On Unique Decodability, McMillan's Theorem and the Expected Length of Codes

Technical Report R.T. 200801-58,
Department of Electronics for Automation, University of Brescia,
Via Branze 38 - 25123, Brescia, Italy.

Marco Dalai & Riccardo Leonardi,
Email: {marco.dalai, riccardo.leonardi}@ing.unibs.it

***Abstract*—In this paper we propose a revisitation of the topic of unique decodability and of some of the related fundamental theorems. It is widely believed that, for any discrete source X, every "uniquely decodable" block code satisfies**

$$E[l(X_1, X_2, \cdots, X_n)] \geq H(X_1, X_2, \ldots, X_n),$$

where $X_1, X_2, \ldots, X_n$ are the first n symbols of the source, $E[l(X_1, X_2, \cdots, X_n)]$ is the expected length of the code for those symbols and $H(X_1, X_2, \ldots, X_n)$ is their joint entropy. We show that, for certain sources with memory, the above inequality only holds if a limiting definition of *"uniquely decodable code"* is considered. In particular, the above inequality is usually assumed to hold for any "practical code" due to a debatable application of McMillan's theorem to sources with memory. We thus propose a clarification of the topic, also providing extended versions of McMillan's theorem and of the Sardinas Patterson test to be used for Markovian sources. This work terminates also with the following interesting remark: both McMillan's original theorem and ours are equivalent to Shannon's theorem on the capacity of noiseless channels.

I. Introduction

The problem of lossless encoding of information sources has been intensively studied over the years (see [1, Sec. II] for a detailed historical overview of the key developments in this field). Shannon initiated the mathematical formulation of the problem in his major work [2] and provided the first results on the average number of bits per source symbol that must be used *asymptotically* in order to represent an information source.

For a random variable X with alphabet $\mathcal{X}$ and probability mass function $p_X(\cdot)$, he defined the *entropy* of X as the quantity

$$H(X) = \sum_{x \in \mathcal{X}} p_X(x) \log \frac{1}{p_X(x)}$$

On another hand, Shannon focused his attention on finite state Markov sources $X = \{X_1, X_2, \ldots\}$, for which he defined the *entropy* as

$$H(X) = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n),$$

a quantity that is now usually called *entropy rate* of the source. Based on these definitions, he derived the fundamental results for fixed length and variable length codes. In particular, he showed that, by encoding sufficiently large blocks of symbols, the average number of bits per symbol used by fixed length codes can be made as close as desired to the entropy rate of the source while maintaining the probability of error as small as desirable. If variable length codes are allowed, furthermore, he showed that the probability of error can be reduced to zero without increasing the asymptotically achievable average rate. Shannon also proved the converse theorem for the case of fixed length codes, but he did not explicitly consider the converse theorem for variable length codes (see [1, Sec. II.C]).

An important contribution in this direction came from McMillan [3], who showed that every *"uniquely decodable"* code using a D-ary alphabet must satisfy Kraft's inequality, $\sum_i D^{-l_i} \leq 1$, l_i being the codeword lengths [4]. Based on this result, he was able to prove that the expected length of a uniquely decodable code for a random variable X is not smaller than its entropy, $E[l(X)] \leq H(X)$. This represents a strong converse result in coding theory. However, while the initial work by Shannon was explicitly referring to finite state Markov sources, McMillan's results basically considered only the encoding of a random variable. This leads to immediate conclusions on the problem of encoding memoryless sources, but an ad hoc study is necessary for the case of sources with memory. The application of McMillan's theorem to these type of sources can be found in [5, Sec. 5.4] and [6, Sec. 3.5]. In these two well-known references, McMillan's result is used not only to derive a converse theorem on the asymptotic average number of bits per symbol needed to represent an information source, but also to deduce a non-asymptotic strong converse to the coding theorem. In particular, the famous result obtained (see [6, Th. 3.5.2], [5, Th. 5.4.2], [7, Sec. II, p. 2047]) is that, for every source with memory, any uniquely decodable code satisfies

$$E[l(X_1, X_2, \cdots, X_n)] \geq H(X_1, X_2, \ldots, X_n), \tag{1}$$

where $X_1, X_2, \ldots, X_n$ are the first n symbols of the source, $E[l(X_1, X_2, \cdots, X_n)]$ is the expected length of the code for those symbols and $H(X_1, X_2, \ldots, X_n)$ represents their joint entropy.

In this paper we want to clarify that the above equation is only valid if a limiting definition of "uniquely decodable code" is assumed. In particular, we show that there are information sources for which a reversible encoding operation exists that produces a code for which equation (1) does not hold any longer for every n. This is demonstrated through a simple example in Section II. In Section III we revisit the topic of unique decodability, consequently providing an extension of McMillan's theorem and of the Sardinas-Patterson test [8] for the case of first order Markov sources. Finally, in Section IV, some interesting findings are reported regarding McMillan's original theorem and on the proposed one, demonstrating their mathematical equivalence to Shannon's theorem on the capacity of constrained noiseless channels [2, Th. 1].

II. A Meaningful Example

Let $X = \{X_1, X_2, \ldots\}$ be a first order Markov source with alphabet $\mathcal{X} = \{A, B, C, D\}$ and with transition probabilities shown by the graph of Fig. 1. Its transition probability matrix is thus

$$\mathbf{P} = \begin{bmatrix} 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix},$$

where rows and columns are associated to the natural alphabetical order of the symbol values A, B, C and D.

It is not difficult to verify that the stationary distribution associated with this transition probability matrix is the uniform distribution. Let X_1 be uniformly distributed, so that the source X is stationary and, in addition, ergodic.

Let us now examine possible binary encoding techniques for this source and possibly find an optimal one. In order to evaluate the performance of different codes we determine the entropy of the sequences of symbols that can be produced by this source. By stationarity of the source, one easily proves that

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= H(X_1) + \sum_{i=2}^{n} H(X_i|X_{i-1}) \\ &= 2 + \frac{3}{2}(n-1), \end{aligned}$$

where $H(X_i|X_{i-1})$ is the conditional entropy of X_i given X_{i-1}, that is

$$H(X_i|X_{i-1}) = \sum_{x,y \in \mathcal{X}} p_{X_i X_{i-1}}(x,y) \log \frac{1}{p_{X_i|X_{i-1}}(x|y)}.$$

Let us now consider the following binary codes to represent sequences produced by this source.

Classic code

We call this first code "classic" as it is the most natural way to encode the source given its particular structure. Since the first symbol is uniformly distributed between four choices, 2 bits are used to uniquely identify it, in an obvious way. For the next symbols we note that we always have dyadic conditional probabilities. So, we apply a state-dependent code. For encoding the k-th symbol we use, again in an obvious way, 1 bit if symbol $k-1$ was an A or a B, and we use 2 bits if symbol $k-1$ was a C or a D. This code seems to perfectly fulfill the source as the number of used bits always corresponds to the uncertainty. Indeed, the average length of the code for the first n symbols is given by

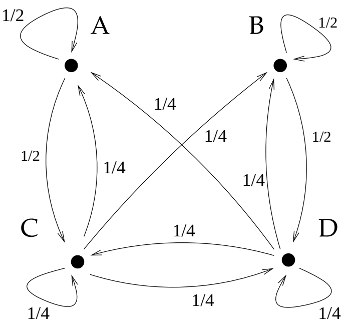

Fig. 1. Graph, with transition probabilities, for the Markov source use in the example.

$$\begin{aligned} E[l(X_1, X_2, \ldots, X_n)] &= E[l(X_1)] + \sum_{i=2}^{n} E[l(X_i)] \\ &= 2 + \frac{3}{2}(n-1). \end{aligned}$$

So, the expected number of bits used for the first n symbols is exactly the same as their entropy, which would let us declare that this encoding technique is optimal.

Alternative code

Let us consider a different code, obtained by applying the following fixed mapping from symbols to bits: $A \to 0$, $B \to 1$, $C \to 01$, $D \to 10$. It will be easy to see that this code maps different sequences of symbols into the same codeword. For example, the sequences AB and C are both coded to 01. This is usually expressed, see for example [5], by saying that the code is not *uniquely decodable*, an expression which suggests the idea that the code cannot be inverted, different sequences being associated to the same code. It is however easy to notice that, for the source considered in this example, the code does not introduce any ambiguity. Different sequences that are producible by the source are in fact mapped into different codes. Thus it is possible to "decode" any sequence of bits without ambiguity. For example the code 01 can only be produced by the single symbol C and not by the sequence AB, since our source cannot produce such sequence (the transition from A to B being impossible). It is not difficult to verify that it is indeed possible to decode any sequence of bits by

Encoding	$A \to 0$ $B \to 1$ $C \to 01$ $D \to 10$	
Decoding	more bits left	one bit left
	$00\ldots \to A+0\ldots$	$0 \to A$
	$01\ldots \to C\ldots$	$1 \to B$
	$10\ldots \to D\ldots$	
	$11\ldots \to B+1\ldots$	

TABLE I
Table of encoding and decoding operations of the proposed alternative code for the Markov source of Figure 1.

operating in the following way. Consider first the case when there are still two or more bits to decode. In such a case, for the first pair of encountered bits, if a 00 (respectively a 11) is observed then clearly this corresponds to an A symbol followed by a code starting with a 0 (respectively a B symbol followed by a code starting with a 1). If, instead, a 01 pair is observed (respectively a 10) then a C must be decoded (respectively a D). Finally, if there is only one bit left to decode, say a 0 or a 1, the decoded symbol is respectively an A or a B. Such coding and decoding operations are summarized in Table I.

Now, what is the performance of this code? The expected number of bits in coding the first n symbols is given by:

$$\begin{aligned} E[l(X_1, X_2, X_3, \cdots, X_n)] &= \sum_{i=1}^{n} E[l(X_i)] \\ &= \frac{3}{2}n \end{aligned}$$

Unexpectedly, the average number of bits used by the code is strictly smaller than the entropy of the symbols. So, the performance of this code is better than what would have been traditionally considered the "optimal" code, that is the classical code. Let us mention that this code is not only more efficient on average, but it is at least as efficient as the classic code for every possible sequence which remains compliant with the source characteristics. For each source sequence, indeed, the number of decoded symbols after reading the first m bits of the alternative code is always larger than or equal to the number of symbols decoded with the first m bits of the classic code. Hence, the proposed alternative code is more efficient than the classic code in all respects. The obtained gain *per symbol* obviously goes to zero asymptotically, as imposed by the Asymptotic Equipartition Property. However, in practical cases we are usually interested in coding a finite number of symbols. Thus, this simple example reveals that the problem of finding an optimal code is not yet well understood for the case of sources with memory. The obtained results may thus have interesting consequences not only from a theoretical point of view, but even for practical purposes in the case of sources exibiting constraints imposing high order dependencies.

Commenting on the "alternative code", one may object that it is not fair to use the knowledge on impossible transitions in order to design the code. But probably nobody would object to the design of what we called the "classic code". Even in that case, however, the knowledge that some transitions are impossible was used, in order to construct a state-dependent "optimal" code.

It is important to point out that we have just shown a fixed to variable length code for a stationary ergodic source that maps sequences of n symbols into strings of bits that can be decoded and such that **the average code length is smaller than the entropy of those n symbols**. Furthermore, this holds for every n, and not for an *a priori* fixed n. In a sense we could say that the given code has a negative *redundancy*. Note that there is a huge difference between the considered setting and that of the so called *one-to-one codes* (see for example [9] for details). In the case of one-to-one codes, it is assumed that only one symbol, or a given known amount of symbols, must be coded, and codes are studied as maps from symbols to binary strings without considering the decodability of concatenation of codewords. Under those hypotheses, Wyner [10] first pointed out that the average codeword length can always be made lower than the entropy, and different authors have studied bounds on the expected code length over the years [11], [12]. Here, instead, we have considered a fixed-to-variable length code used to compress sequences of symbols of whatever length, concatenating the code for the symbols one by one, as in the most classic scenario.

III. Unique decodability for constrained sources

In this section we briefly survey the literature on unique decodability and we then propose an adequate treatment of the particular case of *constrained sources* defined as follows.

Definition 1: A source $X = \{X_1, X_2, \ldots\}$ with symbols in a discrete alphabet $\mathcal{X}$ is a *constrained source* if there exists a finite sequence of symbols from $\mathcal{X}$ that cannot be obtained as output of the source X.

A. Classic definitions and revisitation

It is interesting to consider how the topic of unique decodability has been historically dealt with in the literature and how the results on unique decodability are used to deduce results on the expected length of codes. Taking [6] and [5] as representative references for what can be viewed as the classic approach to lossless source coding, we note some common structures between them in the development of the theory, but also some interesting differences. The most important fact to be noticed is the use, in both references with only marginal differences, of the following chain of deductions:

(a) McMillan's theorem asserts that all uniquely decodable codes satisfy Kraft's inequality;

(b) If a code for a random variable X satisfies Kraft's inequality, then $E[l(X)] \geq H(X)$;

(c) Thus any uniquely decodable code for a random variable X satisfies $E[l(X)] \geq H(X)$;

(d) For sources with memory, by considering sequences of n symbols as *super-symbols*, we deduce that any uniquely decodable code satisfies $E[l(X_1, X_2, \ldots, X_n)] \geq H(X_1, X_2, \ldots, X_n)$.

In the above flow of deductions there is an implicit assumption which is not obvious and, in a certain way, not clearly supported. It is implicitly assumed that the definition of *uniquely decodable code* used in McMillan's theorem is also appropriate for sources with memory. Of course, by definition of "definition", one can freely choose to define "uniquely decodable code" in any preferred way. However, as shown by the code of Table I in the previous section, the definition of *uniquely decodable code* used in McMillan's theorem does not coincide with the intuitive idea of "decodable" for certain sources with memory. To our knowledge, this ambiguity has never been reported previously in the literature, and for this reason it has been erroneously believed that the result $E[l(X_1, X_2, \ldots, X_n)] \geq H(X_1, X_2, \ldots, X_n)$ holds for every "practically usable" code. As shown by the Markov source example presented, this interpretation is incorrect.

In order to better understand the confusion associated to the meaning of "uniquely decodable code", it is interesting to focus on a small difference between the formal definitions given by the authors in [5] and in [6]. We start by rephrasing for notational convenience the definition given by Cover and Thomas in [5].

> *Definition 2:* [5, Sec. 5.1, pp. 79-80] A code is said to be uniquely decodable if no finite sequence of code symbols can be obtained in two or more different ways as a concatenation of codewords.

Note that this definition is the same used in McMillan's paper [3], and it considers a property of the codebook without any reference to sources. It is however difficult to find a clear motivation for such a source independent definition. After all, a code is always designed for a given source, not for a given alphabet. Indeed, right after giving the formal definitions, the authors comment

> *"In other words, any encoded string in a uniquely decodable code has only one possible* source string *producing it."*

So, a reference to sources is introduced. What is not noticed is that the condition given in the formal definition coincides with the phrased one only if the source at hand can produce any possible combination of symbols as output. Conversely, the two definitions are not equivalent, the first one being stronger, the second one being instead "more intuitive".

With respect to formal definitions, Gallager proceeds in a different way with the following:

> *Definition 3:* [6, Sec. 3.2, pg. 45] *"A code is uniquely decodable if for each source sequence of finite length, the sequence of code letters corresponding to that source sequence is different from the sequence of code letters corresponding to any other source sequence."*

Note that this is a formal definition of unique decodability of a code with respect to a given source. Gallager states this definition while discussing memoryless sources[1]. In that case, the definition is clearly equivalent to Definition 2 but, unfortunately, Gallager implicitly uses Definition 2 instead of Definition 3 when dealing with sources with memory.[2]

In order to avoid the above discussed ambiguity, we propose to adopt the following explicit definition.

Definition 4: A code C is said to be *uniquely decodable for the source X* if no two different finite sequences of source symbols producible by X have the same code.

With this definition, not all *uniquely decodable codes for a given source* satisfy Kraft's inequality. So, the chain of deductions (a)-(d) listed at the beginning of this section cannot be used for constrained sources, as McMillan's theorem uses Definition 2 of unique decodability.

The alternative code of Table I thus immediately gives:

Lemma 1: There exists at least one source X and a uniquely decodable code for X such that, for every $n \geq 1$,

$$E[l(X_1, X_2, \dots, X_n)] < H(X_1, X_2, \dots, X_n).$$

B. Extension of McMillan's theorem to Markov sources

In Section II, the proposed alternative code demonstrates that McMillan's theorem does not apply in general to uniquely decodable codes for a constrained source X as defined in Definition 4. In this section a modified version of Kraft's inequality is proposed which represents a necessary condition for the unique decodability of a code for a first order Markov source.

Let X be a Markov source with alphabet $\mathcal{X} = \{1, 2, \dots, m\}$ and transition probability matrix $\mathbf{P}$. Let $W = \{w_1, w_2, \dots, w_m\}$ be a set of D-ary codewords for the alphabet $\mathcal{X}$ and let, $l_i = l(w_i)$ be the length of codeword w_i. McMillan's original theorem can be stated in the following way:

Theorem 1 (McMillan, [3]): If the set of codewords W is uniquely decodable (in the sense of Definition 2) then

$$\sum_{i=1}^{m} D^{-l_i} \leq 1.$$

Comment: It is interesting to consider the proof given by Karush [13] of this theorem. Karush notices that for every $k > 0$, in order for the code to be uniquely decodable, the following inequality must be satisfied

$$\left(\sum_{i=1}^{m} D^{-l_i} \right)^k \leq k\, l_{\max} \qquad (2)$$

where $l_{\max}$ is the largest of the l_i, $i = 1, \dots, m$. Indeed, the term on the left hand side of (2) can be expanded as the sum of m^k terms each of them being a product of factors D^{-l_i} in a different combination. The way the possible combinations of products are constructed is exactly the same as the way the symbols of the source are concatenated in all possible combinations to obtain sequences of k symbols. For example, a sequence starting with '$1, 3, 2 \dots$' translates into $D^{-l_1} D^{-l_3} D^{-l_2} \cdots$ in the expansion of the left hand side of (2). In order to have only one sequence assigned to every code the above inequality must be satisfied for every k. But the right hand side of (2) grows linearly with k, while the left hand side grows exponentially with it if the Kraft inequality is not satisfied. Thus, when (1) does not hold, (2) cannot be satisfied for every k, and the code is not uniquely decodable.

As we said, the expansion of the left hand side of (2) contains terms associated with every possible combinations of symbols of the source alphabet, and is thus appropriate for the case of unconstrained sources. If the source is constrained, however, only some combinations of symbols should be considered. For example, consider again the Markov chain used in the Section II, with l_1, l_2, l_3 and l_4 the lengths of codewords assigned respectively to symbols A, B, C and D. In this case,

[1] See [6, pg. 45] *"We also assume, initially, [...] that successive letters are independent"*

[2] In fact, in [6], the proof of Theorem 3.5.2, on page 58, is based on Theorem 3.3.1, on page 50, the proof of which states: *"...follows from Kraft's inequality, [...] which is valid for any uniquely decodable code"*. But Kraft's inequality is valid for uniquely decodable codes defined as in Definition 2 and not Definition 3.

the terms in the expansion on the left hand side of (2) that contain $\cdots D^{-l_1} D^{-l_2} \cdots$ should be discarded, since B cannot follow A for any source compliant sequence. Consider thus the vector $\mathbf{L} = [D^{-l_1}, D^{-l_2}, D^{-l_3}, D^{-l_4}]'$ and the matrix

$$\mathbf{Q}(D) = \begin{bmatrix} D^{-l_1} & 0 & D^{-l_3} & 0 \\ 0 & D^{-l_2} & 0 & D^{-l_4} \\ D^{-l_1} & D^{-l_2} & D^{-l_3} & D^{-l_4} \\ D^{-l_1} & D^{-l_2} & D^{-l_3} & D^{-l_4} \end{bmatrix}. \qquad (3)$$

It is not difficult to verify that a correct reformulation of eq. (2) for our source should be written, for $k > 0$, as

$$\mathbf{L}'\mathbf{Q}(D)^{k-1}\mathbf{1}_4 \leq k\, l_{\text{max}}, \qquad (4)$$

where $\mathbf{1}_4 = [1, 1, 1, 1]$. It is possible to show that a necessary condition for this inequality to be satisfied for every k is that the matrix $\mathbf{Q}(D)$ has spectral radius[3] at most equal to 1. We will state and prove this fact in the general case, hereafter.

Let X, $\mathbf{P}$ and W be as specified before.

Theorem 2: If the set of codewords W is uniquely decodable for the Markov source X, then the matrix $\mathbf{Q}(D)$ defined by

$$\mathbf{Q}_{ij}(D) = \begin{cases} 0 & \text{if} \quad P_{ij} = 0 \\ D^{-l_j} & \text{if} \quad P_{ij} > 0 \end{cases}$$

has spectral radius at most 1.

Proof: We follow Karush's proof of McMillan theorem. Set $\mathbf{Q} = \mathbf{Q}(D)$ for simplicity. Let $\mathcal{X}^{(k)}$ be the set of all sequences of k symbols that can be produced by the source and let $\mathbf{L} = [D^{-l_1}, D^{-l_2}, \ldots, D^{-l_m}]'$.

We now define, for $k > 0$, the row vector

$$\mathbf{V}^{(k)} = \mathbf{L}'\mathbf{Q}^{k-1}. \qquad (5)$$

Then it is easy to see by induction that the i-th component of $\mathbf{V}^{(k)}$ is written as

$$\mathbf{V}_i^{(k)} = \sum_{h_1, h_2, \ldots, h_k} D^{-l_{h_1} - l_{h_2} \cdots - l_{h_k}} \qquad (6)$$

where the sum runs over all sequences of indices $(h_1, h_2, \ldots, h_k)$ in $\mathcal{X}^{(k)}$ with varying $h_1, h_2, \ldots, h_{k-1}$ and $h_k = i$. So, if we call $\mathbf{1_m}$ the length m vector composed of m 1's, we have

$$\mathbf{L}'\mathbf{Q}^{k-1}\mathbf{1}_m = \sum_{(h_1, h_2, \ldots, h_k) \in \mathcal{X}^{(k)}} D^{-l_{h_1} - l_{h_2} \cdots - l_{h_k}}. \qquad (7)$$

Reindexing the sum with respect to the total length $r = l_{h_1} + l_{h_2} + \cdots + l_{h_k}$ and calling $N(r)$ the number of sequences of $\mathcal{X}^{(k)}$ to which a code of length r is assigned, we have

$$\mathbf{L}'\mathbf{Q}^{k-1}\mathbf{1}_m = \sum_{r = k l_{\text{min}}}^{k l_{\text{max}}} N(r) D^{-r} \qquad (8)$$

where l_{min} and l_{max} are respectively the minimum and the maximum of the values $l_i, i = 1, 2, \ldots, m$. Since the code is uniquely decodable, there are at most D^r sequences with a code of length r. This implies that, for every $k > 0$, we must have

$$\mathbf{L}'\mathbf{Q}^{k-1}\mathbf{1}m \leq \sum_{r = k l_{\text{min}}}^{k l_{\text{max}}} D^r D^{-r} = k(l_{\text{max}} - l_{\text{min}} + 1) \qquad (9)$$

Now, note that the irreducible matrix $\mathbf{Q}$ is also nonnegative. Thus, by the Perron-Frobenius theorem (see [14] for details), its spectral radius $\rho(\mathbf{Q})$ is also an eigenvalue[4], with algebraic multiplicity 1 and with positive associated left eigenvector. Let $\mathbf{w}$ be such eigenvector; then, as $\mathbf{L}$ is positive, there exists a positive constant α such that $\mathbf{z} = \mathbf{L} - \alpha\mathbf{w}$ is a nonnegative vector. Thus, setting $\mathbf{L} = \alpha\mathbf{w} + \mathbf{z}$, we can write the left hand side of (9) as

$$\begin{aligned} \mathbf{L}'\mathbf{Q}^{k-1}\mathbf{1}_m &= \alpha\mathbf{w}'\mathbf{Q}^{k-1}\mathbf{1}_m + \mathbf{z}'\mathbf{Q}^{k-1}\mathbf{1}_m \\ &= \alpha\rho(\mathbf{Q})^{k-1}\mathbf{w}'\mathbf{1}_m + \mathbf{z}'\mathbf{Q}^{k-1}\mathbf{1}_m \\ &= \beta\rho(\mathbf{Q})^{k-1} + \gamma \end{aligned}$$

where $\beta = \alpha\mathbf{w}'\mathbf{1}_m$ is positive and γ is nonnegative. So, if $\rho(\mathbf{Q}) > 1$, the term on the left hand side of eq. (9) asymptotically grows at least as $\rho(\mathbf{Q})^{k-1}$. On the contrary, the right hand side term only grows linearly with k and for large enough k equation (9) could not be verified. We conclude that $\rho(\mathbf{Q}) \leq 1$. ■

Note that if the $\mathbf{P}$ matrix has all strictly positive entries, the matrix $\mathbf{Q}(D)$ is positive with all equal rows. It is known (see again [14]) that the spectral radius of such a matrix is given by the sum of the elements in a row, which in this case is $\sum_i D^{-l_i}$. Thus, for non-constrained sequences, we obtain classic Kraft's inequality.

Furthermore, the case when $\rho(\mathbf{Q}(D)) = 1$ corresponds to a limit situation in terms of $\mathbf{P}$ and $l_1, \ldots, l_m$. This is due to the fact that the spectral radius of a nonnegative positive matrix increases if any of the elements increases. So, if for a given matrix $\mathbf{P}$ there is a decodable code with codeword lengths $l_i, i = 1, \ldots, m$ such that $\rho(\mathbf{Q}(D)) = 1$, then there is no decodable code with lengths l_i' if $l_i' \leq l_i$ for all i with strict inequality for some i. Also, for the same codeword lengths, it is not possible to remove constraints from the Markov chain while keeping unique decodability property, since one of the elements of the matrix $\mathbf{Q}(D)$ would increase from zero to a positive value.

The above presented discussion is focusing on the case of constrained sources that are modeled with Markov chains "in the Moore form", as considered for example in [5]. In other words, we have modeled information sources as Markov chains by assigning an output source symbol to every state. This way we have considered only sources that have a memory of one symbol, because transitions in the Markov chains are always considered to be independent. In order to deal with more general sources we can consider the Markov source model with output symbols associated to transitions between states rather than to states (which corresponds to the Markov source model used by Shannon in [2] or, for example, by Gallager in [6]). We may say that this Markov

[3] Recall that the spectral radius of a matrix is defined as the greatest modulus of its eigenvalues.

[4] Note that in general the spectral radius is not an eigenvalue as it is defined as the maximum of $|\lambda|$ over all eigenvalues λ.

chain representation is in the “Mealy form”. Theorem 2 can be easily extended to Theorem 3 below to deal with this more general type of sources, as it will now be shown. It may be of interest to consider that, for this type of sources, the initial state has to be known or encoded. However, by considering the asymptotic reasoning used in the proof of Theorem 3, it is easy to realize that it does not make any difference to consider whether the initial state is known or not, since it is possible to embed the encoding of the initial state with a prefix free code, without substantially changing the proof of the theorem.

Theorem 3: Let X be a finite state source, with possible states $S_1, S_2, \ldots, S_q$ and with output symbols in the alphabet $\mathcal{X} = \{1, 2, \ldots, m\}$. Let $W = \{w_1, \ldots, w_m\}$ be a set of codewords for the symbols in $\mathcal{X}$ with lengths $l_1, l_2, \ldots, l_m$. Let $O_{i,j}$ be the subsets of $\mathcal{X}$ of possible symbols output by the source when transiting from state S_i to state S_j, O_{ij} being the empty set if transition from S_i to S_j is impossible. If the code is uniquely decodable for the source X, then the matrix $\mathbf{Q}(D)$ defined by

$$\mathbf{Q}_{ij}(D) = \sum_{h \in O_{i,j}} D^{-l_h}$$

has spectral radius at most 1.

Proof: The proof is not substantially different from the proof of Theorem 2. In this case, set again $\mathbf{Q} = \mathbf{Q}(D)$, we need to define $\mathbf{L}$ so that $\mathbf{L}_i = \sum_h D^{-l_h}$ where h runs over all the elements of the set $\cup_r O_{ri}$. Defining again, for $k > 0$, $\mathbf{V}^{(k)} = \mathbf{L}'\mathbf{Q}^{k-1}$, one can verify that

$$\mathbf{V}_i^{(k)} = \sum_{h_1, h_2, \ldots, h_k} D^{-l_{h_1} - l_{h_2} \cdots - l_{h_k}} \tag{10}$$

where the sum now runs over all sequences of indices $(h_1, h_2, \ldots, h_k)$ such that there exists a path in the graph of the Markov chain ending in state S_i which produces the sequence of symbols $(h_1, h_2, \ldots, h_k) \in \mathcal{X}^{(k)}$. The proof then follows as in Theorem 2. ■

As an example, consider again the source used in the preview example. We can represent the same source using only three states with a Mealy representation as indicated in Figure 2. The source is in state α if the last output symbol is an A, it is in state β if the last output symbol is a B, and it is in state γ if the last output symbol is a C or a D. Then, symbols are output at the transition from one state to the other as indicated on the edges in Figure 2. Using this representation, the matrix $\mathbf{Q}(D)$ defined in Theorem 3 is the 3×3 matrix given by

$$\mathbf{Q}(D) = \begin{bmatrix} D^{-l_1} & 0 & D^{-l_3} \\ 0 & D^{-l_2} & D^{-l_4} \\ D^{-l_1} & D^{-l_2} & D^{-l_3} + D^{-l_4} \end{bmatrix}. \tag{11}$$

Coherently, this matrix has the same spectral radius as the matrix defined in equation (3), which for this example is exactly 1, when $D = 2$, if $(l_1, l_2, l_3, l_4) = (1, 1, 2, 2)$ as in the “alternative code”.

As a further remark, we note that from a combinatorial point of view, i.e. distinguishing only between possible and impossible source sequences, unconstrained sources can be modeled with only one state S, every symbol being a possible output when moving form state S to itself. The matrix $\mathbf{Q}(D)$ defined in Theorem 3 is in this case a 1×1 matrix, i.e. a scalar value, which equals $\sum_i D^{-l_i}$. So again one has the classic Kraft inequality.

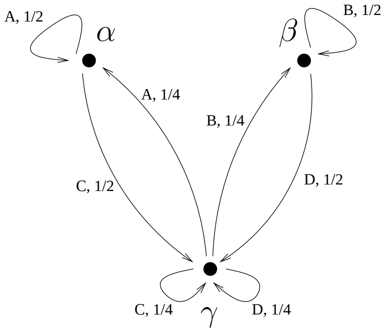

Fig. 2. Markov chain, in the Mealy form, associated to the source of figure 1. Here every arc is labeled with the associated output symbol and the probability of the transition.

It is worth noticing that, with the considered Mealy form representation, one can consider coding techniques that associate different codewords to the same symbol depending on the state of the source. This is precisely the way symbols $X_2, X_3, \ldots$ have been encoded in the “classic code” used in Section II. It is possible to adapt Theorem 3 to this type of encoding techniques by constructing an adequate matrix $\mathbf{Q}(D)$ in an obvious way, by considering in the generic element $\mathbf{Q}_{ij}(D)$, for the different output symbols, the lengths of the codewords used when transiting from state S_i to state S_j. For example, the matrix associated with the “classic code” used in Section II is easily seen to be

$$\mathbf{Q}(2) = \begin{bmatrix} 2^{-1} & 0 & 2^{-1} \\ 0 & 2^{-1} & 2^{-1} \\ 2^{-2} & 2^{-2} & 2^{-2} + 2^{-2} \end{bmatrix}. \tag{12}$$

which has spectral radius equal to 1.

It is important at this point to note that Theorems 2 and 3 only provide a necessary condition for the unique decodability of a given code, while the classic Kraft inequality is a necessary and sufficient condition for a set of integers to be codeword lengths of some uniquely decodable code in the classic sense. It is possible to find examples that show that the conditions given in Theorems 2 and 3 are only necessary and not sufficient. It seems to be difficult to find a necessary and sufficient “closed formula” condition for a set of integers to be codeword lengths of a uniquely decodable code for a constrained source. It is possible, however, to test the unique decodability of a given set of codewords for a given source, as shown in the next section.

C. *Extended Sardinas-Patterson test*

It is well known that the unique decodability, in the classic sense, of a set of codewords can be tested using the Sardinas-Patterson algorithm [8]. In this section we aim at showing how the original algorithm can be easily adapted to the case of constrained sources. The generalization is straightforward, so that it is not necessary to give a formal proof of the correctness, we refer to [15, th. 2.2.1] for the proof in the classic case.

For simplicitly, we consider here only the case of Markov sources modeled in the Moore form.

Let the source alphabet be $\mathcal{X} = \{x_1, x_2, \dots, x_m\}$ and let $W = \{w_i\}_{i=1,\dots,m}$ the set of codewords, where w_i is the code for x_i. For $i = 1, 2, \dots, m$ we call $F_i = \{w_j | P_{ij} > 0\}$ the subset of W containing all codewords that can follow w_i in a source sequence. We construct a sequence of sets $U_1, U_2, \dots$ in the following way. To form U_1 we consider all pairs of codewords of W; if a codeword w_i is a prefix of another codeword w_j, i.e. $w_j = w_i A$ we put the suffix A into U_1. In order to consider only the possible sequences, we have to keep trace of the codewords that have generated every suffix; thus, let us say that we mark the obtained suffix A with the two labels i and j, and we thus write it as ${}_iA_j$. We do this operation for every pair of words w_i and w_j from W, i.e. for $i, j = 1, \dots, m$, so obtaining U_1. Then, for $k > 1$, U_k is constructed by comparing elements of U_{k-1} and elements of W. For a generic element ${}_iB_j$ of U_{k-1} we consider the subset F_i of W:

a) If ${}_iB_j$ is equal to a codeword in F_i, the algorithm stops and the code is not decodable;
b) if ${}_iB_j$ is a prefix of a codeword w_r in F_i, say $w_r = {}_iB_jC$, we put the labelled ${}_jC_r$ suffix into U_k;
c) if instead a codeword w_r in F_i is prefix of ${}_iB_j$, say ${}_iB_j = w_rD$, we place the labelled suffix ${}_rD_j$ into U_k.

The code is uniquely decodable if and only if item a) is never reached.

Note that the algorithm can be stopped after a finite number of steps; there are in fact only a finite number of possible different sets U_i and so the sequence $U_i, i = 1, 2, \dots$ is either finite (i.e., the U_i are empty sets from sufficiently high i) or periodic. We note that the code is uniquely decodable with *finite delay* if the sequence $\{U_i\}$ is finite and uniquely decodable with *infinite delay* if the sequence is periodic. In this case the code is still decodable, since finite strings of code symbols can always be uniquely decoded, but the required delay is not bounded. This means that, for any positive n, there are at least two source sequences that produce codes that require more than n symbols delay in order to be disambiguated.

As an example of SP test for constrained sequences we consider the transition graphs shown in Fig. 3. For both cases we use codewords 0, 1, 01 and 10 for A, B, C and D respectively. For the graph of fig. 3(a) we obtain $U_1 = \{{}_A1_C, {}_B0_D\}$, $U_2 = \emptyset$. Thus the code is finite delay uniquely decodable and we can indeed verify that we need to wait at most two bits for decoding a symbol (this code is indeed the code used for the example of Section 1). For the graph of fig. 3(b), instead, we have $U_1 = \{{}_A1_C, {}_B0_D\}$, $U_2 = \{{}_C0_D, {}_D1_C\}$ and then $U_i = S_2$ for every other $i \geq 3$. So, the code is still uniquely decodable but with infinite delay; in fact it is not possible to distinguish the sequences $ADDD\cdots$ and $CCC\cdots$ until they are finished, so that the delay may be as long as we want.

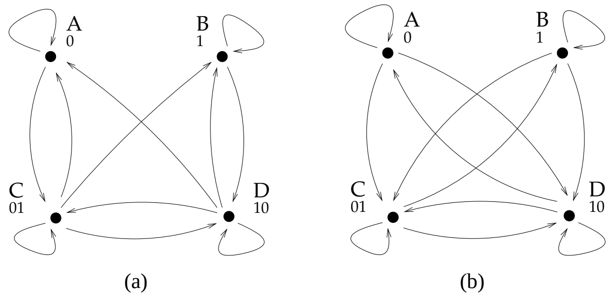

Fig. 3. Two examples of transition graphs with codewords associated to symbols. In both cases $\rho(\mathbf{Q}) = 1$; for source 3(a) the obtained code is uniquely decodable with finite delay, while for source 3(b) the obtained code is uniquely decodable but with infinite delay.

IV. On McMillan-like theorems and a proof by Shannon

In this section we want to provide an analysis of McMillan's theorem from a historical point of view, comparing different proofs and in particular by showing that both the original proof by McMillan [3] and Karush's one [13] are essentially mathematically equivalent to a proof used by Shannon [2] for the evaluation of the capacity of certain channels. In a sense, we can say that McMillan theorem was "almost" already proved in Shannon's paper. Even more interestingly, also our extension of McMillan's theorem was almost already present in Shannon's original paper, hidden in the evaluation of the capacity of finite state channels such as the telegraph [2].

Consider first the original proof by McMillan of his own theorem [3]. Let $l_{\max}$ be the maximum of the lengths $l_1, l_2, \dots, l_m$ and let $w(r)$ the number of words of length r; the Kraft inequality can thus be written as

$$\sum_{r=1}^{l_{\max}} w(r) D^{-r} \leq 1. \tag{13}$$

Let then $\tilde{Q}(x)$ be the polynomial defined by

$$\tilde{Q}(x) = \sum_{r=1}^{l_{\max}} w(r) x^r. \tag{14}$$

The proof is based on the study of $\tilde{Q}(x)$ as a function of a complex variable x and leads to a stronger result than the Kraft inequality, namely to the result that $\tilde{Q}(x) - 1$ has no zeros in the circle $|xD| < 1$ of the complex plane. As $\tilde{Q}(x)$ is continue and monotone for real $x \geq 0$, the Kraft inequality easily follows.

By removing from the original proof the parts that are not strictly important for the proof of the simple Kraft inequality, we obtain approximately the following flow. Let $N(k)$ be the number of sequences of source symbols whose code has total length k. Since the code is uniquely decodable, there are at most D^k such sequences, i.e., $N(k) \leq D^k$. It is thus clear that the series $1 + N(1)x + N(2)x^2 + \cdots$ converges for values of $x < D^{-1}$; let $F(x)$ be the value of this series. Now, the fundamental step in the proof is to consider how the possible $N(k)$ sequences of k letters are obtained. McMillan uses the following reasoning. For every $r \leq l_{\max}$, let C_r be the set of sequences of length k with a first word of length

r. The obtained C_r sets are disjoint because of the unique decodability. For the first r letters of C_r there are exactly $w(r)$ different possibilities, the number of words of r letters, while for the remaining $k-r$ letters there are exactly $N(k-r)$ different combinations. So, we have

$$N(k) = w(1)N(k-1) + w(2)N(k-2) + \cdots \\ + w(l_{\max})N(k-l_{\max}) \quad (15)$$

The above equation holds for every k if one defines $N(r) = 0$ for negative r.

Now, take $x < 1/D$, multiply the above equation by x^k and sum for k from one to infinity. We have

$$F(x) - 1 = F(x)\tilde{Q}(x). \quad (16)$$

But as $F(x)$ is positive, $\tilde{Q}(x)$ must be smaller than one. By continuity one clearly sees that $\tilde{Q}(1/D)$ is at most 1, which is Kraft's inequaliy.

It is interesting to focus the attention on the key point of this proof, which is essentially the combination of eq. (15) with the requirement that $N(k) \leq D^k$. In particular it is implicitly established that the value of $\tilde{Q}(1/D)$ determines how fast $N(k)$ would need to grow in order to have a lossless code. So, by imposing $N(k) \leq D^k$, a constraint on $\tilde{Q}(1/D)$ is obtained as a consequence.

This basic idea is also used in the proof given by Karush, but in an easier way. Instead of considering the set of code strings of length k, Karush considers the sequences of k symbols of the source as explained in the previous section. After an accurate analysis it is not difficult to realize that the proof given by Karush "only" has the advantage of relating the asymptotic behavior[5] of the sum $1 + N(1)D^{-1} + N(2)D^{-2} + ..N(kl_{\max})D^{-kl_{\max}}$ to the value of $\tilde{Q}(1/D)$ in a more direct way. Thus, the two proofs both use the convergence, or the order of magnitude, of the sum $1+N(1)D^{-1}+N(2)D^{-2}+\cdots$ in order to study the asymptotic behavior of $N(k)$. We could then say that both proofs are based on a combinatorial *counting method* for the evaluation of $N(k)$ and by imposing the constraint that $N(k) \leq D^k$

It is interesting to find that the very same technique had already been used by Shannon in Part I, Section 1 of [2] while computing the capacity of discrete noiseless channels. Shannon considers a device which is used to communicate symbols over a channel and wants to study the number of messages that can be communicated per unit of time. He says:

> "*Suppose all sequences of the symbols $S_1, \ldots, S_n$ are allowed and these symbols have durations $t_1, \ldots, t_n$. What is the channel capacity? If $N(t)$ represents the number of sequences of duration t we have*
>
> $$N(t) = N(t-t_1)+N(t-t_2)+\cdots+N(t-t_n). \quad (17)$$
>
> *The total number is the sum of the numbers of sequences ending in $S_1, S_2, \ldots, S_n$ and these are $N(t-t_1), N(t-t_2), \ldots, N(t-t_n)$, respectively. According to a well known result in finite differences, $N(t)$ is then asymptotic for large t to X_0^t where X_0 is the largest real solution of the characteristic equation:*
>
> $$X^{-t_1} + X^{-t_2} + \cdots + X^{-t_n} = 1 \quad (18)$$
>
> *and therefore*
>
> $$C = \log X_0\text{''}. \quad (19)$$

It is not difficult to note that the result obtained by Shannon, if reinterpreted in a source coding setting, is essentially equivalent to McMillan theorem. Indeed, suppose the device considered by Shannon is a discrete time device, emitting a symbol from a D-ary alphabet at every time instant, so that the symbols $S_1, S_2, \ldots, S_n$ are just D-ary words. First note that Shannon's tacit assumption is that the device produces messages that can be decoded at the receiving point. We can thus rewrite this implicit assumption by saying that symbols $S_1, S_2, \ldots, S_n$ form a uniquely decipherable code. Let us now focus on the capacity of the considered device. As the device sends one symbol from a D-ary alphabet at every instant, it is clear, and it was surely obvious for Shannon, that the channel capacity is in this case at most $\log D$. This means that the obtained value of X_0 above satisfies $X_0 \leq D$. But X_0 is a solution to (18), and the left hand side of (18) is nonincreasing in X. So, setting $X = D$ in (18), the Kraft inequality is easily deduced.

In other words, McMillan's theorem was already "proved" in the Shannon paper, but it was not explicitly stated in the source coding formulation. It is clear that the formulation in the source coding setting, rather than in the channel coding one, is of great importance by its own from an information theoretic point of view. From the mathematical point of view, instead, it is very interesting to note that MacMillan proof is only a more rigorous and detailed description of the counting argument used by Shannon. Mathematically speaking, we can say that not only Shannon had already proved McMillan result, but that he had proved it in few lines, in a simple and elegant way, using exactly the same technique used by McMillan.

Now, note that Shannon did not state the above result as a theorem. In fact, he considered the result only as a particular case, used as an example. He indeed started the discussion with the clarification *"Suppose all sequences of the symbols $S_1, \ldots, S_n$ are allowed"*, because his main interest was in the general case where the sequences of symbols are produced with some given constraints, as for example in the case of the detailed study of the telegraph in Section I.1 of his paper. The model used by Shannon for constraints is the following.

> *"We imagine a number of possible states $a_1, a_2, \ldots, a_m$. For each state only certain symbols from the set $S_1, S_2, \ldots, S_n$ can be transmitted [...]. When one of these has been transmitted the state changes to a new state depending both on the old state and the particular symbol transmitted".*

Note that this is exactly the type of constraint that we have indicated as a Markov model in the Mealy form, earlier in this chapter. The general result obtained by Shannon and stated as Theorem 1 in [2] is the following:

[5] More precisely, in the expansion of (2) the coefficient of D^{-r} is, in general, smaller than $N(r)$ for values of r larger than $r/l_{\min}$, but this does not affect the asymptotic behavior of the sum for large k.

Theorem 4 (Shannon): Let $b_{ij}^{(s)}$ be the duration of the s^{th} symbol which is allowable in state i and leads to state j. Then the channel capacity C is equal to $\log W_0$ where W_0 is the largest real root of the determinant equation:

$$\left| \sum_s W^{-b_{ij}^{(s)}} - \delta_{ij} \right| = 0. \tag{20}$$

This theorem is well known in the field of coding for constrained systems (see for example [16], [17]) and can be considered as the channel coding precursor of the Mealy-form of Theorem 3 exactly in the same way as the result obtained by eqs. (18) and (19) is the precursor of McMillan theorem. We now prove that Theorem 4 can indeed be used to mathematically deduce Theorem 3. We prove this fact using Theorem 4 to show that, if the matrix $\mathbf{Q}(D)$ defined in Theorem 3 has spectral radius larger than 1, then the associated code cannot be uniquely decodable. In order to do that, we show that if such a code was decodable, then we could construct a channel using a D-ary alphabet with a capacity larger than $\log D$, which is clearly impossible.

Coherently with the notation of Theorem 3, let $\mathbf{Q}(W) = \sum_s W^{-b_{ij}^{(s)}}$ be the matrix considered in the determinant equation (20). Suppose now that there exists a uniquely decodable code for a constrained source such that the spectral radius of the matrix $\mathbf{Q}(D)$ in Theorem 3 is larger than 1. Then, as the code is uniquely decodable, we can construct a discrete-time D-ary channel with channel symbols exactly equal to the codewords of the given code. Then for this channel, with the above definitions, we have $\rho(\mathbf{Q}(D)) > 1$. Consider now the capacity of such a channel. The largest solution W_0 of the determinant equation (20) can also be considered as the largest positive value of W such that $\mathbf{Q}(W)$ has an eigenvalue equal to 1. Consider thus the largest eigenvalue of $\mathbf{Q}(W)$, i.e. the spectral radius $\rho(\mathbf{Q}(W))$. As the spectral radius of a nonnegative matrix decreases if any of the elements of the matrix decreases, $\rho(\mathbf{Q}(W))$ is a decreasing function of W. Furthermore, it is clear that $\rho(\mathbf{Q}(W)) \to 0$ when $W \to \infty$. Then clearly, since $\rho(\mathbf{Q}(D)) > 1$, there exists a $W > D$ such that $\rho(\mathbf{Q}(W)) = 1$. But this means that we have constructed a D-ary channel with capacity larger than $\log D$, which is clearly impossible. So, the initial hypothesis was wrong, and thus any decodable code for a constrained source is such that the spectral radius of the matrix $\mathbf{Q}(D)$ in Theorem 3 is not larger than 1.

This shows that the results obtained by Shannon for the channel capacity evaluation in his paper [2], actually correspond to very interesting results in the source coding setting, which hide a generalized form of Kraft-McMillan theorem.

V. Conclusions

In this paper we have proposed a revisitation of the foundations of noiseless source coding. In particular, a revisitation of the topic of unique decodability has been provided by properly treating the particular case of constrained sources. For this type of sources, it has been shown that the classic approach to unique decodabiliy leads to misleading results on the average length of codes for finite sequences of symbols. More in detail, we have shown that, contrarily to what has been so far accepted, the first n symbols of a source can be encoded with a lossless variable length code that uses an average number of bits strictly smaller than the entropy of such source symbols.

Based on this observation, we have revisited the topic of unique decodability by providing an extension of McMillan's theorem and of the Sardinas-Patterson test to deal with constrained sources. Finally, it has been clarified that both McMillan's original theorem and our own extension can be mathematically derived from the results obtained by Shannon in his original 1948 paper [2]. An interesting concern remains: what is the lower bound for encoding a finite sequence of symbols?

References

[1] S. Verdú, "Fifty years of shannon theory," *IEEE Trans. on Inform. Theory*, vol. 44, no. 6, pp. 2057–2078, 1998.

[2] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. Journal*, vol. 27, pp. 379–423,623–656, 1948.

[3] B. McMillan, "Two inequalities implied by unique decipherability," *IEEE Trans. Inform. Theory*, vol. IT-2, pp. 115–116, 1956.

[4] L.G. Kraft, "A device for quanitizing, grouping and coding amplitude modulated pulsese," M.S. thesis, Dept. of Electrical Engg., MIT, Cambridge, Mass., 1949.

[5] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley, New York, 1990.

[6] R.G. Gallager, *Information Theory and Reliable Communication*, Wiley, New York, 1968.

[7] A. D. Wyner, J. Ziv, and A. J. Wyner, "On the role of pattern matching in information theory," *IEEE Trans. on Inform. Theory*, vol. 44, no. 6, pp. 2045–2056, 1998.

[8] A.A. Sardinas and G.W. Patterson, "A necessary and sufficient condition for the unique decomposition of coded messages," in *IRE Convention Record, Part 8*, 1953, pp. 104–108.

[9] N. Alon and A. Orlitsky, "A lower bound on the expected length of one-to-one codes," *IEEE Trans. on Inform. Theory*, vol. 40, pp. 1670–1772, 1994.

[10] A. D. Wyner, "An upper bound on the entropy series," *Inform. and Control*, vol. 20, pp. 176–181, 1972.

[11] C. Blundo and R. De Prisco, "New bounds on the expected length of one-to-one codes," *IEEE Trans. on Inform. Theory*, vol. 42, no. 1, pp. 246–250, 1996.

[12] S. A. Savari and A. Naheta, "Bounds on the expected cost of one-to-one codes," in *Proc. IEEE Intern. Symp. on Inform. Theory*, 2004, p. 92.

[13] J. Karush, "A simple proof of an inequality of McMillan," *IRE Trans. Inform. Theory*, vol. IT-7, pp. 118, 1961.

[14] H. Minc, *Nonnegative Matrices*, Wiley, 1988.

[15] R.B. Ash, *Information Theory*, Interscience, New York, 1965.

[16] K.E. Schouhamer Immink, P.H. Siegel, and J.K. Wolf, "Codes for digital recorders," *IEEE Trans. on Inform. Theory*, vol. 44, no. 6, pp. 2260–2299, 1998.

[17] D. Lind and B. Marcus, *An introduction to Symbolic Dynamics and Coding*, Cambridge University Press., 1996.