Universal Decoding for Channels with Memory

Meir Feder, *Senior Member, IEEE*, and Amos Lapidoth, *Member, IEEE*

***Abstract*— A universal decoder for a parametric family of channels is a decoder whose structure depends on the family but not on the individual channel over which transmission takes place, and it yet attains the same random-coding error exponent as the maximum-likelihood receiver tuned to the channel in use. The existence and structure of such decoders is demonstrated under relatively mild conditions of continuity of the channel law with respect to the parameter indexing the family. It is further shown that under somewhat stronger conditions on the family of channels, the convergence of the performance of the universal decoder to that of the optimal decoder is uniform over the set of channels. Examples of families for which universal decoding is demonstrated include the family of finite-state channels and the family of Gaussian intersymbol interference channels.**

***Index Terms*—Compound channel, error exponent, finite-state channel, Gilbert–Elliott channel, intersymbol interference, random coding, universal decoding.**

I. Introduction and Definitions

THIS paper addresses the problem of designing a receiver for digital communication over an unknown channel. The channel over which transmission is to be carried out is unknown to the receiver designer, and the designer only knows that the channel belongs to some family of channels

$$\mathcal{F} = \{p_\theta(\boldsymbol{y} \mid \boldsymbol{x}), \theta \in \Theta\} \tag{1}$$

where Θ is some index set. Had the channel been known in advance, the designer could have used the maximum-likelihood (ML) decoding rule to minimize the average probability of error. This rule, however, cannot be used in our scenario as it typically depends on the channel law, and the ML decoding rule is thus typically different for different members of the family $\mathcal{F}$.

In spite of the above, we shall show in this paper that under fairly mild conditions on the family of channels $\mathcal{F}$, there exists a universal decoder for $\mathcal{F}$ that performs asymptotically as well as the ML decoder and yet does not require knowledge of the channel over which transmission is carried out. The proposed decoder thus not only competes favorably with other detectors that are ignorant of the channel over which transmission is carried out, but even performs asymptotically as well as the best decoder that could have been designed had the channel law been known.

It should be stressed that no prior distribution is assumed on $\theta \in \Theta$, and the universal decoder is required to perform asymptotically as well as the ML decoder on any channel $\theta \in \Theta$.

Before we define asymptotic performance and in order to motivate the definition, we shall first briefly describe the use of training sequences to facilitate communication over an unknown channel, a use which is very common in many wireless systems [1], [2]. In order to help the receiver identify the channel in use, the transmitter sends a known sequence of symbols over the channel. This known input sequence is called "training sequence." Since the sequence is known at the receiver, the receiver can estimate the channel law by studying the statistics of the received symbols corresponding to the known input sequence. The receiver then typically decodes the rest of the transmission by performing ML decoding with respect to the estimated channel law. It should be stressed that the transmitter itself does not know the channel law and cannot therefore convey this information to the receiver.

The use of training sequences has some drawbacks. First, there is a mismatch penalty. Because the training sequences are of limited length, the channel estimate formed at the receiver is imprecise, and the data sequence is thus decoded according to an incorrect likelihood function. This results in an increase in error rates [3], [4] and in a decrease in capacity [5]–[10]. Secondly, there is a penalty in throughput, because the training sequences carry no information. This penalty is of course worse the longer the training sequence is as compared to the length of the data sequence. We thus see that increasing the length of the training sequences results in a hit in throughput, whereas decreasing its length reduces the accuracy of the channel estimation and thus results in a more severe loss in error rates and in the capacity due to the decoding mismatch.

To overcome this tradeoff one might wish to choose the length of the sequence sufficiently large to ensure precise channel estimation, and then choose the data block sufficiently long so as to make the loss in throughput small. This approach, however, seldom works due to delay constraints, as it results in a large delay that the data symbols suffer. This tradeoff between delay and error rates motivates the definition of a universal decoder as one that attains the same asymptotic tradeoff between delay and error rates as the optimal ML receiver.

For most channels of interest, including memoryless channels and indecomposable finite-state channels [11], the best tradeoff between achievable error rates and delay (as measured

Manuscript received December 11, 1996; revised January 28, 1998. The work of M. Feder was supported in part under a grant from the Israeli Science Foundation. The work of A. Lapidoth was supported in part by the Advanced Concepts Committee, Lincoln Laboratory, and by the NSF Faculty Early Career Development (CAREER) Program.The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Ulm, Germany, June 29–July 4, 1997.

M. Feder is with the Department of Electrical Engineering–Systems, Tel-Aviv University, Tel-Aviv 69978, Israel.

A. Lapidoth is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139-4307 USA.

Publisher Item Identifier S 0018-9448(98)05123-2.

by blocklength) when ML decoding is employed is exponential, with the error rate decreasing exponentially with the delay (blocklength) n, where the exponent depends on the channel law and on the rate of transmission, and is typically positive for rates below channel capacity. While finding codes that achieve this performance is typically very difficult, one can often demonstrate their existence by a random-coding argument, i.e., by showing that the average (over codebooks and messages) probability of error of a randomly chosen codebook can exhibit a good exponential tradeoff between error rates and delay.

With these observations in mind, we define a universal sequence of decoders as a sequence of decoders that achieves the same random-coding error exponent as the ML decoder, for every channel in the family. To make this more precise we need the following setup.

Consider a family of channels (1) defined over the common input alphabet $\mathcal{X}$ and the common output alphabet $\mathcal{Y}$. For any $\theta \in \Theta$ the law $p_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ maps every input sequence

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

to a corresponding probability law on $\mathcal{Y}^n$. Notice that we are omitting the dependence on the blocklength n: strictly speaking, $p_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ is thus a sequence of mappings, one for each blocklength n.

Given a rate-R blocklength-n codebook

$$\mathcal{C} = \{\boldsymbol{x}(1), \cdots, \boldsymbol{x}(\lfloor 2^{nR} \rfloor)\} \subset \mathcal{X}^n \tag{2}$$

a decoder ϕ is a mapping

$$\phi : \mathcal{Y}^n \to \{1, \cdots, \lfloor 2^{nR} \rfloor\}$$

that maps every received sequence $\boldsymbol{y} \in \mathcal{Y}^n$ to an index i of some codeword. Strictly speaking, the mapping ϕ depends, of course, not only on the received sequence but also on the codebook, but to avoid cumbersome notation we do not make this explicit. It should however be noted that throughout this paper we assume that the codebook, even when drawn at random, is known to both transmitter and receiver, and that the decoding is allowed, and indeed should, depend on the codebook.

If all the codewords of a code $\mathcal{C}$ are used equiprobably (as we shall assume throughout) then the average (over messages) probability of error $P_{\theta,\phi}(\text{error} \mid \mathcal{C})$ incurred when the codebook $\mathcal{C}$ is used over the channel $p_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ with the decoder ϕ, is given by

$$P_{\theta,\phi}(\text{error} \mid \mathcal{C}) = \frac{1}{\lfloor 2^{nR} \rfloor} \sum_{i=1}^{\lfloor 2^{nR} \rfloor} \sum_{\{\boldsymbol{y}: \phi(\boldsymbol{y}) \neq i\}} p_\theta(\boldsymbol{y} \mid \boldsymbol{x}(i)). \tag{3}$$

When random coding is considered, the codebook $\mathcal{C}$ is drawn at random by choosing its codewords independently and uniformly[1] over some set $B_n \subset \mathcal{X}^n$. The set B_n will be referred to as the *input set*. We shall let $\bar{P}_{\theta,\phi}(\text{error})$ denote the average (over messages and codebooks) probability of error that is incurred when such a random codebook is used over the channel $p_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ and is decoded using the decoder ϕ. In other words, $\bar{P}_{\theta,\phi}(\text{error})$ is just the average of $P_{\theta,\phi}(\text{error} \mid \mathcal{C})$ over the choice of codebooks $\mathcal{C}$.

Given a known channel $p_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ and a codebook $\mathcal{C}$, the decoder that minimizes the average probability of error is the ML decoder [13]. A decoder ϕ is said to be ML for the channel $p_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ if

$$\phi(\boldsymbol{y}) = i \Rightarrow p_\theta(\boldsymbol{y} \mid \boldsymbol{x}(i)) = \max_{1 \le j \le \lfloor 2^{nR} \rfloor} p_\theta(\boldsymbol{y} \mid \boldsymbol{x}(j)). \tag{4}$$

Notice that the ML decoder is not unique as different ML receivers may resolve ties in the likelihood function in different ways. All ML receivers, however, give rise to the same average probability of error for any code $\mathcal{C}$. We denote this average probability of error by $P_{\theta,\theta}(\text{error} \mid \mathcal{C})$. Thus $P_{\theta,\theta}(\text{error} \mid \mathcal{C})$ is the average (over messages) probability of error incurred when the codebook $\mathcal{C}$ is used over the channel $p_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ and ML decoding tuned to θ is employed. We similarly use $\bar{P}_{\theta,\theta}(\text{error})$ to denote the analogous expression for the average (over messages and codebooks) probability of error for a randomly chosen codebook.

We are now in a position to define weak random-coding universality, and to make precise the notion that the universal decoder performs asymptotically as well as the ML receiver tuned to the channel in use.

Definition 1: A sequence of decoders $\{u_n\}$ is said to be *random-coding universal* (or *random-coding weakly universal*) for the family $\{p_\theta(\boldsymbol{y} \mid \boldsymbol{x}), \theta \in \Theta\}$ and the input-set sequence $\{B_n\}$, $B_n \subset \mathcal{X}^n$, if

$$\lim_{n \to \infty} \frac{1}{n} \log \left(\frac{\bar{P}_{\theta,u_n}(\text{error})}{\bar{P}_{\theta,\theta}(\text{error})} \right) = 0, \qquad \forall \theta \in \Theta. \tag{5}$$

Notice that in our definition of a weak random-coding universal decoder we do not require that the decoder attain the same asymptotic performance as the ML decoder for *any* code. This requirement is too restrictive, as there are some codes that cannot be decoded universally even in well-behaved families of channels. For example, if $\mathcal{F}$ is the family of all binary-symmetric channels (BSC) with crossover probability $\theta \in [0, 1]$then, as we shall show later, a weak random-coding universal decoder can be found, and yet there are some singular codes that are not amenable to universal decoding. Indeed, any binary code that is closed under Hamming complement (component-wise negation) is not amenable to reliable universal decoding.

We will, however, show that while not every code is amenable to universal decoding, there are some very good codes that are. More specifically, we will show that under relatively mild regularity conditions on the family of channels one can approach the random-coding error exponent (error-rate versus delay) with sequences of (deterministic) codes that are amenable to universal decoding. This motivates the following definition of weak deterministic-coding universal decoders.

Definition 2: A sequence of decoders u_n is said to be *deterministic-coding universal* (or *deterministic-coding weakly*

[1] Throughout this paper we restrict ourselves to random coding where the codewords are drawn uniformly over the input set B_n, thus excluding independent and identically distributed (i.i.d.) random coding. However, since B_n can be arbitrary and could, for example, be the set of all sequences of a given type, there is no loss in optimality in this restriction; see [12].

universal) for the family $\{p_\theta(\boldsymbol{y} \mid \boldsymbol{x}), \theta \in \Theta\}$ and the input-set sequence $\{B_n\}$ if there exists a sequence of rate-R blocklength-n codebooks $\{\mathcal{C}_n\}$, $\mathcal{C}_n \subset B_n$, such that

$$\lim_{n\to\infty} \frac{1}{n} \log\left(\frac{P_{\theta,u_n}(\text{error} \mid \mathcal{C}_n)}{\bar{P}_{\theta,\theta}(\text{error})}\right) = 0, \qquad \forall \theta \in \Theta. \tag{6}$$

It is interesting to note that even for very simple families of channels, the training sequence approach is not universal. For example, it is shown in Appendix I that even if the family of channels consists of only two channels, say a binary-symmetric channel with crossover probability 0.25 and a binary-symmetric channel with crossover probability 0.75, the training sequence approach is not universal. The reason is that unless the receiver correctly identifies the channel in use, it is almost bound to err, and for the receiver to identify the channel with exponentially small probability of error the length of training sequence must be linear in the blocklength, resulting in a loss in the error exponent.

The issue of universal decoding is intimately related to the problem of determining the compound channel capacity of a family of channels [14]–[17]. A rate R is said to be achievable for the family of channels $\mathcal{F}$ if for any given $\epsilon > 0$ and every sufficiently large blocklength n there exists a blocklength-n rate-R codebook $\mathcal{C}_n$ and a decoder ϕ_n such that

$$\sup_{\theta\in\Theta} P_{\theta,\phi_n}(\text{error} \mid \mathcal{C}_n) < \epsilon.$$

The compound channel capacity $C(\mathcal{F})$ of the family $\mathcal{F}$ is defined as the supremum of all achievable rates.

In a certain sense, finding the sequence of decoders ϕ_n for the compound channel is easier than finding a sequence of universal decoders because in the definition of the compound channel capacity no attention is paid to error exponents: for example, if the family of channels $\mathcal{F}$ is a subset of the class of discrete memoryless channels (DMC) then a training sequence approach to the problem will probably work. On the other hand, the requirements on the decoders for the compound channel are more stringent since ϕ_n must have uniformly good performance over all channels in the family. With the compound channel in mind we thus define the notion of strong universality. The adjective "strong" refers to the uniformity of the convergence. Once again we distinguish between random-coding universality and deterministic-coding universality:

Definition 3: A sequence of decoders $\{u_n\}$ is said to be *random-coding strongly universal* for the family $\{P_\theta(y|x), \theta\in\Theta\}$ and the input sets $\{B_n\}$ if the convergence (5) is uniform over Θ, i.e., if

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \frac{1}{n} \log\left(\frac{\bar{P}_{\theta,u_n}(\text{error})}{\bar{P}_{\theta,\theta}(\text{error})}\right) = 0. \tag{7}$$

Definition 4: The sequence of decoders $\{u_n\}$ is said to be *deterministic-coding strongly universal* for the family $\{P_\theta(y|x), \theta \in \Theta\}$ and the input sets $\{B_n\}$ if there exists a sequence of rate-R blocklength-n codebooks $\{\mathcal{C}_n\}$, $\mathcal{C}_n \subset B_n$, for which the convergence in (6) is uniform over θ, i.e.,

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \frac{1}{n} \log\left(\frac{P_{\theta,u_n}(\text{error} \mid \mathcal{C}_n)}{\bar{P}_{\theta,\theta}(\text{error})}\right) = 0. \tag{8}$$

We shall demonstrate in Theorem 2 that under fairly mild conditions on the family of channels, one can demonstrate strong deterministic-coding universality. Once such universality is established, the achievability of a rate R for the compound channel $\mathcal{F}$ can be demonstrated by showing that

$$\liminf_{n\to\infty} -\frac{1}{n} \log \sup_{\theta\in\Theta} \bar{P}_{\theta,\theta}(\text{error}) > 0.$$

Notice that the above expression involves only random coding (and not specific codes), and more importantly, it only involves optimal ML decoding.

This approach to the compound channel is explored in [12] where it is used to compute the compound channel capacity of a class of finite-state channels (FSC), a class of channels that, as we shall show, admits strong deterministic-coding universality.

Note that a receiver need not be strongly universal in order to achieve the compound channel capacity of a family. For example, if $\mathcal{F}$ is a convex family of memoryless channels, then the compound channel capacity of the family can be achieved using the ML receiver tuned to the channel that achieves the saddle-point for the mutual information functional [7], [18]. On other channels in the family, however, this decoder does not typically attain the same random-coding error exponent as the ML decoder, and this decoder is thus not universal by our definition.

Our various definitions of universal decoding and our approach to the problem have been influenced by previous work on the problem, and particularly by [16] and [19]. In the former work the problem of universal decoding is studied for memoryless channels over finite input and output alphabets, and the definition of universality is very close in nature to what we refer to as "strong deterministic-coding universality." It is shown there that the maximum (empirical) mutual information (MMI) decoding rule, first suggested by Goppa [20], is strongly deterministic-coding universal for any family of memoryless channels defined over finite input and output alphabets. If the family $\mathcal{F}$ consists of the family of *all* discrete memoryless channels over the alphabets $\mathcal{X}$ and $\mathcal{Y}$, then the MMI algorithm is equivalent to a generalized ML decoding rule where given a received sequence $\boldsymbol{y}$, the codeword $\boldsymbol{x}(i)$ receives the score $\sup_{\theta\in\Theta} p_\theta(\boldsymbol{y} \mid \boldsymbol{x}(i))$.

In [19], Ziv studied universal decoding for the class of finite-state channels where the next state is a deterministic—but unknown—function of the previous state, input, and output. For this family of channels Ziv proved that if random coding is carried out by choosing the codewords independently and uniformly over the set of sequences of a given composition (type), then one can find a strong random-coding universal decoder for the family[2]. The decoder that Ziv proposed is based on the Lempel-Ziv algorithm from source coding. Additional work on universal decoding appeared in [21] where weak random-coding universality was demonstrated for a family of memoryless Gaussian channels with an unknown deterministic interference of a special parametric form.

[2] Ziv only claimed weak random-coding universality, but his proof demonstrates strong random-coding universality.

Our work extends the previous work on universal decoding in several ways. First, we study universal decoding not only for DMC's, as in [16], but also for channels with memory. Our results are fairly general and include the family of all finite-state channels [11], [22], and not only those with deterministic transitions, which were studied in [19]. In particular, our results are valid for the class of all Gilbert–Elliott channels [23]–[26], which have random transitions and are often used to model time-varying channels. In addition, we do not require that the benchmark random coding be done over the set of sequences of a given type as in [19]: as long as the codewords are chosen uniformly over some set B_n, this set can be arbitrary. This generalization can be important for channels for which the input distribution that achieves capacity is not independent and identically distributed (i.i.d.). Also, the universality that we demonstrate is not only strong random-coding universality as in [19] but also strong deterministic-coding universality. Our results also extend to more general families of channels, including those with infinite input and output alphabets. For example, we show that the set of all additive Gaussian noise intersymbol interference (ISI) channels with a fixed number of ISI terms of bounded L_2 norm admits strong universal decoding; this problem was posed in [21].

Notice that as in [19] we only consider random coding in which the codewords are drawn independently and uniformly over some input set. In this respect our analysis excludes the classical random-coding approach where the components of each codeword are drawn independently according to some marginal distribution $Q(x)$, [11]. For most applications this is not a serious drawback as the random-coding error exponents that are achieved by choosing the codewords uniformly over a type are usually no worse than those achieved by choosing the codewords according to the product distribution corresponding to that type, see [27] for the Gaussian case and [12] for the more general case.

In some sense, the problem of universal channel decoding is dual to the problem of universal coding for sources of unknown law. It should, however, be noted that no feedback link is assumed in our problem, and the transmitter cannot therefore use a signaling scheme that depends on the channel in use. That is why we cannot typically hope to communicate at channel capacity (of the channel in use), since different channels in the family will typically have different capacities and different capacity-achieving input distributions.

The rest of the paper is organized as follows. In the next section we state the paper's main results. In Section III we discuss how ML decoders can be described using ranking functions and how every ranking naturally defines a decoder. The main result of that section is a description of how a finite number of different decoders (ranking functions) can be merged to obtain a new decoder that performs almost as well as each of those decoders, see Lemma 1. This construction plays a crucial role in the proof of the existence of weak universal decoders, which are treated in Section IV. Strong universal decoders are studied in Section V. All these sections deal with the finite-alphabet case, and in Section VI we extend these results to the infinite-alphabet case. Section VII contains some applications of the results to specific families of channels, particularly the family of DMC's, finite-state channels, and intersymbol interference channels. That section also describes an example of a family of channels that admits weak universal decoding but not strong universal decoding. The paper is concluded with a brief summary and discussion in Section VIII.

II. The Main Results

Before we can state the main result on weak universality we need the following definition of a separable family. Loosely speaking, a family is separable if there exists a countable set $\{\theta_k\}$ that is "dense" in Θ in a sense that is made precise next.

Definition 5: We shall say that the family of channels (1) is (weakly) *separable* for the input sets $\{B_n\}$, $B_n \subseteq \mathcal{X}^n$, if there exists a sequence $\{\theta_k\}_{k=1}^{\infty} \subseteq \Theta$ that is "dense" in the family in the sense that

$$\inf_k \limsup_{n\to\infty} \sup_{(\boldsymbol{x},\boldsymbol{y})\in B_n\times\mathcal{Y}^n} \frac{1}{n}\left|\log\left(\frac{p_\theta(\boldsymbol{y}\mid\boldsymbol{x})}{p_{\theta_k}(\boldsymbol{y}\mid\boldsymbol{x})}\right)\right| = 0, \quad \forall\theta\in\Theta. \tag{9}$$

In other words, for every $\theta \in \Theta$ and every $\epsilon > 0$, there exists some k^* and some n_0 so that for all $n \geq n_0$

$$\sup_{(\boldsymbol{x},\boldsymbol{y})\in B_n\times\mathcal{Y}^n} \frac{1}{n}\left|\log\left(\frac{p_\theta(\boldsymbol{y}\mid\boldsymbol{x})}{p_{\theta_k}(\boldsymbol{y}\mid\boldsymbol{x})}\right)\right| < \epsilon.$$

The following theorem demonstrates that if the family of channels is separable for the input sets $\{B_n\}$, then there exist weak random-coding and weak deterministic-coding universal decoders for the family.

Theorem 1: If a family of channels (1) defined over common finite input and output alphabets $\mathcal{X}, \mathcal{Y}$ is separable for the input sets $\{B_n\}$, then there exists a sequence of decoders $\{u_n\}$ that are random-coding and deterministic-coding universal for the family. Thus

$$\lim_{n\to\infty}\frac{1}{n}\log\left(\frac{\bar{P}_{\theta,u_n}(\text{error})}{\bar{P}_{\theta,\theta}(\text{error})}\right) = 0, \qquad \forall\theta\in\Theta$$

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

$$\lim_{n\to\infty}\frac{1}{n}\log\left(\frac{P_{\theta,u_n}(\text{error}\mid\mathcal{C}_n)}{\bar{P}_{\theta,\theta}(\text{error})}\right) = 0, \qquad \forall\theta\in\Theta.$$

The separability condition is not enough to guarantee the existence of strong universal decoders, as demonstrated in Section VII-D. For this we need a stronger notion, which we have termed "strong separability." Loosely speaking, a family is strongly separable if for any blocklength n there exists a subexponential number $K(n)$ of channels such that the law of any channel in the family can be approximated by one of these channels. The approximation is in the sense that except for rare sequences, the normalized log-likelihood of an output sequence given any input sequence is similar under the two channels. More precisely

Definition 6: A family of channels $\{p_\theta(\boldsymbol{y} \mid \boldsymbol{x}) \ \theta \in \Theta\}$ defined over common finite input and output alphabets $\mathcal{X}, \mathcal{Y}$ is said to be *strongly separable* for the input sets $\{B_n\}$, $B_n \subseteq \mathcal{X}^n$, if there exists some $M > 0$ that upper-bounds the error exponents in the family, i.e., that satisfies

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta} -\frac{1}{n} \log \bar{P}_{\theta,\theta}(\text{error}) < M \tag{10}$$

such that for every $\epsilon > 0$ and blocklength n, there exists a subexponential number $K(n)$ (that may depend on M and on ϵ) of channels $\{\theta_k^{(n)}\}_{k=1}^{K(n)} \subseteq \Theta$

$$\lim_{n\to\infty} \frac{1}{n} \log K(n) = 0 \tag{11}$$

that well approximate any $\theta \in \Theta$ in the following sense: For any $\theta \in \Theta$ there exists $\theta_{k^*}^{(n)} \in \Theta, 1 \le k^* \le K(n)$, so that

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{x}) \le 2^{n\epsilon} p_{\theta_{k^*}^{(n)}}(\boldsymbol{y} \mid \boldsymbol{x}), \quad \forall \boldsymbol{x}, \boldsymbol{y} : p_\theta(\boldsymbol{y} \mid \boldsymbol{x}) > 2^{-n(M+\log|\mathcal{Y}|)} \tag{12}$$

and

$$p_{\theta_{k^*}^{(n)}}(\boldsymbol{y} \mid \boldsymbol{x}) \le 2^{n\epsilon} p_\theta(\boldsymbol{y} \mid \boldsymbol{x}), \quad \forall \boldsymbol{x}, \boldsymbol{y} : p_{\theta_{k^*}^{(n)}}(\boldsymbol{y} \mid \boldsymbol{x}) > 2^{-n(M+\log|\mathcal{Y}|)}. \tag{13}$$

A good candidate for M is $1+\log|\mathcal{X}|$ as $\bar{P}_{\theta,\theta}(\text{error})$ is lower-bounded by the random-coding pairwise error probability (the probability of error corresponding to the case where the codebook consists of only two codewords) and the latter is lower-bounded by $|\mathcal{X}|^{-n}$ corresponding to the probability that the two codewords are identical. Note that we assume throughout that if the transmitted codeword and some other codeword are identical then an error results.

Theorem 2: If a family of channels (1) defined over common finite input and output alphabets $\mathcal{X}, \mathcal{Y}$ is strongly separable for the input sets $\{B_n\}$, then there exists a sequence of decoders $\{u_n\}$ that are random-coding and deterministic-coding strongly universal for the family. Thus

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \frac{1}{n} \log\left(\frac{\bar{P}_{\theta,u_n}(\text{error})}{\bar{P}_{\theta,\theta}(\text{error})}\right) = 0$$

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

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \frac{1}{n} \log\left(\frac{P_{\theta,u_n}(\text{error} \mid \mathcal{C}_n)}{\bar{P}_{\theta,\theta}(\text{error})}\right) = 0.$$

Many of the families of channels arising in digital communications are strongly separable, and thus admit strong universal decoding. We shall, for example, show that in addition to the class of all discrete memoryless channels over finite alphabets, the set of all finite-state channels [11] defined over finite common input, output, and state alphabets $\mathcal{X}, \mathcal{Y}, \mathcal{S}$, respectively, is strongly separable. We shall thus deduce from Theorem 2 the following

Theorem 3: The set of all finite-state channels defined over common finite input, output, and state alphabets $\mathcal{X}, \mathcal{Y}, \mathcal{S}$, and parameterized by the pair of stochastic matrices $P_\theta(y, s \mid x, s')$ and initial states $s_0 \in \mathcal{S}$ where

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{x}, s_0) = \sum_{\boldsymbol{s}\in\mathcal{S}^n} p_\theta(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0)$$

and

$$p_\theta(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0) = \prod_{t=1}^{n} P_\theta(y_t, s_t \mid s_{t-1}, x_t)$$

admits strong deterministic-coding and random-coding universal decoding. Here $\boldsymbol{s} = (s_1, \cdots, s_n)$ and it is important to note that the receiver is assumed ignorant of the state sequence.

If the number of states is finite but unknown, we can guarantee weak random-coding and deterministic-coding universal decoding.

Our results can be extended to infinite alphabets; see Section VI where we prove a theorem analogous to Theorem 2 for infinite alphabets. As a corollary, we can prove, for example, the following

Theorem 4: Consider the discrete-time Gaussian intersymbol interference (ISI) channel where the output Y_t at time t is given by

$$Y_t = \sum_{j=0}^{J} h_j X_{t-j} + Z_t$$

where X_t is the input at time t, the sequence $\{Z_t\}$ is a sequence of i.i.d. Normal random variables of mean zero and unit variance, and $(h_0, \cdots, h_J)$ are the ISI coefficients. Suppose that the ISI coefficients are unknown to the receiver, but that their number[3] $J+1$ and an upper bound H on their norm are known, i.e.,

$$\sum_{j=0}^{J} h_j^2 \le H. \tag{14}$$

If the input sets $\{B_n\}$ from which the codewords are drawn satisfy an average power constraint

$$\boldsymbol{x} \in B_n \Rightarrow \sum_{t=1}^{n} x_t^2 \le nP \tag{15}$$

then a strong random-coding and deterministic-coding universal decoder exists. If the number of ISI coefficients J or an upper bound on their norm H is unknown then we can only guarantee weak random-coding and deterministic-coding universality.

III. Merging Decoders

The ML decoder is not unique since ties in the likelihood function can be resolved in different ways without changing the average probability of error. Condition (4) does not therefore completely specify the decoding function. A more precise description of the ML decoder that also specifies the manner

[3] Since we do not require that h_J be nonzero, J may be overestimated, and in this sense the receiver only needs an upper bound on the ISI memory.

by which ties are resolved is as follows. Assume that all the codewords are in some set $B_n \subseteq \mathcal{X}^n$ of size $|B_n|$

$$\boldsymbol{x}(i) \in B_n, \qquad \forall 1 \le i \le \lfloor 2^{nR} \rfloor$$

and consider a ranking function

$$M_\theta \colon B_n \times \mathcal{Y}^n \to \{1, \cdots, |B_n|\}$$

that given every received sequence $\boldsymbol{y}$ maps the sequence $\boldsymbol{x} \in B_n$ to its ranking among all the sequences in B_n. The mapping $M_\theta(\cdot, \boldsymbol{y})$ thus specifies a complete order from 1 to $|B_n|$ on all the sequences in B_n, i.e., for any $\boldsymbol{y} \in \mathcal{Y}^n$ we have that $M_\theta(\cdot, \boldsymbol{y})$ is a one-to-one mapping of B_n onto $\{1, \cdots, |B_n|\}$. It is further assumed that $M_\theta(\boldsymbol{x}, \boldsymbol{y})$ ranks the sequences according to decreasing order of likelihood, i.e.,

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{x}) > p_\theta(\boldsymbol{y} \mid \boldsymbol{x}') \Rightarrow M_\theta(\boldsymbol{x}, \boldsymbol{y}) < M_\theta(\boldsymbol{x}', \boldsymbol{y}) \tag{16}$$

where the sequence most likely (given the received sequence $\boldsymbol{y}$) is ranked highest, i.e., its rank is 1. Given a codebook $\mathcal{C} \subset B_n$ the ML decoder ϕ_θ that is determined by the ranking function $M_\theta(\cdot, \cdot)$ and defined by

$$\phi_\theta(\boldsymbol{y}) = i \quad \text{iff } M_\theta(\boldsymbol{x}(i), \boldsymbol{y}) < M_\theta(\boldsymbol{x}(j), \boldsymbol{y}), \qquad \forall j \ne i. \tag{17}$$

(If no such i exists, as can only happen if some of the codewords are identical, we declare an error.) Thus given a received sequence $\boldsymbol{y}$, the ML receiver determined by $M_\theta(\cdot, \cdot)$ declares that the transmitted codeword was $\boldsymbol{x}(i)$ if $\boldsymbol{x}(i)$ maximizes $p_\theta(\boldsymbol{y} \mid \boldsymbol{x}(j))$ among all the codewords $\boldsymbol{x}(j)$ in $\mathcal{C}$, and in the case that this maximum is achieved by several codewords, it prefers the one that is ranked highest by $M_\theta(\cdot, \boldsymbol{y})$.

It should be noted that any ranking function $M_u(\boldsymbol{x}, \boldsymbol{y})$, i.e., any function

$$M_u \colon B_n \times \mathcal{Y}^n \to \{1, \cdots, |B_n|\}$$

such that for any $\boldsymbol{y} \in \mathcal{Y}^n$ the function $M_u(\cdot, \boldsymbol{y})$ is one-to-one and onto $\{1, \cdots, |B_n|\}$, defines a decoder u in a manner completely analogous with (17). Thus given a codebook $\mathcal{C} \subset B_n$ and given a received sequence $\boldsymbol{y} \in \mathcal{Y}^n$

$$u(\boldsymbol{y}) = i \quad \text{iff } M_u(\boldsymbol{x}(i), \boldsymbol{y}) < M_u(\boldsymbol{x}(j), \boldsymbol{y}), \qquad \forall j \ne i. \tag{18}$$

We shall find it important to study the performance that results when a codebook $\mathcal{C}$ is used over a channel $p_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ and is decoded using a mismatched ML receiver that is tuned to a different channel, say $p_{\theta'}(\boldsymbol{y} \mid \boldsymbol{x})$. Strictly speaking, the resulting average probability of error should, by (3), be denoted by $P_{\theta, \phi_{\theta'}}(\text{error} \mid \mathcal{C})$, however, to simplify notation, we denote this average probability of error by $P_{\theta,\theta'}(\text{error} \mid \mathcal{C})$ and the corresponding average probability of error averaged over randomly selected codebooks by $\bar{P}_{\theta,\theta'}(\text{error})$. Thus $P_{\theta,\theta'}(\text{error} \mid \mathcal{C})$ denotes the average (over messages) probability of error incurred when the codebook $\mathcal{C}$ is used over the channel $p_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ and is decoded using an ML decoder tuned to the channel $p_{\theta'}(\boldsymbol{y} \mid \boldsymbol{x})$; $\bar{P}_{\theta,\theta'}(\text{error})$ is similarly defined.

The following construction will play a central role in this study. Given K decoders $\phi_1, \cdots, \phi_K$ that are based on the ranking functions $M_{\phi_1}, \cdots, M_{\phi_K}$, as in (18), we can define the merged decoder u_K by constructing its ranking function $M_{u_K}(\cdot, \boldsymbol{y})$ in the following way: Given a received sequence $\boldsymbol{y}$ the ranking function $M_{u_K}(\cdot, \boldsymbol{y})$ ranks number one the sequence in B_n that $M_{\phi_1}(\cdot, \boldsymbol{y})$ ranks highest. It then ranks second the sequence that $M_{\phi_2}(\cdot, \boldsymbol{y})$ ranks highest (unless it is equal to the sequence ranked highest by $M_{\phi_1}(\cdot, \boldsymbol{y})$ in which case it skips to consider the sequence that $M_{\phi_3}(\cdot, \boldsymbol{y})$ ranks highest), followed by the sequence that $M_{\phi_3}(\cdot, \boldsymbol{y})$ ranks highest, etc. After the first rankings of all the decoders $M_{\phi_1}(\cdot, \boldsymbol{y}), \cdots, M_{\phi_K}(\cdot, \boldsymbol{y})$ have been considered we return to $M_{\phi_1}(\cdot, \boldsymbol{y})$ and consider the sequence in B_n ranked second, followed by the sequence that $M_{\phi_2}(\cdot, \boldsymbol{y})$ ranks second, etc. In all cases, if we encounter a sequence that has already been ranked we simply skip it and move on to the next decoder.

This construction guarantees that if a sequence $\boldsymbol{x} \in B_n$ is ranked jth by the kth decoder $M_{\phi_k}(\cdot, \boldsymbol{y})$ then $\boldsymbol{x}$ is ranked $(j-1)K + k$ or higher by $M_{u_K}(\cdot, \boldsymbol{y})$, i.e.,

$$M_{\phi_k}(\boldsymbol{x}, \boldsymbol{y}) = j \text{ implies } M_{u_K}(\boldsymbol{x}, \boldsymbol{y}) \le (j-1)K + k, \quad \forall \boldsymbol{x} \in B_n, \ \forall 1 \le k \le K. \tag{19}$$

Equation (19) can actually serve as a definition for the merging operation, i.e., the construction of $M_{u_K}(\cdot, \boldsymbol{y})$ from $M_{\phi_1}(\cdot, \boldsymbol{y}), \cdots, M_{\phi_K}(\cdot, \boldsymbol{y})$.

Crucial to our analysis is the observation that with this construction

$$M_{u_K}(\boldsymbol{x}, \boldsymbol{y}) \le K M_{\phi_k}(\boldsymbol{x}, \boldsymbol{y}), \quad \forall (\boldsymbol{x}, \boldsymbol{y}) \in B_n \times \mathcal{Y}^n, \ \forall 1 \le k \le K \tag{20}$$

which follows immediately from (19). The following lemma demonstrates that on any channel $p_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ the performance of the merged decoded u_K cannot be much worse than the performance of each of the decoders $\phi_1, \cdots, \phi_K$.

Lemma 1: Given K decoders $\phi_1, \cdots, \phi_K$ there exists a decoder u_K (which can be taken as the merging of these decoders) such that

$$\bar{P}_{\theta, u_K}(\text{error}) \le K \bar{P}_{\theta, \phi_k}(\text{error}), \quad \forall 1 \le k \le K, \ \forall \theta \in \Theta, \ \forall n \ge 1.$$

Proof: If the codewords of a codebook are drawn independently and uniformly over the set $B_n \subseteq \mathcal{X}^n$, and if a decoder ϕ that is based on the ranking function $M_\phi(\cdot, \cdot)$ is used, then the average probability of error $\bar{P}_{\theta,\phi}(\text{error})$ incurred over the channel $p_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ is given by [19]

$$\bar{P}_{\theta,\phi}(\text{error}) = \sum_{\boldsymbol{x} \in B_n} \sum_{\boldsymbol{y} \in \mathcal{Y}^n} \frac{1}{|B_n|} p_\theta(\boldsymbol{y} \mid \boldsymbol{x}) \Pr(\text{error} \mid \boldsymbol{x}, \boldsymbol{y}, \phi) \tag{21}$$

where

$$\Pr(\text{error} \mid \boldsymbol{x}, \boldsymbol{y}, \phi) = 1 - \left(1 - \frac{M_\phi(\boldsymbol{x}, \boldsymbol{y})}{|B_n|}\right)^{\lfloor 2^{nR} \rfloor - 1} \tag{22}$$

is the conditional probability of error given that the transmitted codeword is $\boldsymbol{x}$, the received sequence is $\boldsymbol{y}$, and the decoder being used is ϕ. Equation (22) follows from the observation that the codewords are drawn independently and uniformly over B_n and that if $\boldsymbol{x}$ is the correct codeword and $\boldsymbol{y}$ is the

received sequence then an error occurs only if some other codeword $\boldsymbol{x}'$ is ranked higher than $\boldsymbol{x}$, i.e., if $M_\phi(\boldsymbol{x}',\boldsymbol{y}) \le M_\phi(\boldsymbol{x},\boldsymbol{y})$. Notice that $\Pr(\text{error} \mid \boldsymbol{x},\boldsymbol{y},\phi)$ does not depend on the channel $p_\theta(\cdot \mid \cdot)$ over which transmission is carried out, but only on the correct codeword $\boldsymbol{x}$, the received sequence $\boldsymbol{y}$ and the decoder ϕ.

To continue with our proof we need the following technical lemma, which is proved in Appendix II.

Lemma 2: The following inequalities hold:

1) The function
$$f(z) = 1 - (1-z)^N, \qquad 0 \le z \le 1$$
satisfies
$$\frac{f(s)}{f(t)} \le \max\left\{1, \frac{s}{t}\right\}, \qquad \forall s,t \in [0,\ 1]$$
where throughout this paper $0/0 = 1$.
2) If $\{a_l\}_{l=1}^L$ and $\{b_l\}_{l=1}^L$ are two nonnegative sequences then
$$\frac{a_1 + \cdots + a_L}{b_1 + \cdots + b_L} \le \max_{1 \le l \le L} \frac{a_l}{b_l} \tag{23}$$
where $a/0 = \infty$ for $a > 0$, and $0/0 = 1$.
3) If U and V are nonnegative random variables then
$$E[U] \le E[V] \max \frac{U}{V}$$
where $a/0 = \infty$, unless $a = 0$ in which case $0/0 = 1$.

To continue with the proof of Lemma 1 consider two decoders, ϕ and ϕ', that are based on the ranking functions $M_\phi(\cdot,\cdot)$ and $M_{\phi'}(\cdot,\cdot)$, respectively. It follows from (22) and from the first part of Lemma 2 that
$$\frac{P(\text{error} \mid \boldsymbol{x},\boldsymbol{y},\phi')}{P(\text{error} \mid \boldsymbol{x},\boldsymbol{y},\phi)} \le \max\left\{1, \frac{M_{\phi'}(\boldsymbol{x},\boldsymbol{y})}{M_\phi(\boldsymbol{x},\boldsymbol{y})}\right\} \tag{24}$$
and hence
$$\begin{aligned}
&\frac{\bar{P}_{\theta,\phi'}(\text{error})}{\bar{P}_{\theta,\phi}(\text{error})} \\
&\quad = \frac{\sum_{\boldsymbol{x} \in B_n} \sum_{\boldsymbol{y} \in \mathcal{Y}^n} p_\theta(\boldsymbol{y} \mid \boldsymbol{x}) \Pr(\text{error} \mid \boldsymbol{x},\boldsymbol{y},\phi')}{\sum_{\boldsymbol{x} \in B_n} \sum_{\boldsymbol{y} \in \mathcal{Y}^n} p_\theta(\boldsymbol{y} \mid \boldsymbol{x}) \Pr(\text{error} \mid \boldsymbol{x},\boldsymbol{y},\phi)} \\
&\quad \le \max_{\boldsymbol{x} \in B_n, \boldsymbol{y} \in \mathcal{Y}^n} \frac{P(\text{error} \mid \boldsymbol{x},\boldsymbol{y},\phi')}{P(\text{error} \mid \boldsymbol{x},\boldsymbol{y},\phi)} \\
&\quad \le \max_{\boldsymbol{x} \in B_n, \boldsymbol{y} \in \mathcal{Y}^n} \frac{M_{\phi'}(\boldsymbol{x},\boldsymbol{y})}{M_\phi(\boldsymbol{x},\boldsymbol{y})}.
\end{aligned} \tag{25}$$

The equality follows from (21), the first inequality follows by the third part of Lemma 2, and the last inequality follows from (24) by noting that
$$\max_{\boldsymbol{x} \in B_n, \boldsymbol{y} \in \mathcal{Y}^n} \frac{M_{\phi'}(\boldsymbol{x},\boldsymbol{y})}{M_\phi(\boldsymbol{x},\boldsymbol{y})} \ge 1$$
since for any $\boldsymbol{y} \in \mathcal{Y}^n$ the functions $M_\phi(\cdot,\boldsymbol{y})$ and $M_{\phi'}(\cdot,\boldsymbol{y})$ are both one-to-one mappings onto $\{1,\cdots,|B_n|\}$. Inequality (25) is a refined version of an inequality given in [19]. Its importance is that it relates differences in ranking functions to differences in random-coding error performance.

The proof is now concluded by noting that if u_K is obtained by merging the decoders $\phi_1,\cdots,\phi_K$ then by (20)
$$\max_{\boldsymbol{x} \in B_n, \boldsymbol{y} \in \mathcal{Y}^n} \frac{M_{u_K}(\boldsymbol{x},\boldsymbol{y})}{M_{\phi_k}(\boldsymbol{x},\boldsymbol{y})} \le K, \qquad \forall 1 \le k \le K. \qquad \square$$

As pointed out in [28], the problems of universal decoding and universal ordering are in some sense dual. In this sense Lemma 1 is the dual of [28, Proposition 1].

To prove Lemma 1 we have introduced the notion of merging decoders. An alternative approach might have been to consider the generalized likelihood ratio decoder that given K channels $\theta_1,\cdots,\theta_K$ and a received sequence $\boldsymbol{y}$ declares that codeword i was transmitted only if
$$\max_{1 \le k \le K} p_{\theta_k}(\boldsymbol{y} \mid \boldsymbol{x}(i)) \ge \max_{1 \le k \le K} p_{\theta_k}(\boldsymbol{y} \mid \boldsymbol{x}(j)), \forall 1 \le j \le \lfloor 2^{nR} \rfloor.$$
It turns out, however, that this approach, in general, fails. For a counterexample see [29].

Lemma 1 can be used to demonstrate the existence of a weak (or strong) random-coding universal decoder for the case where the family $\mathcal{F}$ is finite, i.e., when $\Theta = \{\theta_1,\cdots,\theta_K\}$, by choosing the universal decoder u to be the decoder that is obtained by merging the ML decoders corresponding to $\theta_1,\cdots,\theta_K$. This approach can even demonstrate weak universality (but not strong universality) when Θ is countable: one can order Θ and consider the sequence of decoders $\{u_n\}$ where u_n is the merging of the ML decoders of the first n (or any integer-valued subexponential function of the blocklength n that is increasing monotonically to infinity) channels in $\mathcal{F}$. The loss in performance is at most a factor of n (i.e., subexponential) for all n sufficiently large (to guarantee that the true channel is among the first n channels in $\mathcal{F}$). In the next section we shall demonstrate how this approach can be applied to noncountable families of channels.

IV. Weak Universality

In this section we shall build on Lemma 1 to construct a universal decoder for families that are not countable. The idea is to construct the decoder for blocklength n by merging the first n ML decoders for the channels $\theta_1,\cdots,\theta_n$ where $\theta_1,\cdots,\theta_n$ are the first n channels in a countable sequence of channels $\{\theta_k\}_{k=1}^\infty$ that is dense in Θ in the sense of (9).

A key role will be played by the following lemma that demonstrates that if $p_{\theta'}(\boldsymbol{y} \mid \boldsymbol{x})$ is close to $p_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ then $\bar{P}_{\theta,\theta}(\text{error}) \approx \bar{P}_{\theta,\theta'}(\text{error})$. While the proof of the lemma is not complicated, the lemma is not entirely trivial because even if $p_{\theta'}(\boldsymbol{y} \mid \boldsymbol{x})$ is close to $p_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ the ML decoder corresponding to θ can be very different from the one corresponding to θ'. This can be seen by considering the case of the family of binary-symmetric channels (BSC) parameterized by their crossover probability. If θ corresponds to crossover probability $0.5 + \epsilon$ and θ' corresponds to a crossover probability of $0.5 - \epsilon$ then even though p_θ and $p_{\theta'}$ are close, the two ML decoders are very different: one ML decoder decodes according to minimum Hamming distance and the other according to maximum Hamming distance. Nevertheless, $\bar{P}_{\theta,\theta'}(\text{error})$ is a continuous function of $p_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ and $p_{\theta'}(\boldsymbol{y} \mid \boldsymbol{x})$ so the result is to be expected.

Lemma 3: If

$$\frac{1}{n}\left|\log\frac{p_{\theta''}(\boldsymbol{y}\mid\boldsymbol{x})}{p_{\theta'}(\boldsymbol{y}\mid\boldsymbol{x})}\right|\le\epsilon,\qquad\forall(\boldsymbol{x},\boldsymbol{y})\in\mathcal{X}^n\times\mathcal{Y}^n$$

then

$$\bar{P}_{\theta',\theta''}(\text{error})\le 2^{2n\epsilon}\bar{P}_{\theta',\theta'}(\text{error})$$

and

$$\bar{P}_{\theta'',\theta''}(\text{error})\le 2^{n\epsilon}\bar{P}_{\theta'',\theta''}(\text{error}).$$

Proof: To make the proof of the lemma more transparent, let us break up the assumptions of the lemma into two separate assumptions.

$$p_{\theta''}(\boldsymbol{y}\mid\boldsymbol{x})\le p_{\theta'}(\boldsymbol{y}\mid\boldsymbol{x})2^{n\epsilon},\qquad\forall(\boldsymbol{x},\boldsymbol{y})\in\mathcal{X}^n\times\mathcal{Y}^n \quad (26)$$

and

$$p_{\theta''}(\boldsymbol{y}\mid\boldsymbol{x})\ge p_{\theta'}(\boldsymbol{y}\mid\boldsymbol{x})2^{-n\epsilon},\qquad\forall(\boldsymbol{x},\boldsymbol{y})\in\mathcal{X}^n\times\mathcal{Y}^n. \quad (27)$$

We now have

$$\begin{aligned}
&\bar{P}_{\theta',\theta''}(\text{error})\\
&\quad=\sum_{\boldsymbol{x}\in B_n}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\frac{1}{|B_n|}p_{\theta'}(\boldsymbol{y}\mid\boldsymbol{x})\Pr(\text{error}\mid\boldsymbol{x},\boldsymbol{y},\phi_{\theta''})\\
&\quad\le 2^{n\epsilon}\sum_{\boldsymbol{x}\in B_n}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\frac{1}{|B_n|}p_{\theta''}(\boldsymbol{y}\mid\boldsymbol{x})\Pr(\text{error}\mid\boldsymbol{x},\boldsymbol{y},\phi_{\theta''})\\
&\quad=2^{n\epsilon}\bar{P}_{\theta'',\theta''}(\text{error}) \qquad (28)\\
&\quad\le 2^{n\epsilon}\bar{P}_{\theta'',\theta'}(\text{error})\\
&\quad=2^{n\epsilon}\sum_{\boldsymbol{x}\in B_n}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\frac{1}{|B_n|}p_{\theta''}(\boldsymbol{y}\mid\boldsymbol{x})\Pr(\text{error}\mid\boldsymbol{x},\boldsymbol{y},\phi_{\theta'})\\
&\quad\le 2^{2n\epsilon}\sum_{\boldsymbol{x}\in B_n}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\frac{1}{|B_n|}p_{\theta'}(\boldsymbol{y}\mid\boldsymbol{x})\Pr(\text{error}\mid\boldsymbol{x},\boldsymbol{y},\phi_{\theta'})\\
&\quad=2^{2n\epsilon}\bar{P}_{\theta',\theta'}(\text{error}) \qquad (29)
\end{aligned}$$

which completes the proof of the first claim of the lemma. The first inequality follows from (27), the second inequality follows from the optimality of the ML decoder, and the third inequality follows from (26). All equalities follow from (21) and the fact that the conditional error probability, which is defined in (22), depends on $\boldsymbol{x},\boldsymbol{y}$, and ϕ but not on the channel $p_{\theta'}(\cdot|\cdot)$.

The second claim of the lemma follows from (28) by noting that by the optimality of the ML rule

$$\bar{P}_{\theta',\theta'}(\text{error})\le\bar{P}_{\theta',\theta''}(\text{error}).\qquad\square$$

We are now in a position to prove Theorem 1.

Proof: Let $\{\theta_k\}_{k=1}^{\infty}$ be the sequence of channels that satisfies (9), and let $\theta\in\Theta$ be arbitrary but fixed. It follows from (9) that for every $\epsilon>0$ there exists some positive integer k^* (which depends on θ,ϵ) and some n_0 (which also depends on θ,ϵ) such that

$$\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{X}^n\times\mathcal{Y}^n}\frac{1}{n}\left|\log\left(\frac{p_\theta(\boldsymbol{y}\mid\boldsymbol{x})}{p_{\theta_{k^*}}(\boldsymbol{y}\mid\boldsymbol{x})}\right)\right|<\epsilon,\qquad\forall n\ge n_0.$$

Let the decoder u_n be constructed by merging the first $K(n)$ ML decoders corresponding to $\theta_1,\cdots,\theta_{K(n)}$ where for now $K(n)=n$. For all sufficiently large blocklength n we have that $K(n)\ge k^*$ and the ML decoder $\phi_{\theta_{k^*}}$ is among the decoders $\phi_{\theta_1},\cdots,\phi_{\theta_{K(n)}}$ from which u_n is constructed. It, therefore, follows from Lemma 1 that for such sufficiently large n

$$\bar{P}_{\theta,u_{K(n)}}(\text{error})\le K(n)\bar{P}_{\theta,\theta_{k^*}}(\text{error}). \quad (30)$$

If, in addition, n is sufficiently large so that $n\ge n_0$ then by Lemma 3

$$\bar{P}_{\theta,\theta_{k^*}}(\text{error})\le 2^{2n\epsilon}\bar{P}_{\theta,\theta}(\text{error}). \quad (31)$$

Combining (30) and (31) we have that for all sufficiently large n

$$\bar{P}_{\theta,u_{K(n)}}(\text{error})\le K(n)2^{2n\epsilon}\bar{P}_{\theta,\theta}(\text{error}) \quad (32)$$

and the first part of the theorem involving random-coding universality now follows by noting that $K(n)=n$ is subexponential.

The second part of the theorem establishing deterministic-coding universality will now follow once we show that if the family of channels is separable then random-coding weak universality implies deterministic-coding weak universality, which is the content of the following lemma, Lemma 4. $\square$

Inspecting the proof we see that some of the conditions of Theorem 1 can be weakened. First we can replace the separability condition with a weaker form that requires that there exist a sequence $\{\theta_k\}\subseteq\Theta$ and a subexponential integer-valued monotonically increasing function $K(n)$ such that for any $\theta\in\Theta$

$$\limsup_{n\to\infty}\min_{1\le k\le K(n)}\sup_{(\boldsymbol{x},\boldsymbol{y})\in B_n\times\mathcal{Y}^n}\frac{1}{n}\left|\log\left(\frac{p_\theta(\boldsymbol{y}\mid\boldsymbol{x})}{p_{\theta_k}(\boldsymbol{y}\mid\boldsymbol{x})}\right)\right|=0.$$

Such a weaker condition could be useful when studying channels with infinitely many internal states where the number and effect of the internal states grows moderately with the blocklength n. This approach could be also useful when the family of channels is more naturally parameterized with an infinite number of parameters as would, for example, be the case if a natural parameter is the autocorrelation function of some random process.

Secondly, if the random-coding error exponents of the channels in the family are uniformly bounded then we may exclude some sets of pairs $(\boldsymbol{x},\boldsymbol{y})$ from the supremum in (9) provided that the sets have a probability that is negligible with respect to the best error exponent in the family. We adopt this approach in dealing with strong separability.

Lemma 4: If the family of channels $\mathcal{F}$ is separable then random-coding weak universality implies deterministic-coding weak universality.

Proof: Let $\{u_n\}$ be random-coding weakly universal for the family $\{p_\theta(\boldsymbol{y}\mid\boldsymbol{x})\}$ and input-sets $\{B_n\}$, and let $\{p_{\theta_k}(\boldsymbol{y}\mid\boldsymbol{x})\}_{k=1}^{\infty}$ be a sequence of laws that is dense in the sense of (9). It follows from the weak random-coding universality of the sequence $\{u_n\}$ that for any $K>1$ and any $\epsilon>0$ there exists some $n(K,\epsilon)$ such that for all $n\ge n(K,\epsilon)$

$$\bar{P}_{\theta_k,u_n}(\text{error})\le 2^{n\epsilon}\bar{P}_{\theta_k,\theta_k}(\text{error}),\qquad\forall 1\le k\le K. \quad (33)$$

Let $\mathcal{A}_k$ denote the event that a rate-R blocklength-n randomly chosen codebook $\mathcal{C}_n$ whose codewords are drawn independently and uniformly over the set B_n satisfies

$$P_{\theta_k,u_n}(\text{error} \mid \mathcal{C}_n) \geq K^2 2^{n\epsilon} \bar{P}_{\theta_k,\theta_k}(\text{error}).$$

It follows from (33) and Markov's inequality that

$$\Pr(\mathcal{A}_k) \leq \frac{1}{K^2}, \quad \forall 1 \leq k \leq K, \quad \forall n \geq n(K,\epsilon). \tag{34}$$

We thus conclude from (34) and the union of events bound that

$$\begin{aligned} \Pr \bigcap_{k=1}^{K} \mathcal{A}_k^c &= 1 - \Pr \bigcup_{k=1}^{K} \mathcal{A}_k \\ &\geq 1 - K^{-1} \\ &> 0 \end{aligned}$$

where we use D^c to denote the set complement of the set D. We can thus conclude that for $n \geq n(K,\epsilon)$ there exists a codebook $\mathcal{C}_n^*$ such that

$$P_{\theta_k,u_n}(\text{error} \mid \mathcal{C}_n^*) < K^2 2^{n\epsilon} \bar{P}_{\theta_k,\theta_k}(\text{error}), \quad \forall 1 \leq k \leq K. \tag{35}$$

Choosing $\epsilon = \epsilon(K) \to 0$ and letting $K \to \infty$ we can construct a sequence of codebooks $\{\mathcal{C}_n^*\}$ so that

$$\lim_{n\to\infty} \frac{1}{n} \log\left(\frac{P_{\theta_k,u_n}(\text{error} \mid \mathcal{C}_n^*)}{\bar{P}_{\theta_k,\theta_k}(\text{error})}\right) = 0, \qquad \forall k \geq 1. \tag{36}$$

To conclude the proof we show that the validity of (8) for the dense sequence $\{\theta_k\}$, i.e., (36), implies its validity for any θ. This can be seen by noting that if

$$\sup_{(\boldsymbol{x},\boldsymbol{y}) \in B_n \times \mathcal{Y}^n} \frac{1}{n} \left| \log\left(\frac{p_\theta(\boldsymbol{y} \mid \boldsymbol{x})}{p_{\theta_{k^*}}(\boldsymbol{y} \mid \boldsymbol{x})}\right)\right| < \epsilon \tag{37}$$

then by Lemma 3

$$\bar{P}_{\theta,\theta}(\text{error}) \geq 2^{-n\epsilon} \bar{P}_{\theta_{k^*},\theta_{k^*}}(\text{error}) \tag{38}$$

and by noting that (37) also implies that

$$P_{\theta,u_n}(\text{error} \mid \mathcal{C}_n^*) \leq 2^{n\epsilon} P_{\theta_{k^*},u_n}(\text{error} \mid \mathcal{C}_n^*).$$

Indeed, for any $\mathcal{C} = \{\boldsymbol{x}(1), \cdots, \boldsymbol{x}(\lfloor 2^{nR} \rfloor)\} \subset B_n$ and decoder ϕ

$$\begin{aligned} P_{\theta,\phi}(\text{error} \mid \mathcal{C}) &= \frac{1}{\lfloor 2^{nR} \rfloor} \sum_{i=1}^{\lfloor 2^{nR} \rfloor} \sum_{\boldsymbol{y} \in \mathcal{D}_i^c} p_\theta(\boldsymbol{y} \mid \boldsymbol{x}(i)) \\ &\leq \frac{2^{n\epsilon}}{\lfloor 2^{nR} \rfloor} \sum_{i=1}^{\lfloor 2^{nR} \rfloor} \sum_{\boldsymbol{y} \in \mathcal{D}_i^c} p_{\theta_{k^*}}(\boldsymbol{y} \mid \boldsymbol{x}(i)) \qquad (39) \\ &= 2^{n\epsilon} P_{\theta_{k^*},\phi}(\text{error} \mid \mathcal{C}) \qquad (40) \end{aligned}$$

where $\mathcal{D}_i = \phi^{-1}(i)$, i.e., the sequences in $\mathcal{Y}^n$ that are decoded by ϕ to the ith message, and $\mathcal{D}_i^c$ is its complement. □

V. Strong Universality

The following Lemma will be useful in the study of strong universality.

Lemma 5: Let $p_{\theta'}(\boldsymbol{y} \mid \boldsymbol{x})$ and $p_{\theta''}(\boldsymbol{y} \mid \boldsymbol{x}), \theta', \theta'' \in \Theta$ be two channels that satisfy

$$p_{\theta'}(\boldsymbol{y} \mid \boldsymbol{x}) \leq 2^{n\epsilon} p_{\theta''}(\boldsymbol{y} \mid \boldsymbol{x}), \quad \forall \boldsymbol{x}, \boldsymbol{y} : p_{\theta'}(\boldsymbol{y} \mid \boldsymbol{x}) > 2^{-n(M+\log|\mathcal{Y}|)}$$

then for any code $\mathcal{C}$ and decoder ϕ

$$P_{\theta',\phi}(\text{error} \mid \mathcal{C}) \leq 2^{n\epsilon} P_{\theta'',\phi}(\text{error} \mid \mathcal{C}) + 2^{-nM}$$

and

$$\bar{P}_{\theta',\theta'}(\text{error}) \leq 2^{n\epsilon} \bar{P}_{\theta'',\theta''}(\text{error}) + 2^{-nM}.$$

Proof: Given a codeword $\boldsymbol{x}(i) \in \mathcal{C}$ let

$$F_{\boldsymbol{x}(i)} = \{\boldsymbol{y} : p_{\theta'}(\boldsymbol{y} \mid \boldsymbol{x}(i)) > 2^{-n(M+\log|\mathcal{Y}|)}\}$$

and let $\mathcal{D}_i = \phi^{-1}(i)$ be the set of all output sequences that the decoder ϕ decodes to the codeword $\boldsymbol{x}(i)$, and $\mathcal{D}_i^c$ the set complement of $\mathcal{D}_i$. We now have

$$\begin{aligned} &P_{\theta',\phi}(\text{error} \mid \mathcal{C}) \\ &\quad = \frac{1}{\lfloor 2^{nR} \rfloor} \sum_{i=1}^{\lfloor 2^{nR} \rfloor} \sum_{\boldsymbol{y} \in \mathcal{D}_i^c} p_{\theta'}(\boldsymbol{y} \mid \boldsymbol{x}(i)) \\ &\quad \leq \frac{1}{\lfloor 2^{nR} \rfloor} \sum_{i=1}^{\lfloor 2^{nR} \rfloor} \left(\sum_{\boldsymbol{y} \in \mathcal{D}_i^c \cap F_{\boldsymbol{x}(i)}} p_{\theta'}(\boldsymbol{y} \mid \boldsymbol{x}(i)) \right. \\ &\qquad \left. + \sum_{\boldsymbol{y} \in F_{\boldsymbol{x}(i)}^c} 2^{-n(M+\log|\mathcal{Y}|)} \right) \\ &\quad \leq \frac{1}{\lfloor 2^{nR} \rfloor} \sum_{i=1}^{\lfloor 2^{nR} \rfloor} \left(\sum_{\boldsymbol{y} \in \mathcal{D}_i^c \cap F_{\boldsymbol{x}(i)}} 2^{n\epsilon} p_{\theta''}(\boldsymbol{y} \mid \boldsymbol{x}(i)) + 2^{-nM} \right) \\ &\quad \leq \frac{1}{\lfloor 2^{nR} \rfloor} \sum_{i=1}^{\lfloor 2^{nR} \rfloor} \left(\sum_{\boldsymbol{y} \in \mathcal{D}_i^c} 2^{n\epsilon} p_{\theta''}(\boldsymbol{y} \mid \boldsymbol{x}(i)) + 2^{-nM} \right) \\ &\quad = 2^{n\epsilon} P_{\theta'',\phi}(\text{error} \mid \mathcal{C}) + 2^{-nM}. \end{aligned}$$

It now follows by choosing ϕ to be the ML decoder with respect to the law θ'' and by averaging over the codebook $\mathcal{C}$ that

$$\bar{P}_{\theta',\theta''}(\text{error}) \leq 2^{n\epsilon} \bar{P}_{\theta'',\theta''}(\text{error}) + 2^{-nM}$$

from which the second part of the lemma follows by noting that by the optimality of the ML decoder

$$\bar{P}_{\theta',\theta'}(\text{error}) \leq \bar{P}_{\theta',\theta''}(\text{error}).$$ □

With this lemma we can now prove Theorem 2.

Proof: Let $\epsilon > 0$ be arbitrary but sufficiently small to guarantee that

$$\limsup_{n\to\infty} \sup_\theta -\frac{1}{n} \log \bar{P}_{\theta,\theta}(\text{error}) < M - \epsilon$$

where M is the constant appearing in Definition 6 (strong separability), and thus satisfies (10). Let n_0 be sufficiently large to guarantee that

$$2^{-nM} \leq \frac{1}{2} \inf_\theta \bar{P}_{\theta,\theta}(\text{error}) 2^{-n\epsilon}, \qquad \forall n > n_0. \tag{41}$$

Let $\theta_1^{(n)},\cdots,\theta_{K(n)}^{(n)}$ be the channels that demonstrate the strong separability of Θ, see Definition 6. Letting u_n denote the merging of the ML decoders corresponding to $\theta_1^{(n)},\cdots,\theta_{K(n)}^{(n)}$ we have by Lemma 1 that

$$\bar{P}_{\theta,u_n}(\text{error}) \le K(n)\bar{P}_{\theta,\theta_k^{(n)}}(\text{error}), \quad \forall\theta\in\Theta,\ \forall 1\le k\le K(n). \tag{42}$$

Given some $\theta\in\Theta$ let $\theta_{k^*}^{(n)}$ be a channel that satisfies (12) and (13) with $1\le k^*\le K(n)$. We now have

$$\begin{aligned}\bar{P}_{\theta,u_n}(\text{error}) &\le K(n)\bar{P}_{\theta,\theta_{k^*}^{(n)}}(\text{error})\\ &\le K(n)\big(2^{n\epsilon}\bar{P}_{\theta_{k^*}^{(n)},\theta_{k^*}^{(n)}}(\text{error})+2^{-nM}\big)\\ &\le 2K(n)2^{n\epsilon}\bar{P}_{\theta_{k^*}^{(n)},\theta_{k^*}^{(n)}}(\text{error})\\ &\le 2K(n)2^{n\epsilon}(2^{n\epsilon}\bar{P}_{\theta,\theta}(\text{error})+2^{-nM})\\ &\le 4K(n)2^{2n\epsilon}\bar{P}_{\theta,\theta}(\text{error}).\end{aligned}$$

The first inequality follows from (42); the second inequality follows from the first part of Lemma 5 by choosing $\theta'=\theta$, $\theta''=\theta_{k^*}^{(n)}$, ϕ to be the ML decoder with respect to $\theta_{k^*}^{(n)}$, and by averaging over the codebook $\mathcal{C}$; the third inequality follows from (41); the fourth from the second part of Lemma 5 with $\theta'=\theta_{k^*}^{(n)}$ and $\theta''=\theta$; and the last inequality from (41). It thus follows that

$$\frac{\bar{P}_{\theta,u_n}(\text{error})}{\bar{P}_{\theta,\theta}(\text{error})}\le 4K(n)2^{2n\epsilon}$$

and the first part of the theorem follows by noting that $K(n)$ is subexponential and by choosing $\epsilon=\epsilon(n)\to 0$.

The second part of the theorem follows by noting that if Θ is strongly separable then any random-coding strong universal decoder is also a deterministic-coding strong universal decoder, as the next lemma demonstrates. □

Lemma 6: If the family of channels $\{p_\theta(\boldsymbol{y}\mid\boldsymbol{x}),\theta\in\Theta\}$ is strongly separable (see Definition 6) then random-coding strong universality implies deterministic-coding strong universality.

Proof: Let $\epsilon>0$ be arbitrary but sufficiently small to guarantee that

$$\limsup_{n\to\infty}\sup_\theta -\frac{1}{n}\bar{P}_{\theta,\theta}(\text{error})<M-\epsilon$$

where M is the constant appearing in the definition of strong separability (Definition 6), and that thus satisfies (10). Let $n_0(\epsilon)$ be sufficiently large to guarantee that

$$\sup_\theta\frac{\bar{P}_{\theta,u_n}(\text{error})}{\bar{P}_{\theta,\theta}(\text{error})}\le 2^{n\epsilon}.\qquad \forall n\ge n_0(\epsilon) \tag{43}$$

and

$$2^{-nM}<\frac{1}{2}\inf_\theta\bar{P}_{\theta,\theta}(\text{error}),\qquad \forall n\ge n_0(\epsilon) \tag{44}$$

where u_n is the sequence of random-coding strong universal decoders. Given a blocklength n let $\theta_1^{(n)},\cdots,\theta_{K(n)}^{(n)}$ be the channels that demonstrate the strong separability of Θ. Thus for every $\theta\in\Theta$ there exists $\theta_{k^*}^{(n)}$ such that (12) and (13) hold, and the function $K(n)$ is subexponential.

Denoting by $\mathcal{A}_k, k=1,\cdots,K(n)$, the event that a rate-$R$ blocklength-n random codebook $\mathcal{C}_n$, whose codewords are drawn independently and uniformly over B_n, satisfies

$$\bar{P}_{\theta_k^{(n)},u_n}(\text{error})\ge K^2(n)2^{n\epsilon}\bar{P}_{\theta_k^{(n)},\theta_k^{(n)}}(\text{error})$$

we have by (43) and Markov's inequality that

$$\Pr(\mathcal{A}_k)\le\frac{1}{K^2(n)},\qquad \forall 1\le k\le K(n)$$

and thus by the union of events bound

$$\Pr\bigcap_{k=1}^{K(n)}\mathcal{A}_k^c>0$$

and there thus exists a codebook $\mathcal{C}_n^*$ satisfying

$$\frac{P_{\theta_k^{(n)},u_n}(\text{error}\mid\mathcal{C}_n^*)}{\bar{P}_{\theta_k^{(n)},\theta_k^{(n)}}(\text{error})}<K^2(n)2^{n\epsilon},\qquad \forall 1\le k\le K(n). \tag{45}$$

Given $\theta\in\Theta$, let $\theta_{k^*}^{(n)}$ be such that (12) and (13) both hold. We now have

$$\begin{aligned}\frac{P_{\theta,u_n}(\text{error}\mid\mathcal{C}_n^*)}{\bar{P}_{\theta,\theta}(\text{error})} &\le \frac{P_{\theta_{k^*}^{(n)},u_n}(\text{error}\mid\mathcal{C}_n^*)2^{n\epsilon}+2^{-nM}}{\bar{P}_{\theta,\theta}(\text{error})}\\ &\le \frac{P_{\theta_{k^*}^{(n)},u_n}(\text{error}\mid\mathcal{C}_n^*)2^{n\epsilon}+2^{-nM}}{2^{-n\epsilon}\big(\bar{P}_{\theta_{k^*}^{(n)},\theta_{k^*}^{(n)}}(\text{error})-2^{-nM}\big)}\\ &\le \frac{4P_{\theta_{k^*}^{(n)},u_n}(\text{error}\mid\mathcal{C}_n^*)2^{n\epsilon}}{2^{-n\epsilon}\bar{P}_{\theta_{k^*}^{(n)},\theta_{k^*}^{(n)}}(\text{error})}\\ &\le 4K^2(n)2^{2n\epsilon}\end{aligned}$$

and the proof is concluded by recalling that $K(n)$ is subexponential and by choosing $\epsilon=\epsilon(n)\to 0$. Note that the first inequality follows from the first part of Lemma 5 by taking $\theta'=\theta,\theta''=\theta_{k^*}^{(n)}$, and $\phi=u_n$. The second inequality follows from the second part of Lemma 5 with $\theta'=\theta_{k^*}^{(n)},\theta''=\theta$, and that last inequality follows from (44). □

VI. Infinite Alphabets

We next consider some extensions of the results presented in previous sections to the case where the input and output alphabets are not necessarily finite. Once again we restrict ourselves to parametric families

$$\{p_\theta(\boldsymbol{y}\mid\boldsymbol{x}),\theta\in\Theta\} \tag{46}$$

where for any $\theta\in\Theta$ the channel p_θ is a mapping that maps any input sequence $\boldsymbol{x}\in\mathcal{X}^n$ to a probability measure $p_\theta(\cdot\mid\boldsymbol{x})$ defined on a common σ-algebra on $\mathcal{Y}^n$.

As before, we shall discuss random coding in which codewords are drawn independently and uniformly over a set $B_n\subseteq\mathcal{X}^n$. We are implicitly assuming that B_n is endowed with a σ-algebra, and we denote the uniform measure on B_n by μ^x (making the blocklength n implicit).

We shall assume throughout that $\mathcal{X}$ and $\mathcal{Y}$ are complete separable metric spaces (i.e., Polish), that the σ-algebra on B_n

is the restriction of the product Borel σ-algebra on $\mathcal{X}^n$ to B_n, and that the σ-algebra on $\mathcal{Y}^n$ is the product Borel σ-algebra.

We shall endow the set of distributions on $\mathcal{Y}^n$ with the weak topology and assume that for every $\theta \in \Theta$ the mapping $\boldsymbol{x} \mapsto p_\theta(\cdot \mid \boldsymbol{x})$ is Borel measurable. This assumption is equivalent to the assumption that for any $\theta \in \Theta$ and any Borel set $B \in \mathcal{Y}^n$ the function $\boldsymbol{x} \mapsto p_\theta(B \mid \boldsymbol{x})$ from B_n to $\mathbb{R}$ is measurable, see [30]. We can thus define the product measure $\mu_\theta^{x,y}$ on $B_n \times \mathcal{Y}^n$ as the measure that satisfies

$$\mu_\theta^{x,y}(A \times B) = \int_A p_\theta(B \mid \boldsymbol{x})\, d\mu^x(\boldsymbol{x}) \tag{47}$$

for any Borel sets $A \subset B^n,\ B \subset \mathcal{Y}^n$.

An additional assumption that greatly simplifies the analysis is that for every blocklength n there exists a measure ν on $\mathcal{Y}^n$ with respect to which all the measures

$$\{p_\theta(\cdot \mid \boldsymbol{x}), \boldsymbol{x} \in B_n, \theta \in \Theta\}$$

are absolutely continuous. We shall denote by $f_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ the Radon–Nykodim derivative of the measure $p_\theta(\cdot \mid \boldsymbol{x})$ with respect to ν at $\boldsymbol{y}$, i.e.,

$$f_\theta(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{dp_\theta(\boldsymbol{y} \mid \boldsymbol{x})}{d\nu}.$$

This assumption is somewhat restrictive as it rules out channels like the channel where the input X and output Y are real and $Y = X + Z$ where Z is independent noise that takes value in the integers. We shall later remark on how such channels can be treated.

The final assumption we make is that $\mathcal{X}$ admits a measurable total ordering, i.e., a total ordering $\preceq$ such that the set of all predecessors of x is a measurable set. A typical input alphabet that satisfies this assumption is $\mathbb{R}^d$ with the ordering taken to be lexicographical with the standard ordering in $\mathbb{R}$ (i.e., $x \preceq x' \Leftrightarrow x \le x'$).

We can define ranking functions in much the same way that we did for finite alphabets, except that if the input sets B_n are infinite then we prefer to deal with canonical ranking functions. We define a canonical ML decoder ϕ_θ for the channel $p_\theta(\cdot \mid \cdot)$ as a decoder that given a received sequence $\boldsymbol{y}$ and a codebook $\mathcal{C}$ declares that the transmitted codeword is $\boldsymbol{x}(i)$, i.e., $\phi_\theta(\boldsymbol{y}) = i$, if

$$M_\theta(\boldsymbol{x}(i), \boldsymbol{y}) < M_\theta(\boldsymbol{x}(j), \boldsymbol{y}), \qquad \forall j \neq i$$

where the ranking function $M_\theta(\cdot, \cdot)$ satisfies the following conditions:

$$M_\theta : B_n \times \mathcal{Y}^n \to [0, 1]; \tag{48}$$

for any $\boldsymbol{y} \in \mathcal{Y}^n$ the mapping $M(\cdot, \boldsymbol{y})$ is measurable;

$$\mu^x(M^{-1}((0, \alpha), \boldsymbol{y})) = \alpha, \quad \forall \alpha \in [0, 1], \quad \forall \boldsymbol{y} \in \mathcal{Y}^n; \tag{49}$$

and

$$f_\theta(\boldsymbol{y} \mid \boldsymbol{x}) > f_\theta(\boldsymbol{y} \mid \boldsymbol{x}') \Rightarrow M_\theta(\boldsymbol{x}, \boldsymbol{y}) < M_\theta(\boldsymbol{x}', \boldsymbol{y}). \tag{50}$$

Notice that there always exists an optimal decoder which is canonical. Indeed, if $\preceq$ is the total ordering on $\mathcal{X}$ extended to B_n lexicographically then we can define

$$M_\theta(\boldsymbol{x}, \boldsymbol{y}) = \mu^x(\{\boldsymbol{x}' \in B_n : f_\theta(\boldsymbol{y} \mid \boldsymbol{x}') > f_\theta(\boldsymbol{y} \mid \boldsymbol{x}) \text{ or } f_\theta(\boldsymbol{y} \mid \boldsymbol{x}') = f_\theta(\boldsymbol{y} \mid \boldsymbol{x}), \boldsymbol{x}' \preceq \boldsymbol{x}\}). \tag{51}$$

We can now state the continuous alphabet counterpart of Lemma 1. Notice that if B_n is finite then Lemma 1 holds even if $\mathcal{Y}$ is infinite: we did not assume that $\mathcal{Y}$ is finite in proving that lemma. If, however, B_n is infinite then the proof needs some modification as follows.

Lemma 7: Given K canonical decoders that are based on the ranking functions $M_{\theta_1}, \cdots, M_{\theta_K}$, and given any arbitrarily large number $L > 0$, there exists a decoder u_K such that

$$\bar{P}_{\theta, u_K}(\text{error}) \le K \bar{P}_{\theta, \theta_k}(\text{error}) + K 2^{-nL}, \quad \forall 1 \le k \le K, \quad \forall \theta \in \Theta, \quad \forall n \ge 1.$$

Note that when we apply this lemma we typically choose $L = L(n)$ with $L(n) \to \infty$ in order to avoid a loss in the error exponent.

Proof: To prove the lemma divide the unit interval $[0, 1]$ into $2^{n(L+R)}$ disjoint intervals of length $2^{-n(L+R)}$

$$I_0 = [0, 2^{-n(L+R)}]$$
$$I_m = (m2^{-n(L+R)}, (m+1)2^{-n(L+R)}], \quad 1 \le m < 2^{n(L+R)}$$

where R is the code rate. Consider now the merged decoder that forms its decision based on the observation $\boldsymbol{y}$ in the following way: It first considers $M_{\theta_1}^{-1}(I_0, \boldsymbol{y}) \cap \mathcal{C}$. If this is nonempty, it declares that the transmitted codeword was the codeword that ranks highest (according to M_{θ_1}) among $M_{\theta_1}^{-1}(I_0, \boldsymbol{y}) \cap \mathcal{C}$. Otherwise, if $M_{\theta_1}^{-1}(I_0, \boldsymbol{y}) \cap \mathcal{C} = \emptyset$, the decoder considers $M_{\theta_2}^{-1}(I_0, \boldsymbol{y}) \cap \mathcal{C}$. If this is nonempty, it chooses the highest ranking codeword according to M_{θ_2}, and otherwise considers $M_{\theta_3}^{-1}(I_0, \boldsymbol{y}) \cap \mathcal{C}$, etc. If a decision has not been reached after considering $M_{\theta_K}^{-1}(I_0, \boldsymbol{y}) \cap \mathcal{C}$, the decoders considers $M_{\theta_1}^{-1}(I_1, \boldsymbol{y}) \cap \mathcal{C}$ followed by $M_{\theta_2}^{-1}(I_1, \boldsymbol{y}) \cap \mathcal{C}$ etc.

Assume now that transmission is carried out over the channel $p_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ and let $1 \le k \le K$ be arbitrary. We shall now compare the performance of the merged decoder u_K with that of ϕ_{θ_k}, the ML decoder tuned to $p_{\theta_k}(\boldsymbol{y} \mid \boldsymbol{x})$. We thus need to compare $\bar{P}_{\theta, u_K}(\text{error})$ with $\bar{P}_{\theta, \theta_k}(\text{error})$. Hold the received sequence $\boldsymbol{y}$ and the correct codeword $\boldsymbol{x}$ fixed, and assume that given $\boldsymbol{y}$ the decoder ϕ_{θ_k} ranks $\boldsymbol{x}$ in I_m, i.e.,

$$\boldsymbol{x} \in M_{\theta_k}^{-1}(I_m, \boldsymbol{y}).$$

The decoder u_K makes an error only if some codeword $\boldsymbol{x}'$ lies in an interval higher than $\boldsymbol{x}$ in one of the lists $M_{\theta_1}, \cdots, M_{\theta_K}$, i.e., if

$$\boldsymbol{x}' \in \bigcup_{m'=0}^{m-1} \bigcup_{k'=1}^{K} M_{\theta_{k'}}^{-1}(I_{m'}, \boldsymbol{y}) \tag{52}$$

or if some codeword $\boldsymbol{x}'$ lies in the same interval as $\boldsymbol{x}$ in one of the lists $M_{\theta_1}, \cdots, M_{\theta_K}$, i.e.,

$$\boldsymbol{x}' \in \bigcup_{k'=1}^{K} M_{\theta_{k'}}^{-1}(I_m, \boldsymbol{y}).$$

We denote the former event by E_1 and the latter by E_2. Notice that $E_1 \cup E_2$ is a necessary condition for an error but not

sufficient, because of the order in which the decoders are merged.

We first analyze the probability of the event E_2 by noting that

$$\begin{aligned}\Pr(E_2 \mid \boldsymbol{x},\boldsymbol{y}) &\le \sum_{k'=1}^{K}(\lfloor 2^{nR}\rfloor - 1)\mu^x\big(M^{-1}_{\theta_{k'}}(I_m,\boldsymbol{y})\big)\\ &= K\big(\lfloor 2^{nR}\rfloor - 1\big)2^{-n(R+L)}\\ &\le K2^{-nL}\end{aligned}$$

and hence

$$\Pr(E_2) \le K2^{-nL} \tag{53}$$

where the first inequality follows from the union of events bound and the second from the fact that all the ranking functions under consideration are canonical.

As to the event E_1 we note that the probability that $\boldsymbol{x}'$ satisfies (52) is, by the union of events bound and the fact that the rankings are canonical, upper-bounded by

$$\begin{aligned}&\Pr\left(\boldsymbol{x}' \in \bigcup_{m'=0}^{m-1}\bigcup_{k'=1}^{K} M^{-1}_{\theta_{k'}}(I_{m'},\boldsymbol{y})\right)\\ &\quad\le Km2^{-n(L+R)}\\ &\quad\le K\Pr\big(M_{\theta_k}(\boldsymbol{x}',\boldsymbol{y}) < M_{\theta_k}(\boldsymbol{x},\boldsymbol{y})\big)\end{aligned}$$

where all probabilities are, or course, conditional on $\boldsymbol{x}$ and $\boldsymbol{y}$, and where the second inequality follows from the assumption that ϕ_{θ_k} ranks $\boldsymbol{x}$ in I_m. Noting that

$$\begin{aligned}&\Pr\big(E_1 \mid \boldsymbol{x},\boldsymbol{y}\big)\\ &\quad= 1 - \left(1 - \Pr\left(\boldsymbol{x}' \in \bigcup_{m'=0}^{m-1}\bigcup_{k'=1}^{K} M^{-1}_{\theta_{k'}}(I_{m'},\boldsymbol{y})\right)\right)^{2^{nR-1}}\end{aligned}$$

and

$$\begin{aligned}&\Pr\big(\text{error} \mid \boldsymbol{x},\boldsymbol{y},\phi_{\theta_k}\big)\\ &\quad= 1 - \big(1 - \Pr\big(M_{\theta_k}(\boldsymbol{x}',\boldsymbol{y}) < M_{\theta_k}(\boldsymbol{x},\boldsymbol{y})\big)\big)^{2^{nR-1}}\end{aligned}$$

we can use Lemma 2 to conclude that

$$\Pr(E_1) \le K\bar{P}_{\theta,\theta_k}(\text{error}). \tag{54}$$

Inequalities (53) and (54) now prove the lemma. □

Note: We used the assumption that there was a measure ν with respect to which all the measures $\{p_\theta(\cdot \mid \boldsymbol{x})\}$ are absolutely continuous to demonstrate that every ML decoder is equivalent to a decoder that is based on a canonical ranking function. In the more general situation when we do not have an underlying measure ν with respect to which all output distributions are absolutely continuous, one can often define an ML decoder for the channel θ in the following way. To every $\boldsymbol{y} \in \mathcal{Y}^n$ one assigns a measurable set $N_\theta(\boldsymbol{y}) \subset B_n$ with measure $\mu^x(N_\theta(\boldsymbol{y})) = 0$ such that the ML decoder operates as follows. If $\mathcal{C} \cap N_\theta(\boldsymbol{y}) \neq \emptyset$ it declares that the codeword in $\mathcal{C} \cap N_\theta(\boldsymbol{y})$ was transmitted. Otherwise, if $\mathcal{C} \cap N_\theta(\boldsymbol{y}) = \emptyset$ the decoding is performed using a canonical ranking function. Since $N_\theta(\boldsymbol{y})$ has measure zero, the probability of an incorrect codeword being in $N_\theta(\boldsymbol{y})$ is zero. If this is indeed the structure of the optimal receiver then merging of the receivers corresponding to $\theta_1,\cdots,\theta_K$ can be performed by first checking whether there is a codeword in $\cup_{k=1}^{K} N_{\theta_k}$, and then proceeding to merge the canonical ranking functions. A good candidate for $N_\theta(\boldsymbol{y})$ is the singular part of the decomposition of the *a posteriori* probability on B_n given $\boldsymbol{y}$ with respect to the uniform measure μ^x on B_n.

To study strong universality for infinite alphabets we need the following lemma which is the continuous alphabet counterpart of Lemma 5:

Lemma 8: Let $\theta', \theta'' \in \Theta$, and let $\mathcal{C}$ be a rate-R, block-length-n codebook such that for every codeword $\boldsymbol{x} \in \mathcal{C}$ there exists a set $F_{\boldsymbol{x}} \subset \mathcal{Y}^n$ such that

$$\int_{\boldsymbol{y}\notin F_{\boldsymbol{x}}} f_{\theta'}(\boldsymbol{y} \mid \boldsymbol{x})\,d\nu < 2^{-nM}$$

and

$$f_{\theta'}(\boldsymbol{y} \mid \boldsymbol{x}) \le 2^{n\epsilon} f_{\theta''}(\boldsymbol{y} \mid \boldsymbol{x}), \qquad \forall \boldsymbol{y} \in F_{\boldsymbol{x}}.$$

Then for any decoder ϕ

$$P_{\theta',\phi}(\text{error} \mid \mathcal{C}) \le 2^{n\epsilon}P_{\theta'',\phi}(\text{error} \mid \mathcal{C}) + 2^{-nM}.$$

Also,

$$\bar{P}_{\theta',\theta'}(\text{error}) \le 2^{n\epsilon}\bar{P}_{\theta'',\theta''}(\text{error}) + 2^{-nM}.$$

Proof: Let $\mathcal{D}_i = \phi^{-1}(i)$ be the set of received sequences that are decoded by ϕ to message i, where $i = 1,\cdots,\lfloor 2^{nR}\rfloor$. We then have

$$\begin{aligned}&P_{\theta',\phi}(\text{error} \mid \mathcal{C})\\ &\quad= \frac{1}{\lfloor 2^{nR}\rfloor}\sum_{i=1}^{\lfloor 2^{nR}\rfloor}\int_{\mathcal{D}_i^c} f_{\theta'}(\boldsymbol{y} \mid \boldsymbol{x}(i))\,d\nu\\ &\quad\le \frac{1}{\lfloor 2^{nR}\rfloor}\sum_{i=1}^{\lfloor 2^{nR}\rfloor}\left[\int_{\mathcal{D}_i^c\cap F_{\boldsymbol{x}(i)}} f_{\theta'}(\boldsymbol{y} \mid \boldsymbol{x}(i))\,d\nu\right.\\ &\qquad\qquad\left.+ \int_{F^c_{\boldsymbol{x}(i)}} f_{\theta'}(\boldsymbol{y} \mid \boldsymbol{x}(i))\,d\nu\right]\\ &\quad\le 2^{-nM} + \frac{1}{\lfloor 2^{nR}\rfloor}\sum_{i=1}^{\lfloor 2^{nR}\rfloor}\int_{\mathcal{D}_i^c\cap F_{\boldsymbol{x}(i)}} 2^{n\epsilon} f_{\theta''}(\boldsymbol{y} \mid \boldsymbol{x}(i))\,d\nu\\ &\quad\le 2^{-nM} + \frac{1}{\lfloor 2^{nR}\rfloor}\sum_{i=1}^{\lfloor 2^{nR}\rfloor}\int_{\mathcal{D}_i^c} 2^{n\epsilon} f_{\theta''}(\boldsymbol{y} \mid \boldsymbol{x}(i))\,d\nu\\ &\quad= 2^{-nM} + 2^{n\epsilon}P_{\theta'',\phi}(\text{error} \mid \mathcal{C})\end{aligned}$$

which proves the first part of the lemma. The second part follows from the first part by choosing ϕ to be the ML decoder for θ'', by noting that by the optimality of the ML rule

$$P_{\theta',\theta'}(\text{error} \mid \mathcal{C}) \le P_{\theta'\theta''}(\text{error} \mid \mathcal{C})$$

and by averaging over the codebook $\mathcal{C}$. □

We can now define strong separability for general alphabets. Notice that, when applied to finite alphabets, this new definition of strong separability is slightly more inclusive than Definition 6.

Definition 7: A family of channels $\{p_\theta(\boldsymbol{y} \mid \boldsymbol{x})\ \theta \in \Theta\}$ defined over common general input and output alphabets $\mathcal{X},\mathcal{Y}$ is said to be *strongly separable* for the input sets $B_n \subseteq \mathcal{X}^n$ if there exists some $M > 0$ that upper-bounds the error

exponents in the family, i.e., that satisfies

$$\limsup_{n\to\infty}\sup_{\theta\in\Theta}-\frac{1}{n}\log \bar{P}_{\theta,\theta}(\text{error}) < M \tag{55}$$

such that for any $\epsilon > 0$ and blocklength n, there exists a subexponential number $K(n)$ (that depends on M and on ϵ) of channels $\{\theta_k^{(n)}\}_{k=1}^{K(n)} \subseteq \Theta$, such that for any $\theta \in \Theta$ there exists $\theta_{k^*}^{(n)} \in \Theta, 1 \le k^* \le K(n)$ that approximates θ in the following sense.

- For every $\boldsymbol{x} \in B_n$ there exists a measurable set $F_{\boldsymbol{x},\theta} \subset \mathcal{Y}^n$ such that

$$\int_{\boldsymbol{y}\notin F_{\boldsymbol{x},\theta}} f_\theta(\boldsymbol{y} \mid \boldsymbol{x})\, d\nu < 2^{-nM} \tag{56}$$

and

$$f_\theta(\boldsymbol{y} \mid \boldsymbol{x}) \le 2^{n\epsilon} f_{\theta_{k^*}^{(n)}}(\boldsymbol{y} \mid \boldsymbol{x}). \tag{57}$$

- For every $\boldsymbol{x} \in B_n$ there exists a measurable set $F_{\boldsymbol{x},\theta_{k^*}^{(n)}} \subset \mathcal{Y}^n$ such that

$$\int_{\boldsymbol{y}\notin F_{\boldsymbol{x},\theta_{k^*}^{(n)}}} f_{\theta_{k^*}^{(n)}}(\boldsymbol{y} \mid \boldsymbol{x})\, d\nu < 2^{-nM} \tag{58}$$

and

$$f_{\theta_{k^*}^{(n)}}(\boldsymbol{y} \mid \boldsymbol{x}) \le 2^{n\epsilon} f_\theta(\boldsymbol{y} \mid \boldsymbol{x}). \tag{59}$$

We now state the main result on universal decoding for general alphabets:

Theorem 5: If the family of channels $\{p_\theta(\boldsymbol{y} \mid \boldsymbol{x}), \theta \in \Theta\}$ is strongly separable in the sense of Definition 7 and if it satisfies the assumptions preceding Lemma 7 then it admits a random-coding and deterministic-coding strong universal decoder. If Θ can be written as a countable union of strongly separable families then the family admits a random-coding and deterministic-coding weak universal decoder.

Proof: The first part of the theorem follows from Lemmas 7 and 8 in much the same way that Theorem 2 follows from Lemmas 1 and 5. To prove the second part of the theorem note that if

$$\Theta = \bigcup_{m=1}^{\infty} \Theta^{(m)}$$

and $\{u_n^{(m)}\}_{m=1}^{\infty}$ is a sequence of strong random-coding universal decoders for $\Theta^{(m)}$ then the decoder u_n, the results from merging $u_n^{(1)}, \cdots, u_n^{(n)}$, is random-coding weakly universal for Θ. Deterministic-coding universality can be proved by methods similar to those employed in the proof of Lemma 4 by enumerating the union of all approximating channels, where the union is over the blocklengths n, and over the spaces $\Theta^{(m)}$. □

VII. Examples

In this section we shall consider different families of channels and study their separability properties. We shall also demonstrate by example that there are some families of channels that admit weak universal decoding but not strong universal decoding.

A. Discrete Memoryless Channels

Consider the case where the family of channels $\mathcal{F}$ is the family of all discrete memoryless channels (DMC's) over the finite input alphabet $\mathcal{X}$ of size $|\mathcal{X}|$ and the finite output alphabet $\mathcal{Y}$ of size $|\mathcal{Y}|$. This family of channels is parameterized naturally by the set of all $|\mathcal{X}|$ by $|\mathcal{Y}|$ stochastic matrices. We shall thus take this set of matrices as our parameter space Θ and have

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{x}) = \prod_{t=1}^{n} \theta(y_t \mid x_t)$$

where $\theta(y \mid x)$ denotes the entry in row x and column y of the matrix θ, and where $\boldsymbol{x} = (x_1, \cdots, x_n)$, and $\boldsymbol{y} = (y_1, \cdots, y_n)$. To simplify notation we are thus identifying the set $\mathcal{X}$ with the set $\{1, \cdots, |\mathcal{X}|\}$ and likewise for $\mathcal{Y}$.

Lemma 9: The family of all discrete memoryless channels over the finite input and output alphabets $\mathcal{X}, \mathcal{Y}$ is separable in the sense of Definition 5 for any sequence of input sets B_n.

Proof: Since the channels in the family are memoryless we have

$$\frac{p_\theta(\boldsymbol{y} \mid \boldsymbol{x})}{p_{\theta'}(\boldsymbol{y} \mid \boldsymbol{x})} = \prod_{t=1}^{n} \frac{\theta(y_t \mid x_t)}{\theta'(y_t \mid x_t)} \le \left(\max_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \frac{\theta(y \mid x)}{\theta'(y \mid x)}\right)^n.$$

We thus conclude that

$$\frac{1}{n}\left|\log \frac{p_\theta(\boldsymbol{y} \mid \boldsymbol{x})}{p_{\theta'}(\boldsymbol{y} \mid \boldsymbol{x})}\right| \le \max_{x,y}\left|\log \frac{\theta(y \mid x)}{\theta'(y \mid x)}\right|$$

and the required separability now follows by considering the countable set of all stochastic matrices with rational nonnegative (but including zero!) entries. □

Lemma 10: The family of all discrete memoryless channels over finite input and output alphabets $\mathcal{X}, \mathcal{Y}$ is strongly separable in the sense of Definition 6 for any input sets $\{B_n\}$.

Proof: Let $M > 0$ be a strict upper bound on the random-coding error exponents of all the channels in the family of DMC's over the alphabet $\mathcal{X}, \mathcal{Y}$, e.g., $M = 1 + \log|\mathcal{X}|$. By the discussion following Definition 6 this choice of M guarantees that (10) holds. Let

$$M' = M + \log|\mathcal{Y}|$$

and $\epsilon > 0$ be given, and assume for the simplicity of notation that M'/ϵ is an integer. Let the blocklength under consideration $n \ge 1$ be fixed. The idea of the proof is to quantize the set of all stochastic matrices by quantizing each component logarithmically.[4] Some cells will be empty, i.e., contain no stochastic matrices. From those cells that are not empty we choose an arbitrary representative. Special care must be taken in treating cells in which one of the components contains the element 0. The details follow.

[4]The proposed quantization is different from the uniform quantization that is often used to prove capacity results [16, p. 216], [14]. The finer analysis is required because of our interest in error exponents.

Divide the interval $[0, 1]$ into $1+nM'/\epsilon$ disjoint intervals, $I_0, \cdots, I_{nM'/\epsilon}$, where

$$I_0 = [0, 2^{-nM'}] \tag{60}$$

$$I_l = (2^{-((nM'/\epsilon)-l+1)\epsilon}, 2^{-((nM'/\epsilon)-l)\epsilon}], \qquad 1 \le l \le nM'/\epsilon. \tag{61}$$

Notice that except for the interval I_0 all the other intervals have the same ratio between their endpoints, and this ratio is 2^{ϵ}. Thus

$$\alpha, \beta \in I_l \quad \text{and} \quad l \neq 0 \Rightarrow \left|\log \frac{\alpha}{\beta}\right| \le \epsilon. \tag{62}$$

Consider now the component-wise quantization induced by the partition (60) and (61) on the set of all $|\mathcal{X}|$ by $|\mathcal{Y}|$ matrices with elements in $[0, 1]$. This quantization results in $(1+nM'/\epsilon)^{|\mathcal{X}||\mathcal{Y}|}$ cells, some of which contain stochastic matrices and some of which do not. Let $K(n)$ be the number of cells that contain stochastic matrices, and let $\{\theta_1^{(n)}, \cdots, \theta_{K(n)}^{(n)}\}$ be a set of $K(n)$ stochastic matrices representing those cells containing stochastic matrices, one from each cell. Since the total number of cells is polynomial in the blocklength n it follows that $K(n)$ is subexponential.

Given any stochastic matrix θ, let $\theta_{k^*}^{(n)}$ be the stochastic matrix that represents the cell in which θ lies. It follows from (62) and (60) that for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ at least one of the following conditions holds:

$$\left|\log \frac{\theta(y \mid x)}{\theta_{k^*}^{(n)}(y \mid x)}\right| < \epsilon$$

or

$$\max\left\{\theta(y \mid x), \theta_{k^*}^{(n)}(y \mid x)\right\} \le 2^{-nM'}$$

depending on whether $\theta(y \mid x)$ (and hence $\theta_{k^*}^{(n)}(y \mid x)$) lies in I_0 or not. Notice that this condition is symmetric in θ and $\theta_{k^*}^{(n)}$.

We shall next verify that this condition implies (12). By symmetry, this will also imply (13). Let $\boldsymbol{x} = (x_1, \cdots, x_n)$ and $\boldsymbol{y} = (y_1, \cdots, y_n)$ be given. If $\theta(y_{t^*} \mid x_{t^*}) \in I_0$ for some $1 \le t^* \le n$ then

$$\begin{aligned} p_\theta(\boldsymbol{y} \mid \boldsymbol{x}) &= \prod_{t=1}^{n} \theta(y_t \mid x_t) \\ &\le \theta(y_{t^*} \mid x_{t^*}) \\ &\le 2^{-n(M+\log|\mathcal{Y}|)} \end{aligned}$$

and we have nothing further to check, as (12) is satisfied trivially. If, however, $\theta(y_t \mid x_t) \notin I_0$, for for every $1 \le t \le n$, then by (62)

$$\theta(y_t \mid x_t) \le 2^{\epsilon}\theta_{k^*}^{(n)}(y_t \mid x_t), \qquad \forall 1 \le t \le n$$

and hence

$$\begin{aligned} p_\theta(\boldsymbol{y} \mid \boldsymbol{x}) &= \prod_{t=1}^{n} \theta(y_t \mid x_t) \\ &\le \prod_{t=1}^{n} 2^{\epsilon}\theta_{k^*}^{(n)}(y_t \mid x_t) \\ &= 2^{n\epsilon} p_{\theta_{k^*}^{(n)}}(\boldsymbol{y} \mid \boldsymbol{x}) \end{aligned}$$

and (12) holds. □

B. Finite-State Channels

We next consider the family of all finite-state channel that are defined over common finite input, output, and state alphabets $\mathcal{X}, \mathcal{Y}, \mathcal{S}$, respectively. The probability law of any channel in this family is characterized by a conditional probability assignment

$$P_\theta(y, s' \mid x, s), \qquad y \in \mathcal{Y},\ x \in \mathcal{X}, s,\ s' \in \mathcal{S}$$

and an initial state $s_0 \in \mathcal{S}$. Operationally, if at time $t-1$ the state of the channel is s_{t-1} and the input to the channel at time t is x_t, then the output of the channel y_t at time t and the state s_t of the channel at time t are determined according to the distribution

$$P_\theta(y_t, s_t \mid x_t, s_{t-1}).$$

For any input sequence $\boldsymbol{x}$ and output sequence $\boldsymbol{y}$ of length n we have that conditional on the initial state s_0

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{x}, s_0) = \sum_{\boldsymbol{s} \in \mathcal{S}^n} p_\theta(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0) \tag{63}$$

where

$$p_\theta(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0) = \prod_{t=1}^{n} P_\theta(y_t, s_t \mid x_t, s_{t-1}) \tag{64}$$

and $\boldsymbol{s} = (s_1, \cdots, s_n) \in \mathcal{S}^n$. It is helpful to think of the family of finite-state channels as being parameterized by $(\theta, s_0) \in \Theta \times \mathcal{S}$ because the probability of an output sequence $\boldsymbol{y}$ given an input sequence $\boldsymbol{x}$ is determined by the initial state s_0 and by the probability assignment $P_\theta(y, s' \mid x, s)$.

Lemma 11: The family of all finite-state channels over the finite input, output, and state alphabets $\mathcal{X}, \mathcal{Y}, \mathcal{S}$, respectively, is separable in the sense of Definition 5 for any sequence of input sets B_n.

Proof: It follows from Lemma 2 and from (63) and (64) that for any input sequence $\boldsymbol{x}$, output sequence $\boldsymbol{y}$, and initial state $s_0 \in \mathcal{S}$

$$\begin{aligned} \frac{p_\theta(\boldsymbol{y} \mid \boldsymbol{x}, s_0)}{p_{\theta'}(\boldsymbol{y} \mid \boldsymbol{x}, s_0)} &= \frac{\sum_{\boldsymbol{s} \in \mathcal{S}^n} p_\theta(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0)}{\sum_{\boldsymbol{s} \in \mathcal{S}} p_{\theta'}(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0)} \\ &\le \max_{\boldsymbol{s}} \frac{\prod_{t=1}^{n} P_\theta(y_t, s_t \mid x_t, s_{t-1})}{\prod_{t=1}^{n} P_{\theta'}(y_t, s_t \mid x_t, s_{t-1})} \\ &\le \left(\max_{s', s'', x', y'} \frac{P_\theta(y, s'' \mid x', s')}{P_{\theta'}(y, s'' \mid x', s')}\right)^n. \end{aligned}$$

Taking the logarithm of the above equation and considering the same argument applied to θ and θ' in reverse roles we obtain

$$\max_{\boldsymbol{x}, \boldsymbol{y}} \frac{1}{n}\left|\log \frac{p_\theta(\boldsymbol{y} \mid \boldsymbol{x}, s_0)}{p_{\theta'}(\boldsymbol{y} \mid \boldsymbol{x}, s_0)}\right| \le \max_{s', s'', x', y'} \left|\log \frac{P_\theta(y, s'' \mid x', s')}{P_{\theta'}(y, s'' \mid x', s')}\right| \tag{65}$$

The separability of the family now follows by considering the countable family of channels $(P_{\theta_k}(y, s' \mid x, s), s_0)$ consisting of conditional distributions with (nonnegative) rational components and all $s_0 \in \mathcal{S}$. □

Lemma 12: 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 in the sense of Definition 6 for any input sets $\{B_n\}$.

Proof: We shall fix the initial state s_0 and show the existence of a subexponential number of approximating channels for that initial state. Since the number of states is finite, the general result will follow by taking the union of the approximating channels and initial states. Let $M > 0$ upper-bound the error exponents in the family, say $M = 1+\log|\mathcal{X}|$, and set

$$M' = M + \log|\mathcal{Y}| + \log|\mathcal{S}| + 1. \tag{66}$$

Let the blocklength n be fixed, as well as some $\epsilon' > 0$, and let $0 < \epsilon < \epsilon'$ be such that $2 \cdot 2^{n\epsilon} \le 2^{n\epsilon'}$. To simplify notation assume that M'/ϵ is an integer.

Any conditional probability assignments $P_\theta(y, s \mid s', x)$ can be represented by a matrix of $|\mathcal{X}||\mathcal{S}|$ rows and $|\mathcal{Y}||\mathcal{S}|$ columns. To simplify notation we shall use the matrix notation $\theta(y, s \mid s', x)$ for $P_\theta(y, s \mid s', x)$. As in the proof of the strong separability of the family of DMC's, we shall quantize this set of matrices component-wise on a logarithmic scale, as in (60) and (61). Choosing stochastic matrices to represent the cells (of which there are a polynomial number) that contain stochastic matrices as in the proof of the strong separability of the family of DMC's, we can conclude that for any $\theta \in \Theta$ there exists some $\theta^{(n)}_{k^*}$ such that

$$\log\left|\frac{\theta(y, s \mid s', x)}{\theta^{(n)}_{k^*}(y, s \mid s', x)}\right| < \epsilon, \qquad \forall (y, s, s', x) \in G \tag{67}$$

and

$$\max\left\{\theta(y, s \mid s', x), \theta^{(n)}_{k^*}(y, s \mid s', x)\right\} \le 2^{-nM'}, \qquad \forall (y, s, s', x) \notin G \tag{68}$$

where the set G corresponds to components of the matrix θ that do not fall in the interval I_0, i.e.,

$$G = \{(y, s, s', x) : \theta(y, s \mid s', x) \notin I_0\}.$$

Notice that because θ and $\theta^{(n)}_{k^*}$ are in the same cell we also have

$$G = \left\{(y, s, s', x) : \theta^{(n)}_{k^*}(y, s \mid s', x) \notin I_0\right\}.$$

Conditions (67) and (68) are thus completely symmetric with respect to interchanging θ and $\theta^{(n)}_{k^*}$ and thus it suffices to show that these conditions imply (12), because (13) will then follow by symmetry.

Given an input sequence $\boldsymbol{x} = (x_1, \cdots, x_n)$, an output sequence $\boldsymbol{y} = (y_1, \cdots, y_n)$, and an initial state s_0, we define

$$\mathcal{G} = \{\boldsymbol{s} \in \mathcal{S}^n : (y_t, s_t, s_{t-1}, x_t) \in G, \quad \forall 1 \le t \le n\}.$$

Thus $\mathcal{G}$ is the set of "good" state sequences in the sense that for every component t we have that

$$\left|\log\frac{\theta(y_t, s_t \mid s_{t-1}, x_t)}{\theta^{(n)}_{k^*}(y_t, s_t \mid s_{t-1}, x_t)}\right| < \epsilon$$

and hence

$$\frac{1}{n}\left|\log\frac{p_\theta(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0)}{p_{\theta^{(n)}_{k^*}}(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0)}\right| < \epsilon, \qquad \forall \boldsymbol{s} \in \mathcal{G}. \tag{69}$$

Invoking Lemma 2 we have from (69) that

$$\frac{1}{n}\left|\log\frac{\sum_{\boldsymbol{s}\in\mathcal{G}} p_\theta(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0)}{\sum_{\boldsymbol{s}\in\mathcal{G}} p_{\theta^{(n)}_{k^*}}(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0)}\right| < \epsilon. \tag{70}$$

The complement of $\mathcal{G}$, denoted $\mathcal{G}^c$, is referred to as the set of "bad" sequences. Since

$$p_\theta(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0) \le \theta(y_t, s_t \mid s_{t-1}, x_t), \qquad \forall 1 \le t \le n$$

we have by (68)

$$\max\left\{p_\theta(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0), p_{\theta^{(n)}_{k^*}}(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0)\right\} \le 2^{-nM'}, \quad \forall \boldsymbol{s} \notin \mathcal{G}. \tag{71}$$

and since the number of state sequences is $|\mathcal{S}|^n$ it follows from (66) and (71) that

$$\max\left\{\sum_{\boldsymbol{s}\in\mathcal{G}^c} p_\theta(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0), \sum_{\boldsymbol{s}\in\mathcal{G}^c} p_{\theta^{(n)}_{k^*}}(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0)\right\} \le 2^{-n(M+\log|\mathcal{Y}|+1)}. \tag{72}$$

To show that (12) holds for all sequence $\boldsymbol{x}$, and $\boldsymbol{y}$ we treat two cases:

Case 1:

$$\sum_{\boldsymbol{s}\in\mathcal{G}} p_\theta(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0) < 2^{-n(M+1+\log|\mathcal{Y}|)}. \tag{73}$$

In this case, it follows from (72) that

$$\begin{aligned} p_\theta(\boldsymbol{y} \mid \boldsymbol{x}, s_0) &= \sum_{\boldsymbol{s}\in\mathcal{G}^c} p_\theta(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0) + \sum_{\boldsymbol{s}\in\mathcal{G}} p_\theta(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0) \\ &\le 2^{-n(M+1+\log|\mathcal{Y}|)} + 2^{-n(M+1+\log|\mathcal{Y}|)} \\ &\le 2^{-n(M+\log|\mathcal{Y}|)} \end{aligned}$$

and for such $\boldsymbol{x}, \boldsymbol{y}$ (12) holds in the trivial sense.

Case 2: The sequences $\boldsymbol{x}$ and $\boldsymbol{y}$ are such that

$$\sum_{\boldsymbol{s}\in\mathcal{G}} p_\theta(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0) \ge 2^{-n(M+1+\log|\mathcal{Y}|)} \tag{74}$$

and hence, by (70)

$$\sum_{\boldsymbol{s}\in\mathcal{G}} p_{\theta^{(n)}_{k^*}}(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0) \ge 2^{-n\epsilon} 2^{-n(M+1+\log|\mathcal{Y}|)}. \tag{75}$$

For such sequences (12) holds because

$$\begin{aligned} &\frac{p_\theta(\boldsymbol{y} \mid \boldsymbol{x}, s_0)}{p_{\theta^{(n)}_{k^*}}(\boldsymbol{y} \mid \boldsymbol{x}, s_0)} \\ &\quad= \frac{\sum_{\boldsymbol{s}\in\mathcal{G}} p_\theta(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0) + \sum_{\boldsymbol{s}\in\mathcal{G}^c} p_\theta(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0)}{\sum_{\boldsymbol{s}\in\mathcal{G}} p_{\theta^{(n)}_{k^*}}(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0) + \sum_{\boldsymbol{s}\in\mathcal{G}^c} p_{\theta^{(n)}_{k^*}}(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0)} \\ &\quad\le \frac{\sum_{\boldsymbol{s}\in\mathcal{G}} p_\theta(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0) + 2^{-n(M+1+\log|\mathcal{Y}|)}}{\sum_{\boldsymbol{s}\in\mathcal{G}} p_{\theta^{(n)}_{k^*}}(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0)} \\ &\quad\le 2^{n\epsilon} + \frac{2^{-n(M+1+\log|\mathcal{Y}|)}}{\sum_{\boldsymbol{s}\in\mathcal{G}} p_{\theta^{(n)}_{k^*}}(\boldsymbol{y}, \boldsymbol{s} \mid \boldsymbol{x}, s_0)} \\ &\quad\le 2^{n\epsilon} + 2^{n\epsilon} \\ &\quad\le 2^{n\epsilon'} \end{aligned}$$

where the first inequality follows from (72), the second from (70), and the third from (75). □

In the above derivation we have assumed that the number of states $|\mathcal{S}|$ is known to the receiver designer. In fact, only an upper bound $\tilde{S}$ on the number of states is required, as every finite-state channel with $|\mathcal{S}|$ states can be described as a finite-state channel with $\tilde{S} > |\mathcal{A}|$ states by duplicating some of the states. Note, however, that the rate of convergence of the universal decoder depends significantly on the number of states, and designing the receiver to account for more states than the channel really has results in poor rates of convergence. This problem can be solved by designing a "double-universal" decoder. Here we design $\tilde{S}+1$ universal decoders $\{u_n^{(\nu)}\}_{n=1}^{\infty}$ for each of the possible number of states $\nu = 0, \cdots, \tilde{S}$, and then merge the $\tilde{S}+1$ decoders to obtain the double-universal decoder. The double-universal decoder now has a rate of convergence which is at most $\tilde{S}+1$ worse than that of the universal decoder that could have been designed had the number of states been known in advance.

This approach is the dual of the twice-universal source coding approach of [32] and [33].

If the number of states is completely arbitrary, then strong universality cannot be guaranteed, and we can only guarantee weak universality. The latter can be guaranteed by merging $u_n^{(0)}, \cdots, u_n^{(\nu(n))}$ decoders where $\nu(n)$ is subexponentially increasing in the blocklength, and $\{u_n^{(\nu)}\}$ is universal for a finite-state channel with ν states.

C. Intersymbol Interference

In [21] Merhav posed the problem of designing a universal decoder for the discrete-time Gaussian channel with unknown intersymbol interference (ISI) coefficients. The input and output alphabets are both the real line, and

$$Y_t = \sum_{j=0}^{J} h_j X_{t-j} + Z_t$$

where $\boldsymbol{h} = (h_0, \cdots, h_J)$ is the vector of unknown ISI coefficients, and $\{Z_t\}$ is independent of the input and is a sequence of independent Gaussian random variables of zero mean and unit variance. We shall next demonstrate that if the ISI coefficients satisfy an energy constraint of the form (14), and if the input set B_n satisfy an average power constraint (15), where J, H, and P are all known, then the family is strongly separable and a strong universal decoder exists by Theorem 5. If J and H are unknown then we can consider the countable union of ISI channels over all integers J and H to obtain a weak universal decoder for the case where J and H are unknown (but finite).

In this problem the output distribution corresponding to any input $\boldsymbol{x}$ and any ISI sequence $\boldsymbol{h}$ is absolutely continuous with respect to the Lebesgue measure with density

$$f_{\boldsymbol{h}}(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}} \prod_{t=1}^{n} \exp\left\{ -\frac{1}{2} \left(y_t - \sum_{j=0}^{J} h_j x_{t-j} \right)^2 \right\} \tag{76}$$

where we are defining $x_t = 0$ for $t \le 0$, and we are using $\boldsymbol{h}$ rather than θ to parameterize the family.

To establish strong separability first note that by analyzing the two-codewords case one can determine that

$$\bar{P}_{\boldsymbol{h},\boldsymbol{h}}(\text{error}) \ge \mathcal{Q}(\sqrt{nH(J+1)P}) \tag{77}$$

where, see [34, eq. (2.3.18)]

$$\begin{aligned} \mathcal{Q}(\beta) &= \frac{1}{\sqrt{2\pi}} \int_{\beta}^{\infty} e^{-\xi^2/2}\, d\xi \\ &\ge \left(1 - \frac{1}{\beta^2}\right) \frac{e^{-\beta^2/2}}{\sqrt{2\pi}\beta}. \end{aligned} \tag{78}$$

This follows from a simple energy calculation and the Cauchy–Schwartz inequality by noting that

$$\begin{aligned} \sum_{t=1}^{n} \left(\sum_{j=0}^{J} h_j x_{t-j} \right)^2 &\le \sum_{t=1}^{n} \left(\sum_{j=0}^{J} h_j^2 \sum_{j=0}^{J} x_{t-j}^2 \right) \\ &\le nH(J+1)P. \end{aligned}$$

It follows from (77) and (78) that the error exponents of the channels in the family are bounded and

$$M = \frac{1}{2} H(J+1)P \log_2(e) + 1 \tag{79}$$

satisfies (55).

The following lemma, which is proved in Appendix III, will be useful in establishing strong separability.

Lemma 13: Given two sets of ISI coefficients

$$\begin{aligned} \boldsymbol{h}' &= (h_0', \cdots, h_J') \\ \boldsymbol{h}'' &= (h_0'', \cdots, h_J'') \end{aligned}$$

and some $\boldsymbol{x}$ satisfying $\sum x_t^2 \le nP$

$$\begin{aligned} &\left| \frac{1}{n} \sum_{t=1}^{n} \left[\left(y_t - \sum_{j=0}^{J} h_j' x_{t-j} \right)^2 - \left(y_t - \sum_{j=0}^{J} h_j'' x_{t-j} \right)^2 \right] \right| \\ &\le \sqrt{P(J+1)} \|\boldsymbol{h} - \boldsymbol{h}'\|_2 (2\sqrt{Q} + \|\boldsymbol{h} - \boldsymbol{h}'\|_2 \sqrt{P(J+1)}) \end{aligned}$$

where

$$\begin{aligned} \|\boldsymbol{h} - \boldsymbol{h}'\|_2 &= \left(\sum_{j=0}^{J} (h_j' - h_j'')^2 \right)^{1/2} \\ Q &= \min\{Q', Q''\} \\ Q' &= \frac{1}{n} \sum_{t=1}^{n} \left(y_t - \sum_{j=0}^{J} h_j' x_{t-j} \right)^2 \end{aligned} \tag{80}$$

and

$$Q'' = \frac{1}{n} \sum_{t=1}^{n} \left(y_t - \sum_{j=0}^{J} h_j'' x_{t-j} \right)^2. \tag{81}$$

We are now in a position of prove the strong separability of the family. Given M as in (79) we can find, by the Large Deviations principle [35], some sufficiently large Q so that

$$\Pr\left\{ \frac{1}{n} \sum_{t=1}^{n} Z_t^2 > Q \right\} < 2^{-nM} \tag{82}$$

where $\{Z_t\}$ are i.i.d. Normal random variables of zero mean and unit variance. Given any $\epsilon > 0$ we can find by Lemma 13 and (76) some sufficiently small $\delta > 0$ (which depends on ϵ, Q, J, P, and H) so that

$$\|\boldsymbol{h}' - \boldsymbol{h}''\|_2 < \delta \tag{83}$$

implies

$$\frac{1}{n}\left|\log \frac{f_{\boldsymbol{h}'}(\boldsymbol{y} \mid \boldsymbol{x})}{f_{\boldsymbol{h}''}(\boldsymbol{y} \mid \boldsymbol{x})}\right| \leq \epsilon \tag{84}$$

whenever $\sum x_t^2 \leq nP$ and

$$\min\left\{\frac{1}{n}\sum_{t=1}^{n}\left(y_t - \sum_{j=0}^{J} h'_j x_{t-j}\right)^2, \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \sum_{j=0}^{J} h''_j x_{t-j}\right)^2\right\} \leq Q.$$

We now choose the grid $\boldsymbol{h}_1^{(n)}, \cdots, \boldsymbol{h}_{K(n)}^{(n)}$ to guarantee that for every $\boldsymbol{h}$ satisfying (14) there exists some $\boldsymbol{h}_{k^*}^{(n)}, 1 \leq k^* \leq K(n)$, such that

$$\left\|\boldsymbol{h} - \boldsymbol{h}_{k^*}^{(n)}\right\|_2 < \delta$$

with $K(n)$ subexponential. This can be clearly done because any ball of radius H in $\mathbb{R}^{J+1}$ can be covered by

$$\left(\left\lceil \frac{H\sqrt{J+1}}{\delta} \right\rceil\right)^{J+1}$$

balls of radius δ as can be easily verified by considering the size of the smallest cube containing the H-ball, and the largest cube contained in the δ-ball.

Given any $\|\boldsymbol{h}\|_2 \leq H$ let $\boldsymbol{h}_{k^*}^{(n)}$ be such that

$$\left\|\boldsymbol{h} - \boldsymbol{h}_{k^*}^{(n)}\right\|_2 < \delta.$$

For any $\boldsymbol{x}$ satisfying $\sum_{t=1}^{n} x_t^2 \leq nP$ let

$$F_{\boldsymbol{x},\boldsymbol{h}} = \left\{\boldsymbol{y} : \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \sum_{j=0}^{J} h_j x_{t-j}\right)^2 \leq Q\right\}.$$

This choice guarantees that (56) holds by (82), and that (57) holds by (84). The second requirement of Definition 7 follows by a similar argument. This establishes the strong separability of this class of ISI channels, and Theorem 4 is thus proved.

It is interesting to note that the number of δ-balls required to cover the H-ball does not grow with the blocklength n. This leads us to suspect that for this family the convergence of the performance of the universal decoder to that of the ML decoder is very good.

The convergence does, however, depend significantly on the number of ISI coefficients. Using the previously discussed "double-universality" approach (see Section VII-B), one can, however, guarantee that the rate of convergence be essentially determined by the number of ISI coefficients, even if only an upper bound on that number is given.

D. A Pathological Example

The following is an example that demonstrates that some families admit weak universal decoding but not a strong one. The example is really the binary-added arbitrarily varying channel (AVC) in disguise; see [36, p. 189, Example 1] and references therein.

Consider the family of channels with binary inputs and binary outputs (i.e., $\mathcal{X} = \mathcal{Y} = \{0,1\}$) that is parameterized by Θ, where Θ is the countable set of all half-infinite binary sequences that have a finite number of ones. Let $\theta^{(1)}, \theta^{(2)}, \cdots$ denote the binary sequence corresponding to $\theta \in \Theta$, and let

$$p_\theta(\boldsymbol{y} \mid \boldsymbol{x}) = \begin{cases} 1, & \text{if } \boldsymbol{y} = \boldsymbol{x} \oplus \theta \\ 0, & \text{otherwise.} \end{cases}$$

Thus if the sequence $\boldsymbol{x} = (x_1, \cdots, x_n) \in \mathcal{X}^n$ is transmitted through the channel of parameter $\theta = \theta^{(1)}, \theta^{(2)}, \cdots$ then the resulting output is $\boldsymbol{y} \in \mathcal{Y}^n$ where

$$\boldsymbol{y} = x_1 \oplus \theta^{(1)}, \cdots, x_n \oplus \theta^{(n)},$$

and $\oplus$ denotes mod-2 addition (exclusive or).

Every channel $p_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ has capacity 1 bit, and if random coding is carried out uniformly over the set of input sequences with an equal number of zeros and ones[5] then the resulting error exponent is $1-R$, for $0 \leq R \leq 1$ (see [11], [16]), because if θ is known then for all practical purposes the channel $p_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ behaves like a noiseless binary-symmetric channel. Since the parameter space Θ is countable it is separable, and Theorem 1 guarantees the existence of a deterministic-coding and random-coding weak universal decoder for the family.

Yet one can easily show using standard techniques from the theory of arbitrarily varying channels [31], [37], [36, p. 185, eq. (3.15); p. 189, Example 1] that for any code $\mathcal{C}$ (with more than one codeword) and any decoder ϕ that is ignorant of the channel over which transmission is carried out, the average probability of error, maximized over the parameter θ, is bounded from below by $1/4$. There is thus no way to achieve uniformly good performance over all the channels in the family, and a strong deterministic-coding universal decoder does not exist for this family. In fact, by [36, p. 189, Example

[5] This is the choice for even blocklength n. For odd n we can take those sequences where the number of ones exceeds the number of zeros by 1.

1] there does not exists a strong random-coding universal decoder for this family either.

VIII. Summary and Conclusions

In this paper we have demonstrated that for many of the families of channels that are of interest in wireless communication the ignorance of the receiver of the channel in use is not a fundamental impediment for reliable communication. The receiver can employ the proposed universal decoding algorithm to asymptotically perform as well as the maximum-likelihood decoder tuned to the channel in use.

These results easily extend to multiple-access channels. Consider an I-to-one multiple-access channel (MAC) where user ι draws its codewords independently and uniformly over a set $B_n^{(\iota)} \subset \mathcal{X}_{(\iota)}^n$, where $\mathcal{X}_{(\iota)}$ is user-ι's (finite) input alphabet, $1 \le \iota \le I$. A receiver for such a channel can be described by specifying a ranking of $B_n^{(1)} \times \cdots \times B_n^{(I)}$ for each possible received sequence $\boldsymbol{y} \in \mathcal{Y}^n$, where $\mathcal{Y}$ is the output alphabet. The idea of merging decoders extends straightforwardly to the MAC, as do most of the results of the paper. With these tools we can thus demonstrate the existence of universal decoders for fairly general multiple-access channels with memory, thus extending the results of [38] and [39] on universal decoding for memoryless multiple-access channels.

The penalty for not knowing the channel seems to be in complexity. The universal decoder proposed in this paper might, and often does, have a complexity that is much higher than that of the maximum-likelihood decoder. Particularly, since the universal decoder is based on ranking functions and the idea of merging, it is required, for any given received sequence $\boldsymbol{y}$, to compute the ranking of each codeword among all the possible sequences $\boldsymbol{x} \in B_n$ according to each of a polynomial number of channel laws. This can result in formidable complexity particularly if the cost of evaluating $p_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ is high, as is the case for finite-state channels where it is exponential in the blocklength (63).

The existence of universal decoders motivates the search for decoders that are not only universal but also computationally efficient. Some promising results in this direction have been recently reported in [29] and [40].

Appendix I

In this appendix we demonstrate by a simple example that the naive training sequence approach to communicating over unknown channels does not, in general, yield a universal decoder. Consider the simple case where the family of channels $\mathcal{F}$ consists of only two channels, a BSC with crossover probability 0.25 and a BSC with crossover probability 0.75. We denote the first law by $p_1(\boldsymbol{y} \mid \boldsymbol{x})$ and the latter by $p_2(\boldsymbol{y} \mid \boldsymbol{x})$. Clearly, the ML decoding rule for the first channel is minimum Hamming distance decoding, whereas the rule for the second is maximum Hamming distance decoding. Assuming that random-coding is performed so that $\lfloor 2^{nR} \rfloor$ codewords are drawn independently and uniformly over the set of all n-length sequences with an equal number of zeros and ones, we have that the resulting random-coding error exponents are identical [16], and we denote them by $E(R)$. Thus

$$\bar{P}_{1,1}(\text{error}) = \bar{P}_{2,2}(\text{error}) \approx e^{-nE(R)}.$$

Consider now a training sequence approach to the problem where each block of length n begins with a training sequence of length m followed by $n-m$ unknown symbols that constitute a codeword of length $n-m$ from a random codebook with $\lfloor 2^{nR} \rfloor$ codewords. The resulting code, consisting of the training sequence and unknown symbols is thus of rate R and blocklength n. The decoder decides which channel in the family is in use by counting the number of bit inversions in the training sequence, and subsequently uses minimum or maximum Hamming distance decoding for the unknown symbols accordingly, depending on whether more than a half of the training bits were flipped or not.

To analyze the performance of the training sequence approach, let us break up the overall probability of error depending on whether the decoder correctly identifies the channel or not. By Bayes' rule

$$\begin{aligned} P_e = {} & \Pr(\text{correct id.}) \Pr(\text{error} \mid \text{matched dec.}) \\ & + \Pr(\text{incorrect id.}) \Pr(\text{error} \mid \text{mismatched dec.}). \end{aligned}$$

It is fairly straightforward to see that as $n-m$ tends to infinity the probability of error under mismatch conditions tends to one [6]. Likewise, as m tends to infinity, the probability of correct identification $\Pr(\text{correct id.})$ tends to one. Thus

$$P_e \approx \Pr(\text{error} \mid \text{matched dec.}) + \Pr(\text{incorrect id.})$$

and the fact that the training sequence approach does not yield a universal decoder now follows by noting that, by the large deviations principle [35], for the probability of incorrect identification to decrease exponentially in n, the length of the training sequence m must grow linearly in n.

Appendix II

In this appendix we give a Proof of Lemma 2. We start with the first claim of the lemma. First note that the function $f(z)$ is monotonically increasing in the interval $[0, 1]$, and the case $s \le t$ is thus proved. Consider now the case $s \ge t$. Observe that for any $N \ge 1$ the function $f(z)$ is concave in z for $0 \le z \le 1$, and that $f(0) = 0$. Thus by Jensen's inequality, for any $0 \le \alpha \le 1$ and any $0 \le z \le 1$

$$\begin{aligned} f(\alpha z) &= f(\alpha z + (1-\alpha)0) \\ &\ge \alpha f(z) + (1-\alpha) f(0) \\ &= \alpha f(z). \end{aligned}$$

Choosing $\alpha = t/s$ and $z = s$ now concludes the proof of this part.

A proof of the second part of this lemma can be found in [41, Lemma 1]. Note, however, that we have chosen to define $0/0 = 1$ whereas it is defined as 0 in [41]. It is, however, straightforward to verify that the results still hold.

The third claim of the lemma is trivial because it holds point-wise and must therefore also hold in expectation.

Appendix III

The Proof of Lemma 13 is based on repeated application of the Cauchy–Schwartz inequality: Let

$$\begin{aligned} A &= \left| \frac{1}{n} \sum_{t=1}^{n} \left[\left(y_t - \sum_{j=0}^{J} h'_j x_{t-j} \right)^2 - \left(y_t - \sum_{j=0}^{J} h''_j x_{t-j} \right)^2 \right] \right| \\ &= \left| \frac{1}{n} \sum_{t=1}^{n} \left(\sum_{j=0}^{J} (h''_j - h'_j) x_{t-j} \right) \times \left(2y_t - \sum_{j=0}^{J} (h''_j + h'_j) x_{t-j} \right) \right| \\ &\le \frac{1}{n} \left(\sum_{t=1}^{n} \alpha_t^2 \right)^{1/2} \left(\sum_{t=1}^{n} \beta_t^2 \right)^{1/2} \end{aligned}$$

where the last step follows from the Cauchy–Schwartz inequality with

$$\begin{aligned} \alpha_t^2 &= \left(\sum_{j=0}^{J} (h''_j - h'_j) x_{t-j} \right)^2 \\ &\le \left(\sum_{j=0}^{J} (h''_j - h'_j)^2 \right) \sum_{j=0}^{J} x_{t-j}^2 \end{aligned} \tag{85}$$

and

$$\begin{aligned} \beta_t^2 &= \left(2y_t - \sum_{j=0}^{J} (h''_j + h'_j) x_{t-j} \right)^2 \\ &= \left(2 \left(y_t - \sum_{j=0}^{J} h'_j x_{t-j} \right) + \sum_{j=0}^{J} (h'_j - h''_j) x_{t-j} \right)^2. \end{aligned} \tag{86}$$

We thus have from (85)

$$\begin{aligned} \left(\sum_{t=1}^{n} \alpha_t^2 \right)^{1/2} &\le \sqrt{J+1} \|\boldsymbol{h}'' - \boldsymbol{h}'\|_2 \left(\sum_{t=1}^{n} x_t^2 \right)^{1/2} \\ &= \sqrt{J+1} \|\boldsymbol{h}'' - \boldsymbol{h}'\|_2 \sqrt{nP} \end{aligned} \tag{87}$$

and by the triangle inequality

$$\left(\sum_{t=1}^{n} \beta_t^2 \right)^{1/2} \le 2 \left[\sum_{t=1}^{n} \left(y_t - \sum_{j=0}^{J} h'_j x_{t-j} \right)^2 \right]^{1/2} \tag{88}$$

$$+ \left[\sum_{t=1}^{n} \left(\sum_{j=0}^{J} (h'_j - h''_j) x_{t-j} \right)^2 \right]^{1/2}$$

$$\le 2 \left[\sum_{t=1}^{n} \left(y_t - \sum_{j=0}^{J} h'_j x_{t-j} \right)^2 \right]^{1/2} \tag{89}$$

$$+ \sqrt{J+1} \|\boldsymbol{h}'' - \boldsymbol{h}'\|_2 \sqrt{nP}. \tag{90}$$

Recalling the Definition (80) of Q' we have

$$\begin{aligned} A &\le \sqrt{P(J+1)} \cdot \|\boldsymbol{h}'' - \boldsymbol{h}'\|_2 \\ &\quad \cdot (2\sqrt{Q'} + \|\boldsymbol{h}'' - \boldsymbol{h}'\|_2 \sqrt{P(J+1)}). \end{aligned} \tag{91}$$

By symmetry we also have

$$\begin{aligned} A &\le \sqrt{P(J+1)} \cdot \|\boldsymbol{h}'' - \boldsymbol{h}'\|_2 \\ &\quad \cdot (2\sqrt{Q''} + \|\boldsymbol{h}'' - \boldsymbol{h}'\|_2 \sqrt{P(J+1)}) \end{aligned} \tag{92}$$

where Q'' is defined in (81). Inequalities (91) and (92) conclude the proof of the lemma. □

Acknowledgment

Stimulating discussions with R. G. Gallager, N. Merhav, P. Narayan, M. D. Trott, and J. Ziv are gratefully acknowledged.

References

[1] "GSM recommendations series 05, especially 05.03."
[2] M. R. L. Hodges, "The GSM radio interface," *Brit. Telecom. Technol. J.*, vol. 8, pp. 31–43, Jan. 1990.
[3] J. K. Omura and B. K. Levitt, "Coded error probability evaluation for antijam communication systems," *IEEE Trans. Commun.*, vol. COM-30, pp. 896–903, May 1982.
[4] A. Lapidoth and S. Shamai (Shitz), "A lower bound on the bit-error-rate resulting from mismatched Viterbi decoding," *Europ. Trans. Telecommun.*, 1998, to be published.
[5] I. Csiszár and J. Körner, "Graph decomposition: A new key to coding theorems," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 5–12, Jan. 1981.
[6] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai (Shitz), "On information rates for mismatched decoders," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1953–1967, Nov. 1994.
[7] I. Csiszár and P. Narayan, "Channel capacity for a given decoding metric," *IEEE Trans. Inform. Theory*, vol. 41, pp. 35–43, Jan. 1995.
[8] V. B. Balakirsky, "A converse coding theorem for mismatched decoding at the output of binary-input memoryless channels," *IEEE Trans. Inform. Theory*, vol. 41, pp. 1889–1902, Nov. 1995.
[9] A. Lapidoth, "Nearest-neighbor decoding for additive non-Gaussian noise channels," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1520–1529, Sept. 1996.
[10] ———, "Mismatched decoding and the multiple-access channel," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1439–1452, Sept. 1996.
[11] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
[12] A. Lapidoth and İ. E. Telatar, "The compound channel capacity of a class of finite state channels," *IEEE Trans. Inform. Theory*, vol. 44, pp. 973–983, May 1998.
[13] E. L. Lehmann, *Testing Statistical Hypotheses*, 2nd ed. Pacific Grove, CA: Wadsworth & Brooks, 1991.
[14] D. Blackwell, L. Breiman, and A. J. Thomasian, "The capacity of a class of channels," *Ann. Math. Stat.*, vol. 30, pp. 1229–1241, Dec. 1959.
[15] J. Wolfowitz, *Coding Theorems of Information Theory*, 3rd ed. Berlin, Germany: Springer-Verlag, 1978.
[16] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.
[17] W. L. Root and P. P. Varaiya, "Capacity of classes of Gaussian channels," *SIAM J. Appl. Math.*, vol. 16, pp. 1350–1393, Nov. 1968.
[18] I. G. Stiglitz, "Coding for a class of unknown channels," *IEEE Trans. Inform. Theory*, vol. IT-12, pp. 189–195, Apr. 1966.

[19] J. Ziv, "Universal decoding for finite-state channels," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 453–460, July 1985.
[20] V. D. Goppa, "Nonprobabalistic mutual information without memory," *Probl. Contr. Inform. Theory*, vol. 4, pp. 97–102, 1975.
[21] N. Merhav, "Universal decoding for memoryless Gaussian channels with a deterministic interference," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1261–1269, July 1993.
[22] A. J. Goldsmith and P. P. Varaiya, "Capacity, mutual information, and coding for finite-state Markov channels," *IEEE Trans. Inform. Theory*, vol. 42, pp. 868–886, May 1996.
[23] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell Syst. Tech. J.*, vol. 39, pp. 1253–1266, Sept. 1960.
[24] E. O. Elliott, "Estimates of error rates for codes on burst-noise channels," *Bell Syst. Tech. J.*, vol. 42, pp. 1977–1997, Sept. 1963.
[25] M. Mushkin and I. Bar-David, "Capacity and coding for the Gilbert-Elliott channel," *IEEE Trans. Inform. Theory*, vol. IT-35, pp. 1277–1290, Nov. 1989.
[26] G. Bratt, "Sequential decoding for the Gilbert-Elliott channel—Strategy and analysis," Ph.D. dissertation, Lund Univ., Lund, Sweden, June 1994.
[27] C. E. Shannon, "Probability of error for optimal codes in a Gaussian channel," *Bell Syst. Tech. J.*, vol. 38, pp. 611–656, 1959.
[28] M. Weinberger, J. Ziv, and A. Lempel, "On the optimal asymptotic performance of universal ordering and of discrimination of individual sequences," *IEEE Trans. Inform. Theory*, vol. 38, pp. 380–385, Mar. 1992.
[29] A. Lapidoth and J. Ziv, "On the universality of the LZ-based decoding algorithm," this issue, pp. 1746–1755.
[30] I. Csiszár, "Arbitrarily varying channel with general alphabets and states," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1725–1742, Nov. 1992.
[31] D. Blackwell, L. Breiman, and A. J. Thomasian, "The capacities of certain channel classes under random coding," *Ann. Math. Stat.*, vol. 31, pp. 558–567, 1960.
[32] B. Y. Ryabko, "Twice-universal coding," *Probl. Inform. Transm.*, pp. 173–177, July–Sept. 1984.
[33] M. Feder and N. Merhav, "Hierarchical universal coding," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1354–1364, Sept. 1996.
[34] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding*. New York: McGraw-Hill, 1979.
[35] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Boston, MA: Jones and Bartlett, 1993.
[36] I. Csiszár and P. Narayan, "The capacity of the arbitrarily varying channel revisited: Capacity, constraints," *IEEE Trans. Inform. Theory*, vol. 34, pp. 181–193, Jan. 1988.
[37] T. Ericson, "Exponential error bounds for random codes in the arbitrarily varying channel," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 42–48, Jan. 1985.
[38] J. Pokorny and H. Wallmeier, "Random coding bound and codes produced by permutations for the multiple-access channel," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 741–750, Nov. 1985.
[39] Y. S. Liu and B. L. Hughes, "A new universal random coding bound for the multiple-access channel," *IEEE Trans. Inform. Theory*, vol. 42, pp. 376–386, Mar. 1996.
[40] A. Lapidoth and J. Ziv, "Universal sequential decoding," presented at the 1998 Information Theory Workshop, Killarney, Co. Kerry, Ireland, June 22–26, 1998.
[41] T. M. Cover and E. Ordentlich, "Universal portfolios with side information," *IEEE Trans. Inform. Theory*, vol. 42, pp. 348–363, Mar. 1996.