May, 1996

LIDS-P 2330

Research Supported By:

No Sponsor

Universal Decoders for Channels with Memory

Feder, M.
Lapidoth, A.

Universal Decoders for Channels with Memory

Meir Feder and Amos Lapidoth *

Abstract

A universal decoder for a parametric family of channels is a decoder that for any channel in the family attains the same random coding error exponent as the best decoder for that particular channel. The existence and structure of such a decoder is demonstrated under relatively mild conditions of continuity of the channel law with respect to the parameter indexing the family.

Keywords: Universal decoding, error-exponent, random coding, robust decoding, Gilbert-Elliot channel.

1 Introduction

Consider a parametric family of channels

$$p_\theta(\mathbf{y}|\mathbf{x}), \quad \theta \in \Theta, \tag{1}$$

defined on a common finite[1] input alphabet $\mathcal{X}$ and a (possibly infinite) output alphabet $\mathcal{Y}$, where $p_\theta(\mathbf{y}|\mathbf{x})$ is the probability (or probability density) of the output sequence $\mathbf{y} \in \mathcal{Y}^n$ given that the sequence $\mathbf{x} \in \mathcal{X}^n$ is transmitted over the channel with parameter $\theta \in \Theta$. The parameter space Θ is assumed to be separable in the sense that

$$\exists \{\theta_k\}_{k=1}^{\infty} \subseteq \Theta \; : \; \inf_k \limsup_{n\to\infty} \sup_{(\mathbf{x},\mathbf{y})\in\mathcal{X}^n\times\mathcal{Y}^n} \frac{1}{n} \left| \log \left(\frac{p_\theta(\mathbf{y}|\mathbf{x})}{p_{\theta_k}(\mathbf{y}|\mathbf{x})} \right) \right| = 0 \;\; \forall \theta \in \Theta, \tag{2}$$

*Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139-4307. Prof. Feder is on leave from the Department of Electrical Engineering—Systems, Tel Aviv University, Tel Aviv 69978, Israel.

[1] Our results can be extended to the case where the input alphabet is not finite but to simplify notations we choose to restrict ourselves to a finite input alphabet.

where throughout the paper we use the convention

$$\log \frac{0}{0} = 0, \quad \log \frac{a}{0} = -\log \frac{0}{a} = \infty \ \ \forall a > 0. \tag{3}$$

In words, condition (2) says that there exists a countable subset $\{\theta_k\}_{k=1}^{\infty} \subseteq \Theta$ that is dense in Θ with respect to the semi-metric

$$d(\theta, \theta') = \limsup_{n \to \infty} \sup_{(\mathbf{x},\mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n} \frac{1}{n} \left| \log \left(\frac{p_\theta(\mathbf{y}|\mathbf{x})}{p_{\theta_k}(\mathbf{y}|\mathbf{x})} \right) \right|, \tag{4}$$

i.e., that for any $\theta \in \Theta$ and any $\epsilon > 0$ there exists some positive integer k such that $d(\theta, \theta_k) < \epsilon$. The separability condition is a rather mild condition, and as we shall see in Section 3 many of the parametric families arising in wireless communications are separable. Some examples of separable families include the family of discrete memoryless channels, and the family of finite memory channels (which includes the Gilbert-Elliot family of channels as a special case).

Given a codebook

$$\mathcal{C} = \{\mathbf{x}(1), \ldots, \mathbf{x}(2^{nR})\} \subset \mathcal{X}^n \tag{5}$$

of blocklength n and rate R, we define a decoder $\phi : \mathcal{Y}^n \to \{1, \ldots, 2^{nR}\}$ as a mapping that maps every output sequence $\mathbf{y}$ to an index i of some codeword in $\mathcal{C}$. If all codewords are used equiprobably (as we shall assume throughout) then the average probability of error $P_{\theta,\phi}(\text{error}|\mathcal{C})$ incurred when the codebook $\mathcal{C}$ is used over the channel $p_\theta(\mathbf{y}|\mathbf{x})$ with the decoder ϕ, is given by

$$P_{\theta,\phi}(\text{error}|\mathcal{C}) = 2^{-nR} \sum_{i=1}^{2^{nR}} \sum_{\{\mathbf{y}:\phi(\mathbf{y}) \neq i\}} p_\theta\left(\mathbf{y}|\mathbf{x}(i)\right). \tag{6}$$

For a given channel $p_\theta(\mathbf{y}|\mathbf{x})$ and a given codebook $\mathcal{C}$, the optimal decoder (in the sense of minimizing the average probability of error) is the maximum-likelihood decoder ϕ_θ for which $\phi_\theta(\mathbf{y}) = i$ only if

$$\log p_\theta(\mathbf{y}|\mathbf{x}(i)) \geq \log p_\theta(\mathbf{y}|\mathbf{x}(j)) \ \ \forall 1 \leq j \leq 2^{nR}. \tag{7}$$

Notice that (7) does not completely define the maximum-likelihood decoder because it does not specify how ties in the likelihood function are to be resolved. The manner in which such ties are resolved does not affect the average

probability of error, and we shall assume some arbitrary but fixed deterministic mechanism. A more precise description of the maximum-likelihood decoder that also specifies this mechanism is as follows. Assume that all the codewords are in some set $B_n \subseteq \mathcal{X}^n$ of size $|B_n|$,

$$\mathbf{x}(i) \in B_n \ \ \forall 1 \leq i \leq 2^{nR},$$

and consider a ranking function

$$M_\theta \ : B_n \times \mathcal{Y}^n \longrightarrow \{1, \ldots, |B_n|\},$$

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

$$p_\theta(\mathbf{y}|\mathbf{x}) > p_\theta(\mathbf{y}|\mathbf{x}') \Rightarrow M_\theta(\mathbf{x}, \mathbf{y}) < M_\theta(\mathbf{x}', \mathbf{y}), \tag{8}$$

where the sequence most likely (given the received sequence $\mathbf{y}$) is ranked highest, i.e., 1. Given a codebook $\mathcal{C}$ as in (5) the maximum-likelihood decoder ϕ_θ that is determined by the ranking function $M_\theta(\cdot, \cdot)$ is defined by

$$\phi_\theta(\mathbf{y}) = i \ \text{ iff } M_\theta(\mathbf{x}(i), \mathbf{y}) < M_\theta(\mathbf{x}(j), \mathbf{y}) \ \forall j \neq i, \ 1 \leq j \leq 2^{nR}. \tag{9}$$

(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 $\mathbf{y}$, the maximum-likelihood receiver determined by $M_\theta(\cdot, \cdot)$ declares that the transmitted codeword was $\mathbf{x}(i)$ if $\mathbf{x}(i)$ maximizes $p_\theta(\mathbf{y}|\mathbf{x}(j))$ among all the codewords $\mathbf{x}(j)$ in $\mathcal{C}$, and in the case that this maximum is achieved by several codewords, it prefers the one that is ranked highest by $M_\theta(\cdot, \mathbf{y})$

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

$$M_u \ : B_n \times \mathcal{Y}^n \longrightarrow \{1, \ldots, |B_n|\},$$

such that for any $\mathbf{y} \in \mathcal{Y}^n$ the function $M_u(\cdot, \mathbf{y})$ is one-to-one and onto $\{1, \ldots, |B_n|\}$, defines a decoder u in a manner completely analogous with (9). Thus given a codebook $\mathcal{C}$ as in (5) and given a received sequence $\mathbf{y} \in \mathcal{Y}^n$

$$u(\mathbf{y}) = i \text{ iff } M_u(\mathbf{x}(i), \mathbf{y}) < M_u(\mathbf{x}(j), \mathbf{y}) \ \forall j \neq i, \ 1 \leq j \leq 2^{nR}. \tag{10}$$

Slightly abusing the notation introduced in (6) we denote by $P_{\theta,\theta'}(\text{error}|\mathcal{C})$ the average probability of error that is incurred when code $\mathcal{C}$ is used over the channel $p_\theta(\mathbf{y}|\mathbf{x})$ and maximum-likelihood decoding is performed according to the (possibly different) law $p_{\theta'}(\mathbf{y}|\mathbf{x})$. This situation is referred to as mismatched decoding, and its effect on the achievable rates for memoryless channels has been studied in [1], [2],[3], [4], [5], and references therein. Clearly if $\theta = \theta'$ then there is is no mismatch and $P_{\theta,\theta}(\text{error}|\mathcal{C})$ thus denotes the average probability of error incurred when the code $\mathcal{C}$ is used over the channel of parameter θ and is optimally decoded using the maximum-likelihood rule corresponding to that channel.

Consider now the case where the codebook $\mathcal{C}$ is drawn at random. We assume that the codewords are drawn independently of each other, each being drawn uniformly over the set $B_n \subset \mathcal{X}^n$. The set B_n is often taken as the set of all n-length sequences of some given type, but for our purposes, B_n can be arbitrary. We define $\bar{P}_{\theta,\phi}(\text{error})$ to be the average of $P_{\theta,\phi}(\text{error}|\mathcal{C})$ over this random choice of the codebook $\mathcal{C}$. Once again we slightly abuse this notation and use $\bar{P}_{\theta,\theta'}(\text{error})$ to denote $\bar{P}_{\theta,\phi_{\theta'}}(\text{error})$ where $\phi_{\theta'}$ is the maximum-likelihood decoder corresponding to the channel of parameter θ'. In particular $\bar{P}_{\theta,\theta}(\text{error})$ is the ensemble average of $P_{\theta,\theta}(\text{error}|\mathcal{C})$ over all codebooks $\mathcal{C}$ whose codewords are drawn independently and uniformly over the set B_n.

Our main result, which is proved in Section 2, can be now stated as follows:

Theorem 1 *Consider a parametric family of channels $p_\theta(\mathbf{y}|\mathbf{x}) \quad \theta \in \Theta$ defined over a common finite input alphabet $\mathcal{X}$ and a possibly infinite output alphabet $\mathcal{Y}$. If the parameter space is separable in the sense of (2) then there exists a universal sequence of decoders u_n such that*

$$\lim_{n\to\infty} \frac{1}{n} \log \left(\frac{\bar{P}_{\theta,u_n}(\text{error})}{\bar{P}_{\theta,\theta}(\text{error})} \right) = 0 \ \ \forall \theta \in \Theta, \tag{11}$$

where the probabilities of error are averaged over the ensemble of random codebooks whose codewords are drawn independently and uniformly over the set B_n.

We thus claim that there exists a universal decoder that for any channel in the family attains the same random coding error exponent as the optimal decoder designed for that specific channel.

The existence of a universal decoder for the family of memoryless channels was demonstrated in [6] where it was shown that the maximum empirical mutual information (MMI) decoder is universal. For discrete memoryless channels it can be shown that the MMI is equivalent to a generalized maximum-likelihood decoder that ranks the codeword $\mathbf{x}$ according to $\max_{\theta\in\Theta} p_\theta(\mathbf{y}|\mathbf{x})$.

In [7] Ziv proposed a different approach to universal decoding, one that is based on the Lempel-Ziv approach to universal data compression, and showed that a universal decoder exists for the special subclass of finite state channels that satisfy that at any time ν the law of the output y_ν given the input x_ν is determined by the state of the channel s_ν, and where the next state $s_{\nu+1}$ is a *deterministic* function of the previous state s_ν and the previous input and output x_ν and y_ν.

The family of channels for which Ziv showed the existence of a universal decoder, while richer than the family of memoryless channel, is not sufficiently rich for many applications arising in wireless communications. For example, the Gilbert-Elliot channel in which the distribution of y_ν given x_ν is determined by the state s_ν where the sequence of states forms a Markov process, is not covered in Ziv's model. As we shall see in section 3 the families of channels for which Theorem 1 demonstrates the existence of a universal decoder includes the memoryless channels, the families studied by Ziv, the Gilbert-Elliot family of channels, and many more. It should, however, be noted that Ziv's universal decoder seems to have a simpler implementation than the decoder that we propose, particularly for convolutional codes with a long memory where sequential decoding is attractive.

2 A Proof of Theorem 1

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 $p_\theta(\mathbf{y}|\mathbf{x})$ is given by [7],

$$\bar{P}_{\theta,\phi}(\text{error}) = \sum_{\mathbf{x}\in B_n} \sum_{\mathbf{y}\in\mathcal{Y}^n} \frac{1}{|B_n|} p_\theta(\mathbf{y}|\mathbf{x}) \Pr\left(\text{error}|\mathbf{x},\mathbf{y},\phi\right), \tag{12}$$

where

$$\Pr\left(\text{error}|\mathbf{x},\mathbf{y},\phi\right) = 1 - \left(1 - \frac{M_\phi(\mathbf{x},\mathbf{y})}{|B_n|}\right)^{2^{nR}-1}, \tag{13}$$

is the conditional probability of error given that the transmitted codeword is $\mathbf{x}$, the received sequence is $\mathbf{y}$, and the decoder being used is ϕ. Equation (13) follows from the observation that the codewords are drawn independently and uniformly over B_n and that if $\mathbf{x}$ is the correct codeword and $\mathbf{y}$ is the received sequence then an error occurs only if some other codeword $\mathbf{x}'$ is ranked higher than $\mathbf{x}$, i.e., if $M_\phi(\mathbf{x}', \mathbf{y}) \leq M_\phi(\mathbf{x}, \mathbf{y})$. Notice that $\Pr(\text{error}|\mathbf{x}, \mathbf{y}, \phi)$ does not depend on the channel $p_\theta(\cdot|\cdot)$ over which transmission is carried out, but only on the correct codeword $\mathbf{x}$, the received sequence $\mathbf{y}$, and the decoder ϕ.

Let ϕ and ϕ' be two decoders based on the ranking functions $M_\phi(\mathbf{x}, \mathbf{y})$ and $M_{\phi'}(\mathbf{x}, \mathbf{y})$ respectively. In the appendix we prove that

$$\frac{P(\text{error}|\mathbf{x}, \mathbf{y}, \phi')}{P(\text{error}|\mathbf{x}, \mathbf{y}, \phi)} \leq \max\left\{1, \frac{M_{\phi'}(\mathbf{x}, \mathbf{y})}{M_\phi(\mathbf{x}, \mathbf{y})}\right\}, \tag{14}$$

and hence,

$$\begin{aligned}
\frac{\bar{P}_{\theta,\phi'}(\text{error})}{\bar{P}_{\theta,\phi}(\text{error})} &= \frac{\sum_{\mathbf{x}\in B_n}\sum_{\mathbf{y}\in\mathcal{Y}^n}\frac{1}{|B_n|}p_\theta(\mathbf{y}|\mathbf{x})\Pr(\text{error}|\mathbf{x}, \mathbf{y}, \phi')}{\sum_{\mathbf{x}\in B_n}\sum_{\mathbf{y}\in\mathcal{Y}^n}\frac{1}{|B_n|}p_\theta(\mathbf{y}|\mathbf{x})\Pr(\text{error}|\mathbf{x}, \mathbf{y}, \phi)} \\
&\leq \max_{\mathbf{x}\in B_n, \mathbf{y}\in\mathcal{Y}^n} \frac{P(\text{error}|\mathbf{x}, \mathbf{y}, \phi')}{P(\text{error}|\mathbf{x}, \mathbf{y}, \phi)} \\
&\leq \max_{\mathbf{x}\in B_n, \mathbf{y}\in\mathcal{Y}^n} \frac{M_{\phi'}(\mathbf{x}, \mathbf{y})}{M_\phi(\mathbf{x}, \mathbf{y})}.
\end{aligned} \tag{15}$$

The equality follows from (12), the first inequality follows by noting that if U and V are non-negative random variables then

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

and the last inequality follows from (14) by noting that

$$\max_{\mathbf{x}\in B_n, \mathbf{y}\in\mathcal{Y}^n} \frac{M_{\phi'}(\mathbf{x}, \mathbf{y})}{M_\phi(\mathbf{x}, \mathbf{y})} \geq 1,$$

since for any $\mathbf{y} \in \mathcal{Y}^n$ the functions $M_\phi(\cdot, \mathbf{y})$ and $M_{\phi'}(\cdot, \mathbf{y})$ are both one-to-one mappings onto $\{1, \ldots, |B_n|\}$. Inequality (15) is a refined version of an inequality given in [7]. Its importance is that it relates differences in ranking functions to differences in random coding error performance.

Consider now K maximum-likelihood decoders $\phi_{\theta_1}, \ldots, \phi_{\theta_K}$ with corresponding ranking functions $M_{\theta_1}(\mathbf{x}, \mathbf{y}), \ldots, M_{\theta_K}(\mathbf{x}, \mathbf{y})$. For any received sequence $\mathbf{y}$ we define the merged decoder $u_k(\mathbf{y})$ via its ranking function $M_{u_K}(\cdot, \mathbf{y})$

as in (10) where $M_{u_K}(\cdot, \mathbf{y})$ is defined as follows: Given a received sequence $\mathbf{y}$ the ranking function $M_{u_K}(\cdot, \mathbf{y})$ ranks number 1 the sequence in B_n that $M_{\theta_1}(\cdot, \mathbf{y})$ ranks highest. It then ranks second the sequence that $M_{\theta_2}(\cdot, \mathbf{y})$ ranks highest (unless it is equal to the sequence ranked highest by $M_{\theta_1}(\cdot, \mathbf{y})$ in which case it skips to consider the sequence that $M_{\theta_3}(\cdot, \mathbf{y})$ ranks highest), followed by the sequence that $M_{\theta_3}(\cdot, \mathbf{y})$ ranks highest, etc. After the first ranking of all the decoders $M_{\theta_1}(\cdot, \mathbf{y}), \cdots, M_{\theta_K}(\cdot, \mathbf{y})$ have been considered we return to $M_{\theta_1}(\cdot, \mathbf{y})$ and consider the sequence in B_n ranked second, followed by the sequence that $M_{\theta_2}(\cdot, \mathbf{y})$ ranks second etc. In all cases if we encounter a sequence that has already been ranked we simply skip it. This construction guarantees that if a sequence $\mathbf{x} \in B_n$ is ranked j-th by the k-th decoder $M_{\theta_k}(\cdot, \mathbf{y})$ then $\mathbf{x}$ is ranked $(j-1)K + k$ or higher by $M_{u_K}(\cdot, \mathbf{y})$, i.e.,

$$M_{\theta_k}(\mathbf{x}, \mathbf{y}) = j \Rightarrow M_{u_K}(\mathbf{x}, \mathbf{y}) \leq (j-1)K + k \ \ \forall \mathbf{x} \in B_n, \ \ \forall 1 \leq k \leq K. \quad (16)$$

Equation (16) can actually serve as a definition for the merging operation, i.e., the construction of $M_{u_K}(\cdot, \mathbf{y})$ from $M_{\theta_1}(\cdot, \mathbf{y}), \ldots, M_{\theta_K}(\cdot, \mathbf{y})$. Crucial to our analysis is the observation that with this construction

$$M_{u_K}(\mathbf{x}, \mathbf{y}) \leq K M_{\theta_k}(\mathbf{x}, \mathbf{y}) \ \ \forall (\mathbf{x}, \mathbf{y}) \in B_n \times \mathcal{Y}^n, \ \ \forall 1 \leq k \leq K, \quad (17)$$

which follows immediately from (16).

It now follows from and (15) and (17) that for every $\theta \in \Theta$

$$\bar{P}_{\theta, u_K}(\text{error}) \leq K \bar{P}_{\theta, \theta_k}(\text{error}) \ \ \forall \theta \in \Theta \ \ \forall 1 \leq k \leq K, \quad (18)$$

so that on any channel $p_\theta(\cdot|\cdot)$ the merged decoder performs, up to a factor of K, as well as the best of the decoders $\phi_{\theta_1}, \ldots, \phi_{\theta_K}$. Actually, the merged decoder satisfies

$$\frac{P(\text{error}|\mathbf{x}, \mathbf{y}, u_K)}{P(\text{error}|\mathbf{x}, \mathbf{y}, \theta_k)} \leq K$$

for all $\mathbf{x}, \mathbf{y}$.

The next step is to study how on a given channel $p_\theta(\cdot|\cdot)$ the merged decoder compares with the maximum-likelihood decoder ϕ_θ for that channel, where ϕ_θ is typically not one of the decoders $\phi_{\theta_1}, \ldots, \phi_{\theta_K}$. Notice that even if θ and θ' are very close in the sense that $p_\theta(\mathbf{y}|\mathbf{x})$ and $p_{\theta'}(\mathbf{y}|\mathbf{x})$ are very similar, still the maximum likelihood decoder for θ and θ' can be very different. This can be seen by considering the family of binary symmetric channels parameterized by their crossover probability and considering $\theta = 0.5 - \epsilon$ and $\theta' = 0.5 + \epsilon$ where

$\epsilon > 0$ is arbitrarily small. For this case the maximum likelihood decoder ϕ_θ minimizes Hamming distance whereas the maximum-likelihood decoder $\phi_{\theta'}$ maximizes Hamming distance. We cannot, therefore hope for a continuity of the decoding rule with respect to the channel parameter, but what we can show is the continuity with respect to the parameter of the mismatched error exponent. This is formalized in the following lemma.

Lemma 1 *If*

$$\frac{1}{n}\left|\log\frac{p_{\theta'}(\mathbf{y}|\mathbf{x})}{p_\theta(\mathbf{y}|\mathbf{x})}\right| \leq \epsilon, \quad \forall(\mathbf{x},\mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n,$$

then

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

where $\bar{P}_{\theta,\theta'}(\text{error})$ *is the mismatch ensemble averaged probability of error corresponding to channel* $p_\theta(\cdot|\cdot)$ *and a maximum-likelihood decoder for the law* $p_{\theta'}(\cdot|\cdot)$*, and* $\bar{P}_{\theta,\theta}(\text{error})$ *is the ensemble averaged matched probability of error corresponding to the channel* $p_\theta(\cdot|\cdot)$*.*

Proof of Lemma 1: To make the proof of the lemma more transparent, let us break up the assumptions of the lemma into two separate assumptions:

$$p_{\theta'}(\mathbf{y}|\mathbf{x}) \leq p_\theta(\mathbf{y}|\mathbf{x})2^{n\epsilon}, \quad \forall(\mathbf{x},\mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n, \tag{19}$$

and

$$p_{\theta'}(\mathbf{y}|\mathbf{x}) \geq p_\theta(\mathbf{y}|\mathbf{x})2^{-n\epsilon}, \quad \forall(\mathbf{x},\mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n. \tag{20}$$

We now have

$$\begin{aligned}
\bar{P}_{\theta,\theta'}(\text{error}) &= \sum_{\mathbf{x}\in B_n}\sum_{\mathbf{y}\in\mathcal{Y}^n}\frac{1}{|B_n|}p_\theta(\mathbf{y}|\mathbf{x})\Pr\left(\text{error}|\mathbf{x},\mathbf{y},\phi_{\theta'}\right) \\
&\leq 2^{n\epsilon}\sum_{\mathbf{x}\in B_n}\sum_{\mathbf{y}\in\mathcal{Y}^n}\frac{1}{|B_n|}p_{\theta'}(\mathbf{y}|\mathbf{x})\Pr\left(\text{error}|\mathbf{x},\mathbf{y},\phi_{\theta'}\right) \\
&= 2^{n\epsilon}\bar{P}_{\theta',\theta'}(\text{error}) \\
&\leq 2^{n\epsilon}\bar{P}_{\theta',\theta}(\text{error}) \\
&= 2^{n\epsilon}\sum_{\mathbf{x}\in B_n}\sum_{\mathbf{y}\in\mathcal{Y}^n}\frac{1}{|B_n|}p_{\theta'}(\mathbf{y}|\mathbf{x})\Pr\left(\text{error}|\mathbf{x},\mathbf{y},\phi_{\theta}\right) \\
&\leq 2^{2n\epsilon}\sum_{\mathbf{x}\in B_n}\sum_{\mathbf{y}\in\mathcal{Y}^n}\frac{1}{|B_n|}p_{\theta}(\mathbf{y}|\mathbf{x})\Pr\left(\text{error}|\mathbf{x},\mathbf{y},\phi_{\theta}\right) \\
&= 2^{2n\epsilon}\bar{P}_{\theta,\theta}(\text{error}),
\end{aligned}$$

which completes the proof of the Lemma. Notice that the first inequality follows from (19), the second inequality from the optimality of the maximum-likelihood decoder, and the third inequality from (20). All equalities follow from (12) and the fact that the conditional error probability, which is defined in (13), depends on $\mathbf{x}, \mathbf{y}$, and ϕ but not on the channel $p_\theta(\cdot|\cdot)$.

Proof of Theorem 1: Consider the decoder $u_{K(n)}$ that is defined by merging the maximum-likelihood decoders $\phi_{\theta_1}, \ldots, \phi_{\theta_{K(n)}}$ where for now $K(n) = n$, and $\theta_1, \ldots, \theta_{K(n)}$ are the first $K(n)$ parameters in a sequence $\{\theta_k\}_{k=1}^{\infty}$ which is dense in Θ, see (2). Let $\theta^* \in \Theta$ be arbitrary. By the separability assertion it follows that for any $\epsilon > 0$ there exists some k^* and some n_0 such that for all $n \geq n_0$

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

For all sufficiently large blocklength n we have that $K(n) \geq k^*$ and the maximum-likelihood decoder $\phi_{\theta_{k^*}}$ is among the decoders $\phi_{\theta_1}, \ldots, \phi_{\theta_{K(n)}}$ from which $u_{K(n)}$ is constructed. It thus follows from (18) that for such sufficiently large n

$$\bar{P}_{\theta^*, u_{K(n)}}(\text{error}) \leq K(n) \bar{P}_{\theta^*, \theta_{k^*}}(\text{error}). \tag{21}$$

If, in addition, n is sufficiently large so that $n \geq n_0$ then by Lemma 1

$$\bar{P}_{\theta^*, \theta_{k^*}}(\text{error}) \leq 2^{2n\epsilon} \bar{P}_{\theta^*, \theta^*}(\text{error}). \tag{22}$$

Combining (21) and (22) we have that for all sufficiently large n,

$$\bar{P}_{\theta^*, u_{K(n)}}(\text{error}) \leq K(n) 2^{2n\epsilon} \bar{P}_{\theta^*, \theta^*}(\text{error}), \tag{23}$$

and the theorem now follows by letting $\epsilon = \epsilon_n$ tend to zero, and by noting that $K(n) = n$ is sub-exponential. □

Note: Inspecting the proof we see that some of the conditions of Theorem 1 can be actually weakened. First, we can limit $\mathbf{x}$ to B_n in condition (2). Secondly, we can replace the separability condition with a weaker form that requires that there exist a sequence $\{\theta_k\} \subseteq \Theta$ and a sub-exponential function $K(n)$ such that for any $\theta \in \Theta$

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

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.

Finally, if the random coding error exponents of the channels in the family are uniformly bounded then we may exclude some subset of pairs $(\mathbf{x}, \mathbf{y})$ from the supremum in (2) provided that the subset has a probability that is negligible with respect to the best error exponent in the family. This may be useful if the output alphabet is not finite.

3 Some Examples

Example 1: Consider the case where the family of channels is the family of all discrete memoryless channels 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

$$p_\theta(\mathbf{y}|\mathbf{x}) = \prod_{\nu=1}^{n} \theta(y_\nu|x_\nu),$$

where $\theta(y|x)$ denotes the entry in row x and column y of the matrix θ, and where $\mathbf{x} = (x_1, \ldots, x_n)$, and $\mathbf{y} = (y_1, \ldots, y_n)$. Since the channels in the family are memoryless we have

$$\begin{aligned} \frac{p_\theta(\mathbf{y}|\mathbf{x})}{p_{\theta'}(\mathbf{y}|\mathbf{x})} &= \prod_{\nu=1}^{n} \frac{\theta(y_\nu|x_\nu)}{\theta'(y_\nu|x_\nu)} \\ &\leq \left(\max_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \frac{\theta(y|x)}{\theta'(y|x)} \right)^n \end{aligned}$$

and hence

$$\frac{1}{n}\left|\log \frac{p_\theta(\mathbf{y}|\mathbf{x})}{p_{\theta'}(\mathbf{y}|\mathbf{x})}\right| \leq \max_{x,y}\left|\log \frac{\theta(y|x)}{\theta'(y|x)}\right|.$$

The required separability now follows by considering the countable set of all stochastic matrices with rational entries. Theorem 1 therefore demonstrates the existence of a universal decoder for the family of all memoryless channels with finite input and output alphabets. This results is due to Csiszár and

Körner [6]. In fact, the result stated in [6] is somewhat stronger as it shows the existence of a single sequence of universal codes that achieves the random coding exponent of any channel over which it is used, see also example 3.

Example 2: Consider now the family of all finite memory channels defined over a common finite input alphabet $\mathcal{X}$ and a common finite output alphabet $\mathcal{Y}$, with the set of channel states denoted $\mathcal{S}$ and its cardinality denoted $|\mathcal{S}|$. The law of such a channel is determined by a probability distribution $\pi_\theta(s_1)$ on the initial state s_1, a conditional distribution $q_\theta(s_\nu|s_{\nu-1}, x_{\nu-1}, y_{\nu-1})$ of the next state s_ν given the previous state $s_{\nu-1}$ the previous channel input $x_{\nu-1}$ and the previous output $y_{\nu-1}$, and by the conditional distribution $r_\theta(y_\nu|x_\nu, s_\nu)$ on the channel output y_ν given the channel input x_ν and the channel state s_ν. The channel law is thus given by

$$p_\theta(\mathbf{y}|\mathbf{x}) = \sum_{\mathbf{s}\in\mathcal{S}^n} p_\theta(\mathbf{y}, \mathbf{s}|\mathbf{x}), \tag{24}$$

where $\mathbf{x} = (x_1, \ldots, x_n), \mathbf{y} = (y_1, \ldots, y_n)$, $\mathbf{s} = (s_1, \ldots, s_n)$, and where

$$p_\theta(\mathbf{y}, \mathbf{s}|\mathbf{x}) = \pi_\theta(s_1) r_\theta(y_1|x_1, s_1) \prod_{\nu=2}^{n} \left(q_\theta(s_\nu|s_{\nu-1}, x_{\nu-1}, y_{\nu-1}) r_\theta(y_\nu|x_\nu, s_\nu)\right). \tag{25}$$

We now wish to use Theorem 1 to demonstrate the existence of a universal decoder for this family. First note that since a countable union of separable spaces is separable, we may assume without loss of generality that the set of states $\mathcal{S}$ is of some fixed size, say $|\mathcal{S}| = m$. Using the fact [8, Lemma 1] that if $\{a_l\}_{l=1}^L$ and$\{b_l\}_{l=1}^L$ are two non-negative sequences then[2]

$$\frac{a_1 + \ldots + a_L}{b_1 + \ldots + b_L} \leq \max_{1\leq l\leq L} \frac{a_l}{b_l}, \tag{26}$$

where $a/0 = \infty$ for $a > 0$, and $0/0 = 1$, we conclude from (24) that

$$\begin{aligned}
\max_{\mathbf{x},\mathbf{y}} \frac{p_\theta(\mathbf{y}|\mathbf{x})}{p_{\theta'}(\mathbf{y}|\mathbf{x})} &= \max_{\mathbf{x},\mathbf{y}} \frac{\sum_{\mathbf{s}\in\mathcal{S}} p_\theta(\mathbf{y}, \mathbf{s}|\mathbf{x})}{\sum_{\mathbf{s}\in\mathcal{S}} p_{\theta'}(\mathbf{y}, \mathbf{s}|\mathbf{x})} \\
&\leq \max_{\mathbf{x},\mathbf{y},\mathbf{s}} \frac{\pi_\theta(s_1) r_\theta(y_1|x_1, s_1) \prod_{\nu=2}^n \left(q_\theta(s_\nu|s_{\nu-1}, x_{\nu-1}, y_{\nu-1}) r_\theta(y_\nu|x_\nu, s_\nu)\right)}{\pi_{\theta'}(s_1) r_{\theta'}(y_1|x_1, s_1) \prod_{\nu=2}^n \left(q_{\theta'}(s_\nu|s_{\nu-1}, x_{\nu-1}, y_{\nu-1}) r_{\theta'}(y_\nu|x_\nu, s_\nu)\right)} \\
&\leq \max_{s} \frac{\pi_\theta(s)}{\pi_{\theta'}(s)} \left(\max_{s_1,s_2,x_1,y_1} \frac{q_\theta(s_2|s_1, x_1, y_1)}{q_{\theta'}(s_2|s_1, x_1, y_1)}\right)^{n-1} \max_{x,y,s} \left(\frac{r_\theta(y|x, s)}{r_{\theta'}(y|x, s)}\right)^n.
\end{aligned}$$

[2]In [8, Lemma 1] $0/0$ is defined as 0, whereas we chose to define it as 1. Nevertheless, it is straightforward to verify that inequality (26) still holds.

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

$$\begin{aligned}\max_{\mathbf{x},\mathbf{y}} \frac{1}{n}\left|\log\frac{p_\theta(\mathbf{y}|\mathbf{x})}{p_{\theta'}(\mathbf{y}|\mathbf{x})}\right| \;\leq\; & \max_{s}\left|\log\frac{\pi_\theta(s)}{\pi_{\theta'}(s)}\right| + \\ & \max_{s_1,s_2,x_1,y_1}\left|\log\frac{q_\theta(s_2|s_1,x_1,y_1)}{q_{\theta'}(s_2|s_1,x_1,y_1)}\right| + \\ & \max_{x,y,s}\left|\log\frac{r_\theta(y|x,s)}{r_{\theta'}(y|x,s)}\right|. \qquad (27)\end{aligned}$$

The separability of the family of finite memory channels and hence the existence of a universal decoder for this family now follows by considering the countable set of all tuples $(\pi(s), q(s_2|s_1,x_1,y_1), r(y|x,s))$ taking non-negative rational values.

This result considerably generalizes the result of Ziv [7] who showed the existence of a universal decoder for the special case where the state transition law $q_\theta(s_\nu|s_{\nu-1},x_{\nu-1},y_{\nu-1})$ is deterministic, i.e., when the next state is a deterministic function of the previous state and previous input and output. In particular as opposed to Ziv's results, our result holds for the family of Gilbert-Elliot channels [9],[10] (and references therein) for which the state sequence forms a non-deterministic Markov chain.

For the Gilbert-Elliot channel the input and output alphabet are both binary, as is the state space. The state of the channel forms a Markov process, and when the channel is in state "0" ("1") the output of the channel is related to the input as it would for a binary symmetric channel with crossover probability p_0 (respectively p_1). Our results thus show that one can design a universal decoder for the class of all Gilbert-Elliot channels without knowing the law of the underlying Markov process that describes the channel state and without knowing the crossover probabilities that correspond to each state.

It should be noted that the existence of a universal decoder in the sense of Theorem 1, i.e., in the sense (11), is meaningful only if the average probability of error decreases exponentially with the blocklength for rates below the channel capacity. Fortunately, this is usually the case, and holds for indecomposable finite memory channels [11, Theorem 4.6.4, 5.9.2], and in particular for the Gilbert-Elliot channels in the non-trivial case where the underlying Markov chain is ergodic.

Example 3: The following is a pathological example that demonstrates the subtleties involved with defining universality as in (11). Consider the

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

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

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

$$\mathbf{y} = x^{(1)} \oplus \theta^{(1)}, \ldots, x^{(n)} \oplus \theta^{(n)},$$

and $\oplus$ denotes binary addition (exclusive or).

Every channel $p_\theta(\mathbf{y}|\mathbf{x})$ has capacity 1 bit and an infinite error exponent (for rates below capacity). Since the parameter space Θ is countable it is separable, and Theorem 1 guarantees the existence of a universal decoder that need not know θ and can nevertheless achieve (11).

However, one can easily show using standard techniques from the theory of arbitrarily varying channels [12], [13, Appendix] 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.

Nevertheless, we can show that there exists a deterministic family of codes $\{\mathcal{C}_n\}_{n=1}^{\infty}$ which, when decoded using our universal decoder, gives rise to the right exponential decay of the probability of error for a restricted class of parameters $\mathcal{D}_n$, a class which increases to Θ as n tends to infinity.

Acknowledgment

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

Appendix

In this appendix we give a proof of (14). Observe that for any $N \geq 1$ the function $f(x) = 1-(1-x)^N$ is concave in x for $0 \leq x \leq 1$, and that $f(0) = 0$.

Thus, by Jensen's inequality, for any $0 \leq \alpha \leq 1$ and any $0 \leq x \leq 1$,

$$f(\alpha x) = f(\alpha x + (1-\alpha)0) \geq \alpha f(x) + (1-\alpha)f(0) = \alpha f(x).$$

Letting $K = 1/\alpha \geq 1$ and substituting $y = \alpha x$ so that $x = Ky$, and $0 \leq Ky \leq 1$ we have

$$\frac{f(Ky)}{f(y)} = \frac{1-(1-Ky)^N}{1-(1-y)^N} \leq K, \quad K \geq 1. \tag{28}$$

Also, since $f(\cdot)$ is monotonically increasing in $[0,1]$, we have for $K \leq 1$, $0 \leq y \leq 1$

$$\frac{f(Ky)}{f(y)} = \frac{1-(1-Ky)^N}{1-(1-y)^N} \leq 1, \quad K \leq 1. \tag{29}$$

Inequality (14) now follows from (28) and (29) by substituting $N = 2^{nR} - 1$, $y = M_\phi(\mathbf{x},\mathbf{y})/|B_n|$, and $K = M_{\phi'}(\mathbf{x},\mathbf{y})/M_\phi(\mathbf{x},\mathbf{y})$.

References

[1] V. B. Balakirsky. A converse coding theorem for mismatched decoding at the output of binary-input memoryless channels. *IEEE Trans. on Inform. Theory*, 41(6):1889–1902, Nov. 1995.

[2] I. Csiszár and P. Narayan. Channel capacity for a given decoding metric. *IEEE Trans. on Inform. Theory*, 41(1):35–43, Jan. 1995.

[3] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai(Shitz). On information rates for mismatched decoders. *IEEE Trans. on Inform. Theory.*, 40(6):1953–1967, Nov. 1994.

[4] A. Lapidoth. Nearest-neighbor decoding for additive non-Gaussian noise channels. *To appear in IEEE Trans. on Inform. Theory, Sept. 1996.*

[5] A. Lapidoth. Mismatched decoding and the multiple-access channel. *To appear in IEEE Trans. on Inform. Theory.*

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

[7] J. Ziv. Universal decoding for finite-state channels,. *IEEE Trans. on Inform. Theory.*, 31(4):453–460, July 1985.

[8] T.M. Cover and E. Ordentlich. Universal portfolios with side information. *IEEE Trans. on Inform. Theory*, 42:348–363, March 1996.

[9] M. Mushkin and I. Bar-David. Capacity and coding for the Gilbert-Elliot channel. *IEEE Trans. on Inform. Theory*, 35(6):1277–1290, Nov. 1989.

[10] A.J. Goldsmith and P.P. Varaiya. Capacity, mutual information, and coding for finite-state Markov channels. *IEEE Trans. on Inform. Theory*, 42(3), 1996.

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

[12] L. Breiman D. Blackwell and A.J. Thomasian. The capacities of certain channel classes under random coding. *Ann. Math. Statis.*, 31:558–567, 1960.

[13] I. Csiszár and P. Narayan. Capacity of the Gaussian arbitrarily varying channel. *IEEE Trans. on Inform. Theory*, 37(1):18–26, January 1991.