(n)$ time, with lengths defined as in (4), starting from an optimal one-to-one code. Thus:

$$\begin{aligned} L(C) &= \sum_{i=1}^{n} p_i \ell_i \\ &= \sum_{i=1}^{n} p_i \lfloor \log_k((k-1)\,i+1) \rfloor + \sum_{i=1}^{\frac{k^{h-1}-1}{k-1}-1} p_i + \sum_{i=\frac{k^h+k^{h-1}-2}{k-1}-\lceil n/k \rceil}^{\frac{k^h-1}{k-1}-1} p_i \\ &= L_+ + \sum_{i=1}^{\frac{k^{h-1}-1}{k-1}-1} p_i + \sum_{i=\frac{k^h+k^{h-1}-2}{k-1}-\lceil n/k \rceil}^{\frac{k^h-1}{k-1}-1} p_i \\ &\leq L_+ + \sum_{i=1}^{\lceil n/k \rceil - 1} p_i \quad \text{(since we are adding } \lceil n/k \rceil - 1 \; p_i\text{'s, and the } p_i\text{'s are ordered).} \end{aligned}$$

□

Note that, from Lemma 1, we obtain that the average length of any *optimal* (i.e., of minimum average length) prefix-free code ending with a space is upper bounded by the Formula (2). Furthermore, we have an upper bound on the average length of the optimal prefix-free codes ending with a space in terms of the average length of optimal one-to-one codes.

We can also derive a *lower* bound on the average length of optimal prefix-free codes ending with a space in terms of the average length of optimal one-to-one codes. For such a purpose, we need two intermediate results. We first recall that, given a k-ary code C, its codewords can be represented as nodes in a k-ary tree with labels $0, \dots, k-1$ on the edges. Indeed, for each node v, the k-ary string (codeword) associated with v can be obtained by concatenating all the labels in the path from the root of the tree to v. We also recall that, in prefix-free codes, the codewords correspond to the node leaves of the associated tree, while in one-to-one codes, the codewords may correspond also to the internal nodes of the associated tree.

Lemma 2. *Let $S = \{s_1, \dots, s_n\}$ be the set of source symbols, and let $\mathbf{p} = (p_1, \dots, p_n)$, $p_1 \geq \dots \geq p_n > 0$, be a probability distribution on S. There exists an* optimal prefix-free *code ending with a space $C : S \longmapsto \{0, \dots, k-1\}^+ \cup \{0, \dots, k-1\}^+\sqcup$ such that the following property holds: For any internal node v (except the root) of the tree representation of C, if we denote by w the k-ary string associated with the node v, then the string $w\sqcup$ belongs to the codeword set of C.*

Proof. Let C be an arbitrary optimal prefix-free code ending with a space. Let us assume that, in the tree representation of C, there exists an internal node v whose associated string w is such that $w\sqcup$ *does not* belong to the codeword set of C. Since v is an internal node, there is at least a leaf x, which is a descendant of v, whose associated string is the codeword of some symbol s_j. We modify the encoding, by assigning the codeword $w\sqcup$ to the symbol s_j. The new encoding is still prefix-free, and its average length can only decrease since the length of the newly assigned codeword to s_j cannot be greater than the previous one. We can repeat the argument for all internal nodes that do not satisfy the property stated in the lemma to complete the proof. □

Lemma 3. *Let $C : S \longmapsto \{0, \dots, k-1\}^+ \cup \{0, \dots, k-1\}^+\sqcup$ be an arbitrary prefix-free code, then the code $D : S \longmapsto \{0, \dots, k-1\}^+$ one obtains from C by removing the space $\sqcup$ from each codeword of C is a* one-to-one *code.*

Proof. The proof is straightforward. Since C is prefix-free, it holds that, for any pair $s_i, s_j \in S$, with $s_i \neq s_j$, the codeword $C(s_i)$ is not a prefix of $C(s_j)$ and vice versa. Therefore, since D is obtained from C by removing the space, we have four cases:

1. $C(s_i) = D(s_i)$ and $C(s_j) = D(s_j)$: then $D(s_i) \neq D(s_j)$ since $C(s_i) \neq C(s_j)$;
2. $C(s_i) = D(s_i)\sqcup$ and $C(s_j) = D(s_j)\sqcup$: then $D(s_i) \neq D(s_j)$ since $C(s_i)$ is not a prefix of $C(s_j)$ and vice versa;
3. $C(s_i) = D(s_i)\sqcup$ and $C(s_j) = D(s_j)$: then $D(s_i) \neq D(s_j)$ since $C(s_j)$ is not a prefix of $C(s_i)$;
4. $C(s_i) = D(s_i)$ and $C(s_j) = D(s_j)\sqcup$: then $D(s_i) \neq D(s_j)$ since $C(s_i)$ is not a prefix of $C(s_j)$.

Therefore, for any pair $s_i, s_j \in S$, with $s_i \neq s_j$, $D(s_i) \neq D(s_j)$, and D is a one-to-one code. □

We can now derive a lower bound on the average length of optimal prefix-free codes with space in terms of the average length of optimal one-to-one codes.

Lemma 4. *Let $S = \{s_1, \dots, s_n\}$ be the set of source symbols, and let $\mathbf{p} = (p_1, \dots, p_n)$, $p_1 \geq \dots \geq p_n > 0$, be a probability distribution on S, then the average of an optimal prefix-free code $C : S \longmapsto \{0, \dots, k-1\}^+ \cup \{0, \dots, k-1\}^+ \sqcup$ satisfies*

$$L(C) \geq L_+ + \sum_{i=1}^{\lceil n/k \rceil - 1} p_{n-i+1}, \tag{5}$$

where L_+ is the average length of an optimal k-ary one-to-one code on S.

Proof. From Lemma 2, we know that there exists an optimal prefix-free code C with a space in which exactly $\lceil \frac{n}{k} \rceil - 1$ codewords contain the space character at the end. Let $A \subset \{1, \dots, n\}$ be the set of indices associated with the symbols whose codeword contains the space. Moreover, from Lemma 3, we know that the code D obtained by removing the space from C is a one-to-one code. Putting it all together, we obtain that

$$L(D) = L(C) - \sum_{i \in A} p_i. \tag{6}$$

From (6), we have that

$$\begin{aligned} L(C) &= L(D) + \sum_{i \in A} p_i \\ &\geq L_+ + \sum_{i \in A} p_i \qquad \text{(since } D \text{ is a one-to-one code)} \\ &\geq L_+ + \sum_{i=1}^{\lceil n/k \rceil - 1} p_{n-i+1} \quad \text{(since } A \text{ contains } \lceil \tfrac{n}{k} \rceil - 1 \text{ elements).} \end{aligned}$$

□

We notice that the difference between the expression (2) and the lower bound (5) is, because of (3), less than

$$\sum_{i=1}^{\lceil n/k \rceil - 1} p_i - \sum_{i=1}^{\lceil n/k \rceil - 1} p_{n-i+1} < 1; \tag{7}$$

therefore, the prefix-free codes ending with a space that we construct in Lemma 1 have an average length that differs from the minimum possible by at most one. Moreover, since both the upper bound (3) and the lower bound (5) are easily computable, we can determine the average length of an *optimal* prefix-free code $C : S \longmapsto \{0, \dots, k-1\}^+ \cup \{0, \dots, k-1\}^+ \sqcup$ with a tolerance at most of one. One can also see that the left-hand side of (7) is, often, much smaller than one.

In the following sections, we will focus on providing upper and lower bounds on the average length L_+ of k-ary optimal one-to-one codes in terms of the k-ary Shannon entropy $H_k(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log_k p_i$ of the source distribution $\mathbf{p}$. Because of Lemmas 1 and 4, this will give us the corresponding upper and lower bounds on the average length of optimal prefix-free codes ending with a space.

3. Lower Bounds on the Average Length

In this section, we provide lower bounds on the average length of the optimal one-to-one code and, subsequently, thanks to Lemma 4, on the average length of the optimal prefix-free code with a space. For technical reasons, it will be convenient to consider one-to-one codes that make use of the empty word ϵ, that is one-to-one mappings $D_\epsilon : S \longmapsto \{0, 1, \dots, k-1\}^+ \cup \{\epsilon\}$. It is easy to see (cf. (1)) that the optimal one-to-one code that makes use of the empty word assigns to the i-th symbol $s_i \in S$ a codeword of length ℓ_i given by:

$$\ell_i = \lfloor \log_k((k-1)i) \rfloor. \tag{8}$$

where k is the cardinality of the code alphabet.

Thus, by denoting by L_+ the average length of the optimal one-to-one code that *does* not make use of the empty word and with L_ϵ the average length of the optimal one-to-one code that *does* use it, we obtain the following relation:

$$L_+ = L_\epsilon + \sum_{i=1}^{\lfloor \log_k \lceil \frac{n-1}{k} \rceil \rfloor} p_{\frac{k^i-1}{k-1}}. \tag{9}$$

Our first result is a generalization of the lower bound to the average length of the optimal one-to-one codes presented in [5], from the binary case to the general case of k-ary alphabets, $k \geq 2$. Our proof technique differs from that of [5] since we are dealing with a set of source symbols of *bounded* cardinality (in [5], the authors considered the case of a numerable set of source symbols).

Lemma 5. *Let $S = \{s_1, \ldots, s_n\}$ be the set of source symbols and $\mathbf{p} = (p_1, \ldots, p_n)$ be a probability distribution on S, with $p_1 \geq \ldots \geq p_n$. The average length L_ϵ of the optimal one-to-one code $D : \{s_1, \ldots, s_n\} \to \{0, \ldots, k-1\}^+ \cup \{\epsilon\}$ satisfies*

$$\begin{aligned} L_\epsilon >& H_k(\mathbf{p}) - (H_k(\mathbf{p}) + \log_k(k-1)) \log_k\left(1 + \frac{1}{H_k(\mathbf{p}) + \log_k(k-1)}\right) \\ &- \log_k(H_k(\mathbf{p}) + \log_k(k-1) + 1), \end{aligned}$$

where $H_k(\mathbf{p}) = -\sum_{i=1}^n p_i \log_k p_i$.

Proof. The proof is an adaptation of Alon et al.'s proof [4] from the binary case to the $k \geq 2$-ary case.

We recall that the optimal one-to-one code (i.e., whose average length achieves the minimum L_ϵ) has codeword lengths ℓ_i given by:

$$\ell_i = \lfloor \log_k((k-1)i) \rfloor. \tag{10}$$

For each $j \in \{0, \ldots, \lfloor \log_k n \rfloor\}$, let us define the quantities q_j as

$$q_j = \sum_{i=\frac{k^j-1}{k-1}+1}^{\frac{k^{j+1}-1}{k-1}} p_i.$$

It holds that $\sum_{j=0}^{\lfloor \log_k n \rfloor} q_j = 1$. Let Y be a random variable that takes values in $\{0, \ldots, \lfloor \log_k n \rfloor\}$ according to the probability distribution $\mathbf{q} = (q_0, \ldots, q_{\lfloor \log_k n \rfloor})$, that is

$$\forall j \in \{0, \ldots, \lfloor \log_k n \rfloor\} \quad \Pr\{Y = j\} = q_j.$$

From (10), we have

$$\begin{aligned} L_\epsilon &= \sum_{i=1}^n \lfloor \log_k((k-1)i) \rfloor p_i \\ &= \sum_{j=0}^{\lfloor \log_k n \rfloor} \sum_{i=\frac{k^j-1}{k-1}+1}^{\frac{k^{j+1}-1}{k-1}} \lfloor \log_k((k-1)i) \rfloor p_i \\ &= \sum_{j=0}^{\lfloor \log_k n \rfloor} j q_j = \mathbb{E}[Y]. \end{aligned} \tag{11}$$

By applying the entropy grouping rule ([3], Ex. 2.27) to the distribution $\mathbf{p}$, we obtain

$$\begin{aligned}
H_2(\mathbf{p}) &= H_2(\mathbf{q}) + \sum_{j=0}^{\lfloor \log_k n \rfloor} q_j H_2\left(\frac{p_{\frac{k^j-1}{k-1}+1}}{q_j}, \ldots, \frac{p_{\frac{k^{j+1}-1}{k-1}}}{q_j}\right) \\
&\le H_2(\mathbf{q}) + \sum_{j=0}^{\lfloor \log_k n \rfloor} q_j \log_2 k^j \qquad \left(\text{since } H_2\left(\frac{p_{\frac{k^j-1}{k-1}+1}}{q_j}, \ldots, \frac{p_{\frac{k^{j+1}-1}{k-1}}}{q_j}\right) \le \log_2 k^j\right) \\
&= H_2(\mathbf{q}) + \sum_{j=0}^{\lfloor \log_k n \rfloor} j q_j \log_2 k \\
&= H_2(\mathbf{q}) + \mathbb{E}[Y] \log_2 k. \qquad (12)
\end{aligned}$$

We now derive an upper bound to $H_2(Y) = H_2(q)$ in terms of the expected value $\mathbb{E}[Y]$.

To this end, let us consider an auxiliary random variable Y' with the same distribution of Y, but with values ranging from 1 to $\lfloor \log_k(n) \rfloor + 1$ (instead of from 0 to $\lfloor \log_k(n) \rfloor$). It is easy to verify that $\mu = \mathbb{E}[Y'] = \mathbb{E}[Y] + 1$.

Let α be a positive number, whose value will be chosen later. We obtain that

$$\begin{aligned}
H_k(Y) - \alpha\mu &= \sum_{i=1}^{\lfloor \log_k(n) \rfloor + 1} q_{i-1} \log_k \frac{1}{q_{i-1}} - \alpha \sum_{j=1}^{\lfloor \log_k(n) \rfloor + 1} j q_{j-1} \\
&= \sum_{i=1}^{\lfloor \log_k(n) \rfloor + 1} q_{i-1} \log_k \frac{1}{q_{i-1}} + \sum_{j=1}^{\lfloor \log_k(n) \rfloor + 1} (-\alpha j) q_{j-1} \\
&= \sum_{i=1}^{\lfloor \log_k(n) \rfloor + 1} q_{i-1} \log_k \frac{1}{q_{i-1}} + \sum_{j=1}^{\lfloor \log_k(n) \rfloor + 1} q_{j-1} \log_k (k^{-\alpha j}) \\
&= \sum_{i=1}^{\lfloor \log_k(n) \rfloor + 1} q_{i-1} \log_k \frac{k^{-\alpha i}}{q_{i-1}} \\
&\le \log_k \sum_{i=1}^{\lfloor \log_k(n) \rfloor + 1} k^{-\alpha i} \qquad (\text{ by Jensen's inequality}) \\
&= \log_k \left[\left(\frac{1}{k^\alpha}\right)\left(\frac{1 - k^{-\alpha(\lfloor \log_k(n) \rfloor + 1}}{1 - k^{-\alpha}}\right)\right] \\
&\le \log_k \left(\frac{1 - k^{-\alpha(\log_k(n)+1)}}{k^\alpha - 1}\right) \\
&= \log_k \left(\frac{1 - (kn)^{-\alpha}}{k^\alpha - 1}\right).
\end{aligned}$$

By substituting $\log_k \frac{\mu}{\mu-1}$ with α in the obtained inequality

$$H_k(Y) \le \alpha\mu + \log_k \left(\frac{1 - (kn)^{-\alpha}}{k^\alpha - 1}\right),$$

we obtain

$$H_k(Y) \le \mu \log_k \frac{\mu}{\mu - 1} + \log_k(\mu - 1) + \log_k \left(1 - \left(\frac{1}{kn}\right)^{\log_k \frac{\mu}{\mu-1}}\right). \qquad (13)$$

Since $\left(\frac{1}{kn}\right)^{\log_k \frac{\mu}{\mu-1}}$ is decreasing in μ, and because $\mu = \mathbb{E}[Y] + 1 > 1$, we obtain:

$$H_k(Y) < \mathbb{E}[Y] \log_k \left(1 + \frac{1}{\mathbb{E}[Y]}\right) + \log_k(\mathbb{E}[Y] + 1). \qquad (14)$$

By applying (14) to (12) and since $H_k(Y) = \frac{H_2(Y)}{\log_2 k}$, we obtain

$$H_2(\mathbf{p}) < \mathbb{E}[Y] \log_2 k + \mathbb{E}[Y] \log_2 \left(1 + \frac{1}{E[Y]}\right) + \log_2(E[Y] + 1). \tag{15}$$

From (11), we have that $L_\epsilon = \mathbb{E}[Y]$; moreover, from the inequality (28) of Lemma 7 (proven in the next Section 4), we know that

$$L_\epsilon \leq H_k(\mathbf{p}) + \log_k(k-1). \tag{16}$$

Hence, since the function $f(z) = z \log_k \left(1 + \frac{1}{z}\right)$ is increasing in z, we can apply (16) to upper-bound the term

$$\mathbb{E}[Y] \log_2 \left(1 + \frac{1}{E[Y]}\right),$$

to obtain the following inequality:

$$\begin{aligned} H_2(\mathbf{p}) <& L_\epsilon \log_2 k + (H_k(\mathbf{p}) + \log_k(k-1)) \log_2 \left(1 + \frac{1}{H_k(\mathbf{p}) + \log_k(k-1)}\right) \\ &+ \log_2(H_k(\mathbf{p}) + \log_k(k-1) + 1). \end{aligned} \tag{17}$$

Rewriting (17), we finally obtain

$$\begin{aligned} L_\epsilon >& H_k(\mathbf{p}) - (H_k(\mathbf{p}) + \log_k(k-1)) \log_k \left(1 + \frac{1}{H_k(\mathbf{p}) + \log_k(k-1)}\right) \\ &- \log_k(H_k(\mathbf{p}) + \log_k(k-1) + 1), \end{aligned}$$

and that concludes our proof. $\square$

By bringing into play the size of the largest mass in addition to the entropy, Lemma 5 can be improved, as shown in the following result.

Lemma 6. *Let $S = \{s_1, \ldots, s_n\}$ be the set of source symbols and $\mathbf{p} = (p_1, \ldots, p_n)$, $p_1 \geq \ldots \geq p_n$, be a probability distribution on S. The average length L_ϵ of the optimal one-to-one code $D : \{s_1, \ldots, s_n\} \to \{0, \ldots, k-1\}^+ \cup \{\epsilon\}$ has the following lower bounds:*

1. *If $0 < p_1 \leq 0.5$,*

$$\begin{aligned} L_\epsilon \geq & H_k(\mathbf{p}) - \left(H_k(\mathbf{p}) - p_1 \log_k \frac{1}{p_1} + (1-p_1)\log_k(k-1)\right) \\ & \log_k \left(1 + \frac{1}{H_k(\mathbf{p}) - p_1 \log_k \frac{1}{p_1} + (1-p_1)\log_k(k-1)}\right) \\ & - \log_k\left(H_k(\mathbf{p}) - p_1 \log_k \frac{1}{p_1} + (1-p_1)\log_k(k-1) + 1\right) \\ & - \log_k \left(1 - \left(\frac{1}{kn}\right)^{\log_k\left(1+\frac{1}{1-p_1}\right)}\right), \end{aligned} \tag{18}$$

2. *if $0.5 < p_1 \leq 1$*

$$\begin{aligned} L_\epsilon \geq & H_k(\mathbf{p}) - (H_k(\mathbf{p}) - \mathcal{H}_k(p_1) + (1-p_1)(1 + \log_k(k-1))) \\ & \log_k \left(1 + \frac{1}{H_k(\mathbf{p}) - \mathcal{H}_k(p_1) + (1-p_1)(1+\log_k(k-1))}\right) \\ & - \log_k(H_k(\mathbf{p}) - \mathcal{H}_k(p_1) + (1-p_1)(1+\log_k(k-1)) + 1) \end{aligned}$$

$$-\log_k\left(1-\left(\frac{1}{kn}\right)^{\log_k\left(1+\frac{1}{1-p_1}\right)}\right), \tag{19}$$

where $\mathcal{H}_k(p_1) = -p_1 \log_k p_1 - (1-p_1)\log_k(1-p_1)$.

Proof. The proof is the same as the proof of Lemma 5. However, we change two steps in the demonstration.

First, since

$$\left(\frac{1}{kn}\right)^{\log_k \frac{\mu}{\mu-1}} = \left(\frac{1}{kn}\right)^{\log_k \frac{\mathbb{E}[Y]+1}{\mathbb{E}[Y]}} = \left(\frac{1}{kn}\right)^{\log_k\left(1+\frac{1}{\mathbb{E}[Y]}\right)}$$

is decreasing in μ and $\mathbb{E}[Y] = L_\epsilon = 0p_1 + 1p_2 + \cdots \geq 1 - p_1$, we have

$$\log_k\left(1-\left(\frac{1}{kn}\right)^{\log_k \frac{\mu}{\mu-1}}\right) \leq \log_k\left(1-\left(\frac{1}{kn}\right)^{\log_k\left(1+\frac{1}{1-p_1}\right)}\right). \tag{20}$$

Hence, by applying (20) to the right-hand side of (13), we obtain

$$H_k(Y) \leq \mathbb{E}[Y]\log_k\left(1+\frac{1}{\mathbb{E}[Y]}\right) + \log_k(\mathbb{E}[Y]+1) + \log_k\left(1-\left(\frac{1}{kn}\right)^{\log_k\left(1+\frac{1}{1-p_1}\right)}\right). \tag{21}$$

Now, by applying (21) (instead of (14)) to (12) and since $H_k(Y) = \frac{H_2(Y)}{\log_2 k}$, we obtain

$$\begin{aligned} H_2(\mathbf{p}) \leq & \mathbb{E}[Y]\log_2 k + \mathbb{E}[Y]\log_2\left(1+\frac{1}{E[Y]}\right) + \log_2(E[Y]+1) \\ & + \log_2\left(1-\left(\frac{1}{kn}\right)^{\log_k\left(1+\frac{1}{1-p_1}\right)}\right). \end{aligned} \tag{22}$$

Here, instead of applying the upper bound:

$$L_\epsilon \leq H_k(\mathbf{p}) + \log_k(k-1)$$

of Lemma 7 to the right-hand side of (22), we apply the improved version:

$$L_\epsilon \leq \begin{cases} H_k(\mathbf{p}) - p_1 \log_k \frac{1}{p_1} + (1-p_1)\log_k(k-1) \text{ if } 0 < p_1 \leq 0.5, \\ H_k(\mathbf{p}) - \mathcal{H}_k(p_1) + (1-p_1)\log_k 2(k-1) \text{ if } 0.5 < p_1 \leq 1, \end{cases}$$

proven in Lemma 8 of the Section 4. Then, we simply need to rewrite the inequality, concluding the proof. □

Thanks to Lemma 4 and the Formula (9), the above lower bounds on L_ϵ can be applied to derive our main results for prefix-free codes with a space, as shown in the following theorems.

Theorem 1. *The average length of an optimal prefix-free code with space* $C : S \longmapsto \{0,\ldots,k-1\}^+ \cup \{0,\ldots,k-1\}^+ \sqcup$ *satisfies*

$$\begin{aligned} L(C) > & H_k(\mathbf{p}) - (H_k(\mathbf{p}) + \log_k(k-1))\log_k\left(1+\frac{1}{H_k(\mathbf{p}) + \log_k(k-1)}\right) \\ & - \log_k(H_k(\mathbf{p}) + \log_k(k-1)+1) + \sum_{i=1}^{\lceil \frac{n}{k} \rceil - 1} p_{n-i+1} + \sum_{i=1}^{\lfloor \log_k \lceil \frac{n-1}{k} \rceil \rfloor} p_{\frac{k^i-1}{k-1}}. \end{aligned} \tag{23}$$

Proof. From Lemma 4 and the Formula (9), we have

$$L(C) \geq L_\epsilon + \sum_{i=1}^{\lceil \frac{n}{k} \rceil - 1} p_{n-i+1} + \sum_{i=1}^{\lfloor \log_k \lceil \frac{n-1}{k} \rceil \rfloor} p_{\frac{k^i-1}{k-1}}. \tag{24}$$

By applying the lower bound (10) of Lemma 5 to (24), we obtain (23). □

Analogously, by exploiting (the possible) knowledge of the maximum source symbol probability value, we have the following result.

Theorem 2. *The average length of the optimal prefix-free code with space $C : S \longmapsto \{0, \ldots, k-1\}^+ \cup \{0, \ldots, k-1\}^+ \sqcup$ has the following lower bounds:*

1. *If $0 < p_1 \leq 0.5$:*

$$\begin{aligned} L(C) \geq & H_k(\mathbf{p}) - \left(H_k(\mathbf{p}) - p_1 \log_k \frac{1}{p_1} + (1-p_1)\log_k(k-1)\right) \\ & \log_k\left(1 + \frac{1}{H_k(\mathbf{p}) - p_1 \log_k \frac{1}{p_1} + (1-p_1)\log_k(k-1)}\right) \\ & - \log_k\left(H_k(\mathbf{p}) - p_1 \log_k \frac{1}{p_1} + (1-p_1)\log_k(k-1) + 1\right) \\ & - \log_k\left(1 - \left(\frac{1}{kn}\right)^{\log_k\left(1+\frac{1}{1-p_1}\right)}\right) + \sum_{i=1}^{\lceil \frac{n}{k} \rceil - 1} p_{n-i+1} + \sum_{i=1}^{\lfloor \log_k \lceil \frac{n-1}{k} \rceil \rfloor} p_{\frac{k^i-1}{k-1}}. \end{aligned} \tag{25}$$

2. *If $0.5 < p_1 \leq 1$:*

$$\begin{aligned} L(C) \geq & H_k(\mathbf{p}) - \left(H_k(\mathbf{p}) - \mathcal{H}_k(p_1) + (1-p_1)(1 + \log_k(k-1))\right) \\ & \log_k\left(1 + \frac{1}{H_k(\mathbf{p}) - \mathcal{H}_k(p_1) + (1-p_1)(1 + \log_k(k-1))}\right) \\ & - \log_k\left(H_k(\mathbf{p}) - \mathcal{H}_k(p_1) + (1-p_1)(1 + \log_k(k-1)) + 1\right) \\ & - \log_k\left(1 - \left(\frac{1}{kn}\right)^{\log_k\left(1+\frac{1}{1-p_1}\right)}\right) + \sum_{i=1}^{\lceil \frac{n}{k} \rceil - 1} p_{n-i+1} + \sum_{i=1}^{\lfloor \log_k \lceil \frac{n-1}{k} \rceil \rfloor} p_{\frac{k^i-1}{k-1}}. \end{aligned} \tag{26}$$

Proof. From Lemma 4 and the Formula (9), we have

$$L(C) \geq L_\epsilon + \sum_{i=1}^{\lceil \frac{n}{k} \rceil - 1} p_{n-i+1} + \sum_{i=1}^{\lfloor \log_k \lceil \frac{n-1}{k} \rceil \rfloor} p_{\frac{k^i-1}{k-1}}. \tag{27}$$

By applying the lower bounds (18) and (19) of Lemma 6 to (27), we obtain (25) or (26) according to the value of the maximum source symbol probability. □

4. Upper Bounds on the Average Length

In this section, we will first derive *upper* bounds on the average length of optimal one-to-one codes. Successively, we will provide corresponding upper bounds on the average length of optimal prefix-free codes ending with a space.

We start by extending the result obtained in [13] from the binary case to the k-ary case, $k \geq 2$.

Lemma 7. *Let $S = \{s_1, \ldots, s_n\}$ be the set of source symbols and $\mathbf{p} = (p_1, \ldots, p_n)$, $p_1 \geq \ldots \geq p_n$, be a probability distribution on S. The average length L_ϵ of the optimal one-to-one code $D : \{s_1, \ldots, s_n\} \to \{0, \ldots, k-1\}^+ \cup \{\epsilon\}$ satisfies*

$$L_\epsilon \le H_k(\mathbf{p}) + \log_k(k-1). \tag{28}$$

Proof. Under the standing hypothesis that $p_1 \ge \ldots \ge p_n$, it holds that

$$\forall i = 1, \ldots, n \qquad p_i \le \frac{1}{i}. \tag{29}$$

We recall that the length of the i-th codeword of the optimal one-to-one code D is equal to

$$\ell_i = \lfloor \log_k((k-1)i) \rfloor. \tag{30}$$

Therefore, from (29), we can upper bound each length ℓ_i as

$$\ell_i = \lfloor \log_k((k-1)i) \rfloor \le \log_k((k-1)i) \le \log_k(k-1) + \log_k \frac{1}{p_i}. \tag{31}$$

Hence, by applying (31) to the average length of D, we obtain

$$L_\epsilon = \sum_{i=1}^{n} p_i \ell_i \le \sum_{i=1}^{n} p_i \left(\log_k(k-1) + \log_k \frac{1}{p_i} \right) = H_k(\mathbf{p}) + \log_k(k-1). \tag{32}$$

This concludes our proof. □

By exploiting the knowledge of the maximum probability value of $\mathbf{p}$, we generalize the upper bound in [5] from $k = 2$ to arbitrary $k \ge 2$.

Lemma 8. *Let $S = \{s_1, \ldots, s_n\}$ be the set of source symbols and $\mathbf{p} = (p_1, \ldots, p_n)$, $p_1 \ge \ldots \ge p_n$, be a probability distribution on S. The average length L_ϵ of the optimal one-to-one code $D : \{s_1, \ldots, s_n\} \to \{0, \ldots, k-1\}^+ \cup \{\epsilon\}$ satisfies*

$$L_\epsilon \le \begin{cases} H_k(\mathbf{p}) - p_1 \log_k \frac{1}{p_1} + (1-p_1)\log_k(k-1) & \text{if } 0 < p_1 \le 0.5, \\ H_k(\mathbf{p}) - \mathcal{H}_k(p_1) + (1-p_1)\log_k 2(k-1) & \text{if } 0.5 < p_1 \le 1. \end{cases} \tag{33}$$

Proof. Let us prove first that the length of an optimal one-to-one code satisfies the inequality:

$$L_\epsilon \le \sum_{i=2}^{n} p_i \log_k(i(k-1)) - 0.5 \sum_{j \ge 2 : \frac{k^j-1}{k-1} \le n} p_{\frac{k^j-1}{k-1}}. \tag{34}$$

Indeed, by recalling that $\ell_1 = \lfloor \log_k(k-1) \rfloor = 0$, we can write L_ϵ as follows:

$$\begin{aligned}
L_\epsilon &= \sum_{i=2}^{n} p_i \lfloor \log_k(i(k-1)) \rfloor \\
&= \sum_{j \ge 1 : \frac{k^j-1}{k-1}+1 \le n} \; \sum_{i=\frac{k^j-1}{k-1}+1}^{\min(\frac{k^{j+1}-1}{k-1}-1, n)} p_i \lfloor \log_k(i(k-1)) \rfloor + \sum_{j \ge 2 : \frac{k^j-1}{k-1} \le n} p_{\frac{k^j-1}{k-1}} \left\lfloor \log_k\left(\frac{k^j-1}{k-1}(k-1) \right) \right\rfloor \\
&= \sum_{j \ge 1 : \frac{k^j-1}{k-1}+1 \le n} \; \sum_{i=\frac{k^j-1}{k-1}+1}^{\min(\frac{k^{j+1}-1}{k-1}-1, n)} p_i \lfloor \log_k(i(k-1)) \rfloor + \sum_{j \ge 2 : \frac{k^j-1}{k-1} \le n} p_{\frac{k^j-1}{k-1}} \log_k(k^j-1) \\
&\quad - \sum_{j \ge 2 : \frac{k^j-1}{k-1} \le n} p_{\frac{k^j-1}{k-1}} (\log_k(k^j-1) - \lfloor \log_k(k^j-1) \rfloor) \\
&\le \sum_{i=2}^{n} p_i \log_k(i(k-1)) - \sum_{j \ge 2 : \frac{k^j-1}{k-1} \le n} p_{\frac{k^j-1}{k-1}} (\log_k(k^j-1) - \lfloor \log_k(k^j-1) \rfloor),
\end{aligned}$$

where the last inequality holds since

$$\sum_{j\geq 1:\frac{k^j-1}{k-1}+1\leq n}\ \sum_{i=\frac{k^j-1}{k-1}+1}^{\min(\frac{k^{j+1}-1}{k-1}-1,n)} p_i\lfloor\log_k(i(k-1))\rfloor \leq \sum_{j\geq 1:\frac{k^j-1}{k-1}+1\leq n}\ \sum_{i=\frac{k^j-1}{k-1}+1}^{\min(\frac{k^{j+1}-1}{k-1}-1,n)} p_i\log_k(i(k-1)).$$

We note that the function $f(j)=\log_k(k^j-1)-\lfloor\log_k(k^j-1)\rfloor$ is increasing in j. Therefore, it reaches the minimum at $j=2$, where it takes the value

$$\log_k(k^2-1)-\lfloor\log_k(k^2-1)\rfloor = 1+\log_k\left(1-\frac{1}{k^2}\right) > 0.5,$$

for any $k\geq 2$. Thus, (34) holds as we claimed.

Let us now show that

$$L_\epsilon \leq H_k(\mathbf{p}) - p_1\log_k\frac{1}{p_1} + (1-p_1)\log_k(k-1) - 0.5\sum_{j\geq 2:\frac{k^j-1}{k-1}\leq n} p_{\frac{k^j-1}{k-1}}. \tag{35}$$

Since the distribution $\mathbf{p}$ is ordered in a non-increasing fashion, from (29) and (34), we have

$$\begin{aligned}
L_\epsilon &\leq \sum_{i=2}^{n} p_i\log_k(i(k-1)) - 0.5\sum_{j\geq 2:\frac{k^j-1}{k-1}\leq n} p_{\frac{k^j-1}{k-1}}\\
&\leq \sum_{i=2}^{n} p_i\log_k\frac{1}{p_i}(k-1) - 0.5\sum_{j\geq 2:\frac{k^j-1}{k-1}\leq n} p_{\frac{k^j-1}{k-1}} \qquad (\text{since } i\leq \tfrac{1}{p_i})\\
&= H_k(\mathbf{p}) - p_1\log_k\frac{1}{p_1} + (1-p_1)\log_k(k-1) - 0.5\sum_{j\geq 2:\frac{k^j-1}{k-1}\leq n} p_{\frac{k^j-1}{k-1}}.
\end{aligned}$$

Therefore, (35) holds.

To conclude the proof, it remains to prove that

$$L_\epsilon \leq H_k(\mathbf{p}) - \mathcal{H}_k(p_1) + (1-p_1)(\log_k 2(k-1)) - 0.5\sum_{j\geq 2:\frac{k^j-1}{k-1}\leq n} p_{\frac{k^j-1}{k-1}}. \tag{36}$$

By observing that for any $i\geq 2$, it holds that

$$p_i \leq \frac{2(1-p_1)}{i}, \tag{37}$$

we obtain:

$$\begin{aligned}
L_\epsilon &\leq \sum_{i=2}^{n} p_i\log_k(i(k-1)) - 0.5\sum_{j\geq 2:\frac{k^j-1}{k-1}\leq n} p_{\frac{k^j-1}{k-1}}\\
&\leq \sum_{i=2}^{n} p_i\log_k\left(\frac{2(1-p_1)}{p_i}(k-1)\right) - 0.5\sum_{j\geq 2:\frac{k^j-1}{k-1}\leq n} p_{\frac{k^j-1}{k-1}} \quad (\text{since from (37), we have } i\leq \tfrac{2(1-p_1)}{p_i})\\
&= H_k(\mathbf{p}) - p_1\log_k\frac{1}{p_1} + (\log_k 2 + \log_k(1-p_1) + \log_k(k-1))(1-p_1) - 0.5\sum_{j\geq 2:\frac{k^j-1}{k-1}\leq n} p_{\frac{k^j-1}{k-1}}\\
&= H_k(\mathbf{p}) - \mathcal{H}_k(p_1) + (1-p_1)(\log_k 2(k-1)) - 0.5\sum_{j\geq 2:\frac{k^j-1}{k-1}\leq n} p_{\frac{k^j-1}{k-1}}.
\end{aligned}$$

Therefore, (36) holds as well.

From (35) and (36), since

$$\sum_{j\geq 2:\frac{k^j-1}{k-1}\leq n} p_{\frac{k^j-1}{k-1}} \geq 0,$$

we obtain

$$L_\epsilon \leq H_k(\mathbf{p}) - p_1 \log_k \frac{1}{p_1} + (1-p_1)\log_k(k-1),$$

and

$$L_\epsilon \leq H_k(\mathbf{p}) - \mathcal{H}_k(p_1) + (1-p_1)(\log_k 2(k-1)).$$

Now, it is easy to verify that $p_1 \log_k \frac{1}{p_1} \geq \mathcal{H}_k(p_1) + (1-p_1)\log_k 2$ for $0 < p_1 \leq 0.5$, proving the Lemma. □

Thanks to the result of Lemma 1 and to the Formula (9), the upper bounds obtained above can be used to derive our upper bounds on the average length of optimal prefix-free codes with space, as shown in the following theorems.

Theorem 3. *The average length of an optimal prefix-free code with space* $C : \{s_1, \ldots, s_n\} \to \{0, \ldots, k-1\}^+ \cup \{0, \ldots, k-1\}^+ \sqcup$ *satisfies*

$$\begin{aligned} L(C) \leq & H_k(\mathbf{p}) + \log_k(k-1) + \sum_{i=1}^{\lfloor \log_k \lceil \frac{n-1}{k} \rceil \rfloor} p_{\frac{k^i-1}{k-1}} \\ & + \sum_{i=1}^{\frac{k^{h-1}-1}{k-1}-1} p_i + \sum_{i=\frac{k^h+k^{h-1}-2}{k-1}-\lceil n/k \rceil}^{\frac{k^h-1}{k-1}-1} p_i & (38) \\ \leq & H_k(\mathbf{p}) + \log_k(k-1) + \sum_{i=1}^{\lfloor \log_k \lceil \frac{n-1}{k} \rceil \rfloor} p_{\frac{k^i-1}{k-1}} + \sum_{i=1}^{\lceil \frac{n}{k} \rceil -1} p_i, & (39) \end{aligned}$$

where $h = \lceil \log_k(n - \lceil n/k \rceil) \rceil$.

Proof. From Lemma 1 and the Formula (9), we have

$$\begin{aligned} L(C) \leq & L_\epsilon + \sum_{i=1}^{\lfloor \log_k \lceil \frac{n-1}{k} \rceil \rfloor} p_{\frac{k^i-1}{k-1}} \\ & + \sum_{i=1}^{\frac{k^{h-1}-1}{k-1}-1} p_i + \sum_{i=\frac{k^h+k^{h-1}-2}{k-1}-\lceil n/k \rceil}^{\frac{k^h-1}{k-1}-1} p_i. & (40) \end{aligned}$$

By applying the upper bound (28) on L_ϵ of Lemma 7 to (40), we obtain (38). □

Theorem 4. *The average length of an optimal prefix-free code with space* $C : \{s_1, \ldots, s_n\} \to \{0, \ldots, k-1\}^+ \cup \{0, \ldots, k-1\}^+ \sqcup$ *satisfies*

$$L(C) \leq \begin{cases} H_k(\mathbf{p}) - p_1 \log_k \frac{1}{p_1} + (1-p_1)\log_k(k-1) \\ + \sum_{i=1}^{\lceil \frac{n}{k} \rceil -1} p_i + \sum_{i=1}^{\lfloor \log_k \lceil \frac{n-1}{k} \rceil \rfloor} p_{\frac{k^i-1}{k-1}} & \text{if } 0 < p_1 \leq 0.5, \\ \\ H_k(\mathbf{p}) - \mathcal{H}_k(p_1) + (1-p_1)\log_k 2(k-1) \\ + \sum_{i=1}^{\lceil \frac{n}{k} \rceil -1} p_i + \sum_{i=1}^{\lfloor \log_k \lceil \frac{n-1}{k} \rceil \rfloor} p_{\frac{k^i-1}{k-1}} & \text{if } 0.5 < p_1 \leq 1. \end{cases} \quad (41)$$

Proof. From Lemma 1 and the Formula (9) and by recalling that $h = \lceil \log_k(n - \lceil n/k \rceil) \rceil$, we have

$$
\begin{aligned}
L(C) \leq & L_\epsilon + \sum_{i=1}^{\lfloor \log_k \lceil \frac{n-1}{k} \rceil \rfloor} p_{\frac{k^i-1}{k-1}} \\
& + \sum_{i=1}^{\frac{k^{h-1}-1}{k-1}-1} p_i + \sum_{i=\frac{k^h+k^{h-1}-2}{k-1}-\lceil n/k \rceil}^{\frac{k^h-1}{k-1}-1} p_i \\
\leq & L_\epsilon + \sum_{i=1}^{\lceil \frac{n}{k} \rceil -1} p_i + \sum_{i=1}^{\lfloor \log_k \lceil \frac{n-1}{k} \rceil \rfloor} p_{\frac{k^i-1}{k-1}}.
\end{aligned} \tag{42}
$$

We apply the upper bound (33) on L_ϵ of Lemma 8 to (42). That gives us (41). □

Remark 1. *One can estimate how much the average length of optimal prefix-free codes ending with a space differs from the minimum average length of* unrestricted *optimal prefix-free codes on the alphabet* $\{0, 1, \ldots, k-1, \sqcup\}$, *that is optimal prefix-free codes in which the special symbol* $\sqcup$ *is not constrained to appear at the end of the codewords, only.*

Let $S = \{s_1, \ldots, s_n\}$ *be the set of source symbols and* $\mathbf{p} = (p_1, \ldots, p_n)$ *be a probability distribution on S. Let us denote by* $C_\sqcup : \{s_1, \ldots, s_n\} \to \{0, \ldots, k-1\}^+ \cup \{0, \ldots, k-1\}^+ \sqcup$ *an optimal prefix-free code ending with a space for S and by* $C : \{s_1, \ldots, s_n\} \to \{0, \ldots, k-1, \sqcup\}^+$ *an optimal prefix-free code without the restriction of where the space can occur. Clearly,* $L(C_\sqcup) < H_k(\mathbf{p}) + 1$, *since the more constrained optimal code* $C' : \{s_1, \ldots, s_n\} \to \{0, \ldots, k-1\}^+$ *has an average length less than* $H_k(\mathbf{p}) + 1$. *Therefore,*

$$
\begin{aligned}
L(C_\sqcup) - L(C) < & H_k(\mathbf{p}) + 1 - L(C) \\
\leq & H_k(\mathbf{p}) + 1 - H_{k+1}(\mathbf{p}) \qquad \textit{(since } L(C) \geq H_{k+1}(\mathbf{p})\textit{)} \\
= & H_k(\mathbf{p}) \left(1 - \frac{1}{\log_k(k+1)}\right) + 1.
\end{aligned}
$$

Since $\lim_{k\to\infty} \log_k(k+1) = 1$, *we have that, as the cardinality of the code alphabet increases, the constraint that the space can appear only at the end of codewords becomes less and less influential.*

5. Concluding Remarks

In this paper, we have introduced the class of prefix-free codes where a specific symbol (referred to as a “space") can only appear at the end of codewords. We have proposed a linear-time algorithm to construct “almost"-optimal codes with this characteristic, and we have shown that their average length is at most one unit longer than the *minimum* average length of any prefix-free code in which the space can appear only at the end of codewords. We have proven this result by highlighting a connection between our type of codes and the well-known class of one-to-one codes. We have also provided upper and lower limits of the average length of optimal prefix-free codes ending with a space, expressed in terms of the source entropy and the cardinality of the code alphabet.

We leave open the problem of providing an efficient algorithm to construct *optimal* prefix-free codes ending with a space. It seems that there is no easy way to modify the classical Huffman greedy algorithm to solve our problem. It is possible that the more powerful dynamic programming approach could be useful to provide an optimal solution to the problem, as done in [14] for optimal binary codes ending with ones. This will be the subject of future investigations.

Author Contributions: All authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The authors would like to thank the referees and the Guest Editor I. Sason for their corrections and useful suggestions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Daniels, P.T.; Bright, W. (Eds.) *The World's Writing Systems*; Oxford University Press: Oxford, UK, 1996.
2. Wikipedia: Writing Systems without Word Boundaries. Available online: https://en.wikipedia.org/wiki/Category:Writing_systems_without_word_boundaries (accessed on 25 March 2024).
3. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, USA, 2006.
4. Alon, N.; Orlitsky, A. A lower bound on the expected length of one-to-one codes. *IEEE Trans. Inf. Theory* **1994**, *40*, 1670–1672. [CrossRef]
5. Blundo, C.; De Prisco, R. New bounds on the expected length of one-to-one codes. *IEEE Trans. Inf. Theory* **1996**, *42*, 246–250. [CrossRef]
6. Jaynes, E. Note on unique decipherability. *Ire Trans. Inf. Theory* **1959**, *5*, 98–102. [CrossRef]
7. Shannon, C.E. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana, IL, USA, 1949.
8. Courtade, T.; Verdú, S. Cumulant generating function of codeword lengths in optimal lossless compression. *IEEE Int. Symp. Inf. Theory* **2014**, 2494–2498.
9. Kontoyiannis, I.; Verdú, S. Optimal lossless data compression: Non-asymptotics and asymptotics. *IEEE Trans. Inf. Theory* **2014**, *60*, 777–795. [CrossRef]
10. Kosut, O.; Sankar, L. Asymptotics and non-asymptotics for universal Fixed-to-Variable source coding. *IEEE Trans. Inf. Theory* **2017**, *63*, 3757–3772. [CrossRef]
11. Szpankowski, W.; Verdu, S. Minimum expected length of Fixed-to-Variable lossless compression without prefix constraints. *IEEE Trans. Inf. Theory* **2011**, *57*, 4017–4025. [CrossRef]
12. Knuth, D. *The Art of Computer Programming, Volume 3*, 2nd ed.; Sorting and Searching; Addison Wesley Longman Publishing Co., Inc.: Boston, MA, USA, 1998.
13. Wyner, A. An upper bound on the entropy series. *Inf. Control* **1972**, *20*, 176–181. [CrossRef]
14. Chan, S.-L.; Golin, M.J. A dynamic programming algorithm for constructing optimal 1-ended binary prefix-free codes. *IEEE Trans. Inf. Theory* **2000**, *46*, 1637–1644. [CrossRef]