On Unique Decodability

Marco Dalai, *Associate Member, IEEE*, and Riccardo Leonardi, *Senior Member, IEEE*

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

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

where $X_1, X_2, \ldots, X_n$ are the first n symbols of the source, $E[l(X_1, X_2, \ldots, X_n)]$ 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 when 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 an extended version of McMillan's theorem to be used for Markovian sources.

Index Terms—Constrained sources, lossless source coding, McMillan's theorem, minimum expected code length.

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 types 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 nonasymptotic strong converse to the coding theorem. In particular, the famous result obtained (see [6, Th. 3.5.2], [5, Th. 5.4.2], and [7, Sec. II, p. 2047]) is that, for every source with memory, any uniquely decodable code satisfies

$$E[l(X_1, X_2, \ldots, X_n)] \geq H(X_1, X_2, \ldots, X_n) \qquad (1)$$

where $X_1, X_2, \ldots, X_n$ are the first n symbols of the source, $E[l(X_1, X_2, \ldots, X_n)]$ 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 for the case of first-order Markov sources. Finally, in Section IV, some additional interesting remarks on the considered topic are made.

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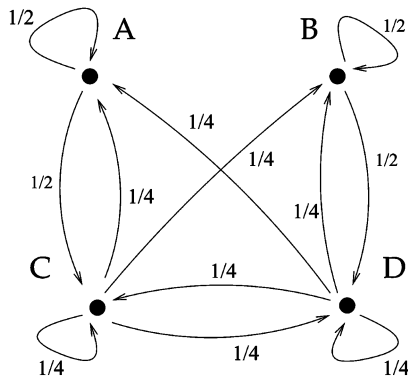
Fig. 1. Graph, with transition probabilities, for the Markov source used in the example.

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. 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

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

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.

A. Classic Code

We call this first code "classic" as it is the most natural way to encode the source given its particular structure. Because 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 kth symbol, we use, again in an obvious way, 1 bit if symbol $k-1$ is an A or a B, and we use 2 bits if symbol $k-1$ is 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

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

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.

B. Alternative Code

Let us consider a different code, obtained by applying the following fixed mapping from symbols to bits: $A \rightarrow 0$, $B \rightarrow 1$, $C \rightarrow 01$, and $D \rightarrow 10$. 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, because 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 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

$$E[l(X_1, X_2, X_3, \ldots, X_n)] = \sum_{i=1}^{n} E[l(X_i)]$$
$$= \frac{3}{2}n.$$

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, which 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 that remains compliant with the source characteristics. For each source sequence, indeed, the number

TABLE I
ENCODING AND DECODING OPERATIONS OF THE PROPOSED ALTERNATIVE
CODE FOR THE MARKOV SOURCE OF FIG. 1

Encoding	$A \rightarrow 0$		
	$B \rightarrow 1$		
	$C \rightarrow 01$		
	$D \rightarrow 10$		

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

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 to design the code. However, 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 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, [8] 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 [9] 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 [10], [11]. 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.

1) McMillan's theorem asserts that all uniquely decodable codes satisfy Kraft's inequality.
2) If a code for a random variable X satisfies Kraft's inequality, then $E[l(X)] \geq H(X)$.
3) Thus, any uniquely decodable code for a random variable X satisfies $E[l(X)] \geq H(X)$.
4) 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 that 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 the 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, 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.

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 as the one 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]

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 1)–4) 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 the following.

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

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

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.

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

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

Let X be a Markov source with alphabet $\mathcal{X} = \{1, 2, \ldots, m\}$ and transition probability matrix $\mathbf{P}$. Let $W = \{w_1, w_2, \ldots, w_m\}$ 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 [3]: If the set of codewords W is uniquely decodable (in the sense of Definition 2), then

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

We propose a modified theorem for considering the unique decodability for the specific source.

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

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

has spectral radius at most 1.

Proof: The proof is very similar to Karush's proof of McMillan's theorem [12]. Let $\mathcal{X}^{(k)}$ be the set of all sequences of k symbols that can be produced by the source and let $\mathbf{L} = [D^{-l_1}, D^{-l_2}, \ldots, D^{-l_m}]'$. For $k > 0$, define the row vector

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

It is easy to see by induction that the ith component of $\mathbf{V}^{(k)}$ can be written as

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

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

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

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 $\mathcal{X}^{(k)}$, which are mapped in a length r code, we have

$$\mathbf{L}'\mathbf{Q}^{k-1}\mathbf{1}_m = \sum_{r=1}^{kl_{\max}} N(r)D^{-r}$$

where $l_{\max}$ is the maximum of the values $l_i, i = 1, 2, \ldots, m$. Because the code is uniquely decodable for the source X, there are at most D^r source-compatible sequences with a code of length r, that is, $N(r) \leq D^r$. Hence, for every $k > 0$

$$\mathbf{L}'\mathbf{Q}^{k-1}\mathbf{1}_m \leq \sum_{r=1}^{kl_{\max}} D^r D^{-r} = kl_{\max}. \qquad (2)$$

Now, note that the irreducible matrix $\mathbf{Q}$ is also nonnegative. Thus, by the Perron–Frobenius theorem (see [13] for details),

its spectral radius $\rho(\mathbf{Q})$ is also an eigenvalue, with algebraic multiplicity 1 and with positive associated left eigenvector.

Suppose now $\rho(\mathbf{Q}) > 1$. Because $\mathbf{L}$ and $\mathbf{1}_m$ are both positive, it is easy to deduce that the term on the left-hand side of (2) asymptotically grows as $\rho(\mathbf{Q})^{k-1}$ when k goes to infinity. On the contrary, the right-hand side term only grows linearly with k and, for large enough k, (2) cannot hold. We conclude that $\rho(\mathbf{Q}) \leq 1$. $\square$

IV. SOME ADDITIONAL REMARKS

Remark 1 (Theorem 2 Generalizes Theorem 1): In the case of unconstrained Markov sources, Theorem 2 is equivalent to Theorem 1. Indeed, the Markov source being not constrained means that its transition probability matrix $\mathbf{P}$ has all strictly positive entries. This implies that the matrix $\mathbf{Q}$ defined in Theorem 2 has all equal rows. The spectral radius of such a matrix equals the sum of the elements in every row, which is $\sum_j D^{-l_j}$, reducing thus to the classic Kraft's inequality.

Remark 2 (Nonsufficiency of the Condition): Kraft's inequality is both a necessary and sufficient condition for the existence of a uniquely decodable code (in the sense of Definition 2) with codeword lengths l_i. Theorem 2, instead, only gives a necessary condition on the lengths l_i for the unique decodability of a code for a given source. It is easy to show that the condition stated in the theorem is not a sufficient condition for the existence of a uniquely decodable code for a source with codeword lengths l_i. Finding a necessary and sufficient condition seems to be a much harder problem.

Remark 3 (Extended Sardinas–Patterson Test): With respect to the previous remark, we point out that it is possible however to test a given code for decodability for a given source by devising a generalization of the Sardinas–Patterson test [14] to deal with constrained sources (see [15]).

Remark 4 (A More General Form of Theorem 2): Theorem 2 was formulated for the case of 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. In order to deal with more general sources, we can consider Markov sources in the "Mealy form," where output symbols are not associated to states but to transitions between states (which corresponds to the Markov source model used by Shannon in [2] or, for example, by Gallager in [6]). Theorem 2 can be extended to this type of Markov sources as follows (see [15]).

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 $\mathbf{Q}$ defined by

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

has spectral radius at most 1.

Remark 5 (Shannon's Insight): A historical analysis reveals that both McMillan's theorem and the proposed generalized one in the form of Theorem 3 are mathematically equivalent to a formulation obtained already by Shannon in [2, Part I, Sec. 1] for the evaluation of the capacity of discrete noiseless channels. In particular, in [2], Shannon established that the capacity of an unconstrained noiseless channel with symbol durations $t_1, t_2, \ldots, t_m$ is given by the value $\log X_0$, where X_0 is the largest real solution of the difference equation

$$X^{-t_1} + X^{-t_2} + \cdots + X^{-t_m} = 1.$$

It is not difficult to show that McMillan's theorem is equivalent to the obvious statement that the capacity of a D-ary channel is at most $\log D$.

Furthermore, Shannon generalized the capacity formula to the case of noiseless finite state channels, by stating the following [2, Th. 1].

Theorem 4 [2]: Let $b_{ij}^{(s)}$ be the duration of the sth 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

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

As for the unconstrained case, it is possible to show that Theorem 3 is equivalent to the statement that every finite state D-ary channel has capacity at most $\log D$.

REFERENCES

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

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

[3] B. McMillan, "Two inequalities implied by unique decipherability," *IRE Trans. Inf. Theory*, vol. IT-2, no. 4, pp. 115–116, Dec. 1956.

[4] L. G. Kraft, "A device for quantizing, grouping and coding amplitude modulated pulsese," M.S. thesis, Dept. Electr. Eng., Massachusetts Inst. Technol., Cambridge, MA, 1949.

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

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

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

[8] N. Alon and A. Orlitsky, "A lower bound on the expected length of one-to-one codes," *IEEE Trans. Inf. Theory*, vol. 40, no. 5, pp. 1670–1772, Sep. 1994.

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

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

[11] S. A. Savari and A. Naheta, "Bounds on the expected cost of one-to-one codes," in *Proc. IEEE Int. Symp. Inf. Theory*, 2004, pp. 92–92.

[12] J. Karush, "A simple proof of an inequality of McMillan," *IRE Trans. Inf. Theory*, vol. IT-7, no. 2, pp. 118–118, Apr. 1961.

[13] H. Minc, *Nonnegative Matrices*. New York: Wiley, 1988.

[14] 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.

[15] M. Dalai and R. Leonardi, "On unique decodability, McMillan's theorem and the expected length of codes," Univ. Brescia, Brescia, Italy, Tech. Rep. R.T. 2008–01-58, 2008.