October, 1997

LIDS- P 2401

Research Supported By:

NSF Faculty Early Career Development
(CAREER) Program

On the Universality of the LZ-based Decoding Algorithm

Lapidoth, A.
Ziv, J.

On the Universality of the LZ-based Decoding Algorithm

Amos Lapidoth and Jacob Ziv *

ABSTRACT

A universal decoder for a family of channels is a decoder that can be designed without prior knowledge of the particular channel over which transmission will be carried out, and it yet attains the same random coding error exponent as the optimal decoder tuned to the channel in use. In this paper we study Ziv's decoding algorithm, which is based on Lempel-Ziv incremental string parsing, and demonstrate that while it was originally proposed as a universal decoder for the family of finite state channels with deterministic (but unknown) transitions, it is in fact universal for the much broader class of all finite state channels.

The complexity of this decoder is substantially smaller than that of the universal decoder recently proposed by Feder and Lapidoth. However, the universality established is somewhat weaker than that established by Feder and Lapidoth as it only holds if the set from which the codewords of the random codebook are drawn is permutation invariant, as is the case if the codewords are chosen independently and uniformly over the set of sequences of a given type.

KEYWORDS: Universal decoding, Lempel-Ziv parsing, Error exponents, Finite state channels, Gilbert-Elliott channel.

*A. Lapidoth is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139-4307. J. Ziv is with the Department of Electrical Engineering, Technion — Israel Institute of Technology, Haifa 32000, Israel. This research was carried out in part while the authors were visiting Lucent Technologies, Murray Hill, NJ 07974. The work of A. Lapidoth was supported in part by the NSF Faculty Early Career Development (CAREER) Program.

1 Introduction

This paper deals with the design of a receiver for coded communication over an unknown channel. The channel is assumed to belong to the family of finite state channels over finite input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$, but is otherwise unknown. We do not even assume that the number of states is known to the receiver designer, let alone the probability law governing the channel behavior. The code being used is, however, known to the receiver designer, and the receiver is expected to decode the received sequences with a low probability of error. Had the channel been known to the receiver it could have employed the maximum-likelihood decoding rule, which minimizes the probability of error. This rule, however cannot be implemented in our scenario because it depends on the channel law, which is unknown at the receiver. In this paper we study a decoding rule that does not require knowing the channel law, and can yet perform asymptotically as well as the maximum-likelihood rule for many good codes.

The decoder we study was first proposed by Ziv in [1] for decoding finite-state channels with *deterministic* (but unknown) state transitions, i.e., finite state channels where the next state is a deterministic function of the present state, input, and output. Ziv showed that if $\bar{P}_{\theta_d,\mathrm{ML}}(\text{error})$ denotes the average (over messages and codebooks) probability of error incurred over the channel θ_d when maximum-likelihood decoding is performed to decode a rate-R blocklength-n codebook whose codewords are drawn independently and uniformly over a permutation invariant set $B_n \subset \mathcal{X}^n$, and if $\bar{P}_{\theta_d,\mathrm{Z}}(\text{error})$ denotes the analogous expression when Ziv's decoder is used instead of the optimal maximum-likelihood rule, then

$$\lim_{n\to\infty} \sup_{\theta_d} \frac{1}{n} \log \frac{\bar{P}_{\theta_d,\mathrm{Z}}(\text{error})}{\bar{P}_{\theta_d,\mathrm{ML}}(\text{error})} = 0,$$

where the supremum is over all finite-state channels with deterministic transitions defined over common finite input, output, and state alphabets $\mathcal{X}, \mathcal{Y}, \mathcal{S}$ respectively. In the terminology of [2] this is referred to as "strong random coding universality". (The adjective "strong" reflects that the convergence of the performance of the sub-optimal decoder to that of the maximum-likelihood decoder is uniform over the family.)

In this paper we extend Ziv's results in two ways. First, and most importantly, we show that if the sets B_n are permutation invariant then his decoder is "strong random coding universal" for the family of *all* finite state channels, and not only for the set of finite state channels with deterministic transitions. Thus

$$\lim_{n\to\infty} \sup_{\theta} \frac{\bar{P}_{\theta,\mathrm{z}}(\mathrm{error})}{\bar{P}_{\theta,\mathrm{ML}}(\mathrm{error})} = 0,$$

where the supremum is over all finite-state channels defined over common finite input, output, and state alphabets $\mathcal{X}, \mathcal{Y}, \mathcal{S}$. Secondly, we establish that there exists a sequence of rate-R blocklength-n codebooks $\mathcal{C}_n \subset B_n$ such that

$$\lim_{n\to\infty} \sup_{\theta} \frac{1}{n} \log \frac{P_{\theta,\mathrm{z}}(\mathrm{error}|\mathcal{C}_n)}{\bar{P}_{\theta,\mathrm{ML}}(\mathrm{error})} = 0,$$

where $P_{\theta,\mathrm{z}}(\mathrm{error}|\mathcal{C}_n)$ denotes the average (over messages) probability of error incurred by Ziv's decoder in decoding the codebook $\mathcal{C}_n$ over the channel θ. This form of universality is referred to in [2] as "strong deterministic coding universality".

It should be noted that prior to Ziv's work, Csiszár and Körner had studied the problem of decoding an unknown *memoryless channel* in [3] and demonstrated that a different decoder, the Maximum empirical Mutual Information (MMI) decoder, is "strong deterministic coding universal" in the sense that if $\bar{P}_{\theta_m,\mathrm{ML}}(\mathrm{error})$ denotes the average (over messages and codebooks) probability of error over the memoryless channel θ_m incurred by the maximum-likelihood decoder in decoding a random codebook whose codewords are drawn independently and uniformly over a type set then there exists a sequence of rate-R blocklength-n codebooks $\mathcal{C}_n$ for which

$$\lim_{n\to\infty} \sup_{\theta_m} \frac{1}{n} \log \frac{P_{\theta_m,\mathrm{MMI}}(\mathrm{error}|\mathcal{C}_n)}{\bar{P}_{\theta_m,\mathrm{ML}}(\mathrm{error})} = 0,$$

where $P_{\theta_m,\mathrm{MMI}}(\mathrm{error}|\mathcal{C}_n)$ is the average probability of error incurred by the MMI decoder in decoding the codebook $\mathcal{C}_n$ over the memoryless channel θ_m, and the supremum is over all memoryless channel defined over common finite input and output alphabets $\mathcal{X}, \mathcal{Y}$.

For the class of memoryless channels the MMI decoder is equivalent to the generalized likelihood ratio test that given a codebook $\mathcal{C}$, a received sequence $\mathbf{y}$, and a family of channels $\mathcal{F}$ declares that the transmitted codeword is $\mathbf{x} \in \mathcal{C}$ only if

$$\max_{\phi\in\mathcal{F}} p_{\phi}(\mathbf{y}|\mathbf{x}) \geq \max_{\phi\in\mathcal{F}} p_{\phi}(\mathbf{y}|\mathbf{x}'), \ \forall \mathbf{x}' \in \mathcal{C}.$$

The universality of the MMI decoder for the family of memoryless channels might lead one to conjecture that the generalized likelihood ratio test is canonical in the sense that it is universal for any family of channels for which a universal decoder exist. As we shall see in Section 2, this is false.

In [2] Feder and Lapidoth introduced yet another universal decoding rule, one that is based on the idea of "merging decoders" and demonstrated the universality of this decoder for fairly general families of channels, including the family of finite-state channels. The results reported there are somewhat more general then the results reported in this paper, as the universality in [2] does not require that the sets B_n from which the codewords are drawn be permutation invariant. However, Ziv's LZ-based universal decoder has significant advantages over the decoder proposed in [2] in terms of complexity.

Given some received sequence $\mathbf{y}$ both decoders associate a score with each of the codewords, and choose the codeword that attains the highest score. However while the complexity of assigning a score to each codewords is linear in the blocklength for Ziv's decoder, this complexity is typically exponential for the decoder proposed by Feder and Lapidoth. To compute the latter score one needs to consider the ranking of the candidate codeword among all *sequences* (not just codewords) in the set B_n, and to compute this ranking for each of a polynomial number channels. To compute this ranking for a given channel, one must typically compute the likelihood of each of the sequences in B_n, of which there are typically an exponential number. Moreover, for a finite-state channel, even to compute the likelihood of a *single* sequence for a *given* channel requires complexity that is exponential in the blocklength as the likelihood needs to be summed over all possible state sequences, see (10). To compute the score assigned by Feder and Lapidoth's decoder to a given codeword thus requires roughly $|B_n||\mathcal{S}|^n$ computations, multiplied by the polynomial number of decoders being merged. This should be contrasted with the linear complexity (per codeword) of Ziv's decoder!

Moreover, as shown in [1], Ziv's decoder is sequential and is thus particularly suitable for sequential decoding. These implementation consideration make Ziv's decoder particularly attractive. However, while the finite state channel with deterministic transition studied in [1] is often useful for modeling intersymbol interference channels [4], this model is ill suited for many wireless applications where the channel time-variations may be independent of the input signal, as in the Gilbert-Elliott channel [5].

The present contribution demonstrates that Ziv's decoder works, without modifications, not only for finite state channels with deterministic transitions, but also for the more general finite state channels encountered in wireless communications. It is hoped that the decoder's low complexity on the one hand, and its universality for such general families of channel on the other, will promote its use in wireless applications.

The rest of the paper is organized as follows. We conclude this section with a precise statement of our main results. A proof of the main result, Theorem 1, is given in Section 2, and the paper is concluded with a discussion in Section 3 where we also demonstrate that the generalized likelihood ratio test is not canonical, and where we discuss the duality between the universal source coding problem and the universal decoding problem.

Precise Statement of the Problem and Main result

We begin by describing Ziv's decoding rule [1] for a channel with finite input alphabet $\mathcal{X}$ and finite output alphabet $\mathcal{Y}$. To implement this decoding rule the receiver must know the codebook being used, but need not know the channel law. Consider then a codebook $\mathcal{C}$ of rate R and blocklength n,

$$\mathcal{C} = \left\{ \mathbf{x}(1), \dots, \mathbf{x}(2^{nR}) \right\} \subset \mathcal{X}^n. \tag{1}$$

(Strictly speaking we should denote the number of codewords by $\lfloor 2^{nR} \rfloor$ but for simplicity we use 2^{nR} instead.) Given a received sequence

$$\mathbf{y} = y_1, \dots, y_n$$

Ziv's decoder declares that the transmitted codeword is $\mathbf{x}(i)$ if

$$u(\mathbf{x}(i), \mathbf{y}) < u(\mathbf{x}(j), \mathbf{y}) \ \ \forall j \neq i,$$

and declares a decoding failure if no such codeword exists, as can only be the case if the minimum of $u(\cdot, \mathbf{y})$ over $\mathcal{C}$ is not unique. The function u, mapping $\mathcal{X}^n \times \mathcal{Y}^n$ to the reals will be described next.

Given the received sequence $\mathbf{y} \in \mathcal{Y}^n$ and any codeword $\mathbf{x} \in \mathcal{X}^n$ let $\mathbf{w} \in \mathcal{X}^n \times \mathcal{Y}^n$ be the sequence of ordered pairs

$$\mathbf{w} = w_1, \dots, w_n, \ \ w_i = (x_i, y_i).$$

Consider the incremental parsing [6] of $\mathbf{w}$ into phrases (strings) such that 1) all the phrases (except for possibly the last one) are distinct, and 2) the prefix of each phrase[1] is identical to some previous phrase. Let g be the number of resulting phrases. Using w_i^j to denote the string of length $j-i+1$ beginning with w_i and terminating in w_j, i.e.,

$$w_i^j = w_i w_{i+1} \cdots w_j,$$

we have that

$$\mathbf{w} = w_1^{l_1} w_{l_1+1}^{l_2} w_{l_2+1}^{l_3} w_{l_3+1}^{l_4} \cdots w_{l_{g-1}+1}^{l_g}, \tag{2}$$

where $w_{l_{i-1}+1}^{l_i}$ is the i-th phrase, $l_1 = 1$, $l_g = n$, and for convenience we set $l_0 = 0$. The first condition on the incremental parsing translates to

$$w_{l_{i-1}+1}^{l_i} \neq w_{l_{k-1}+1}^{l_k} \;\; \forall k \neq l, \;\; 1 \leq i, k < g,$$

and the second condition on the parsing translates to

$$\forall 1 \leq i < g \; \exists k < i \; : \; w_{l_{i-1}+1}^{l_i - 1} = w_{l_{k-1}+1}^{l_k} \tag{3}$$

where (3) need hold only if $l_i - l_{i-1} > 1$, since otherwise $w_{l_{i-1}+1}^{l_i-1}$ is the empty string. We let $c(\mathbf{x}, \mathbf{y})$ denote the number of distinct phrases and note that $g - 1 \leq c(\mathbf{x}, \mathbf{y}) \leq g$ as all but possibly the last phrase are distinct.

The incremental parsing of $\mathbf{w}$, see (2), induces a parsing on $\mathbf{y}$ defined by

$$\begin{aligned} \mathbf{y} &= y_1^{l_1} y_{l_1+1}^{l_2} y_{l_2+1}^{l_3} y_{l_3+1}^{l_4} \cdots y_{l_{g-1}+1}^{l_g} \\ &= \mathbf{y}^{(1)} \mathbf{y}^{(2)} \cdots \mathbf{y}^{(g)} \\ &= y_{b(1)}^{e(1)} y_{b(2)}^{e(2)} \cdots y_{b(g)}^{e(g)} \end{aligned} \tag{4}$$

where we denote the m-th phrase by $\mathbf{y}^{(m)}$, its beginning by $b(m)$ and its end by $e(m)$. Thus

$$\mathbf{y}^{(m)} = y_{l_{m-1}+1}^{l_m} = y_{b(m)}^{e(m)}, \;\; m = 1, \dots, g, \tag{5}$$

is the m-th phrase in the induced parsing on $\mathbf{y}$, a phrase that begins in time index $b(m) = l_{m-1} + 1$ and ends at time $e(m) = l_m$. Similarly

$$\begin{aligned} \mathbf{x} &= x_1^{l_1} x_{l_1+1}^{l_2} x_{l_2+1}^{l_3} x_{l_3+1}^{l_4} \cdots x_{l_{g-1}+1}^{l_g} \\ &= \mathbf{x}^{(1)} \mathbf{x}^{(2)} \cdots \mathbf{x}^{(g)} \\ &= x_{b(1)}^{e(1)} x_{b(2)}^{e(2)} \cdots x_{b(g)}^{e(g)} \end{aligned} \tag{6}$$

[1] The prefix of a phrase is the string that results when the last symbol of the phrase is deleted.

where

$$\mathbf{x}^{(m)} = x_{l_{m-1}+1}^{l_m} = x_{b(m)}^{e(m)}, \quad m = 1, \dots, g. \tag{7}$$

Note that in the induced parsings the phrases are not necessarily distinct, as two phrases of $\mathbf{w}$ could be distinct and yet have identical $\mathbf{y}$ (or $\mathbf{x}$) components. We set $c(\mathbf{y})$ to be the number of distinct phrases in the parsing of $\mathbf{y}$ induced by the incremental parsing of $\mathbf{w}$. We denote by $y(l)$, $1 \leq l \leq c(\mathbf{y})$ the l-th distinct phrase in the induced parsing on $\mathbf{y}$, and set $c_l(\mathbf{x}|\mathbf{y})$ to be the number of distinct x phrases that appear jointly with $y(l)$. We thus have

$$\sum_{l=1}^{c(\mathbf{y})} c_l(\mathbf{x}|\mathbf{y}) = c(\mathbf{x}, \mathbf{y}). \tag{8}$$

In fact, $c_l(\mathbf{x}|\mathbf{y})$ is at most the number of occurrences of the phrase $y(l)$ in the induced parsing of $\mathbf{y}$, and at least this number of occurrences minus one, as all but possibly the last phrase of $\mathbf{w}$ are distinct.

The function $u(\mathbf{x}, \mathbf{y})$ proposed in [1] can now be defined as

$$u(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{l=1}^{c(\mathbf{y})} c_l(\mathbf{x}|\mathbf{y}) \log c_l(\mathbf{x}|\mathbf{y}), \tag{9}$$

thus concluding the description of Ziv's decoding rule.

In this paper we study the performance of Ziv's decoding rule when used over a finite state channel. We thus need some definitions regarding finite state channels. A finite state channel [4] over the finite input alphabet $\mathcal{X}$, finite output alphabet $\mathcal{Y}$, and finite state alphabet $\mathcal{S}$ is specified by a probability law

$$P(y, s'|x, s), \quad y \in \mathcal{Y},\ x \in \mathcal{X},\ s, s' \in \mathcal{S},$$

which specifies the probability law of the channel's current output and current state, given the channel's current input and preceding state. Specifically the probability of an n length output sequence

$$\mathbf{y} = (y_1, \dots, y_n) \in \mathcal{Y}^n,$$

given the channel input

$$\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{X}^n,$$

and the channel's initial state $s_0 \in \mathcal{S}$, is given by

$$P_n(\mathbf{y}|\mathbf{x}, s_0) = \sum_{\mathbf{s} \in \mathcal{S}^n} P_n(\mathbf{y}, \mathbf{s}|\mathbf{x}, s_0), \tag{10}$$

where

$$P_n(\mathbf{y}, \mathbf{s}|\mathbf{x}, s_0) = \prod_{i=1}^{n} P(y_i, s_i|x_i, s_{i-1}). \tag{11}$$

We shall denote the set of all pairs of initial states s_0 and probability laws $P(y, s'|x, s)$ by Θ. For any

$$\theta = (s_0, P(\cdot, \cdot|\cdot, \cdot)) \in \Theta$$

we set

$$p_\theta(\mathbf{y}|\mathbf{x}) = \sum_{s_1, \ldots, s_n} \prod_{i=1}^{n} P(y_i, s_i|x_i, s_{i-1}),$$

to be the corresponding channel law.

We say that the channel has *deterministic state transitions* if the channel state is a deterministic function of its preceding state, i.e., if

$$s_i = q(s_{i-1}),$$

for some deterministic function $q : \mathcal{S} \to \mathcal{S}$. Alternatively, the channel has deterministic state transitions if

$$P(y, s'|x, s)P(\tilde{y}, \tilde{s}'|\tilde{x}, s) > 0 \Rightarrow s' = \tilde{s}'.$$

We shall denote by $\Theta_d \subset \Theta$ the set of all pairs of initial states and deterministic transition laws.

Our definition of finite state channels with deterministic transitions seems to be more restrictive than the definition adopted in [1] where the channel state s_i at time i is allowed to be a deterministic function not only of the preceding state s_{i-1} but also of the previous input x_{i-1} and previous output y_{i-1}, i.e., when $s_i = q(s_{i-1}, x_{i-1}, y_{i-1})$ for some deterministic function $q : \mathcal{S} \times \mathcal{X} \times \mathcal{Y} \to \mathcal{S}$. This ostensibly more general situation can, however, be accounted for by augmenting the state alphabet so that the channel input and output are determined by the channel state: one would consider the state alphabet $\hat{\mathcal{S}}$ defined by

$$\hat{\mathcal{S}} = \mathcal{S} \times \mathcal{X} \times \mathcal{Y},$$

and the law $\hat{P}$, where for any input $x \in \mathcal{X}$, output $y \in \mathcal{Y}$ and any two states $\hat{s} = (s, \hat{x}, \hat{y}) \in \hat{\mathcal{S}}$, and $\hat{s}' = (s', \hat{x}', \hat{y}') \in \hat{\mathcal{S}}$

$$\hat{P}(y, \hat{s}'|x, \hat{s}) = \begin{cases} P(y, s'|x, s) & \text{if } \hat{x}' = x \text{ and } \hat{y}' = y \\ 0 & \text{otherwise} \end{cases}.$$

Having defined the families of channels with which this paper deals, we now turn to the performance measures we adopt. Given a rate-R, blocklength-n codebook $\mathcal{C}$ as in (1) we set $P_{\theta,\theta}(\text{error}|\mathcal{C})$ to be the average (over messages) probability of error that is incurred over the channel $\theta \in \Theta$ when decoding is performed according to the maximum-likelihood rule tuned to the channel θ, i.e., a rule that given a received sequence $\mathbf{y} \in \mathcal{Y}^n$ declares that the transmitted codeword is $\mathbf{x}(i)$ only if

$$p_\theta(\mathbf{y}|\mathbf{x}(i)) \geq p_\theta(\mathbf{y}|\mathbf{x}(j)), \quad \forall j \neq i. \tag{12}$$

We similarly denote by $P_{\theta,\text{Z}}(\text{error}|\mathcal{C})$ the average (over messages) probability of error incurred when the code $\mathcal{C}$ is used over the channel θ and is decoded using Ziv's decoder, i.e.,

$$P_{\theta,\text{Z}}(\text{error}|\mathcal{C}) = 2^{-nR} \sum_{i=1}^{2^{nR}} \sum_{\mathbf{y} \notin \mathcal{D}_i} p_\theta(\mathbf{y}|\mathbf{x}(i)),$$

where

$$\mathcal{D}_i = \{\mathbf{y} : u(\mathbf{x}(i), \mathbf{y}) < u(\mathbf{x}(j), \mathbf{y}), \quad \forall j \neq i\}.$$

Given a set $B_n \subset \mathcal{X}^n$ and a rate R, we can consider a random codebook whose 2^{nR} codewords are drawn independently, each according to a uniform distribution over the set $B_n \subset \mathcal{X}^n$. We shall refer to the set B_n as the *input set*, and denote its cardinality by $|B_n|$. The average (over messages) probability of error for this random codebook when used over the channel θ and when decoded using the maximum-likelihood rule is a random variable, and we denote its expected value by $\bar{P}_{\theta,\text{ML}}(\text{error})$. Thus,

$$\bar{P}_{\theta,\text{ML}}(\text{error}) = |B_n|^{-(2^{nR})} \sum_{\mathcal{C}} P_{\theta,\theta}(\text{error}|\mathcal{C}),$$

where the sum is over all the blocklength-n rate-R codebooks whose codewords are all in B_n. We similarly define

$$\bar{P}_{\theta,\mathrm{Z}}(\text{error}) = |B_n|^{-(2^{nR})} \sum_{\mathcal{C}} P_{\theta,\mathrm{Z}}(\text{error}|\mathcal{C}),$$

to be the average (over codebooks and messages) probability of error that is incurred over the channel θ when a random codebook whose codewords are drawn independently and uniformly over the set B_n is decoded using Ziv's decoder.

To state our results we shall need one more technical term. We shall say that the input set B_n is *permutation invariant* if B_n is closed under permutations, i.e., if

$$(x_1, \dots, x_n) \in B_n,$$

implies

$$(x_{\pi(1)}, \dots, x_{\pi(n)}) \in B_n,$$

for any permutation π on $\{1, \dots, n\}$. The most interesting case is where the set B_n is the set of all n-length sequences of a given type (composition), i.e., when B_n is the smallest permutation invariant set that contains some sequence $\mathbf{x}$.

In [1] Ziv proved that if the sets B_n are permutation invariant then, to use the terminology of [2], Ziv's decoder is "strongly random coding universal" for the family of finite state channels with deterministic transitions, i.e.,

$$\lim_{n\to\infty} \sup_{\theta_d \in \Theta_d} \frac{1}{n} \log \frac{\bar{P}_{\theta_d,\mathrm{Z}}(\text{error})}{\bar{P}_{\theta_d,\mathrm{ML}}(\text{error})} = 0.$$

In this paper we shall strengthen this result in two ways. First, we shall show that this result also holds for the larger family of all finite state channels, and not only for those with deterministic transitions. In addition we shall demonstrate a deterministic coding result that is referred to as "strong deterministic coding universality" in [2]. We shall thus establish the following theorem:

THEOREM 1. *Let Θ denote the set of all pairs of initial states and transition laws of finite state channels defined over common finite input alphabet $\mathcal{X}$, finite state alphabet $\mathcal{S}$ and finite output alphabet $\mathcal{Y}$. Let $\bar{P}_{\theta,ML}$(error) denote the average (over messages and codebooks) probability of error incurred* *over the channel θ when the maximum-likelihood decoding rule is used to decode a random rate-R blocklength-n codebook whose codewords are drawn independently and uniformly over some permutation invariant set $B_n \subset \mathcal{X}^n$. Similarly let $\bar{P}_{\theta,z}(\text{error})$ denote the analogous performance of Ziv's decoder, i.e., the average (over messages and codebooks) probability of error incurred over the channel θ by Ziv's decoder. Finally, for a specific (deterministic) codebook $\mathcal{C}$ let $P_{\theta,z}(\text{error}|\mathcal{C})$ denote the average (over messages) probability of error incurred by Ziv's decoder when the codebook $\mathcal{C}$ is used over the channel θ. Then*

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \frac{1}{n} \log \frac{\bar{P}_{\theta,z}(\text{error})}{\bar{P}_{\theta,ML}(\text{error})} = 0, \tag{13}$$

and there exists a sequence of rate-R blocklength n codebooks $\mathcal{C}_n \subset B_n$ such that

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \frac{1}{n} \log \frac{P_{\theta,z}(\text{error}|\mathcal{C}_n)}{\bar{P}_{\theta,ML}(\text{error})} = 0. \tag{14}$$

2 Proof of Theorem

First note that (14), which is referred to as "strong deterministic coding universality" in [2], follows from the "strong random coding universality" (13) since the family of all finite state channels defined over common finite input, output, and state alphabets $\mathcal{X}, \mathcal{Y}, \mathcal{S}$ is strongly separable; see [2] and particularly Lemma 6 and Lemma 12 there. It thus suffices to prove (13), i.e., the random coding strong universality of Ziv's algorithm.

Rather than comparing the performance of Ziv's decoder and the maximum-likelihood decoder directly, we shall find it easier to demonstrate that each of these decoders performs very similarly to a third decoder, a "threshold decoder", and hence infer that they must have similar performance, thus establishing the universality of Ziv's decoder.

A threshold decoder for the channel $\theta \in \Theta$ with threshold sequence $\alpha_n \geq 1$ is a decoder that given the received sequence $\mathbf{y}$ declares that codeword $\mathbf{x}(i)$ was transmitted only if

$$p_\theta(\mathbf{y}|\mathbf{x}(i)) \geq \alpha_n p_\theta(\mathbf{y}|\mathbf{x}(j)), \quad \forall j \neq i, \tag{15}$$

declaring an error if no such codeword exists. Notice that the threshold decoder is not universal since its implementation requires knowing the channel law. Also, by its definition, it is, in general, inferior to the maximum-likelihood decoder.

The first step in the proof is to demonstrate that even though the threshold decoder is in general inferior to the maximum-likelihood, if the sequence of thresholds α_n is sub-exponential, i.e., satisfies

$$\lim_{n\to\infty} \frac{1}{n} \log \alpha_n = 0, \tag{16}$$

then

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \frac{1}{n} \log \frac{\bar{P}_{\theta,\mathrm{Th}}(\mathrm{error})}{\bar{P}_{\theta,\mathrm{ML}}(\mathrm{error})} = 0, \tag{17}$$

where $\bar{P}_{\theta,\mathrm{Th}}(\mathrm{error})$ is the average (over messages and codebooks) performance of the threshold decoder over the channel θ. This claim follows immediately from Lemma 2 but before we can state and prove this lemma, we need to introduce ranking functions [1], [2] and explain how they relate to the average (over messages and codebooks) probability of error of the various decoders.

Consider first the maximum-likelihood decoder. Condition (12) does not specify uniquely the maximum-likelihood decoding rule, because it does not specify how ties in the likelihood should be resolved. Any deterministic way of resolving such ties will result, however, in the same performance. To be more specific we thus assume that the maximum-likelihood decoding rule is based on a ranking function $M_\theta(\mathbf{x}, \mathbf{y})$. This function from $B_n \times \mathcal{Y}^n$ onto $\{1, \ldots, |B_n|\}$ is assumed to satisfy that for any $\mathbf{y} \in \mathcal{Y}^n$ the function $M_\theta(\cdot, \mathbf{y})$ is one-to-one from B_n onto $\{1, \ldots, |B_n|\}$ and

$$p_\theta(\mathbf{y}|\mathbf{x}) > p_\theta(\mathbf{y}|\mathbf{x}') \Rightarrow M_\theta(\mathbf{x}, \mathbf{y}) < M_\theta(\mathbf{x}', \mathbf{y}).$$

The function $M_\theta(\mathbf{x}, \mathbf{y})$ thus ranks the sequences in B_n according to the likelihood score. The maximum-likelihood rule based on the ranking function $M_\theta(\mathbf{x}, \mathbf{y})$ is defined as the decoding rule that given the received sequence $\mathbf{y}$ declares that the transmitted codewords was $\mathbf{x}(i)$ only if

$$M_\theta(\mathbf{x}(i), \mathbf{y}) < M_\theta(\mathbf{x}(j), \mathbf{y}), \quad \forall j \neq i.$$

If no such codeword exists, as can only be the case if there are two identical codewords in the codebook, an error is declared.

Denoting by $\bar{P}_{\theta,\theta}(\text{error}|\mathbf{x},\mathbf{y})$ the conditional probability of error over the channel θ using a maximum-likelihood rule based on the ranking function $M_\theta(\cdot,\cdot)$ given the correct codeword $\mathbf{x}$ and the received sequence $\mathbf{y}$, we have

$$\bar{P}_{\theta,\mathrm{ML}}(\text{error}) = \sum_{\mathbf{x}\in B_n}\sum_{\mathbf{y}\in\mathcal{Y}^n} \frac{1}{|B_n|} p_\theta(\mathbf{y}|\mathbf{x}) \bar{P}_{\theta,\theta}(\text{error}|\mathbf{x},\mathbf{y}),$$

and

$$\bar{P}_{\theta,\theta}(\text{error}|\mathbf{x},\mathbf{y}) = 1 - \left(1 - \frac{M_\theta(\mathbf{x},\mathbf{y})}{|B_n|}\right)^{2^{nR}-1}. \tag{18}$$

The last relation follows as in [1] by noting that, given the correct codeword $\mathbf{x}$ and the received sequence $\mathbf{y}$, the decoder decodes correctly if and only if all other codewords are ranked lower than $M_\theta(\mathbf{x},\mathbf{y})$, and by noting that the incorrect codewords are drawn independently and uniformly over B_n.

In a similar way we express the probability of error of Ziv's decoder as

$$\bar{P}_{\theta,\mathrm{Z}}(\text{error}) = \sum_{\mathbf{x}\in B_n}\sum_{\mathbf{y}\in\mathcal{Y}^n} \frac{1}{|B_n|} p_\theta(\mathbf{y}|\mathbf{x}) \bar{P}_{\theta,\mathrm{Z}}(\text{error}|\mathbf{x},\mathbf{y}),$$

where

$$\bar{P}_{\theta,\mathrm{Z}}(\text{error}|\mathbf{x},\mathbf{y}) = 1 - \left(1 - \frac{M_{\mathrm{Z}}(\mathbf{x},\mathbf{y})}{|B_n|}\right)^{2^{nR}-1},$$

and where

$$M_{\mathrm{Z}}(\mathbf{x},\mathbf{y}) = |\{\mathbf{x}' : u(\mathbf{x}',\mathbf{y}) \le u(\mathbf{x},\mathbf{y})\}|.$$

The following technical lemma will be useful in relating ranking functions and decoders performance.

LEMMA 1. *The following inequalities hold:*

1. *The function*

$$f(z) = 1-(1-z)^N \quad 0 \le z \le 1,$$

satisfies

$$\frac{f(s)}{f(t)} \le \max\left\{1, \frac{s}{t}\right\}, \quad \forall s,t \in (0,1].$$

2. *If $\{a_l\}_{l=1}^L$ and $\{b_l\}_{l=1}^L$ are two non-negative sequences then*

$$\frac{a_1 + \cdots + a_L}{b_1 + \cdots + b_L} \leq \max_{1 \leq l \leq L} \frac{a_l}{b_l}, \tag{19}$$

where $a/0 = \infty$ for $a > 0$, and $0/0 = 1$.

3. *If U and V are non-negative random variables then*

$$E[U] \leq E[V] \max \frac{U}{V},$$

where $a/0 = \infty$, unless $a = 0$ in which case $0/0 = 1$.

Proof. For a proof of this lemma see [2, Lemma 2]. □

We are now in a position to state and prove Lemma 2, which implies (17).

LEMMA 2. *Let $\bar{P}_{\theta,ML}(\text{error})$ denote the average probability of error incurred by a maximum-likelihood decoder over the channel $p_\theta(\mathbf{y}|\mathbf{x})$ using a random codebook consisting of $N + 1$ codewords that are drawn independently and uniformly over a set $B_n \subset \mathcal{X}^n$. Let $\bar{P}_{\theta,Th}(\text{error})$ denote the analogous expression for a threshold decoder with threshold $\alpha > 1$, i.e., a decoder that declares that the transmitted codeword is $\mathbf{x}$ only if*

$$p_\theta(\mathbf{y}|\mathbf{x}) \geq \alpha p_\theta(\mathbf{y}|\mathbf{x}'),$$

for every codeword $\mathbf{x}' \neq \mathbf{x}$, and declares an error if no such codeword exists. Then

$$\bar{P}_{\theta,Th}(\text{error}) \leq \alpha \ln\bigl(e^2 |B_n|\bigr) \bar{P}_{\theta,ML}(\text{error}).$$

Proof. Fix some received sequence $\mathbf{y}$, and let

$$p_\theta(\mathbf{x}|\mathbf{y}) = \frac{p_\theta(\mathbf{y}|\mathbf{x})}{\sum_{\mathbf{x}' \in B_n} p_\theta(\mathbf{y}|\mathbf{x}')},$$

be the conditional distribution on $\mathbf{x}$ given the received sequence $\mathbf{y}$ for the channel $p_\theta(\mathbf{y}|\mathbf{x})$ assuming that $\mathbf{x}$ is a priori uniformly distributed over B_n. To avoid cumbersome notation we shall omit the dependence of quantities on the received sequence $\mathbf{y}$ and the channel θ. We shall thus denote $p_\theta(\mathbf{x}|\mathbf{y})$ by $p(\mathbf{x})$, and denote $M_\theta(\mathbf{x}, \mathbf{y})$ by $M(\mathbf{x})$.

Given $\mathbf{y}$, the conditional probability of error of the threshold decoder $\bar{P}_{\theta,\mathrm{Th}}(\text{error}|\mathbf{y})$ is given by

$$\bar{P}_{\theta,\mathrm{Th}}(\text{error}|\mathbf{y}) = \sum_{\mathbf{x}\in B_n} p(\mathbf{x}) \left[1 - \left(1 - \frac{M(\mathbf{x}) + L(\mathbf{x})}{|B_n|}\right)^N\right], \tag{20}$$

where

$$L(\mathbf{x}) = \left|\{\mathbf{x}' : M(\mathbf{x}') > M(\mathbf{x}), p(\mathbf{x}') \geq \alpha^{-1} p(\mathbf{x})\}\right| \tag{21}$$

is the number of sequences in B_n that, given that $\mathbf{x}$ is the correct codeword and $\mathbf{y}$ was received, would cause an error in the threshold decoder if they were drawn as codewords, but would not cause an error in the maximum-likelihood decoder. Note that here too we have made the dependence on the channel law and the received sequence implicit.

Let

$$r(\mathbf{x}) = \sum_{\mathbf{x}':M(\mathbf{x}')<M(\mathbf{x})} p(\mathbf{x}'), \tag{22}$$

omitting, once again, the dependence on the channel law and on the received sequence $\mathbf{y}$. We can upper bound $L(\mathbf{x})$ in terms of $r(\mathbf{x})$ by

$$L(\mathbf{x}) \leq \alpha \frac{1 - r(\mathbf{x})}{p(\mathbf{x})}, \tag{23}$$

by noting that

$$\begin{aligned} 1 &= \sum_{\mathbf{x}'\in B_n} p(\mathbf{x}') \\ &\geq \sum_{\mathbf{x}':M(\mathbf{x}')\leq M(\mathbf{x})+L(\mathbf{x})} p(\mathbf{x}') \\ &\geq r(\mathbf{x}) + p(\mathbf{x}) + L(\mathbf{x})\alpha^{-1}p(\mathbf{x}) \\ &\geq r(\mathbf{x}) + L(\mathbf{x})\alpha^{-1}p(\mathbf{x}), \end{aligned}$$

where the second inequality follows from (21).

It follows from Lemma 1 that

$$\left[1 - \left(1 - \frac{M(\mathbf{x}) + L(\mathbf{x})}{|B_n|}\right)^N\right] \leq \frac{M(\mathbf{x}) + L(\mathbf{x})}{M(\mathbf{x})} \left[1 - \left(1 - \frac{M(\mathbf{x})}{|B_n|}\right)^N\right], \tag{24}$$

and we are now in a position to compare the conditional (on the received sequence $\mathbf{y}$) probability of error incurred by the maximum-likelihood decoder to that incurred by the threshold decoder as follows.

$$
\begin{aligned}
&\bar{P}_{\theta,\mathrm{Th}}(\mathrm{error}|\mathbf{y}) - \bar{P}_{\theta,\theta}(\mathrm{error}|\mathbf{y}) \\
&\qquad \overset{(a)}{\leq} \sum_{\mathbf{x}\in B_n} p(\mathbf{x}) \frac{L(\mathbf{x})}{M(\mathbf{x})} \left[1 - \left(1 - \frac{M(\mathbf{x})}{|B_n|}\right)^N\right] \\
&\qquad \overset{(b)}{\leq} \alpha \sum_{\mathbf{x}\in B_n} (1 - r(\mathbf{x})) \frac{1}{M(\mathbf{x})} \left[1 - \left(1 - \frac{M(\mathbf{x})}{|B_n|}\right)^N\right] \\
&\qquad \overset{(c)}{=} \alpha \sum_{\mathbf{x}\in B_n} \left(\sum_{\mathbf{x}':M(\mathbf{x}')\geq M(\mathbf{x})} p(\mathbf{x}') \right) \frac{1}{M(\mathbf{x})} \left[1 - \left(1 - \frac{M(\mathbf{x})}{|B_n|}\right)^N\right] \\
&\qquad \overset{(d)}{=} \alpha \sum_{\mathbf{x}'\in B_n} \sum_{\mathbf{x}:M(\mathbf{x})\leq M(\mathbf{x}')} p(\mathbf{x}') \frac{1}{M(\mathbf{x})} \left[1 - \left(1 - \frac{M(\mathbf{x})}{|B_n|}\right)^N\right] \\
&\qquad \overset{(e)}{\leq} \alpha \sum_{\mathbf{x}'\in B_n} \sum_{\mathbf{x}:M(\mathbf{x})\leq M(\mathbf{x}')} p(\mathbf{x}') \frac{1}{M(\mathbf{x})} \left[1 - \left(1 - \frac{M(\mathbf{x}')}{|B_n|}\right)^N\right] \\
&\qquad \overset{(f)}{\leq} \alpha \left(\sum_{i=1}^{|B_n|} \frac{1}{i} \right) \sum_{\mathbf{x}'} p(\mathbf{x}') \left[1 - \left(1 - \frac{M(\mathbf{x}')}{|B_n|}\right)^N\right] \\
&\qquad \overset{(g)}{\leq} \alpha(\ln(|B_n|) + 1) \bar{P}_{\theta,\theta}(\mathrm{error}|\mathbf{y}),
\end{aligned}
$$

where (a) follows from (18), (20), and (24); (b) follows from (23); (c) follows from the definition of $r(\mathbf{x})$ (see (22)); (d) follows by interchanging the order of summation; (e) from the monotonicity of the function $f(z) = 1 - (1-z)^N$; (f) follows by increasing the range over which $\mathbf{x}$ is summed, and (g) follows by a simple bound on the harmonic sum. It follows from this calculation that

$$\bar{P}_{\theta,\theta}(\mathrm{error}|\mathbf{y}) \leq \Big(1 + \alpha(\ln(|B_n|) + 1)\Big) \bar{P}_{\theta,\theta}(\mathrm{error}|\mathbf{y}),$$

and the lemma now follows by noting that $\alpha \geq 1$ and by taking expectation with respect to $\mathbf{y}$.

Notice that the term $\ln |B_n|$ is at most linear in the blocklength n since $B_n \subseteq \mathcal{X}^n$ and thus,

$$\ln |B_n| \leq \ln |\mathcal{X}^n| = n \ln |\mathcal{X}|.$$

□

We now turn to the second part of the theorem's proof and establish that there exists a sub-exponential threshold sequence α_n such that Ziv's decoder competes favorably with the threshold decoder based on α_n, i.e.,

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta} \frac{1}{n} \log \frac{\bar{P}_{\theta,\mathrm{z}}(\mathrm{error})}{\bar{P}_{\theta,\mathrm{Th}}(\mathrm{error})} \leq 0. \tag{25}$$

This, combined with (17) and the optimality of the maximum-likelihood rule will conclude the proof of the theorem.

Given some sequence of thresholds α_n and a law $p_\theta(\cdot|\cdot)$, $\theta \in \Theta$ we define $\mathcal{N}_\theta(\mathbf{x}, \mathbf{y})$ to be the set of all sequences $\mathbf{x}' \in \mathcal{X}^n$ that are permutations of $\mathbf{x}$ and that satisfy

$$p_\theta(\mathbf{y}|\mathbf{x}') \geq \alpha_n^{-1} p_\theta(\mathbf{y}|\mathbf{x}).$$

The set $\mathcal{N}_\theta(\mathbf{x}, \mathbf{y})$ depends, of course, on the sequence of thresholds α_n, but this dependence is not made explicit in our notation. Notice that if $\mathbf{x}$ is in B_n then $\mathcal{N}_\theta(\mathbf{x}, \mathbf{y}) \subset B_n$ since, by assumption, B_n is permutation invariant. Also note that

$$|\mathcal{N}_\theta(\mathbf{x}, \mathbf{y})| \leq M_\theta(\mathbf{x}, \mathbf{y}) + L(\mathbf{x}), \tag{26}$$

where $L(\mathbf{x})$ is defined in (21), and that this inclusion could be strict since we insist that the sequences in $\mathcal{N}_\theta(\mathbf{x}, \mathbf{y})$ be permutations of $\mathbf{x}$. We now have

$$\begin{aligned}
\frac{\bar{P}_{\theta,\mathrm{z}}(\mathrm{error})}{\bar{P}_{\theta,\mathrm{Th}}(\mathrm{error})} &= \frac{\sum_{\mathbf{x}\in B_n} \sum_{\mathbf{y}\in\mathcal{Y}^n} \frac{1}{|B_n|} \bar{P}_{\theta,\mathrm{z}}(\mathrm{error}|\mathbf{x}, \mathbf{y})}{\sum_{\mathbf{x}\in B_n} \sum_{\mathbf{y}\in\mathcal{Y}^n} \frac{1}{|B_n|} \bar{P}_{\theta,\mathrm{Th}}(\mathrm{error}|\mathbf{x}, \mathbf{y})} \\
&\leq \max_{\mathbf{x}\in B_n, \mathbf{y}\in\mathcal{Y}^n} \frac{\bar{P}_{\theta,\mathrm{z}}(\mathrm{error}|\mathbf{x}, \mathbf{y})}{\bar{P}_{\theta,\mathrm{Th}}(\mathrm{error}|\mathbf{x}, \mathbf{y})} \\
&\leq 1 + \max_{\mathbf{x}\in B_n, \mathbf{y}\in\mathcal{Y}^n} \frac{M_\mathrm{z}(\mathbf{x}, \mathbf{y})}{M_\theta(\mathbf{x}, \mathbf{y}) + L(\mathbf{x})} \\
&\leq 1 + \max_{\mathbf{x}\in B_n, \mathbf{y}\in\mathcal{Y}^n} \frac{M_\mathrm{z}(\mathbf{x}, \mathbf{y})}{|\mathcal{N}_\theta(\mathbf{x}, \mathbf{y})|}
\end{aligned} \tag{27}$$

where the first inequality follows from part 3 of Lemma 1, the second inequality follows from part 1 of Lemma 1, and the last inequality follows from (26). It follows that to prove (25) we need to upper bound $M_\mathrm{z}(\mathbf{x}, \mathbf{y})$ and to find a sub-exponential sequences of thresholds that will yield a good lower bound on $\mathcal{N}_\theta(\mathbf{x}, \mathbf{y})$. The following lemma is used to upper bound $M_\mathrm{z}(\mathbf{x}, \mathbf{y})$.

LEMMA 3. *Given two finite alphabets $\mathcal{X}, \mathcal{Y}$, and any sequence $\mathbf{y} \in \mathcal{Y}^n$,*

$$\frac{1}{n} \log |\{\mathbf{x}' \in \mathcal{X}^n : u(\mathbf{x}', \mathbf{y}) \leq D\}| \leq D + O\left(\frac{\log\log n}{\log n}\right), \tag{28}$$

where the correction term depends only the alphabet sizes $|\mathcal{X}|, |\mathcal{Y}|$, and the function $u(\cdot, \cdot)$ is defined (9). In particular,

$$\begin{aligned}\frac{1}{n} \log M_z(\mathbf{x}, \mathbf{y}) &= \frac{1}{n} \log |\{\mathbf{x}' \in B_n : u(\mathbf{x}', \mathbf{y}) \leq u(\mathbf{x}, \mathbf{y})\}| \\ &\leq \frac{1}{n} \log |\{\mathbf{x}' \in \mathcal{X}^n : u(\mathbf{x}', \mathbf{y}) \leq u(\mathbf{x}, \mathbf{y})\}| \\ &\leq u(\mathbf{x}, \mathbf{y}) + O\left(\frac{\log\log n}{\log n}\right). \end{aligned} \tag{29}$$

Proof. This inequality has nothing to do with the channel model, and is a property of strings. This lemma appears in [1] where a proof can also be found. □

We now turn to lower bounding $|\mathcal{N}_\theta(\mathbf{x}, \mathbf{y})|$. To this end we choose the sequence of thresholds to be

$$\alpha_n = |\mathcal{S}|^{\gamma(n, |\mathcal{X}||\mathcal{Y}|)}, \tag{30}$$

where $|\mathcal{S}|$ is the number of states and $\gamma(n, |\mathcal{X}||\mathcal{Y}|)$ is the maximum number of phrases that can be produced when an n-length sequence over $\mathcal{X} \times \mathcal{Y}$ is incrementally parsed. Notice that α_n is sub-exponential as

$$\gamma(n, |\mathcal{X}||\mathcal{Y}|) = O\left(\frac{n}{\log n}\right), \tag{31}$$

see [6]. For this choice of the thresholds we can now lower bound $|\mathcal{N}_\theta(\mathbf{x}, \mathbf{y})|$ as follows.

LEMMA 4. *For the sequence of thresholds given in (30) and any sequence $\mathbf{x} \in \mathcal{X}^n$ and $\mathbf{y} \in \mathcal{Y}^n$*

$$\frac{1}{n} \log(|\mathcal{N}_\theta(\mathbf{x}, \mathbf{y})|) \geq u(\mathbf{x}, \mathbf{y}) - O\left(\frac{1}{\log n}\right) \log |\mathcal{S}|^2, \tag{32}$$

where the correction term depends only on the cardinalities of the input and output alphabets and not on the channel law or initial state.

Proof. Fix some $\theta \in \Theta$ corresponding to a transition law $P(y, s'|x, s)$ and some initial state s_0. Fix also some $\mathbf{x} \in B_n$ and $\mathbf{y} \in \mathcal{Y}^n$. Let g be the number of distinct phrases that result from the joint parsing of $(\mathbf{x}, \mathbf{y})$ and recall that this joint parsing induces a parsing on $\mathbf{x}$ and $\mathbf{y}$ as in (4), (6). For any g-length sequence $\{\tilde{s}(m)\}_{m=1}^{g} \subset \mathcal{S}^g$ we define

$$\tilde{S} = \{\mathbf{s} = (s_1, \ldots, s_n) \in \mathcal{S}^n : s_{e(m)} = \tilde{s}(m)\},$$

and define

$$p(\mathbf{y}, \tilde{s}|\mathbf{x}, s_0) = \sum_{\mathbf{s} \in \tilde{S}} P_n(\mathbf{y}, \mathbf{s}|\mathbf{x}, s_0). \tag{33}$$

Given the input sequence $\mathbf{x}$ and the initial state s_0, the quantity $p(\mathbf{y}, \tilde{s}|\mathbf{x}, s_0)$ is thus the conditional probability of the channel producing the output sequence $\mathbf{y}$ while following a state trajectory that coincides with $\tilde{s}$ at the sampling times $e(1), \ldots, e(g)$ (the sampling times corresponding to the endings of the different phrases). It now follows from (10) and (33) that

$$p_\theta(\mathbf{y}|\mathbf{x}) = \sum_{\tilde{s} \in \mathcal{S}^g} p(\mathbf{y}, \tilde{s}|\mathbf{x}, s_0),$$

and hence that there exists a sequence $\{\sigma(m)\}_{m=1}^{g} \subset \mathcal{S}^g$ such that

$$p(\mathbf{y}, \sigma|\mathbf{x}, s_0) \geq \frac{1}{|\mathcal{S}|^g} p_\theta(\mathbf{y}|\mathbf{x}). \tag{34}$$

This choice of the sequence σ depends, of course, on the received sequence $\mathbf{y}$, the correct codeword $\mathbf{x}$, the initial state s_0, and the transition law $P(y_k s_k|x_k s_{k-1})$. Henceforth the sequence σ will be held fixed.

Since we have defined $\gamma(n, |\mathcal{X}||\mathcal{Y}|)$ to be the maximal number of phrases into which an n-length sequence over $\mathcal{X} \times \mathcal{Y}$ can be incrementally parsed, it follows that $g \leq \gamma(n, |\mathcal{X}||\mathcal{Y}|)$, and thus, by (34),

$$p(\mathbf{y}, \sigma|\mathbf{x}, s_0) \geq \alpha_n^{-1} p_\theta(\mathbf{y}|\mathbf{x}). \tag{35}$$

We next show that if $\mathbf{x}'$ is a permutation of $\mathbf{x}$ with some properties that will be discussed shortly, then

$$p(\mathbf{y}, \sigma|\mathbf{x}', s_0) = p(\mathbf{y}, \sigma|\mathbf{x}, s_0), \tag{36}$$

and thus $\mathbf{x}' \in \mathcal{N}_\theta(\mathbf{x},\mathbf{y})$ because

$$\begin{aligned} p_\theta(\mathbf{y}|\mathbf{x}') &= \sum_{\bar{s}\in\mathcal{S}^g} p(\mathbf{y},\bar{s}|\mathbf{x}',s_0) \\ &\geq p(\mathbf{y},\sigma|\mathbf{x}',s_0) \\ &= p(\mathbf{y},\sigma|\mathbf{x},s_0) \\ &\geq \alpha_n^{-1} p(\mathbf{y}|\mathbf{x},s_0), \end{aligned}$$

where the second equality follows from (36), and the last inequality follows from (35). The proof of the lemma will be then concluded by counting the number of permutations of $\mathbf{x}$ that satisfy (36).

We shall find it convenient to define $\sigma(0) = s_0$ and to note that

$$p(\mathbf{y},\sigma|\mathbf{x},s_0) = \prod_{m=1}^{g} p(\mathbf{y}^{(m)},\sigma(m)|\mathbf{x}^{(m)},\sigma(m-1)) \tag{37}$$

where $p(\mathbf{y}^{(m)},\sigma(m)|\mathbf{x}^{(m)},\sigma(m-1))$ is the probability that the channel will be at time $e(m)-b(m)+1$ at state $\sigma(m)$ and produce the output $\mathbf{y}^{(m)}$ given that it starts at state $\sigma(m-1)$ and is fed with the input $\mathbf{x}^{(m)}$. Thus

$$p(\mathbf{y}^{(m)},\sigma(m)|\mathbf{x}^{(m)},\sigma(m-1)) = \sum_{\substack{\bar{s}_0,\ldots,\bar{s}_{l(m)} \\ \bar{s}_0=\sigma(m-1) \\ \bar{s}_{l(m)}=\sigma(m)}} \prod_{i=1}^{l(m)} P(y_{b(m)+i-1},\bar{s}_i|x_{b(m)+i-1},\bar{s}_{i-1}),$$

where

$$l(m) = e(m) - b(m) + 1,$$

is the length of the m-th phrase. Suppose now that $1 \leq m < m' \leq g$ are such that:

- $l(m) = l(m')$
- $\mathbf{y}^{(m)} = \mathbf{y}^{(m')}$
- $\sigma(m-1) = \sigma(m'-1)$
- $\sigma(m) = \sigma(m')$.

It follows from (37) that if $\mathbf{x}'$ is produced from $\mathbf{x}$ by exchanging $\mathbf{x}^{(m)}$ with $\mathbf{x}^{(m')}$, i.e,

$$\mathbf{x}' = \mathbf{x}^{(1)} \dots \mathbf{x}^{(m-1)}\mathbf{x}^{(m')}\mathbf{x}^{(m+1)} \dots \mathbf{x}^{(m'-1)}\mathbf{x}^{(m)}\mathbf{x}^{(m'+1)} \dots \mathbf{x}^{(g)},$$

then (36) holds. It thus remains to count how many permutations of $\mathbf{x}$ can be produced with such transpositions. This counting argument is identical to the one appearing in [1] in the proof of Lemma 1. For the sake of completeness we repeat it here.

For any $s, s' \in \mathcal{S}$ and for any $\mathbf{y}$-phrase $y(l)$ we set $c_l(\mathbf{x}|\mathbf{y}, s, s')$ to be the number of distinct $\mathbf{x}$-phrases that appear jointly with $y(l)$, that end in state s, and such that the phrase preceding them ends at state s'. Thus

$$c_l(\mathbf{x}|\mathbf{y}, s, s') = \left|\{\mathbf{x}^{(m)}\ 1 \le m \le g\ :\ \mathbf{y}^{(m)} = y(l), \sigma(m) = s, \sigma(m-1) = s'\}\right|,$$

and

$$\sum_{(s,s')\in\mathcal{S}^2} c_l(\mathbf{x}|\mathbf{y}, s, s') = c_l(\mathbf{x}|\mathbf{y}) \quad . \tag{38}$$

Recalling that $c(\mathbf{y})$ denotes the number of distinct $\mathbf{y}$ phrases we have

$$|\mathcal{N}_\theta(\mathbf{x}, \mathbf{y})| \ge \prod_{l=1}^{c(\mathbf{y})} \prod_{s,s'} c_l(\mathbf{x}|\mathbf{y}, s, s')! \quad .$$

Using the Sterling formula we have

$$\begin{aligned}\log |\mathcal{N}_\theta(\mathbf{x}, \mathbf{y})| &\ge \sum_{l=1}^{c(\mathbf{y})} \sum_{s,s'} c_l(\mathbf{x}|\mathbf{y}, s, s')\big(\log c_l(\mathbf{x}|\mathbf{y}, s, s') - \log e\big) \\ &= -\sum_{l=1}^{c(\mathbf{y})} c_l(\mathbf{x}|\mathbf{y}) \sum_{s,s'} \frac{c_l(\mathbf{x}|\mathbf{y}, s, s')}{c_l(\mathbf{x}|\mathbf{y})} \log \frac{c_l(\mathbf{x}|\mathbf{y}, s, s')}{c_l(\mathbf{x}|\mathbf{y})} \\ &+ \sum_{l=1}^{c(\mathbf{y})} \big(\log c_l(\mathbf{x}|\mathbf{y}) - \log e\big),\end{aligned}$$

where the last equality follows from (38). By the convexity of the logarithmic function, and using (8) we have

$$\begin{aligned}\log|\mathcal{N}_\theta(\mathbf{x},\mathbf{y})| &\geq \sum_{l=1}^{c(\mathbf{y})} c_l(\mathbf{x}|\mathbf{y})\big(\log c_l(\mathbf{x}|\mathbf{y}) - \log|\mathcal{S}|^2 - \log e\big) \\ &= \sum_{l=1}^{c(\mathbf{y})} c_l(\mathbf{x}|\mathbf{y})\log c_l(\mathbf{x}|\mathbf{y}) - c(\mathbf{x},\mathbf{y})\log(|\mathcal{S}|^2 e) \\ &\geq n\left[u(\mathbf{x},\mathbf{y}) - \frac{1}{n}O(\frac{n}{\log n})\log|\mathcal{S}|^2\right],\end{aligned}$$

where the last inequality follows from (31) by noting that $c(\mathbf{x},\mathbf{y}) < \gamma(n, |\mathcal{X}||\mathcal{Y}|)$, and by recalling the definition of the universal decoding function $u(\mathbf{x},\mathbf{y})$, see (9). □

Proof of Theorem 1: We now have all the ingredients needed to complete the proof of Theorem 1. Choose the sequence of thresholds as in (30) and note that by (31) this sequence is sub-exponential, i.e., satisfies (16). It follows, by Lemma 2 that the threshold decoder based on the sequence α_n and the maximum-likelihood decoder have the same asymptotic performance, i.e., that (17) holds. Comparing the performance of Ziv's decoder to that of the threshold decoder we use (27), (29), and (32) to deduce that Ziv's decoder and the threshold decoder perform similarly, i.e., that (25) holds. The theorem is now follows from (17) and (25).

3 Discussion

As Theorem 1 demonstrates, the Lempel-Ziv incremental parsing is a very powerful tool not only in universal source coding, but also in universal channel decoding. It is thus interesting to explore the duality between universal source coding and universal channel decoding, and to investigate whether every universal source code can be used to design a universal channel decoder. After all, any source code can be used to assign probabilities $\mathcal{P}(\mathbf{u})$ to source sequences $\mathbf{u}$ according to the length of the codewords assigned to them using the assignment

$$\mathcal{P}(\mathbf{u}) = 2^{-\mathcal{L}(\mathbf{u})}. \tag{39}$$

For a good universal source code this assignment should approximate the true probability of the sequence, and one could therefore use a universal source code as a channel decoder in the following way: Given a received sequence $\mathbf{y}$ and any candidate codeword $\mathbf{x}(i)$ one would estimate the true probability $p_\theta(\mathbf{x}(i), \mathbf{y})$ using the source code, and pick the codeword that maximizes this probability. Assuming that the universal probability assignment is close to the true probability, this rule should be somewhat similar to the maximum-likelihood rule tuned to the true channel, and there is hope for universality.

The example that will be described in this section demonstrates that the above approach does not always yield a universal decoder, even if the redundancy of the universal source code is small for every sequence [7] and not only on the average. The problem seems to be that while the source coding universality of a code guarantees a lower bound on the probability it assigns to sequences (in terms of the true probability) it does not always guarantee an upper-bound that is sharp enough for the purpose of efficient universal channel decoding.

The example also serves to show that even for finite families of channels $\mathcal{F}$, for which a universal decoder can always be constructed by merging the different maximum-likelihood decoders [2], the generalized maximum-likelihood rule may fail to be universal, where the generalized maximum-likelihood rule decides that message $\mathbf{x}(i)$ was transmitted only if

$$p_{\text{GL}}(\mathbf{y}|\mathbf{x}(i)) \geq p_{\text{GL}}(\mathbf{y}|\mathbf{x}(j)), \quad \forall j,$$

where

$$p_{\text{GL}}(\mathbf{y}|\mathbf{x}) = \frac{\sup_{\phi \in \mathcal{F}} p_\phi(\mathbf{y}|\mathbf{x})}{\sum_{\mathbf{y}' \in \mathcal{Y}^n} \sup_{\phi \in \mathcal{F}} p_\phi(\mathbf{y}'|\mathbf{x})}.$$

Note also that the generalized maximum-likelihood decoder *is* universal for the family of discrete memoryless channels [3] (for which it is equivalent to the maximum empirical mutual information decoder MMI), but, as the example demonstrates, not for all families, not even those with a finite number of channels.

Example: Consider a family of channels $\Theta = \{\theta_1, \theta_2\}$ consisting of two channels defined over the input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ where

$$\mathcal{X} = \mathcal{Y} = \{0, 1, 2, 3\}.$$

Under both channels the output sequence $\mathbf{y} = y_1, \ldots, y_n$ corresponding to the input sequence $\mathbf{x} = x_1, \ldots, x_n$ is given by

$$\mathbf{y} = \mathbf{x} + \mathbf{z} \bmod 4, \tag{40}$$

where $\mathbf{z}$ is a random noise sequence, and the mod four addition in (40) corresponds to componentwise addition, so that

$$y_i = x_i + z_i \bmod 4. \tag{41}$$

The two channels θ_1 and θ_2 differ in the probability law governing the noise sequence $\mathbf{z}$. To specify these laws consider three *disjoint* subsets of $\mathcal{X}^n$, say E, F, and G of cardinalities

$$|E| = 1, \tag{42}$$

$$|F| = \left\lceil \left(\frac{3}{2}\right)^n \right\rceil, \tag{43}$$

$$|G| = 3^n. \tag{44}$$

Under the channel θ_1 the law of $\mathbf{z}$ is given by

$$p_{\theta_1}(\mathbf{z}) = \begin{cases} q & \text{if } \mathbf{z} \in E \\ p & \text{if } \mathbf{z} \in F \\ 0 & \text{otherwise} \end{cases}, \tag{45}$$

where

$$q = 1 - 2^{-n}, \tag{46}$$

$$p = \frac{2^{-n}}{|F|}. \tag{47}$$

A direct calculation demonstrates that

$$\frac{3^{-n}}{1 + (2/3)^n} < p < 3^{-n} \tag{48}$$

which follows because by (43)

$$\left(\frac{3}{2}\right)^n < |F| < 1 + \left(\frac{3}{2}\right)^n. \tag{49}$$

Under the law θ_2 the noise sequences are distributed according to

$$p_{\theta_2}(\mathbf{z}) = \begin{cases} 3^{-n} & \text{if } \mathbf{z} \in G \\ 0 & \text{otherwise} \end{cases}. \tag{50}$$

The generalized likelihood of a sequence $\mathbf{z}$ is defined as

$$p_{\text{GL}}(\mathbf{z}) = \frac{\max_{\theta \in \{\theta_1, \theta_2\}} p_\theta(\mathbf{z})}{\sum_{\mathbf{z}'} \max_{\theta \in \{\theta_1, \theta_2\}} p_\theta(\mathbf{z}')}. \tag{51}$$

Notice that

$$p_{\text{GL}}(\mathbf{z}) \geq \frac{1}{2} p_{\theta_i}(\mathbf{z}), \quad i = 1, 2, \tag{52}$$

since the numerator in (51) is no smaller than $p_{\theta_i}(\mathbf{z})$, and the denominator can be upper bounded by noting that

$$\max_{\theta \in \{\theta_1, \theta_2\}} p_\theta(\mathbf{z}') \leq p_{\theta_1}(\mathbf{z}') + p_{\theta_2}(\mathbf{z}'),$$

and hence

$$\begin{aligned} \sum_{\mathbf{z}'} \max_{\theta \in \{\theta_1, \theta_2\}} p_\theta(\mathbf{z}') &\leq \sum_{i=1}^{2} \sum_{\mathbf{z}'} p_{\theta_i}(\mathbf{z}') \\ &= 2. \end{aligned}$$

Consider a random codebook consisting of two codewords, each drawn uniformly over the set

$$B_n = \mathcal{X}^n,$$

and consider the performance of this random codebook on the channel θ_1. First note that by the symmetry of the problem the average (over codebook and messages) probability of error, conditioned on the received sequence $\mathbf{y}$, does not depend on the received sequence, i.e., the performance of a random codebook is independent of the received sequence, and

$$\bar{P}_{\theta_1, \text{ML}}(\text{error}) = \bar{P}_{\theta_1, \theta_1}(\text{error}|\mathbf{y}), \quad \forall \mathbf{y} \in \mathcal{Y}^n,$$

so that we may assume without loss in generality that the all-zero sequence was received.

Given that $\mathbf{y}$ is all zero, the maximum-likelihood rule tuned to θ_1 ranks the sequence in E highest, followed by the sequences in F. We now have that

$$p_{\theta_1}(\mathbf{x}|\mathbf{y}=\mathbf{0}) = \begin{cases} q & \text{if } \mathbf{x} \in E \\ p & \text{if } \mathbf{x} \in F \end{cases}$$

and thus

$$\bar{P}_{\theta_1,\mathrm{ML}}(\text{error}) \leq q\frac{1}{|\mathcal{X}|^n} + (1-q)\frac{|E|+|F|}{|\mathcal{X}|^n},$$

which follows from the pessimistic assumption that if the correct codewords is not in E then it is ranked lowest by the maximum-likelihood decoder among all the sequences in F. Evaluating this expression we obtain that

$$-\lim_{n\to\infty}\frac{1}{n}\log\bar{P}_{\theta_1,\mathrm{ML}}(\text{error}) = \log(4). \tag{53}$$

(The equality follows by noting that the probability of error is always lower bounded by $|\mathcal{X}|^{-n}$ since this is the probability that the two codewords in the codebook are identical.)

Consider now the performance over the channel θ_1 of the generalized maximum-likelihood decoder. Once again we may assume without loss in generality that the all-zero sequence was received, and we note that the generalized maximum-likelihood ranks the sequence in E highest, followed by the sequences in G, followed finally by the sequences in F. Lower bounding the average probability of error by assuming that if the correct codeword is not in E then it is ranked highest among all the sequences in F we conclude that

$$\bar{P}_{\theta_1,\mathrm{GL}}(\text{error}) \geq q\frac{1}{|\mathcal{X}|^n} + (1-q)\frac{|E|+|G|+1}{|\mathcal{X}|^n},$$

which demonstrates that

$$-\liminf_{n\to\infty}\frac{1}{n}\log\bar{P}_{\theta_1,\mathrm{GL}}(\text{error}) \leq \log\left(\frac{8}{3}\right). \tag{54}$$

Comparing (53) and (54) demonstrates that for this example the generalized maximum likelihood decoder is not universal, even though, as a source code it has low redundancy *for every message*, as seen in (52); the redundancy is always upper bounded by $\log 2/n$. Moreover, the failure of the generalized maximum-likelihood rule is not due to the non-existence of a universal decoder for the family: since the family is finite, the universal decoder derived by merging the maximum-likelihood decoders corresponding to the two channels [2] is universal for the family.

Acknowledgments

Section 2 has greatly benefited from discussions with M. Feder to whom we are indebted. Discussions with S. Shamai and İ. E. Telatar are also gratefully acknowledged.

References

[1] J. Ziv, "Universal decoding for finite-state channels,," *IEEE Trans. on Inform. Theory.*, vol. 31, pp. 453–460, July 1985.

[2] M. Feder and A. Lapidoth, "Universal decoding for channels with memory," *Submitted.*

[3] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Academic Press, 1981.

[4] R. Gallager, *Information Theory and Reliable Communication.* John Wiley & Sons, 1968.

[5] M. Mushkin and I. Bar-David, "Capacity and coding for the Gilbert-Elliot channel," *IEEE Trans. on Inform. Theory*, vol. 35, pp. 1277–1290, Nov. 1989.

[6] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. on Inform. Theory*, vol. 24, pp. 530–536, Sept. 1978.

[7] Y. M. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, pp. 3–17, July-Sept. 1987.