# Universal Decoding Using a Noisy Codebook

Neri Merhav

The Andrew & Erna Viterbi Faculty of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, ISRAEL
E–mail: merhav@ee.technion.ac.il

**Abstract**

We consider the topic of universal decoding with a decoder that does not have direct access to the codebook, but only to noisy versions of the various randomly generated codewords, a problem motivated by biometrical identification systems. Both the source that generates the original (clean) codewords, and the channel that corrupts them in generating the noisy codewords, as well as the main channel for communicating the messages, are all modeled by non–unifilar, finite–state systems (hidden Markov models). As in previous works on universal decoding, here too, the average error probability of our proposed universal decoder is shown to be as small as that of the optimal maximum likelihood (ML) decoder, up to a multiplicative factor that is a sub–exponential function of the block length. It therefore has the same error exponent, whenever the ML decoder has a positive error exponent. The universal decoding metric is based on Lempel–Ziv (LZ) incremental parsing of each noisy codeword jointly with the given channel output vector, but this metric is somewhat different from the one proposed in earlier works on universal decoding for finite–state channels, by Ziv (1985) and by Lapidoth and Ziv (1998). The reason for the difference is that here, unlike in those earlier works, the probability distribution that governs the (noisy) codewords is, in general, not uniform across its support. This non–uniformity of the codeword distribution also makes our derivation more challenging. Another reason for the more challenging analysis is the fact that the effective induced channel between the noisy codeword of the transmitted message and the main channel output is not a finite–state channel in general.

**Index Terms:** Universal decoding, finite–state channel, hidden Markov model, Lempel–Ziv algorithm, error exponent.

# 1 Introduction

The topic of universal decoding under channel uncertainty has received considerable attention in the last four decades. In [9] the *maximum mutual information* (MMI) decoder was first proposed and shown to achieve the capacity for discrete memoryless channels (DMC's). Csiszár and Körner [3] showed that the random coding error exponent of the MMI decoder, associated with a uniform random coding distribution over a given type class, achieves the same random coding error exponent as the maximum likelihood (ML) decoder. Csiszár [2] proved that for any modulo–additive DMC and the uniform random coding distribution over linear codes, the optimum random coding error exponent is universally achieved by a decoder that minimizes the empirical entropy of the difference between the output sequence and the input sequence. In [13], a parallel result was obtained for a certain class of memoryless Gaussian channels with slow fading and an unknown interference signal.

For channels with memory, Ziv [20] considered universal decoding for unknown unifilar finite–state (FS) channels with finite input and output alphabets, i.e., FS channels for which at each time instant, the next channel state is given by an unknown deterministic function of the channel current state, input and output. For ensembles of codes governed by the uniform distribution over a given permutation–invariant set of channel input vectors (namely, a type class or the disjoint union of several type classes), he proved that a decoder based on the Lempel–Ziv (LZ) incremental parsing algorithm asymptotically achieves the same error exponent as the ML decoder. In [11], Lapidoth and Ziv proved that the same universal decoder continues to be universally asymptotically optimum even for the broader class of FS channels with stochastic, rather than deterministic, next–state transitions. They still assumed a random coding distribution which is uniform over a given permutation–invariant set. In [7], Feder and Lapidoth have furnished sufficient conditions for general families of channels with memory to have universal decoders that asymptotically achieve the random coding error exponent of ML decoding. In [8], a competitive minimax criterion was proposed, in the quest for a more general systematic approach to the problem of universal decoding. Two additional related works on general methodologies for universal decoding are those of [12] and [14].

This paper is a further development on [11] and [20]. In particular, here we consider universal decoding in a situation where the decoder that does not have direct access to the codebook of the encoder, but only to noisy versions of the various randomly generated codewords, a problem motivated by applications in biometrical identification systems (see, e.g., [10, Section 5], [17], [18], [19], and many references therein) or other applications where storage, or finite–precision limitations do not enable the decoder to save the exact codewords of all messages, and then they must be quantized and hence distorted. In our model, both the source that generates the original (clean) codewords, and the channel that corrupts them in the process of generating the noisy codewords, as well as the main channel for communicating the messages, are all modeled by non–unifilar, FS systems (hidden Markov models). As in the previous above–mentioned works on universal decoding, here too, the average error probability of our proposed universal decoder is shown to be as small as that of the optimal maximum likelihood (ML) decoder, up to a multiplicative factor that is a sub–exponential function of the block length, $n$. It therefore has the same error exponent, whenever the ML decoder has a positive error exponent. As in [11] and [20], the universal decoding metric is based on Lempel–Ziv (LZ) incremental parsing of each noisy codeword jointly with the given channel output vector, but this metric is somewhat different from that of [11] and [20]. Specifically, it includes an additional term, which is the logarithm of the induced probability of generating the noisy codeword of the message being tested. The reason for this difference is that here, unlike in [11] and [20], the probability distribution which governs the (noisy) codewords is, in general, not uniform across its support. This non–uniformity of the codeword distribution also makes our derivation quite more challenging. Another factor that makes the analysis here more involved is the fact that the effective induced channel between the noisy codeword of the transmitted message and the main channel output is not a FS channel in general.

The outline of the rest of the paper is as follows. In Section 2, we establish the notation conventions, define the problem formally, and spell out the assumptions. Section 3 is devoted to the statement of the main result and a discussion. Finally, in Section 4 the main results is proved.

## 2 Notation Conventions, Problem Formulation and Assumptions

### 2.1 Notation Conventions

Throughout the paper, random variables will be denoted by capital letters, specific values they may take will be denoted by the corresponding lower case letters, and their alphabets will be denoted by calligraphic letters. Random vectors and their realizations will be denoted, respectively, by capital letters and the corresponding lower case letters, both in the bold face font. Their alphabets will be superscripted by their dimensions. For example, the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$, ($n$ – positive integer) may take a specific vector value $\boldsymbol{x} = (x_1, \ldots, x_n)$ in $\mathcal{X}^n$, the $n$–th order Cartesian power of $\mathcal{X}$, which is the alphabet of each component of this vector. The probability of an event $\mathcal{E}$ (with respect to) w.r.t. a probability measure $P$ will be denoted by $P[\mathcal{E}]$, and the expectation operator w.r.t. $P$ will be denoted by $\boldsymbol{E}_P\{\cdot\}$. The subscript will be omitted if the underlying probability distribution is clear from the context. Logarithms and exponents will be defined w.r.t. the natural basis $e$, unless specified otherwise. In particular, $\exp_2(t)$ will sometimes be used to denote $2^t$. The cardinality of a finite set, say, $\mathcal{A}$, will be denoted by $|\mathcal{A}|$.

### 2.2 Problem Formulation and Assumptions

Consider a coded communication system, defined as follows. First, a rate–$R$ block code of length $n$, $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\}$, $M = e^{nR}$, is selected at random, where each $\boldsymbol{x}_m \in \mathcal{X}^n$, $m = 1, 2, \ldots, M$, is drawn independently under a distribution $G(\boldsymbol{x})$. A message $m$ is selected under the uniform distribution over the index set $\{1, 2, \ldots, M\}$, and accordingly, the codeword $\boldsymbol{x}_m$ is transmitted over a vector channel $W(\boldsymbol{z}|\boldsymbol{x})$, henceforth referred to as the *primary channel* (or the *main channel*), and the resulting channel output vector, $\boldsymbol{z} \in \mathcal{Z}^n$, is received at the decoder side. The decoder, however, does not have access to the codebook, $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\}$, used by the encoder, but instead, it has access to a noisy version of that codebook, $\mathcal{C} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M\}$, $\boldsymbol{y}_m \in \mathcal{Y}^n$, $m = 1, 2, \ldots, M$, where each $\boldsymbol{y}_m$ is generated from the corresponding $\boldsymbol{x}_m$ by another channel, $V(\boldsymbol{y}|\boldsymbol{x})$, henceforth referred to as the *secondary channel.* Clearly, this model, which was addressed by Willems *et al.* in [18] and [19] with application to biometrical identification systems (and later, further developed by Tuncel [17] and others), is formally equivalent to the ordinary model of channel random coding, where the codebook $\mathcal{C}$ is selected at random, with each member, $\boldsymbol{y}_m$, being drawn independently under the random coding distribution,

$$P(\boldsymbol{y}) = \sum_{\boldsymbol{x} \in \mathcal{X}^n} G(\boldsymbol{x}) V(\boldsymbol{y}|\boldsymbol{x}), \tag{1}$$

and where upon selecting the index $m$ of the transmitted message, the corresponding codeword, $\boldsymbol{y}_m$, is transmitted over the channel

$$P(\boldsymbol{z}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}, \boldsymbol{z})}{P(\boldsymbol{y})} = \frac{\sum_{\boldsymbol{x} \in \mathcal{X}^n} G(\boldsymbol{x}) V(\boldsymbol{y}|\boldsymbol{x}) W(\boldsymbol{z}|\boldsymbol{x})}{\sum_{\boldsymbol{x} \in \mathcal{X}^n} G(\boldsymbol{x}) V(\boldsymbol{y}|\boldsymbol{x})}. \tag{2}$$

From this point onward, the original codebook $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\}$ no longer plays a role. Accordingly, we henceforth refer to $\{P(\boldsymbol{y}), \boldsymbol{y} \in \mathcal{Y}^n\}$ as the *induced random coding distribution* (or the *effective random coding distribution*), and to $\{P(\boldsymbol{z}|\boldsymbol{y})\ \boldsymbol{y} \in \mathcal{Y}^n,\ \boldsymbol{z} \in \mathcal{Z}^n\}$ – as the *induced channel* (or the *effective channel*). Clearly, if $G$ is a discrete memoryless source (DMS) and $V$ is a discrete memoryless channel (DMC), then $\{P(\boldsymbol{y}), \boldsymbol{y} \in \mathcal{Y}^n\}$ is a DMS as well. If, in addition, $W$ is also a DMC, then so is the channel $\{P(\boldsymbol{z}|\boldsymbol{y})\ \boldsymbol{y} \in \mathcal{Y}^n,\ \boldsymbol{z} \in \mathcal{Z}^n\}$. In this case, the capacity of the system is simply the mutual information, $I(Y;Z)$, pertaining to the single–letter marginal $\{P(y,z),\ y \in \mathcal{Y},\ z \in \mathcal{Z}\}$, see [18], [19]. It should be noted, however, that unlike the traditional model of random coding for channels, where random coding is a technical concept that merely serves the purpose of proving existence of good codes, here, when it comes to biometrical systems applications, the randomness of the code is part of the model setting. As a consequence, both $G$ and $V$, and hence also the induced random coding distribution, $\{P(\boldsymbol{y}), \boldsymbol{y} \in \mathcal{Y}^n\}$, are dictated to us, and are not subjected to our control.[1]

As in [18], [19], here too, it is assumed that all three alphabets, $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$, are finite. In this paper, however, we go considerably beyond the realm of memoryless systems, and allow $G$, $V$ and $W$ to be all non-unifilar, FS systems (hidden Markov models), as follows. The distribution $G$ assumes the form

$$G(\boldsymbol{x}) = \sum_{\boldsymbol{\omega}} \prod_{i=1}^{n} G(x_i, \omega_i|\omega_{i-1}), \tag{3}$$

where $\boldsymbol{x}$ is as before, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$ is the source state vector, whose components take on values in a finite set $\Omega$, and the initial state, $\omega_0$ is assumed fixed. The primary channel, $W$, is modeled as

$$W(\boldsymbol{z}|\boldsymbol{x}) = \sum_{\boldsymbol{\sigma}} \prod_{i=1}^{n} W(z_i, \sigma_i|x_i, \sigma_{i-1}), \tag{4}$$

[1] For this reason, the capacity is simply given by $I(Y;Z)$, without maximizing over the distribution of $Y$.

where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$ is the channel state vector, whose components take on values in a finite set $\Sigma$ and the initial state, $\sigma_0$, is fixed. Likewise, the secondary channel, $V$, is given by

$$V(\boldsymbol{y}|\boldsymbol{x}) = \sum_{\boldsymbol{\theta}} \prod_{i=1}^{n} V(y_i, \theta_i | x_i, \theta_{i-1}), \tag{5}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is the state vector whose components take on values in a finite set $\Theta$ and there is fixed initial state, $\theta_0$.

We consider the problem of universal decoding for the effective channel $P(\boldsymbol{z}|\boldsymbol{y})$ induced by the source (3), the main channel (4) and the secondary channel (5), according to (2). We will assume that $G$, $V$ and $W$ are not known to the decoder, and hence nor is the effective channel $\{P(\boldsymbol{z}|\boldsymbol{y})\ \boldsymbol{y} \in \mathcal{Y}^n, \boldsymbol{z} \in \mathcal{Z}^n\}$. Nonetheless, the effective random coding distribution, $\{P(\boldsymbol{y}),\ \boldsymbol{y} \in \mathcal{Y}^n\}$, will assumed known to the decoder. The rationale behind the latter assumption stems from the fact that the decoder knows the codebook, $\mathcal{C} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M\}$, and so, it has access to an exponential amount of data from which the parameters of this distribution can be estimated very accurately. In particular, note that $P(\boldsymbol{y})$ has a hidden Markov structure,

$$\begin{aligned} P(\boldsymbol{y}) &= \sum_{\boldsymbol{x}} G(\boldsymbol{x}) V(\boldsymbol{y}|\boldsymbol{x}) \\ &= \sum_{\boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{x}} \prod_{i=1}^{n} G(x_i, \omega_i | \omega_{i-1}) V(y_i, \theta_i | x_i, \theta_{i-1}) \\ &= \sum_{\boldsymbol{\theta}, \boldsymbol{\omega}} \prod_{i=1}^{n} \left[ \sum_{x} G(x, \omega_i | \omega_{i-1}) V(y_i, \theta_i | x, \theta_{i-1}) \right] \\ &= \sum_{\boldsymbol{\theta}, \boldsymbol{\omega}} \prod_{i=1}^{n} \pi(y_i, \theta_i, \omega_i | \theta_{i-1}, \omega_{i-1}), \end{aligned} \tag{6}$$

where in the last passage, we have defined the parameters $\pi(y, \theta, \omega|\theta', \omega') \stackrel{\Delta}{=} \sum_x G(x, \omega|\omega') V(y, \theta|x, \theta')$. These parameters can be estimated using well known estimation methods for hidden Markov models.[2] It will be assumed[3] that

$$\pi(y, \theta, \omega|\theta', \omega') > 0 \tag{7}$$

for all $(\omega, \omega', \theta, \theta', y) \in \Omega^2 \times \Theta^2 \times \mathcal{Y}$, and we denote $\pi_{\min} \stackrel{\Delta}{=} \min_{\omega, \omega', \theta, \theta', y} \pi(y, \theta, \omega|\theta', \omega')$.

Like in previous works on universal decoding, our objective is to devise a universal decoding metric whose average error probability is of the same exponential order as that of the ML decoder. As described in the Introduction, the problem of universal decoding for FS channels was considered

[2] The ML estimator for the parameters of a hidden Markov model, is known to be strongly consistent [1], [15]. More practically, one may use the iterative Baum algorithm, which is an instance of the EM algorithm [5] (see also the tutorials [6], [16] and references therein).

[3] Note that this assumption concerns $G$ and $V$ only, it has nothing to do with the primary channel $W$. If the parameters $\{\pi(y, \theta, \omega|\theta', \omega')\}$ are estimated using the ML estimator (referring to footnote 2), then eq. (7) can be imposed as a constraint on the estimator.

first in [20], where it was assumed that the next–state transitions are given by a deterministic function of the current state, the current input and the current output. In [11], the framework was extended to handle general FS channels, where the state transitions were allowed to be stochastic (as in eqs. (4) and (5) above). Also, in both [11] and [20], the random coding distribution was assumed uniform across a given permutation–invariant set.[4] Here the situation is different from both [11] and [20] because of two reasons.

1. The effective random coding distribution $\{P(\boldsymbol{y}),\ \boldsymbol{y}\in\mathcal{Y}^n\}$ is not uniform over a permutation–invariant set, in general.
2. The effective channel $\{P(\boldsymbol{z}|\boldsymbol{y}),\ \boldsymbol{y}\in\mathcal{Y}^n,\ \boldsymbol{z}\in\mathcal{Z}^n\}$ is not a FS channel, in general.

These differences are important, because in [11] and [20], both assumptions were used rather heavily.

For a given noisy code $\mathcal{C}$ and a given channel output vector $\boldsymbol{z}$, let us define (similarly as in [7] and [11]) the ranking of the members of $\mathcal{Y}^n$, according to descending likelihood values, i.e., $P(\boldsymbol{z}|\boldsymbol{y}[1])\geq P(\boldsymbol{z}|\boldsymbol{y}[2])\geq\ldots$, and let us denote by $M_{\rm o}(\boldsymbol{y},\boldsymbol{z})$ the ranking of $\boldsymbol{y}$ given $\boldsymbol{z}$. For a given $\boldsymbol{z}$, the ranking function $M_{\rm o}(\boldsymbol{y},\boldsymbol{z})$ is therefore a one–to–one mapping from $\mathcal{Y}^n$ to the set $\{1,2,\ldots,|\mathcal{Y}|^n\}$ with the property that $P(\boldsymbol{z}|\boldsymbol{y}')>P(\boldsymbol{z}|\boldsymbol{y})$ implies $M_{\rm o}(\boldsymbol{y}',\boldsymbol{z})<M_{\rm o}(\boldsymbol{y},\boldsymbol{z})$. The probability of error associated with the ML decoder for the given code $\mathcal{C}$ and the effective channel, $\{P(\boldsymbol{z}|\boldsymbol{y}),\ \boldsymbol{y}\in\mathcal{Y}^n,\ \boldsymbol{z}\in\mathcal{Z}^n\}$, is given by

$$\mathrm{P}_{\rm e,o}(\mathcal{C})=\frac{1}{M}\sum_{m=1}^{M}P\left[\bigcup_{m'\neq m}\{M_{\rm o}(\boldsymbol{y}_{m'},\boldsymbol{Z})\leq M_{\rm o}(\boldsymbol{y}_m,\boldsymbol{Z})\}\middle|\text{message } m \text{ was sent}\right], \tag{8}$$

where the event $M_{\rm o}(\boldsymbol{y}_{m'},\boldsymbol{Z})=M_{\rm o}(\boldsymbol{y}_m,\boldsymbol{Z})$ accounts for the case where $\boldsymbol{y}_{m'}=\boldsymbol{y}_m$ (which is possible since the members of $\mathcal{C}$ are chosen independently at random). The average probability of error w.r.t. the randomness of $\mathcal{C}$, is then

$$\begin{aligned}\overline{\mathrm{P}}_{\rm e,o} &= \boldsymbol{E}\{\mathrm{P}_{\rm e,o}(\mathcal{C})\} && (9)\\ &= 1-\sum_{\boldsymbol{y},\boldsymbol{z}}P(\boldsymbol{y},\boldsymbol{z})\left(1-P[\mathcal{E}_{\rm o}(\boldsymbol{y},\boldsymbol{z})]\right)^{e^{nR}-1}, && (10)\end{aligned}$$

where

$$\mathcal{E}_{\rm o}(\boldsymbol{y},\boldsymbol{z})\stackrel{\triangle}{=}\{\boldsymbol{y}':\ M_{\rm o}(\boldsymbol{y}',\boldsymbol{z})\leq M_{\rm o}(\boldsymbol{y},\boldsymbol{z})\}. \tag{11}$$

As in [7] and [11], for later use, we define the function

$$f(t)\stackrel{\triangle}{=}1-(1-t)^{e^{nR}-1},\quad t\in[0,1], \tag{12}$$

and so,

$$\overline{\mathrm{P}}_{\rm e,o}=\sum_{\boldsymbol{y},\boldsymbol{z}}P(\boldsymbol{y},\boldsymbol{z})f\left(P[\mathcal{E}_{\rm o}(\boldsymbol{y},\boldsymbol{z})]\right). \tag{13}$$

[4] A permutation–invariant set is a set that is closed under permutations, in other words, a set that can be represented by the disjoint union of type classes.

By the same token, for an arbitrary decoding metric $u(\boldsymbol{y},\boldsymbol{z})$, we define a ranking function $M_{\rm u}(\boldsymbol{y},\boldsymbol{z})$, as any one–to–one mapping $\mathcal{Y}^n :\to \{1,2,\ldots,|\mathcal{Y}|^n\}$ given $\boldsymbol{z}$, such that $\boldsymbol{u}(\boldsymbol{y}',\boldsymbol{z}) < \boldsymbol{u}(\boldsymbol{y},\boldsymbol{z})$ implies $M_{\rm u}(\boldsymbol{y}',\boldsymbol{z}) < M_{\rm u}(\boldsymbol{y},\boldsymbol{z})$. Accordingly, the average error probability associated with $u(\cdot,\cdot)$, is given by

$$\overline{\mathrm{P}}_{\rm e,u} = \sum_{\boldsymbol{y},\boldsymbol{z}} P(\boldsymbol{y},\boldsymbol{z}) f\left(P[\mathcal{E}_{\rm u}(\boldsymbol{y},\boldsymbol{z})]\right), \tag{14}$$

where

$$\mathcal{E}_{\rm u}(\boldsymbol{y},\boldsymbol{z}) \stackrel{\Delta}{=} \{\boldsymbol{y}' :\ M_{\rm u}(\boldsymbol{y}',\boldsymbol{z}) \le M_{\rm u}(\boldsymbol{y},\boldsymbol{z})\}. \tag{15}$$

We are interested in a universal metric $u(\cdot,\cdot)$, that is independent of the unknown effective channel (but possibly dependent on the effective random coding distribution), such that $\overline{\mathrm{P}}_{\rm e,u}$ would not exceed $\overline{\mathrm{P}}_{\rm e,o}$ by more than a sub–exponential function of $n$, i.e.,

$$\overline{\mathrm{P}}_{\rm e,u} \le e^{n\epsilon(n)}\overline{\mathrm{P}}_{\rm e,o}, \tag{16}$$

where $\epsilon(n) \to 0$ as $n \to \infty$.

## 3 Main Result

Given two sequences, $\boldsymbol{y}$ and $\boldsymbol{z}$, both of length $n$, consider the joint incremental parsing [21] of the sequence of pairs

$$(y_1,z_1),(y_2,z_2),\ldots,(y_n,z_n)$$

into $c$ distinct phrases. Specifically, denoting $w_i = (y_i,z_i)$, $i = 1,2,\ldots,n$, we parse $\boldsymbol{w} = (w_1,\ldots,w_n)$, sequentially into the distinct[5] phrases, $w_1^{n_1}, w_{n_1+1}^{n_2},\ldots,w_{n_{c-1}+1}^{n}$, where $n_i+1$ is the starting point of the $i$–th phrase, $i = 1,2,\ldots,c$ $(n_0 = 0)$. According to the incremental parsing procedure of the LZ algorthm, each phrase $w_{n_i+1}^{n_{i+1}}$ is the shortest string that has not been encountered before as a parsed phrase, which means that its prefix, $w_{n_i+1}^{n_{i+1}-1}$, is identical to an earlier phrase, $w_{n_j+1}^{n_{j+1}}$, $j < i$. Let $c \equiv c(\boldsymbol{y},\boldsymbol{z})$ denote the number of distinct phrases. For example,[6] if

$$\begin{aligned}\boldsymbol{y} &= 0 \mid 1 \mid 0\ 0 \mid 0\ 1| \\ \boldsymbol{z} &= 0 \mid 1 \mid 0\ 1 \mid 0\ 1|\end{aligned}$$

then $c(\boldsymbol{y},\boldsymbol{z}) = 4$. Let $c(\boldsymbol{z})$ denote the resulting number of distinct phrases of $\boldsymbol{z}$, and let $\boldsymbol{z}(\ell)$ denote the $\ell$th distinct $\boldsymbol{z}$–phrase, $\ell = 1,2,...,c(\boldsymbol{z})$. In the above example, $c(\boldsymbol{z}) = 3$. Denote by $c_\ell(\boldsymbol{y}|\boldsymbol{z})$ the number of occurrences of $\boldsymbol{z}(\ell)$ in the parsing of $\boldsymbol{z}$, or equivalently, the number of distinct $\boldsymbol{y}$-phrases that jointly appear with $\boldsymbol{z}(\ell)$. Clearly, $\sum_{\ell=1}^{c(\boldsymbol{z})} c_\ell(\boldsymbol{y}|\boldsymbol{z}) = c(\boldsymbol{y},\boldsymbol{z})$. In the above example, $\boldsymbol{z}(1) = 0$, $\boldsymbol{z}(2) = 1$, $\boldsymbol{z}(3) = 01$, $c_1(\boldsymbol{y}|\boldsymbol{z}) = c_2(\boldsymbol{y}|\boldsymbol{z}) = 1$, and $c_3(\boldsymbol{y}|\boldsymbol{z}) = 2$. We next define our universal decoding metric as

$$u(\boldsymbol{y},\boldsymbol{z}) \stackrel{\Delta}{=} \log P(\boldsymbol{y}) + \sum_{\ell=1}^{c(\boldsymbol{z})} c_\ell(\boldsymbol{y}|\boldsymbol{z}) \log c_\ell(\boldsymbol{y}|\boldsymbol{z}), \tag{17}$$

[5] To be more precise, the phrases are all distinct with the possible exception of the last phrase, which might be incomplete.

[6] The same example appears in [20].

which in turn, defines the decoder

$$\hat{m}_{\mathrm{u}} = \arg\ \min_m u(\boldsymbol{y}_m, \boldsymbol{z}), \tag{18}$$

where ties broken according to an arbitrary ranking function $M_{\mathrm{u}}(\cdot, \boldsymbol{z})$ associated with (17).

We are now ready to state our main result, whose proof appears in Section 4.

**Theorem 1** *Under the assumptions of Subsection 2.2, the universal decoder (18) satisfies eq. (16) where $\epsilon(n) = O((\log\log n)/\log n)$, with a leading term*[7] *that is linear in $\log|\mathcal{Y} \times \mathcal{Z}|$.*

It should be noticed that the universal decoding metric (17) is different from the one in [11] and [20], because it includes the term $\log P(\boldsymbol{y})$ in addition to the LZ conditional compressibility term, $\sum_{\ell=1}^{c(\boldsymbol{z})} c_\ell(\boldsymbol{y}|\boldsymbol{z}) \log c_\ell(\boldsymbol{y}|\boldsymbol{z})$ (see also [14]). The reason for this difference is that the effective random coding distribution, $\{P(\boldsymbol{y}),\ \boldsymbol{y} \in \mathcal{Y}^n\}$, is not necessarily uniform over its support, in contrast to the assumption in both [11] and [20]. In a way, the decoder (18) can be seen as an extension of the MMI decoder, which is the well known universal decoder for DMCs [3]. To see this, observe that (18) can be rewritten as

$$\hat{m}_{\mathrm{u}} = \arg\ \max_m \left\{ \frac{1}{n} \log\left[\frac{1}{P(\boldsymbol{y}_m)}\right] - \frac{1}{n} \sum_{\ell=1}^{c(\boldsymbol{z})} c_\ell(\boldsymbol{y}_m|\boldsymbol{z}) \log c_\ell(\boldsymbol{y}_m|\boldsymbol{z}) \right\}, \tag{19}$$

where the term $\frac{1}{n}\log[1/P(\boldsymbol{y}_m)]$ plays a role like the empirical entropy associated with $\boldsymbol{y}_m$ and the term $\frac{1}{n}\sum_{\ell=1}^{c(\boldsymbol{z})} c_\ell(\boldsymbol{y}_m|\boldsymbol{z}) \log c_\ell(\boldsymbol{y}_m|\boldsymbol{z})$ is parallel to the conditional empirical entropy of $\boldsymbol{y}_m$ given $\boldsymbol{z}$. Thus, the difference is analogous to a certain notion of a generalized empirical mutual information. But having said that, we should add a digression that, when confining the discussion to the memoryless case, the first term in (19) gives the empirical entropy of $\boldsymbol{y}_m$ only in the case where $\{P(\boldsymbol{y})\}$ is uniform across a single type class. If instead, it is a product distribution, then the MMI metric should be supplemented with a divergence term between the empirical distribution and the true distribution.[8]

The proof of Theorem 1 contains essentially similar ingredients to those in [11]. There are, however, a few differences that should be pointed out. In the previous paragraph, we mentioned that here, as opposed to those papers, the random coding distribution is not uniform in general. This difference is also responsible for the fact that there are a few non–trivial issues in the extension of the derivations of [11] and [20] to our setting, as in those two earlier papers, the uniformity of the random coding distribution (across its support), was used quite heavily. In particular, the pairwise error probability, $P[\mathcal{E}_{\mathrm{o}}(\boldsymbol{y}, \boldsymbol{z})]$, which plays a central role in the analysis in [11] and [20], is simply proportional to the cardinality of $\mathcal{E}_{\mathrm{o}}(\boldsymbol{y}, \boldsymbol{z})$, namely to $M_{\mathrm{o}}(\boldsymbol{y}, \boldsymbol{z})$, which in turn, can be evaluated using combinatorial considerations. Here, on the other hand, the members of $\mathcal{E}_{\mathrm{o}}(\boldsymbol{y}, \boldsymbol{z})$ have to

[7] The sequence $\epsilon(n)$ depends also on other parameters of the problem, like $|\Theta|$, $|\Omega|$, $|\Sigma|$, and $\pi_{\min}$, but these parameters appear in negligible terms of $\epsilon(n)$, that decay faster than $(\log\log n)/\log n$.

[8] In this context, the author has some doubts concerning the asymptotic optimality of the MMI decoder used in [4].

be weighed by their various probabilities, $\{P(\boldsymbol{y}'),\ \boldsymbol{y}'\in\mathcal{E}_{\rm o}(\boldsymbol{y},\boldsymbol{z})\}$. In particular, in an important technical lemma of [11] (Lemma 2 therein), the last step of the proof is relatively easy, because thanks to the uniformity assumption therein, it is associated with the calculation of the quantity, $\sum_{\boldsymbol{y}} 1/M_{\rm o}(\boldsymbol{y},\boldsymbol{z})$ (in our notation), which is nothing but the harmonic series, $\sum_{i=1}^N 1/i \le \ln N+1$ ($N$ – positive integer), as $M_{\rm o}(\boldsymbol{y},\boldsymbol{z})$ is defined as a ranking function (see, in particular, the last step in the chain of inequalities at the end of page 1751 in [11]). For the non–uniform input considered here, the relevant extension of the above mentioned expression turns out to be $\sum_{\boldsymbol{y}} P(\boldsymbol{y})/P[\mathcal{E}_{\rm o}(\boldsymbol{y},\boldsymbol{z})]$, which is not as straightforward to bound in a useful manner. Fortunately enough, as is shown in Lemma 1 below, this can nevertheless still be done, and in a quite general manner, that is almost completely unrelated to the hidden Markov structure of the model. Another source for some technical challenges is the fact that the induced channel, $\{P(\boldsymbol{z}|\boldsymbol{y})\}$, is not a FS channel, in general. This calls for separate treatment of the numerator and the denominator of $P(\boldsymbol{z}|\boldsymbol{y}) = P(\boldsymbol{y},\boldsymbol{z})/P(\boldsymbol{y})$ (which both obey a hidden Markov model), that in turn, may be dominated by two different sequences of states. Nonetheless, these difficulties can also be circumvented, as will be seen in Section 4.

## 4 Proof of Theorem 1

The idea of the proof is to lower bound $\bar{\rm P}_{\rm e,o}$ and to upper bound $\bar{\rm P}_{\rm e,u}$ by two expressions which are identical up to a multiplicative factor of $e^{n\epsilon(n)}$. We begin with the upper bound to $\bar{\rm P}_{\rm e,u}$.

Let us denote

$$v(\boldsymbol{y},\boldsymbol{z}) \stackrel{\Delta}{=} \sum_{\ell=1}^{c(\boldsymbol{z})} c_\ell(\boldsymbol{y}|\boldsymbol{z})\log c_\ell(\boldsymbol{y}|\boldsymbol{z}), \tag{20}$$

so that $u(\boldsymbol{y},\boldsymbol{z}) = \log P(\boldsymbol{y}) + v(\boldsymbol{y},\boldsymbol{z})$. We will use the fact that $v(\boldsymbol{y},\boldsymbol{z})$ is almost large enough to serve as a legitimate length function for lossless compression of $\boldsymbol{y}$ given $\boldsymbol{z}$, where $\boldsymbol{z}$ serves as side information available to both the encoder and the decoder. In particular, in the proof of Lemma 2 in [20, p. 460], Ziv describes a lossless compression scheme with side information, whose length function, $L(\boldsymbol{y}|\boldsymbol{z})$, satisfies

$$L(\boldsymbol{y}|\boldsymbol{z}) \le v(\boldsymbol{y},\boldsymbol{z}) + n\epsilon_1(n), \tag{21}$$

with

$$\epsilon_1(n) = O\left(\frac{\log\log n}{\log n}\right), \tag{22}$$

whose leading term is linear in $\log|\mathcal{Y}\times\mathcal{Z}|$. Now, let us define

$$\bar{\rm P}_{\rm e,u}(\boldsymbol{y},\boldsymbol{z}) = f(P[\mathcal{E}_{\rm u}(\boldsymbol{y},\boldsymbol{z})]), \tag{23}$$

where $f(\cdot)$ is defined as in (12). Now,

$$P[\mathcal{E}_{\rm u}(\boldsymbol{y},\boldsymbol{z})] \;=\; \sum_{\{\boldsymbol{y}':\ M_{\rm u}(\boldsymbol{y}',\boldsymbol{z})\le M_{\rm u}(\boldsymbol{y},\boldsymbol{z})\}} P(\boldsymbol{y}')$$

$$\begin{aligned}
&\leq \sum_{\{\boldsymbol{y}':\ P(\boldsymbol{y}')\exp_2[v(\boldsymbol{y}',\boldsymbol{z})]\leq P(\boldsymbol{y})\exp_2[v(\boldsymbol{y},\boldsymbol{z})]\}} P(\boldsymbol{y}') \\
&\leq \sum_{\{\boldsymbol{y}':\ P(\boldsymbol{y}')\exp_2[v(\boldsymbol{y}',\boldsymbol{z})]\leq P(\boldsymbol{y})\exp_2[v(\boldsymbol{y},\boldsymbol{z})]\}} P(\boldsymbol{y})\exp_2[v(\boldsymbol{y},\boldsymbol{z})-v(\boldsymbol{y}',\boldsymbol{z})] \\
&\leq P(\boldsymbol{y})\exp_2[v(\boldsymbol{y},\boldsymbol{z})]\sum_{\boldsymbol{y}'\in\mathcal{Y}^n}\exp_2[-v(\boldsymbol{y}',\boldsymbol{z})] \\
&\leq 2^{n\epsilon_1(n)}P(\boldsymbol{y})\exp_2[v(\boldsymbol{y},\boldsymbol{z})]\sum_{\boldsymbol{y}'\in\mathcal{Y}^n}2^{-L(\boldsymbol{y}'|\boldsymbol{z})} \\
&\leq e^{n\epsilon_1(n)}P(\boldsymbol{y})\cdot\exp_2[v(\boldsymbol{y},\boldsymbol{z})] \\
&= e^{n\epsilon_1(n)}\cdot\exp_2[u(\boldsymbol{y},\boldsymbol{z})], \qquad (24)
\end{aligned}$$

where the in the second to the last step, we have used Kraft's inequality and we bounded $2^{n\epsilon_1(n)}$ by $e^{n\epsilon_1(n)}$, simply for convenience in later steps of the proof. It now follows from (24) and the monotonicity of $f$ that

$$\bar{\mathrm{P}}_{\mathrm{e,u}}(\boldsymbol{y},\boldsymbol{z})\leq f\left(e^{n\epsilon_1(n)}\cdot\exp_2[u(\boldsymbol{y},\boldsymbol{z})]\right). \qquad (25)$$

For later use, we also have

$$\begin{aligned}
\bar{\mathrm{P}}_{\mathrm{e,u}}(\boldsymbol{z}) &\stackrel{\triangle}{=} \sum_{\boldsymbol{y}\in\mathcal{Y}^n}P(\boldsymbol{y}|\boldsymbol{z})\bar{\mathrm{P}}_{\mathrm{e,u}}(\boldsymbol{y},\boldsymbol{z}) \qquad (26)\\
&\leq \sum_{\boldsymbol{y}\in\mathcal{Y}^n}P(\boldsymbol{y}|\boldsymbol{z})f\left(e^{n\epsilon_1(n)}\cdot\exp_2[u(\boldsymbol{y},\boldsymbol{z})]\right). \qquad (27)
\end{aligned}$$

We next move on to derive a matching lower bound to $\bar{\mathrm{P}}_{\mathrm{e,o}}$. Similarly, as in [11], we will need to refer to an auxiliary threshold decoder (in the terminology of [11]), which is a slightly more conservative version of the ML decoder. Specifically, for a given threshold parameter, $\alpha>1$, this decoder outputs the message $m$ with the property that $P(\boldsymbol{z}|\boldsymbol{y}_m)>\alpha\cdot P(\boldsymbol{z}|\boldsymbol{y}_{m'})$ for all $m'\neq m$, and declares an error if no such $m$ exists. Accordingly, let $\bar{\mathrm{P}}_{\mathrm{e,t}}(\boldsymbol{y},\boldsymbol{z})$ denote the conditional average error probability of the threshold decoder, given $(\boldsymbol{y},\boldsymbol{z})$, i.e.,

$$\bar{\mathrm{P}}_{\mathrm{e,t}}(\boldsymbol{y},\boldsymbol{z})=f(P[\mathcal{E}_{\mathrm{t}}(\boldsymbol{y},\boldsymbol{z})]), \qquad (28)$$

where

$$\mathcal{E}_{\mathrm{t}}(\boldsymbol{y},\boldsymbol{z})=\{\boldsymbol{y}':\ P(\boldsymbol{z}|\boldsymbol{y}')\geq\alpha^{-1}P(\boldsymbol{z}|\boldsymbol{y})\}. \qquad (29)$$

As in Lemma 2 of [11], here too, the next lemma (proved in the appendix) asserts that the performance of the threshold decoder cannot be much worse than that of the ML decoder, provided that $\alpha$ is not too large. In particular, if $\alpha=\alpha_n$ grows subexponetially with $n$, then the threshold decoder has the same error exponent as that of the ML decoder.

**Lemma 1** *Define*

$$\bar{P}_{e,t}(\boldsymbol{z})=\sum_{\boldsymbol{y}\in\mathcal{Y}^n}P(\boldsymbol{y}|\boldsymbol{z})f(P[\mathcal{E}_t(\boldsymbol{y},\boldsymbol{z})]) \qquad (30)$$

$$\bar{P}_{e,o}(\boldsymbol{z}) = \sum_{\boldsymbol{y}\in\mathcal{Y}^n} P(\boldsymbol{y}|\boldsymbol{z}) f(P[\mathcal{E}_o(\boldsymbol{y},\boldsymbol{z})]). \tag{31}$$

*Then, under the positivity assumption (7),*

$$\bar{P}_{e,t}(\boldsymbol{z}) \leq \left\{\alpha\left[n\ln\left(\frac{1}{\pi_{\min}\cdot|\Theta|\cdot|\Omega|}\right)+1\right]+1\right\}\cdot\bar{P}_{e,o}(\boldsymbol{z}) \tag{32}$$

*for every $\boldsymbol{z}\in\mathcal{Z}^n$.*

It should be noted that assumption (7) is essentially not needed for the above Lemma. What is really needed is that the smallest $P(\boldsymbol{y})$, across all $\boldsymbol{y}\in\mathcal{Y}^n$ with $P(\boldsymbol{y})>0$, would not decay faster than exponentially with $n$. But owing to (6), one can easily see that $P(\boldsymbol{y})\geq\pi_+^n$, where $\pi_+$ is the smallest positive $\pi(y,\theta,\omega|\theta',\omega')$. We are using (7) nonetheless, because we make this assumption anyway (as it is needed elsewhere), and then the upper bound given by the lemma is slightly tighter.

On the basis of Lemma 1, any lower bound on $\bar{\mathrm{P}}_{\mathrm{e,t}}$ in terms of $\bar{\mathrm{P}}_{\mathrm{e,u}}$, would immediately yield a lower bound $\bar{\mathrm{P}}_{\mathrm{e,o}}$ in terms of $\bar{\mathrm{P}}_{\mathrm{e,u}}$, as desired. Accordingly, the next step would be to lower bound $\bar{\mathrm{P}}_{\mathrm{e,t}}$. This in turn will be done by lower bounding $P[\mathcal{E}_{\mathrm{t}}(\boldsymbol{y},\boldsymbol{z})]$ (for a certain choice of the threshold $\alpha$, to be defined) in terms of $P[\mathcal{E}_1(\boldsymbol{y},\boldsymbol{z})]$, for a certain $\mathcal{E}_1(\boldsymbol{y},\boldsymbol{z})\subseteq\mathcal{E}_{\mathrm{t}}(\boldsymbol{y},\boldsymbol{z})$ to be specified shortly.

First observe that, similarly as in eq. (6),

$$\begin{aligned}
P(\boldsymbol{y},\boldsymbol{z}) &= \sum_{\boldsymbol{\theta},\boldsymbol{\sigma},\boldsymbol{\omega},\boldsymbol{x}}\prod_{i=1}^n[G(x_i,\omega_i|\omega_{i-1})V(y_i,\theta_i|x_i,\theta_{i-1})W(z_i,\sigma_i|x_i,\sigma_{i-1})] && (33)\\
&= \sum_{\boldsymbol{\theta},\boldsymbol{\sigma},\boldsymbol{\omega}}\prod_{i=1}^n\sum_x[G(x,\omega_i|\omega_{i-1})V(y_i,\theta_i|x,\theta_{i-1})W(z_i,\sigma_i|x,\sigma_{i-1})] && (34)\\
&= \sum_{\boldsymbol{\theta},\boldsymbol{\sigma},\boldsymbol{\omega}}\prod_{i=1}^n\Pi(y_i,z_i,\theta_i,\sigma_i,\omega_i|\theta_{i-1},\sigma_{i-1},\omega_{i-1}) && (35)
\end{aligned}$$

where we have defined $\Pi(y,z,\theta,\sigma,\omega|\theta',\sigma',\omega')=\sum_x G(x,\omega|\omega')V(y,\theta|x,\theta')W(z,\sigma|x,\sigma')$. We will henceforth use the following notation for two positive integers $i$ and $j$, where $j>i$:

$$\begin{aligned}
&\Pi(y_i^j,z_i^j,\theta_j,\sigma_j,\omega_j|\theta_{i-1},\sigma_{i-1},\omega_{i-1})\\
&= \sum_{\theta_i^{j-1}}\sum_{\sigma_i^{j-1}}\sum_{\omega_i^{j-1}}\prod_{k=i}^j\Pi(y_k,z_k,\theta_k,\sigma_k,\omega_k|\theta_{k-1},\sigma_{k-1},\omega_{k-1}) && (36)
\end{aligned}$$

and

$$\pi(y_i^j,\theta_j,\omega_j|\theta_{i-1},\omega_{i-1}) = \sum_{\theta_i^{j-1}}\sum_{\omega_i^{j-1}}\prod_{k=i}^j\pi(y_k,\theta_k,\omega_k|\theta_{k-1},\omega_{k-1}). \tag{37}$$

Next, define

$$\boldsymbol{t} \;\stackrel{\triangle}{=}\; \{(\theta_i,\sigma_i,\omega_i):\; i=n_0,n_1,\ldots,n_{c-1}\}, \tag{38}$$

$$\boldsymbol{s} \overset{\Delta}{=} \{(\theta_i, \omega_i): \; i = n_0, n_1, \ldots, n_{c-1}\}, \tag{39}$$

where $\{n_i\}$ are phrase boundaries, as defined at the beginning of Section 3, for a given $(\boldsymbol{y}, \boldsymbol{z})$. Now, for the same $(\boldsymbol{y}, \boldsymbol{z})$, let

$$\begin{aligned} \hat{\boldsymbol{t}} &= \arg\max_{\boldsymbol{t}} P(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{t}) = \arg\max_{\boldsymbol{t}} \prod_{i=0}^{c-1} \Pi(y_{n_i+1}^{n_{i+1}}, z_{n_i+1}^{n_{i+1}}, \theta_{n_{i+1}}, \sigma_{n_{i+1}}, \omega_{n_{i+1}} | \theta_{n_i}, \sigma_{n_i}, \omega_{n_i}) & (40) \\ \tilde{\boldsymbol{s}} &= \arg\max_{\boldsymbol{s}} P(\boldsymbol{y}, \boldsymbol{s}) = \arg\max_{\boldsymbol{s}} \prod_{i=0}^{c-1} \pi(y_{n_i+1}^{n_{i+1}}, \theta_{n_{i+1}}, \omega_{n_{i+1}} | \theta_{n_i}, \omega_{n_i}). & (41) \end{aligned}$$

We denote the components of $\hat{\boldsymbol{t}}$ and $\tilde{\boldsymbol{s}}$ by $\{(\hat{\theta}_{n_i}, \hat{\sigma}_{n_i}, \hat{\omega}_{n_i})\}$ and $\{(\tilde{\theta}_{n_i}, \tilde{\omega}_{n_i})\}$, respectively. Denoting $K = |\Theta \times \Sigma \times \Omega|$, it is obvious that $P(\boldsymbol{y}, \boldsymbol{z}, \hat{\boldsymbol{t}}) \geq K^{-c} P(\boldsymbol{y}, \boldsymbol{z})$, and a similar relation holds between $P(\boldsymbol{y}, \tilde{\boldsymbol{s}})$ and $P(\boldsymbol{y})$. For the given pair $(\boldsymbol{y}, \boldsymbol{z})$, let

$$\mathcal{E}_1(\boldsymbol{y}, \boldsymbol{z}) \overset{\Delta}{=} \left\{\boldsymbol{y}' : \; P(\boldsymbol{y}', \boldsymbol{z}, \hat{\boldsymbol{t}}) = P(\boldsymbol{y}, \boldsymbol{z}, \hat{\boldsymbol{t}}), \; P(\boldsymbol{y}', \tilde{\boldsymbol{s}}) = P(\boldsymbol{y}, \tilde{\boldsymbol{s}})\right\}. \tag{42}$$

Owing to assumption (7), it is shown in the appendix (similarly as in [22, eq. (A.7)]) that

$$P(\boldsymbol{y}') \leq P(\boldsymbol{y}', \tilde{\boldsymbol{s}}) \cdot \left(\frac{|\Theta \times \Omega|}{\pi_{\min}^2}\right)^c \leq P(\boldsymbol{y}', \tilde{\boldsymbol{s}}) \cdot \left(\frac{K}{\pi_{\min}^2}\right)^c, \tag{43}$$

and so, for $\boldsymbol{y}' \in \mathcal{E}_1(\boldsymbol{y}, \boldsymbol{z})$, the chain of inequalities,

$$\begin{aligned} \left(\frac{K}{\pi_{\min}^2}\right)^c \cdot P(\boldsymbol{z}|\boldsymbol{y}') &= \left(\frac{K}{\pi_{\min}^2}\right)^c \cdot \frac{P(\boldsymbol{y}', \boldsymbol{z})}{P(\boldsymbol{y}')} & (44) \\ &\geq \left(\frac{K}{\pi_{\min}^2}\right)^c \frac{P(\boldsymbol{y}', \boldsymbol{z}, \hat{\boldsymbol{t}})}{(K/\pi_{\min}^2)^c P(\boldsymbol{y}', \tilde{\boldsymbol{s}})} & (45) \\ &= \frac{P(\boldsymbol{y}', \boldsymbol{z}, \hat{\boldsymbol{t}})}{P(\boldsymbol{y}', \tilde{\boldsymbol{s}})} & (46) \\ &= \frac{P(\boldsymbol{y}, \boldsymbol{z}, \hat{\boldsymbol{t}})}{P(\boldsymbol{y}, \tilde{\boldsymbol{s}})} & (47) \\ &\geq K^{-c} \frac{P(\boldsymbol{y}, \boldsymbol{z})}{P(\boldsymbol{y})} & (48) \\ &= K^{-c} P(\boldsymbol{z}|\boldsymbol{y}), & (49) \end{aligned}$$

implies that

$$\begin{aligned} \mathcal{E}_1(\boldsymbol{y}, \boldsymbol{z}) &\subseteq \{\boldsymbol{y}' : \; P(\boldsymbol{z}|\boldsymbol{y}') \geq (K/\pi_{\min})^{-2c} P(\boldsymbol{z}|\boldsymbol{y})\} & (50) \\ &\subseteq \{\boldsymbol{y}' : \; P(\boldsymbol{z}|\boldsymbol{y}') \geq (K/\pi_{\min})^{-2\bar{c}_n} P(\boldsymbol{z}|\boldsymbol{y})\} & (51) \\ &= \mathcal{E}_{\mathrm{t}}(\boldsymbol{y}, \boldsymbol{z}) \qquad \text{with the choice } \alpha = (K/\pi_{\min})^{2\bar{c}_n} & (52) \end{aligned}$$

where

$$\bar{c}_n \overset{\Delta}{=} \frac{n \log |\mathcal{Y} \times \mathcal{Z}|}{(1 - \varepsilon_n) \log n}, \tag{53}$$

with $\varepsilon_n \to 0$ as $n \to 0$, so that $\bar{c}_n$ serves as a uniform upper bound to $c \equiv c(\boldsymbol{y}, \boldsymbol{z})$ for every $(\boldsymbol{y}, \boldsymbol{z}) \in \mathcal{Y} \times \mathcal{Z}^n$, according to [21, eq. (6)]. Thus,

$$\begin{aligned}
P[\mathcal{E}_{\mathrm{t}}(\boldsymbol{y}, \boldsymbol{z})] &= \sum_{\boldsymbol{y}' \in \mathcal{E}_{\mathrm{t}}(\boldsymbol{y}, \boldsymbol{z})} P(\boldsymbol{y}') && (54)\\
&\geq \sum_{\boldsymbol{y}' \in \mathcal{E}_1(\boldsymbol{y}, \boldsymbol{z})} P(\boldsymbol{y}') && (55)\\
&\geq \sum_{\boldsymbol{y}' \in \mathcal{E}_1(\boldsymbol{y}, \boldsymbol{z})} P(\boldsymbol{y}', \tilde{\boldsymbol{s}}) && (56)\\
&= \sum_{\boldsymbol{y}' \in \mathcal{E}_1(\boldsymbol{y}, \boldsymbol{z})} P(\boldsymbol{y}, \tilde{\boldsymbol{s}}) && (57)\\
&= |\mathcal{E}_1(\boldsymbol{y}, \boldsymbol{z})| \cdot P(\boldsymbol{y}, \tilde{\boldsymbol{s}}) && (58)\\
&\geq K^{-c} \cdot |E_1(\boldsymbol{y}, \boldsymbol{z})| \cdot P(\boldsymbol{y}) && (59)\\
&\geq K^{-\bar{c}_n} \cdot |E_1(\boldsymbol{y}, \boldsymbol{z})| \cdot P(\boldsymbol{y}). && (60)
\end{aligned}$$

Now, let $\mathcal{T}(\boldsymbol{y}|\boldsymbol{z}, \hat{\boldsymbol{t}}, \tilde{\boldsymbol{s}})$ denote the set of all $\boldsymbol{y}' \in \mathcal{Y}^n$ that are obtained from $\boldsymbol{y}$ by permuting $\boldsymbol{y}$–phrases, $\{y_{n_i+1}^{n_{i+1}}\}$, that are: (i) aligned to the same $\boldsymbol{z}$-phrases, $z_{n_i+1}^{n_{i+1}}$, (ii) of the same length, (iii) begin at the same states, of both $\hat{t}_i = (\hat{\theta}_{n_i}, \hat{\sigma}_{n_i}, \hat{\omega}_{n_i})$ and $\tilde{s}_i = (\tilde{\theta}_{n_i}, \tilde{\omega}_{n_i})$, and (iv) end at the same states of both $\hat{t}_{i+1} = (\hat{\theta}_{n_{i+1}}, \hat{\sigma}_{n_{i+1}}, \hat{\omega}_{n_{i+1}})$ and $\tilde{s}_{i+1} = (\tilde{\theta}_{n_{i+1}}, \tilde{\omega}_{n_{i+1}})$. Clearly, $\mathcal{T}(\boldsymbol{y}|\boldsymbol{z}, \hat{\boldsymbol{t}}, \tilde{\boldsymbol{t}}) \subseteq \mathcal{E}_1(\boldsymbol{y}, \boldsymbol{z})$, and so, $P[\mathcal{E}_{\mathrm{t}}(\boldsymbol{y}, \boldsymbol{z})]$ is further lower bounded by

$$P[\mathcal{E}_{\mathrm{t}}(\boldsymbol{y}, \boldsymbol{z})] \geq K^{-\bar{c}_n} |\mathcal{T}(\boldsymbol{y}|\boldsymbol{z}, \hat{\boldsymbol{t}}, \tilde{\boldsymbol{t}})| \cdot P(\boldsymbol{y}). \tag{61}$$

Now, according to Lemma 1 of [20],

$$|\mathcal{T}(\boldsymbol{y}|\boldsymbol{z}, \hat{\boldsymbol{t}}, \tilde{\boldsymbol{t}})| \geq \exp_2\{v(\boldsymbol{y}, \boldsymbol{z}) - n\epsilon_2'(n)\}, \tag{62}$$

where

$$\begin{aligned}
\epsilon_2'(n) &= \frac{\bar{c}_n}{n} \cdot \log(|\Theta|^4 \cdot |\Omega|^4 \cdot |\Sigma|^2 e) && (63)\\
&= \frac{\log(|\mathcal{Y}| \cdot |\mathcal{Z}|)}{(1 - \varepsilon_n) \log n} \cdot \log(|\Theta|^4 \cdot |\Omega|^4 \cdot |\Sigma|^2 e) && (64)\\
&= O\left(\frac{1}{\log n}\right). && (65)
\end{aligned}$$

Thus,

$$P[\mathcal{E}_{\mathrm{t}}(\boldsymbol{y}, \boldsymbol{z})] \geq K^{-\bar{c}_n} P(\boldsymbol{y}) \cdot \exp_2\{v(\boldsymbol{y}, \boldsymbol{z}) - n\epsilon_2'(n)\} \stackrel{\Delta}{=} \exp_2\{u(\boldsymbol{y}, \boldsymbol{z}) - n\epsilon_2(n)\} \tag{66}$$

where

$$\begin{aligned}
\epsilon_2(n) &= \epsilon_2'(n) + \frac{\bar{c}_n \log K}{n} && (67)\\
&\leq \epsilon_2'(n) + \frac{\log(|\mathcal{Y}| \cdot |\mathcal{Z}|) \cdot \log K}{(1 - \varepsilon_n) \log n} && (68)
\end{aligned}$$

$$= O\left(\frac{1}{\log n}\right), \tag{69}$$

and so,

$$P[\mathcal{E}_{\mathrm{t}}(\boldsymbol{y},\boldsymbol{z})] \geq \exp_2\{u(\boldsymbol{y},\boldsymbol{z}) - n\epsilon_2(n)\}. \tag{70}$$

To complete the proof, we use the first part of Lemma 1 of [11], which asserts that for every $a, b \in [0,1]$, $f(a)/f(b) \leq \max\{1, a/b\}$, and so,

$$\begin{aligned}
\frac{f(P[\mathcal{E}_{\mathrm{u}}(\boldsymbol{y},\boldsymbol{z})])}{f(P[\mathcal{E}_{\mathrm{t}}(\boldsymbol{y},\boldsymbol{z})])} &\leq \max\left\{1, \frac{P[\mathcal{E}_{\mathrm{u}}(\boldsymbol{y},\boldsymbol{z})]}{P[\mathcal{E}_{\mathrm{t}}(\boldsymbol{y},\boldsymbol{z})]}\right\} &(71)\\
&\leq \max\left\{1, \frac{\exp_2\{u(\boldsymbol{y},\boldsymbol{z}) + n\epsilon_1(n)\}}{\exp_2\{u(\boldsymbol{y},\boldsymbol{z}) - n\epsilon_2(n)\}}\right\} &(72)\\
&\leq e^{n[\epsilon_1(n)+\epsilon_2(n)]}, &(73)
\end{aligned}$$

where in the second inequality, we have used eqs. (24) and (70). Now, referring to Lemma 1, let us define

$$\begin{aligned}
\epsilon_3(n) &= \frac{1}{n}\log\left\{\left(\frac{K}{\pi_{\min}}\right)^{2\bar{c}_n}\left[n\ln\left(\frac{1}{\pi_{\min}|\Theta\times\Sigma|}\right)+1\right]+1\right\} &(74)\\
&= O\left(\frac{1}{\log n}\right). &(75)
\end{aligned}$$

Then,

$$\begin{aligned}
\bar{\mathrm{P}}_{\mathrm{e,o}}(\boldsymbol{z}) &\geq e^{-n\epsilon_3(n)}\bar{\mathrm{P}}_{\mathrm{e,t}}(\boldsymbol{z}) \qquad \text{(by Lemma 1)} &(76)\\
&= e^{-n\epsilon_3(n)}\sum_{\boldsymbol{y}\in\mathcal{Y}^n} P(\boldsymbol{y}|\boldsymbol{z})f(P[\mathcal{E}_{\mathrm{t}}(\boldsymbol{y},\boldsymbol{z})]) &(77)\\
&\geq e^{-n[\epsilon_1(n)+\epsilon_2(n)+\epsilon_3(n)]}\sum_{\boldsymbol{y}\in\mathcal{Y}^n} P(\boldsymbol{y}|\boldsymbol{z})f(P[\mathcal{E}_{\mathrm{u}}(\boldsymbol{y},\boldsymbol{z})]) &(78)\\
&= e^{-n[\epsilon_1(n)+\epsilon_2(n)+\epsilon_3(n)]}\bar{\mathrm{P}}_{\mathrm{e,u}}(\boldsymbol{z}). &(79)
\end{aligned}$$

Finally, upon averaging both sides over $\{\boldsymbol{z}\}$, we complete the proof of Theorem 1, with

$$\epsilon(n) \stackrel{\Delta}{=} \epsilon_1(n) + \epsilon_2(n) + \epsilon_3(n), \tag{80}$$

which is $O((\log\log n)/\log n)$ since $\epsilon_1(n)$ is such.

# Appendix

## A1. Proof of Lemma 1

Let us define

$$\Delta(\boldsymbol{y},\boldsymbol{z}) \stackrel{\Delta}{=} \{\boldsymbol{y}' :\ M_{\mathrm{o}}(\boldsymbol{y}',\boldsymbol{z}) > M_{\mathrm{o}}(\boldsymbol{y},\boldsymbol{z}),\ P(\boldsymbol{z}|\boldsymbol{y}') \geq \alpha^{-1}P(\boldsymbol{z}|\boldsymbol{y})\} \tag{A.1}$$

$$= \{\boldsymbol{y}':\ M_{\rm o}(\boldsymbol{y}',\boldsymbol{z}) > M_{\rm o}(\boldsymbol{y},\boldsymbol{z}),\ P(\boldsymbol{y})P(\boldsymbol{y}'|\boldsymbol{z}) \geq \alpha^{-1}P(\boldsymbol{y}')P(\boldsymbol{y}|\boldsymbol{z})\}, \tag{A.2}$$

so that $\mathcal{E}_{\rm t}(\boldsymbol{y},\boldsymbol{z})$ is given by the disjoint union of $\mathcal{E}_{\rm o}(\boldsymbol{y},\boldsymbol{z})$ and $\Delta(\boldsymbol{y},\boldsymbol{z})$. Then the average conditional error probabilities given $\boldsymbol{z}$ are

$$\begin{aligned}
\bar{P}_{\rm e,o}(\boldsymbol{z}) &= \sum_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{z})f(P[\mathcal{E}_{\rm o}(\boldsymbol{y},\boldsymbol{z})]) && \text{(A.3)}\\
\bar{P}_{\rm e,t}(\boldsymbol{z}) &= \sum_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{z})f(P[\mathcal{E}_{\rm o}(\boldsymbol{y},\boldsymbol{z})] + P[\Delta(\boldsymbol{y},\boldsymbol{z})]) && \text{(A.4)}\\
&\leq \sum_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{z})\left(\frac{P[\mathcal{E}_{\rm o}(\boldsymbol{y},\boldsymbol{z})] + P[\Delta(\boldsymbol{y},\boldsymbol{z})]}{(P[\mathcal{E}_{\rm o}(\boldsymbol{y},\boldsymbol{z})]}\right)f(P[\mathcal{E}_{\rm o}(\boldsymbol{y},\boldsymbol{z})]), && \text{(A.5)}
\end{aligned}$$

where in the last step, we have used the first part of Lemma 1 from [11] (see also [7]). Now, let us define

$$r(\boldsymbol{y},\boldsymbol{z}) \stackrel{\Delta}{=} \sum_{\boldsymbol{y}'\in\mathcal{E}_{\rm o}(\boldsymbol{y},\boldsymbol{z})} P(\boldsymbol{y}'|\boldsymbol{z}). \tag{A.6}$$

Then,

$$\begin{aligned}
P(\boldsymbol{y}) &= \sum_{\boldsymbol{y}'} P(\boldsymbol{y})P(\boldsymbol{y}'|\boldsymbol{z}) && \text{(A.7)}\\
&\geq \sum_{\boldsymbol{y}'\in\mathcal{E}_{\rm o}(\boldsymbol{y},\boldsymbol{z})} P(\boldsymbol{y})P(\boldsymbol{y}'|\boldsymbol{z}) + \sum_{\boldsymbol{y}'\in\Delta(\boldsymbol{y},\boldsymbol{z})} P(\boldsymbol{y})P(\boldsymbol{y}'|\boldsymbol{z}) && \text{(A.8)}\\
&= P(\boldsymbol{y})r(\boldsymbol{y},\boldsymbol{z}) + \sum_{\boldsymbol{y}'\in\Delta(\boldsymbol{y},\boldsymbol{z})} P(\boldsymbol{y})P(\boldsymbol{y}'|\boldsymbol{z}) && \text{(A.9)}\\
&\geq P(\boldsymbol{y})r(\boldsymbol{y},\boldsymbol{z}) + \frac{1}{\alpha}\sum_{\boldsymbol{y}'\in\Delta(\boldsymbol{y},\boldsymbol{z})} P(\boldsymbol{y}')P(\boldsymbol{y}|\boldsymbol{z}) && \text{(A.10)}\\
&= P(\boldsymbol{y})r(\boldsymbol{y},\boldsymbol{z}) + \frac{P(\boldsymbol{y}|\boldsymbol{z})}{\alpha}P[\Delta(\boldsymbol{y},\boldsymbol{z})], && \text{(A.11)}
\end{aligned}$$

and so,

$$P(\boldsymbol{y}|\boldsymbol{z})P[\Delta(\boldsymbol{y},\boldsymbol{z})] \leq \alpha P(\boldsymbol{y})[1 - r(\boldsymbol{y},\boldsymbol{z})]. \tag{A.12}$$

We then have

$$\begin{aligned}
&\bar{P}_{\rm e,t}(\boldsymbol{z}) - \bar{P}_{\rm e,o}(\boldsymbol{z}) && \text{(A.13)}\\
\leq\ & \sum_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{z})\frac{P[\Delta(\boldsymbol{y},\boldsymbol{z})]}{P[\mathcal{E}_{\rm o}(\boldsymbol{y},\boldsymbol{z})]}f(P[\mathcal{E}_{\rm o}(\boldsymbol{y},\boldsymbol{z})]) && \text{(A.14)}\\
\leq\ & \alpha\cdot\sum_{\boldsymbol{y}} \frac{P(\boldsymbol{y})[1 - r(\boldsymbol{y},\boldsymbol{z})]}{P[\mathcal{E}_{\rm o}(\boldsymbol{y},\boldsymbol{z})]}f(P[\mathcal{E}_{\rm o}(\boldsymbol{y},\boldsymbol{z})]) && \text{(A.15)}\\
=\ & \alpha\cdot\sum_{\boldsymbol{y}}\ \sum_{\{\boldsymbol{y}':\ M_{\rm o}(\boldsymbol{y}',\boldsymbol{z})>M_{\rm o}(\boldsymbol{y},\boldsymbol{z})\}} \frac{P(\boldsymbol{y})P(\boldsymbol{y}'|\boldsymbol{z})}{P[\mathcal{E}_{\rm o}(\boldsymbol{y},\boldsymbol{z})]}f(P[\mathcal{E}_{\rm o}(\boldsymbol{y},\boldsymbol{z})]) && \text{(A.16)}\\
\stackrel{\rm (a)}{=}\ & \alpha\cdot\sum_{\boldsymbol{y}'}\ \sum_{\{\boldsymbol{y}:\ M_{\rm o}(\boldsymbol{y}',\boldsymbol{z})>M_{\rm o}(\boldsymbol{y},\boldsymbol{z})\}} \frac{P(\boldsymbol{y})P(\boldsymbol{y}'|\boldsymbol{z})}{P[\mathcal{E}_{\rm o}(\boldsymbol{y},\boldsymbol{z})]}f(P[\mathcal{E}_{\rm o}(\boldsymbol{y},\boldsymbol{z})]) && \text{(A.17)}
\end{aligned}$$

$$
\begin{aligned}
&\overset{\text{(b)}}{\le} \quad \alpha \cdot \sum_{\boldsymbol{y}'} \sum_{\{\boldsymbol{y}:\ M_{\text{o}}(\boldsymbol{y}',\boldsymbol{z})>M_{\text{o}}(\boldsymbol{y},\boldsymbol{z})\}} \frac{P(\boldsymbol{y})P(\boldsymbol{y}'|\boldsymbol{z})}{P[\mathcal{E}_{\text{o}}(\boldsymbol{y},\boldsymbol{z})]} f(P[\mathcal{E}_{\text{o}}(\boldsymbol{y}',\boldsymbol{z})]) && \text{(A.18)}\\
&\le \quad \alpha \cdot \sum_{\boldsymbol{y}'} P(\boldsymbol{y}'|\boldsymbol{z}) f(P[\mathcal{E}_{\text{o}}(\boldsymbol{y}',\boldsymbol{z})]) \cdot \sum_{\boldsymbol{y}} \frac{P(\boldsymbol{y})}{P[\mathcal{E}_{\text{o}}(\boldsymbol{y},\boldsymbol{z})]} && \text{(A.19)}\\
&= \quad \alpha \cdot \bar{P}_{\text{e,o}}(\boldsymbol{z}) \cdot \sum_{\boldsymbol{y}\in\mathcal{Y}^n} \frac{P(\boldsymbol{y})}{P[\mathcal{E}_{\text{o}}(\boldsymbol{y},\boldsymbol{z})]}, && \text{(A.20)}
\end{aligned}
$$

where in (a) we have interchanged the order of the summation and in (b), we have used the monotonicity of $f$ together with the fact that $\mathcal{E}_{\text{o}}(\boldsymbol{y},\boldsymbol{z}) \subseteq \mathcal{E}_{\text{o}}(\boldsymbol{y}',\boldsymbol{z})$ whenever $M_{\text{o}}(\boldsymbol{y}',\boldsymbol{z}) > M_{\text{o}}(\boldsymbol{y},\boldsymbol{z})$. To complete the proof, it remains to show then that for any $\boldsymbol{z}$,

$$
L_n(\boldsymbol{z}) \stackrel{\Delta}{=} \sum_{\boldsymbol{y}\in\mathcal{Y}^n} \frac{P(\boldsymbol{y})}{P[\mathcal{E}(\boldsymbol{y},\boldsymbol{z})]} = \sum_{\boldsymbol{y}\in\mathcal{Y}^n} \frac{P(\boldsymbol{y})}{\sum_{\{\boldsymbol{y}':\ M_{\text{o}}(\boldsymbol{y}',\boldsymbol{z})\le M_{\text{o}}(\boldsymbol{y},\boldsymbol{z})\}} P(\boldsymbol{y}')} \tag{A.21}
$$

cannot exceed $1 + n\ln[1/(\pi_{\min}|\Theta \times \Omega|)]$. For the given $\boldsymbol{z}$, consider the ordering of all members of $\mathcal{Y}^n$ (not only those in $\mathcal{C}$) according to the ranking function $M_{\text{o}}(\boldsymbol{y},\boldsymbol{z})$, i.e.,

$$
P(\boldsymbol{z}|\boldsymbol{y}[1]) \ge P(\boldsymbol{z}|\boldsymbol{y}[2]) \ge \ldots \ge P(\boldsymbol{z}|\boldsymbol{y}[N]), \qquad N = |\mathcal{Y}|^n \tag{A.22}
$$

and let us denote $a_i = P(\boldsymbol{y}[i])$, $A_i = \sum_{j=1}^{i} a_j$, $i = 1, \ldots, N$. Then, using the facts that $A_1 = a_1 = P(\boldsymbol{y}[1])$ and $A_N = 1$, as well as the inequality

$$
\ln(1+u) \equiv -\ln\left(1 - \frac{u}{1+u}\right) \ge \frac{u}{1+u}, \tag{A.23}
$$

we have

$$
\begin{aligned}
L_n(\boldsymbol{z}) &= \sum_{i=1}^{N} \frac{a_i}{A_i} && \text{(A.24)}\\
&= 1 + \sum_{i=2}^{N} \frac{a_i}{A_{i-1} + a_i} && \text{(A.25)}\\
&= 1 + \sum_{i=2}^{N} \frac{a_i/A_{i-1}}{1 + a_i/A_{i-1}} && \text{(A.26)}\\
&\le 1 + \sum_{i=2}^{N} \ln\left(1 + \frac{a_i}{A_{i-1}}\right) && \text{(A.27)}\\
&= 1 + \sum_{i=2}^{N} \ln\left(\frac{A_{i-1} + a_i}{A_{i-1}}\right) && \text{(A.28)}\\
&= 1 + \sum_{i=2}^{N} \ln\left(\frac{A_i}{A_{i-1}}\right) && \text{(A.29)}\\
&= 1 + \ln\left(\frac{A_N}{A_1}\right) && \text{(A.30)}
\end{aligned}
$$

$$
\begin{aligned}
&= \ln\left[\frac{1}{P(\boldsymbol{y}[1])}\right] + 1 &\text{(A.31)}\\
&\leq \ln\left[\frac{1}{(\pi_{\min}\cdot|\Theta|\cdot|\Omega|)^n}\right] + 1 &\text{(A.32)}\\
&= n\ln\left(\frac{1}{\pi_{\min}\cdot|\Theta|\cdot|\Omega|}\right) + 1, &\text{(A.33)}
\end{aligned}
$$

where we have used the assumption (7), which implies that $P(\boldsymbol{y}) \geq (\pi_{\min}\cdot|\Theta|\cdot|\Omega|)^n$ for all $\boldsymbol{y}$. This completes the proof of Lemma 1.

## A.2 Proof of Eq. (43)

We next show that for every $\boldsymbol{y}$ and $\boldsymbol{s}$,

$$P(\boldsymbol{y}) \leq P(\boldsymbol{y},\boldsymbol{s})\cdot\left(\frac{|\Theta\times\Omega|}{\pi_{\min}^2}\right)^c. \tag{A.34}$$

For the sake of brevity, let us denote $\zeta_i = (\theta_i, \omega_i)$ (so that $s_i = \zeta_{n_i}$). Now,

$$P(\boldsymbol{y},\boldsymbol{s}) = \prod_{i=0}^{c-1}\pi(y_{n_i+1}^{n_{i+1}}, \zeta_{n_{i+1}}|\zeta_{n_i}). \tag{A.35}$$

But

$$
\begin{aligned}
\pi(y_{n_i+1}^{n_{i+1}}, \zeta_{n_{i+1}}|\zeta_{n_i}) &= \sum_{\zeta_{n_i+1}^{n_{i+1}-1}}\prod_{t=n_i+1}^{n_{i+1}}\pi(y_t,\zeta_t|\zeta_{t-1}) &\text{(A.36)}\\
&= \sum_{\zeta_{n_i+1}}\pi(y_{n_i+1},\zeta_{n_i+1}|\zeta_{n_i})\times\\
&\quad \sum_{\zeta_{n_i+2}^{n_{i+1}-2}}\prod_{t=n_i+2}^{n_{i+1}-1}\pi(y_t,\zeta_t|\zeta_{t-1})\times\\
&\quad \sum_{\zeta_{n_{i+1}-1}}\pi(y_{n_{i+1}},\zeta_{n_{i+1}}|\zeta_{n_{i+1}-1}) &\text{(A.37)}\\
&\geq \pi_{\min}^2\sum_{\zeta_{n_i+1}^{n_{i+1}-1}}\prod_{t=n_i+2}^{n_{i+1}-1}\pi(y_t,\zeta_t|\zeta_{t-1}), &\text{(A.38)}
\end{aligned}
$$

where we have assumed that $n_i + 2 \leq n_{i+1} - 1$, which means that the phrase length must be at least three,[9] and where we have lower bounded both $\pi(y_{n_i+1},\zeta_{n_i+1}|\zeta_{n_i})$ and $\pi(y_{n_{i+1}},\zeta_{n_{i+1}}|\zeta_{n_{i+1}-1})$ by $\pi_{\min}$. Similarly, since both $\pi(y_{n_i+1},\zeta_{n_i+1}|\zeta_{n_i})$ and $\pi(y_{n_{i+1}},\zeta_{n_{i+1}}|\zeta_{n_{i+1}-1})$ are upper bounded by unity, we have

$$\pi(y_{n_i+1}^{n_{i+1}}, \zeta_{n_{i+1}}|\zeta_{n_i}) \leq \sum_{\zeta_{n_i+1}^{n_{i+1}-1}}\prod_{t=n_i+2}^{n_{i+1}-1}\pi(y_t,\zeta_t|\zeta_{t-1}). \tag{A.39}$$

[9] This assumption does not affect the generality, as the number of phrases of length shorter than three cannot exceed $|\mathcal{Y}\times\mathcal{Z}| + |\mathcal{Y}\times\mathcal{Z}|^2$, which is fixed and hence negligible compared to the total number of phrases for large $n$.

Since the expression

$$\sum_{\zeta_{n_i+1}^{n_{i+1}-1}} \prod_{t=n_i+2}^{n_{i+1}-1} \pi(y_t, \zeta_t | \zeta_{t-1})$$

depends neither on $\zeta_{n_i}$ nor on $\zeta_{n_{i+1}}$, it follows that for any $\zeta_{n_i}$, $\zeta'_{n_i}$, $\zeta_{n_{i+1}}$, and $\zeta'_{n_{i+1}}$,

$$\pi_{\min}^2 \le \frac{\pi(y_{n_i+1}^{n_{i+1}}, \zeta'_{n_{i+1}} | \zeta'_{n_i})}{\pi(y_{n_i+1}^{n_{i+1}}, \zeta_{n_{i+1}} | \zeta_{n_i})} \le \frac{1}{\pi_{\min}^2}, \tag{A.40}$$

and so,

$$\begin{aligned}
P(\boldsymbol{y}) &= \sum_{\boldsymbol{s}'} P(\boldsymbol{y}, \boldsymbol{s}') && \text{(A.41)}\\
&= P(\boldsymbol{y}, \boldsymbol{s}) \sum_{\boldsymbol{s}'} \frac{P(\boldsymbol{y}, \boldsymbol{s}')}{P(\boldsymbol{y}, \boldsymbol{s})} && \text{(A.42)}\\
&= P(\boldsymbol{y}, \boldsymbol{s}) \sum_{\boldsymbol{s}'} \prod_{i=0}^{c-1} \frac{\pi(y_{n_i+1}^{n_{i+1}}, \zeta'_{n_{i+1}} | \zeta'_{n_i})}{\pi(y_{n_i+1}^{n_{i+1}}, \zeta_{n_{i+1}} | \zeta_{n_i})} && \text{(A.43)}\\
&\le P(\boldsymbol{y}, \boldsymbol{s}) \sum_{\boldsymbol{s}'} \prod_{i=0}^{c-1} \frac{1}{\pi_{\min}^2} && \text{(A.44)}\\
&= P(\boldsymbol{y}, \boldsymbol{s}) \cdot \left( \frac{|\Omega \times \Theta|}{\pi_{\min}^2} \right)^c, && \text{(A.45)}
\end{aligned}$$

which completes the proof of eq. (43).

## References

[1] L. E. Baum and T. Petrie, ”Statistical inference for probabilistic functions of finite state Markov chains,” *Ann. Math. Statist.*, vol. 37, pp. 1554–1563, 1966.

[2] I. Csiszár, ”Linear codes for sources and source networks: error exponents, universal coding,” *IEEE Trans. Inform. Theory*, vol. IT–28, no. 4, pp. 585–592, July 1982.

[3] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Cambridge University Press, 2011.

[4] G. Dasarathy and S. C. Draper, ”On reliability of content identification from databases based on noisy queries,” *The 2011 IEEE Proc. International Symposium on Information Theory (ISIT 2011)*, pp. 1066–1070, St. Petersburg, Russia, July–August 2011.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, ”Maximum likelihood from incomplete data via the EM algorithm,” *J. Roy. Statist. Soc.*, vol. B39, pp. 1–39, 1977.

[6] Y. Ephraim and N. Merhav, ”Hidden Markov processes,” *IEEE Trans. Inform. Theory*, special issue in memory of Aaron D. Wyner, June 2002

[7] M. Feder and A. Lapidoth, ”Universal decoding for channels with memory,” *IEEE Trans. Inform. Theory*, vol. 44, no. 5, pp. 1726–1745, September 1998.

[8] M. Feder and N. Merhav, ”Universal composite hypothesis testing: a competitive minimax approach,” *IEEE Trans. Inform. Theory*, special issue in memory of Aaron D. Wyner, vol. 48, no. 6, pp. 1504–1517, June 2002.

[9] V. D. Goppa, ”Nonprobabilistic mutual information without memory,” *Probl. Cont. Information Theory*, vol. 4, pp. 97–102, 1975.

[10] T. Ignatenko and F. M. J. Willems, ”Biometric security from an information–theoretical perspective,” *Foundations and Trends in Communications and Information Theory*, vol. 7, nos. 2–3, pp. 135–316.

[11] A. Lapidoth and J. Ziv, ”On the universality of the LZ–based noisy channels decoding algorithm,” *IEEE Trans. Inform. Theory*, vol. 44, no. 5, pp. 1746–1755, September 1998.

[12] Y. Lomnitz and M. Feder, ”Communication over individual channels – a general framework,” *IEEE Trans. Inform. Theory*, vol. 57, no. 11, pp. 7333–7358, November 2011.

[13] N. Merhav, ”Universal decoding for memoryless Gaussian channels with a deterministic interference,” *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1261–1269, July 1993.

[14] N. Merhav, ”Universal decoding for arbitrary channels relative to a given family of decoding metrics,” *IEEE Trans. Inform. Theory*, vol. 59, no. 9, pp. 5566–5576, September 2013.

[15] T. Petrie, ”Probabilistic functions of finite state Markov chains,” *Ann. Math. Statist.*, vol. 40, no. 1, pp. 97–115, 1969.

[16] L. R. Rabiner, ”A tutorial of hidden Markov models and selected applications in speech recognition”, *Proceedings of the IEEE*, vol. 77, no. 2, February 1989.

[17] E. Tuncel, ”Capacity/storage tradeoff in high–dimensional identification systems,” *IEEE Trans. Inform. Theory*, vol. 55, no. 5, pp. 2097–2106, May 2009.

[18] F. Willems, T. Kalker, J. Goseling, and J.-P. Linnartz, ”On the capacity of a biometrical identification system,” *The 2003 IEEE Proc. International Symposium on Information Theory (ISIT 2003)*, p. 82, Yokohama, Japan, June–July 2003.

[19] F. Willems, T. Kalker, S. Baggen, and J.-P. Linnartz, ”On the capacity of a biometrical identification system,” (unknown year) available on–line at:
`http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.74.9512&rep=rep1&type=pdf`

[20] J. Ziv, ”Universal decoding for finite–state channels,” *IEEE Trans. Inform. Theory*, vol. IT–31, no. 4, pp. 453–460, July 1985.

[21] J. Ziv and A. Lempel, “Compression of individual sequences via variable-rate coding,” *IEEE Trans. Inform. Theory*, vol. IT–24, no. 5, pp. 530–536, September 1978.

[22] J. Ziv and N. Merhav, ”Estimating the number of states of a finite–state source,” *IEEE Trans. Inform. Theory*, vol. 38, no. 1, pp. 61–65, January 1992.