Universal Decoding With an Erasure Option *

Neri Merhav[†]   Meir Feder[‡]

December 21, 2006

Abstract

Motivated by applications of rateless coding, decision feedback, and automatic repeat request (ARQ), we study the problem of universal decoding for unknown channels, in the presence of an erasure option. Specifically, we harness the competitive minimax methodology developed in earlier studies, in order to derive a universal version of Forney's classical erasure/list decoder, which in the erasure case, optimally trades off between the probability of erasure and the probability of undetected error. The proposed universal erasure decoder guarantees universal achievability of a certain fraction ξ of the optimum error exponents of these probabilities (in a sense to be made precise in the sequel). A single–letter expression for ξ, which depends solely on the coding rate and the Neyman–Pearson threshold (to be defined), is provided. The example of the binary symmetric channel is studied in full detail, and some conclusions are drawn.

Index Terms: rateless codes, erasure, error exponent, universal decoding, generalized likelihod ratio test, channel uncertainty, competitive minimax.

[*] This research was supported by the Israel Science Foundation, grant no. 223/05.

[†] Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel. E–mail: [merhav@ee.technion.ac.il].

[‡] Department of Electrical Engineering – Systems, Tel Aviv University, Tel Aviv 69978, Israel. E-mail: [meir@eng.tau.ac.il].

1 Introduction

When communicating across an unknown channel, classical channel coding at any fixed rate, however small, is inherently problematic since this fixed rate might be larger than the unknown capacity of the underlying channel. It makes sense then to try to adapt the coding rate to the channel conditions, which can be learned on–line at the transmitter whenever a feedback link, from the receiver to the transmitter, is available.

One of the recent promising approaches to this end is rateless coding proposed in [17],[18] (see also [5],[6], [7], [14], [20], and references therein). Independently, rateless codes were also proposed in a networking scenario for the packet erasure channel [2],[3],[15], where they have been referred to as *fountain codes*. Fountain codes also have a low density structure that allow computationally efficient decoding. In rateless coding, there is a fixed number M of messages, each one being represented by a codeword of unlimited length, in principle. A possible receiver for a rateless code examines, after each symbol has been received, whether it can decode the message, with "reasonably good confidence," or alternatively, to request, via the feedback link, an additional symbol.[1] Upon receiving the new channel output, again, the receiver either makes a decision, or requests another symbol from the transmitter, and so on. The coding rate is then defined by $\log M$ divided by the expected number of symbols transmitted before the decoder makes a decision. Clearly, at every time instant, the receiver of a rateless communication system operates just like an *erasure decoder* [10],[2] which partitions the space of channel output vectors into $(M+1)$ regions, M for each one of the possible messages, and an additional region for "erasure," which, in the rateless regime, is used for requesting an additional symbol. Keeping the erasure probability small is then motivated by the desire to keep the expected transmission time, for each message, small. Although these two criteria are not completely equivalent, they are strongly related.

When the channel is unknown at the decoder, it was suggested in some of the references above, to use a universal decoder, which is inspired by the maximum mutual information (MMI) decoder [4]: by using a certain threshold, the receiver decides whether to make a decision or ask for another symbol. While this approach works fairly well, there is no evidence of optimality.

[1] Alternatively, the receiver can use the feedback link only to notify the transmitter when it reached a decision regarding the current message (and keep silent at all other times). In network situations, this would not load the network much as it is done only once per each message.

[2] See also [21], [1], [13], [12] and references therein for later studies.

These observations, as well as techniques such as automatic repeat request (ARQ) and decision feedback, motivate us to study the problem in a more systematic manner. Specifically, we consider the problem of universal decoding with an erasure option, for the class of discrete memoryless channels (DMC's) indexed by an unknown parameter vector θ (e.g., the set of channel transition probabilities). We harness the competitive minimax methodology proposed in [9], in order to derive a universal version of Forney's classical erasure/list decoder. For a given DMC with parameter θ, a given coding rate R, and a given threshold parameter T (all to be formally defined later), Forney's erasure/list decoder optimally trades off between the exponent $E_1(R,T,\theta)$ of the probability of the erasure event, $\mathcal{E}_1$, and the exponent, $E_2(R,T,\theta) = E_1(R,T,\theta) + T$, of the probability of undetected error event, $\mathcal{E}_2$, in the random coding regime.

The universal erasure decoder, proposed in this paper, guarantees universal achievability of an erasure exponent, $\hat{E}_1(R,T,\theta)$, which is at least as large as $\xi \cdot E_1(R,T,\theta)$ for all θ, for some constant $\xi \in (0,1]$, that is independent of θ (but does depend on R and T), and at the same time, an undetected error exponent $\hat{E}_2(R,T,\theta) \geq \xi \cdot E_1(R,T,\theta) + T$ for all θ (in the random coding sense). At the very least this guarantees that whenever the probabilities of $\mathcal{E}_1$ and $\mathcal{E}_2$ decay exponentially for a known channel, so they do even when the channel is unknown, using the proposed universal decoder. The question is, of course: what is the largest value of ξ for which the above statement holds? We partially answer this question by deriving a single–letter expression for a lower bound to the largest value of ξ, denoted henceforth by $\xi^*(R,T)$, that is guaranteed to be attainable by this decoder. While $\xi^*(R,T)$ is only a lower bound to the universally achievable fraction of the error exponent, and $\xi^*(R,T)$ may, in general, be strictly less than unity (as we show in examples), it is conjectured that the true universally achievable fraction of the error exponent may still be less than unity as well. If this conjecture is true, then it means that there is a major difference between ordinary universal decoding and universal erasure decoding: While for the former, it is well known that optimum[3] random coding error exponents are fully universally achievable (at least for some classes of channels and certain random coding distributions [4],[22],[8]), in the latter, when the erasure option is available, this is may no longer be the case, in general. Explicit results, including numerical values of $\xi^*(R,T)$, are derived for the example of the binary symmetric channel (BSC), parameterized

[3] Optimum exponents – corresponding to optimum maximum likelihood decoding.

by the crossover probability θ, and some conclusions are drawn.

The outline of the paper is as follows. In Section 2, we establish the notation conventions and we briefly review some known results about erasure decoding. In Section 3, we formulate the problem of universal decoding with erasures. In Section 4, we present the proposed universal erasure decoder and prove its asymptotic optimality in the competitive minimax sense. In Section 5, we present the main results concering the performance of the proposed universal decoder. Section 6 is devoted to the example of the BSC. Finally, in Section 7, we summarize our conclusions.

2 Notation and Preliminaries

Throughout this paper, scalar random variables (RV's) will be denoted by capital letters, their sample values will be denoted by the respective lower case letters, and their alphabets will be denoted by the respective calligraphic letters. A similar convention will apply to random vectors of dimension n and their sample values, which will be denoted with same symbols in the bold face font. The set of all n–vectors with components taking values in a certain alphabet, will be denoted as the same alphabet superscripted by n. Thus, for example, a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ may assume a specific vector value $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ as each component takes values in $\mathcal{X}$. Channels will be denoted generically by the letter P, or P_θ, when we wish to emphasize that the channel is indexed or parametrized by a certain scalar or vector θ, taking on values in some set Θ. Information theoretic quantities like entropies and conditional entropies, will be denoted following the usual conventions of the information theory literature, e.g., $H(X)$, $H(X|Y)$, and so on. With a slight abuse of notation, when we wish to emphasize the dependence of the entropy on the underlying probability distribution P, we denote it by $H(P)$. The cardinality of a finite set $\mathcal{A}$ will be denoted by $|\mathcal{A}|$.

Consider a discrete memoryless channel (DMC) with a finite input alphabet $\mathcal{X}$, finite output alphabet $\mathcal{Y}$, and single–letter transition probabilities $\{P(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$. As the channel is fed by an input vector $\boldsymbol{x} \in \mathcal{X}^n$, it generates an output vector $\boldsymbol{y} \in \mathcal{Y}^n$ according to the sequence conditional probability distributions (cf. [16]):

$$P(y_i|x_1, \ldots, x_i, y_1, \ldots, y_{i-1}) = P(y_i|x_i), \quad i = 1, 2, \ldots, n \tag{1}$$

where for $i = 1$, $(y_1, \ldots, y_{i-1})$ is understood as the null string. A rate–R block code of length n consists of $M = e^{nR}$ n–vectors $\boldsymbol{x}_m \in \mathcal{X}^n$, $m = 1, 2, \ldots, M$, which represent M different messages. We will assume that all possible messages are a–priori equiprobable, i.e., $P(m) = 1/M$ for all $m = 1, 2, \ldots, M$.

A decoder with an erasure option is a partition of $\mathcal{Y}^n$ into $(M+1)$ regions, $\mathcal{R}_0, \mathcal{R}_1, \ldots, \mathcal{R}_M$. Such a decoder works as follows: If $\boldsymbol{y}$ falls into $\mathcal{R}_m$, $m = 1, 2, \ldots, M$, then a decision is made in favor of message number m. If $\boldsymbol{y} \in \mathcal{R}_0$, no decision is made and an erasure is declared. We will refer to $\mathcal{R}_0$ as the *erasure event.*

Given a code $\mathcal{C} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ and a decoder$\mathcal{R} = (\mathcal{R}_0, \mathcal{R}_1, \ldots, \mathcal{R}_m)$, let us now define two additional undesired events. The event $\mathcal{E}_1$ is the event of not making the right decision. This event is the disjoint union of the erasure event and the event $\mathcal{E}_2$, which is the *undetected error* event, namely, the event of making the wrong decision. The probabilities of all three events are defined as follows:

$$\begin{aligned}
\Pr\{\mathcal{E}_1\} &= \sum_{m=1}^{M} \sum_{\boldsymbol{y} \in \mathcal{R}_m^c} P(\boldsymbol{x}_m, \boldsymbol{y}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{\boldsymbol{y} \in \mathcal{R}_m^c} P(\boldsymbol{y}|\boldsymbol{x}_m) \qquad (2)\\
\Pr\{\mathcal{E}_2\} &= \sum_{m=1}^{M} \sum_{\boldsymbol{y} \in \mathcal{R}_m} \sum_{m' \neq m} P(\boldsymbol{x}_{m'}, \boldsymbol{y}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{\boldsymbol{y} \in \mathcal{R}_m} \sum_{m' \neq m} P(\boldsymbol{y}|\boldsymbol{x}_{m'}) \qquad (3)\\
\Pr\{\mathcal{R}_0\} &= \Pr\{\mathcal{E}_1\} - \Pr\{\mathcal{E}_2\}. \qquad (4)
\end{aligned}$$

Forney [10] assumes that the DMC is known to the decoder, and shows, using the Neyman–Pearson methodology, that the best tradeoff between $\Pr\{\mathcal{E}_1\}$ and $\Pr\{\mathcal{E}_2\}$ (or, equivalently, between $\Pr\{\mathcal{R}_0\}$ and $\Pr\{\mathcal{E}_2\}$) is attained by the decoder $\mathcal{R}^* = (\mathcal{R}_0^*, \mathcal{R}_1^*, \ldots, \mathcal{R}_M^*)$ defined by

$$\begin{aligned}
\mathcal{R}_m^* &= \left\{\boldsymbol{y} : \frac{P(\boldsymbol{y}|\boldsymbol{x}_m)}{\sum_{m' \neq m} P(\boldsymbol{y}|\boldsymbol{x}_{m'})} \geq e^{nT}\right\}, \quad m = 1, 2, \ldots, M\\
\mathcal{R}_0^* &= \bigcap_{m=1}^{M} (\mathcal{R}_m^*)^c, \qquad (5)
\end{aligned}$$

where $(\mathcal{R}_m^*)^c$ is the complement of $\mathcal{R}_m^*$, and where $T \geq 0$ is a parameter, henceforth referred to as the *threshold*, which controls the balance between the probabilities of $\mathcal{E}_1$ and $\mathcal{E}_2$.

Forney devotes the remaining part of his paper [10] to derive lower bounds to the random coding exponents (associated with $\mathcal{R}^*$), $E_1(R, T)$ and $E_2(R, T)$, of $\overline{\Pr}\{\mathcal{E}_1\}$ and $\overline{\Pr}\{\mathcal{E}_2\}$, the average[4] probabilities of $\mathcal{E}_1$ and $\mathcal{E}_2$, respectively, and to investigate their properties. Specifically, Forney shows, among other things, that for the ensemble of randomly chosen

[4] Here, "average" means with respect to (w.r.t.) the ensemble of randomly selected codes.

codes, where each codeword is chosen independently under an i.i.d. distribution $Q_n(\boldsymbol{x}) = \prod_{i=1}^n Q(x_i)$,

$$E_1(R,T) = \max_{0\le s\le\rho\le 1}\max_Q[E_0(s,\rho,Q) - \rho R - sT] \tag{6}$$

where

$$E_0(s,\rho,Q) = -\ln\left[\sum_{y\in\mathcal{Y}}\left(\sum_{x\in\mathcal{X}} Q(x)P^{1-s}(y|x)\right)\cdot\left(\sum_{x'\in\mathcal{X}} Q(x')P^{s/\rho}(y|x')\right)^\rho\right], \tag{7}$$

and

$$E_2(R,T) = E_1(R,T) + T. \tag{8}$$

A simple observation that we will need, before passing to the case of an unknown channel, is that the same decision rule $\mathcal{R}^*$ would be obtained if rather than adopting the Neyman–Pearson approach, one would consider a Lagrange function,

$$\Gamma(\mathcal{C},\mathcal{R}) \stackrel{\triangle}{=} \Pr\{\mathcal{E}_2\} + e^{-nT}\Pr\{\mathcal{E}_1\}, \tag{9}$$

for a given code $\mathcal{C} = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_M\}$ and a given threshold T, as the figure of merit, and seek a decoder $\mathcal{R}$ that minimizes it. To see that this is equivalent, let us rewrite $\Gamma(\mathcal{C},\mathcal{R})$ as follows:

$$\Gamma(\mathcal{C},\mathcal{R}) = \frac{1}{M}\sum_{m=1}^M\left[\sum_{\boldsymbol{y}\in\mathcal{R}_m}\sum_{m'\ne m} P(\boldsymbol{y}|\boldsymbol{x}_{m'}) + \sum_{\boldsymbol{y}\in\mathcal{R}_m^c} e^{-nT}P(\boldsymbol{y}|\boldsymbol{x}_m)\right], \tag{10}$$

and it is now clear that for each m, the bracketed expression (which has the form of weighted error of a binary hypothesis testing problem) is minimized by $\mathcal{R}_m^*$ as defined above. Since this decision rule is identical to Forney's one, it is easy to see that the resulting exponential decay of the ensemble average

$$\boldsymbol{E}\{\Gamma(\mathcal{C},\mathcal{R}^*)\} = \overline{\Pr}\{\mathcal{E}_2\} + e^{-nT}\overline{\Pr}\{\mathcal{E}_1\}$$

is $E_2(R,T)$, as $\overline{\Pr}\{\mathcal{E}_1\}$ decays according to $e^{-nE_1(R,T)}$, $\overline{\Pr}\{\mathcal{E}_2\}$ decays according to $e^{-nE_2(R,T)}$, and $E_2(R,T) = E_1(R,T) + T$, as mentioned earlier. This Largrangian approach will be more convenient to work with, when we next move on to the case of an unknown DMC, because it allows as to work with one figure of merit instead of a trade–off between two.

3 Unknown Channel – Problem Description

We now move on to the case of an unknown channel. While our techniques can be applied to quite general classes of channels, here, for the sake of concreteness and conceptual simplicity, and following in [10], we confine attention to DMC's. Consider then a family of DMC's $\{P_\theta(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y},\ \theta \in \Theta\}$, where θ is the parameter, or the index of the channel in the class, taking values in some set Θ. For example, θ may be a positive integer, denoting the index of the channel within a finite or a countable index set. As another example, θ may simply represent the set of all $|\mathcal{X}| \cdot (|\mathcal{Y}| - 1)$ single–letter transition probabilties that define the DMC, and if there are some symmetries (like in the BSC), these reduce the dimensionality of θ. The basic questions are now the following:

1. How to devise a good erasure decoder when the underlying channel is known to belong to the class $\{P_\theta(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y},\ \theta \in \Theta\}$, but θ is unknown?

2. What are the resulting error exponents of $\mathcal{E}_1$ and $\mathcal{E}_2$ and how do they compare to Forney's exponents for known θ?

In the quest for universal schemes for decoding with an erasure option, two difficulties[5] are encountered in light of [10]. The first difficulty is that here we have two figures of merits, the probabilities of $\mathcal{E}_1$ and $\mathcal{E}_2$. But this difficulty can be alleviated by adopting the Lagrangian approach, described at the end of the previous section. The second difficulty is somewhat deeper: Classical derivations of universal decoding rules for ordinary decoding (without erasures) over the class of DMC's, like the maximum mutual information (MMI) decoder [4] and its variants, were based on ideas that are deeply rooted in considerations of joint typicality between the channel output $\boldsymbol{y}$ and each hypothesized codeword $\boldsymbol{x}_m$. These considerations were easy to apply in ordinary decoding, where the score function (or, the "metric") associated with the optimum maximum likelihood (ML) decoding, $\log P_\theta(\boldsymbol{y}|\boldsymbol{x}_m)$, involves only *one* codeword at a time, and that this function depends on $\boldsymbol{x}_m$ and $\boldsymbol{y}$ only via their joint empirical distribution, or, in other words, their joint type. Moreover, in the case of decoding without erasures, given the true transmitted codeword $\boldsymbol{x}_m$ and the resulting channel output $\boldsymbol{y}$, the scores associated with all other randomly chosen codewords, are independent of each other, a fact that facilitates the analysis to a great extent. This is very different from the situation in erasure decoding, where Forney's optimum score function for each codeword,

$$\frac{P_\theta(\boldsymbol{y}|\boldsymbol{x}_m)}{\sum_{m' \neq m} P_\theta(\boldsymbol{y}|\boldsymbol{x}_{m'})},$$

[5] These difficulties may also be related to the observation discussed in the Introduction, that optimum error exponents may not be universally achievable in the erasure decoding setting.

depends on *all* codewords at the same time. Consequently, in a random coding analysis, it is rather complicated to apply joint typicality considerations, or to analyze the statistical behavior of this expression, let alone the statistical dependency between the score functions associated with the various codewords.

This difficulty is avoided if the competitive minimax methodology, proposed and developed in [9], is applied. Specifically, let $\Gamma_\theta(\mathcal{C},\mathcal{R})$ denote the above defined Lagrangian, where we now emphasize the dependence on the index of the channel, θ. Let us also define $\bar{\Gamma}_\theta^* = \boldsymbol{E}\{\min_{\mathcal{R}} \Gamma_\theta(\mathcal{C},\mathcal{R})\}$, i.e., the ensemble average of the minimum of the above Lagrangian (achieved by Forney's optimum decision rule) with respect to (w.r.t.) the channel $\{P_\theta(y|x)\}$ for a given θ. Note that the exponential order of $\bar{\Gamma}_\theta^*$ is $e^{-n[E_1(R,T,\theta)+T]} = e^{-nE_2(R,T,\theta)}$, where $E_1(R,T,\theta)$ and $E_2(R,T,\theta)$ are the new notations for $E_1(R,T)$ and $E_2(R,T)$, respectively, with the dependence on the channel index θ, made explicit. In principle, we would have been interested in a decision rule $\mathcal{R}$ that achieves

$$\min_{\mathcal{R}} \max_{\theta\in\Theta} \frac{\Gamma_\theta(\mathcal{C},\mathcal{R})}{\bar{\Gamma}_\theta^*}, \tag{11}$$

or, equivalently,

$$\min_{\mathcal{R}} \max_{\theta\in\Theta} \frac{\Gamma_\theta(\mathcal{C},\mathcal{R})}{e^{-n[E_1(R,T,\theta)+T]}}. \tag{12}$$

However, as is discussed in [9] (in the analogous context of ordinary decoding, without erasures), such an ambitious minimax criterion of competing with the optimum performance may be too optimisitic: If $[E_1(R,T,\theta)+T]$ is not universally achievable, then the value of the above minimax may grow exponentially with n, and then there might be values of θ for which the numerator does not tend to zero at all, whereas the denominator still does. A better approach would be to compete with a similar expression of the exponential behavior, but where the term $E_1(R,T,\theta)$ is being multiplied by a constant $\xi \in (0,1]$, which we would like to choose as large as possible. In other words, we are interested in the competitive minimax criterion

$$K_n(\mathcal{C}) \stackrel{\triangle}{=} \min_{\mathcal{R}} \max_{\theta\in\Theta} \frac{\Gamma_\theta(\mathcal{C},\mathcal{R})}{e^{-n[\xi E_1(R,T,\theta)+T]}}. \tag{13}$$

Similarly as in [9], we wish to find the largest value of ξ such that the ensemble average $\bar{K}_n \stackrel{\triangle}{=} \boldsymbol{E}\{K_n(\mathcal{C})\}$ would *not* grow exponentially fast, i.e.,

$$\limsup_{n\to\infty} \frac{1}{n} \log \bar{K}_n \le 0. \tag{14}$$

The rationale behind this is the following: If $\bar{K}_n$ is sub–exponential in n, for some ξ, then this guarantees that there exists a code $\hat{\mathcal{C}}$ and a universal erasure decoder $\hat{\mathcal{R}}$, such that for *every* $\theta \in \Theta$, the exponential order of $\Gamma_\theta(\hat{\mathcal{C}}, \hat{\mathcal{R}})\}$ is no worse than $e^{-n[\xi E_1(R,T,\theta)+T]}$. This, in turn, implies that *both* terms of $\Gamma_\theta(\hat{\mathcal{C}}, \hat{\mathcal{R}})$ decay at least as $e^{-n[\xi E_1(R,T,\theta)+T]}$, which means that for the decoder $\hat{\mathcal{R}}$, the exponent of $\overline{\Pr}\{\mathcal{E}_1\}$ is at least $\xi \cdot E_1(R,T,\theta)$ and the exponent of $\overline{\Pr}\{\mathcal{E}_2\}$ is at least $\xi \cdot E_1(R,T,\theta) + T$, both for every $\theta \in \Theta$. Thus, the difference between the two (guaranteed) exponents remains T as before (as the weight of the term $\overline{\Pr}\{\mathcal{E}_1\}$ in $\Gamma(\mathcal{R}, \mathcal{C})$ is e^{-nT}), but the other term, $E_1(R,T,\theta)$, is now scaled by a factor of ξ.

The remaining parts of this paper focus on deriving a universal decoding rule that asymptotically achieves $\bar{K}_n(\mathcal{C})$ for a given ξ, and on analyzing its performance, i.e., finding the maximum value of ξ such that $\bar{K}_n$ still grows sub–exponentially rapidly.

4 Derivation of a Universal Erasure Decoder

For a given $\xi \in (0,1]$, let us define

$$f(\boldsymbol{x}_m, \boldsymbol{y}) \stackrel{\triangle}{=} \max_{\theta \in \Theta} \left\{ e^{n[\xi E_1(R,T,\theta)+T]} P_\theta(\boldsymbol{y}|\boldsymbol{x}_m) \right\} \tag{15}$$

and consider the decoder

$$\begin{aligned} \hat{\mathcal{R}}_m &= \left\{ \boldsymbol{y} :\ \frac{f(\boldsymbol{x}_m, \boldsymbol{y})}{\sum_{m' \neq m} f(\boldsymbol{x}_{m'}, \boldsymbol{y})} \geq e^{nT} \right\}, \quad m = 1, 2, \ldots, M \\ \hat{\mathcal{R}}_0 &= \bigcap_{m=1}^{M} \hat{\mathcal{R}}_m^c. \end{aligned} \tag{16}$$

Note that this can be thought of as an extension of a decoder in the spirit of the generalized–likelihood–ratio–test (GLRT), where the unknown parameter θ is estimated by the maximum likelihood estimator for each term $P_\theta(\boldsymbol{y}|\boldsymbol{x}_i)$ individually. While this GLRT–like decoder is a special case of the above, corresponding to $\xi = 0$, the more general decoder, proposed here, assigns higher weights to good channels, as discussed in [9]. Denoting

$$K_n(\mathcal{C}, \mathcal{R}) = \max_{\theta \in \Theta} \frac{\Gamma_\theta(\mathcal{C}, \mathcal{R})}{e^{-n[\xi E_1(R,T,\theta)+T]}}, \tag{17}$$

for a given encoder $\mathcal{C} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ and decoder $\mathcal{R}$, our first main result estabilishes the asymptotic optimality of $\hat{\mathcal{R}}$ in the competitive minimax sense, namely, that $K_n(\mathcal{C}, \hat{\mathcal{R}})$ is within a sub–exponential factor as small as $K_n(\mathcal{C}) = \min_{\mathcal{R}} K_n(\mathcal{C}, \mathcal{R})\}$, and therefore, $\boldsymbol{E}\{K_n(\mathcal{C}, \hat{\mathcal{R}})\}$ is within the same sub–exponential factor as small as $\bar{K}_n = \boldsymbol{E}\{K_n(\mathcal{C})\}$.

Theorem 1 *For every code $\mathcal{C}$,*

$$K_n(\mathcal{C},\hat{\mathcal{R}}) \le (n+1)^{|\mathcal{X}|\cdot|\mathcal{Y}|-1} K_n(\mathcal{C}). \tag{18}$$

Proof. The result and the proof technique is similar to those of [9]. As $\boldsymbol{x}$ and $\boldsymbol{y}$ exhaust their spaces, $\mathcal{X}^n$ and $\mathcal{Y}^n$, let Θ_n denote set of values of θ that achieve $\{f(\boldsymbol{x},\boldsymbol{y}),\ \boldsymbol{x}\in\mathcal{X}^n,\ \boldsymbol{y}\in\mathcal{Y}^n\}$. Observe that for every θ, the expression $[e^{n[\xi E_1(R,T,\theta)+T]}P_\theta(\boldsymbol{y}|\boldsymbol{x})]$ depends on $(\boldsymbol{x},\boldsymbol{y})$ only via their joint empirical distribution (or, the joint type). Consequently, the value of θ that achieves $f(\boldsymbol{x},\boldsymbol{y})$ also depends on $(\boldsymbol{x},\boldsymbol{y})$ only via their joint empirical distribution. Since the number of joint empirical distributions of $(\boldsymbol{x},\boldsymbol{y})$ never exceeds $(n+1)^{|\mathcal{X}|\cdot|\mathcal{Y}|-1}$ (see [4]), then obviously

$$|\Theta_n| \le (n+1)^{|\mathcal{X}|\cdot|\mathcal{Y}|-1} \tag{19}$$

as well. Now, for every encoder $\mathcal{C}$ and decoder $\mathcal{R}$,

$$\begin{aligned}
K_n(\mathcal{C},\mathcal{R}) &= \max_{\theta\in\Theta}\frac{\Gamma_\theta(\mathcal{C},\mathcal{R})}{e^{-n[\xi E_1(R,T,\theta)+T]}}\\
&= \max_{\theta\in\Theta}\frac{1}{M}\sum_{m=1}^{M}\left[\sum_{\boldsymbol{y}\in\mathcal{R}_m}\sum_{m'\neq m}\frac{P_\theta(\boldsymbol{y}|\boldsymbol{x}_{m'})}{e^{-n[\xi E_1(R,T,\theta)+T]}}+e^{-nT}\sum_{\boldsymbol{y}\in\mathcal{R}_m^c}\frac{P_\theta(\boldsymbol{y}|\boldsymbol{x}_m)}{e^{-n[\xi E_1(R,T,\theta)+T]}}\right]\\
&\le \frac{1}{M}\sum_{m=1}^{M}\left[\sum_{\boldsymbol{y}\in\mathcal{R}_m}\sum_{m'\neq m}\max_{\theta\in\Theta}\frac{P_\theta(\boldsymbol{y}|\boldsymbol{x}_{m'})}{e^{-n[\xi E_1(R,T,\theta)+T]}}+e^{-nT}\sum_{\boldsymbol{y}\in\mathcal{R}_m^c}\max_{\theta\in\Theta}\frac{P_\theta(\boldsymbol{y}|\boldsymbol{x}_m)}{e^{-n[\xi E_1(R,T,\theta)+T]}}\right]\\
&= \frac{1}{M}\sum_{m=1}^{M}\left[\sum_{\boldsymbol{y}\in\mathcal{R}_m}\sum_{m'\neq m}f(\boldsymbol{x}_{m'},\boldsymbol{y})+e^{-nT}\sum_{\boldsymbol{y}\in\mathcal{R}_m^c}f(\boldsymbol{x}_m,\boldsymbol{y})\right]\\
&\stackrel{\triangle}{=} \hat{K}_n(\mathcal{C},\mathcal{R})\\
&= \frac{1}{M}\sum_{m=1}^{M}\left[\sum_{\boldsymbol{y}\in\mathcal{R}_m}\sum_{m'\neq m}\max_{\theta\in\Theta_n}\frac{P_\theta(\boldsymbol{y}|\boldsymbol{x}_{m'})}{e^{-n[\xi E_1(R,T,\theta)+T]}}+e^{-nT}\sum_{\boldsymbol{y}\in\mathcal{R}_m^c}\max_{\theta\in\Theta_n}\frac{P_\theta(\boldsymbol{y}|\boldsymbol{x}_m)}{e^{-n[\xi E_1(R,T,\theta)+T]}}\right]\\
&\le \frac{1}{M}\sum_{m=1}^{M}\left[\sum_{\boldsymbol{y}\in\mathcal{R}_m}\sum_{m'\neq m}\left(\sum_{\theta\in\Theta_n}\frac{P_\theta(\boldsymbol{y}|\boldsymbol{x}_{m'})}{e^{-n[\xi E_1(R,T,\theta)+T]}}\right)+\right.\\
&\qquad\left.e^{-nT}\sum_{\boldsymbol{y}\in\mathcal{R}_m^c}\left(\sum_{\theta\in\Theta_n}\frac{P_\theta(\boldsymbol{y}|\boldsymbol{x}_m)}{e^{-n[\xi E_1(R,T,\theta)+T]}}\right)\right]\\
&= \sum_{\theta\in\Theta_n}\frac{1}{M}\sum_{m=1}^{M}\left[\sum_{\boldsymbol{y}\in\mathcal{R}_m}\sum_{m'\neq m}\frac{P_\theta(\boldsymbol{y}|\boldsymbol{x}_{m'})}{e^{-n[\xi E_1(R,T,\theta)+T]}}+\right.\\
&\qquad\left.e^{-nT}\sum_{\boldsymbol{y}\in\mathcal{R}_m^c}\frac{P_\theta(\boldsymbol{y}|\boldsymbol{x}_m)}{e^{-n[\xi E_1(R,T,\theta)+T]}}\right]\\
&\le |\Theta_n|\cdot\max_{\theta\in\Theta_n}\frac{1}{M}\sum_{m=1}^{M}\left[\sum_{\boldsymbol{y}\in\mathcal{R}_m}\sum_{m'\neq m}\frac{P_\theta(\boldsymbol{y}|\boldsymbol{x}_{m'})}{e^{-n[\xi E_1(R,T,\theta)+T]}}+\right.
\end{aligned}$$

$$
\begin{aligned}
& e^{-nT}\sum_{\boldsymbol{y}\in\mathcal{R}_m^c}\frac{P_\theta(\boldsymbol{y}|\boldsymbol{x}_m)}{e^{-n[\xi E_1(R,T,\theta)+T]}}\Bigg] \\
\le\ & (n+1)^{|\mathcal{X}|\cdot|\mathcal{Y}|-1}\cdot\max_{\theta\in\Theta}\frac{1}{M}\sum_{m=1}^{M}\Bigg[\sum_{\boldsymbol{y}\in\mathcal{R}_m}\sum_{m'\neq m}\frac{P_\theta(\boldsymbol{y}|\boldsymbol{x}_{m'})}{e^{-n[\xi E_1(R,T,\theta)+T]}}+ \\
& e^{-nT}\sum_{\boldsymbol{y}\in\mathcal{R}_m^c}\frac{P_\theta(\boldsymbol{y}|\boldsymbol{x}_m)}{e^{-n[\xi E_1(R,T,\theta)+T]}}\Bigg] \\
\le\ & (n+1)^{|\mathcal{X}|\cdot|\mathcal{Y}|-1}\cdot K_n(\mathcal{C},\mathcal{R}). \qquad (20)
\end{aligned}
$$

Thus, we have defined $\hat{K}_n(\mathcal{C},\mathcal{R})$ and sandwiched it between $K_n(\mathcal{C},\mathcal{R})$ and $(n+1)^{|\mathcal{X}|\cdot|\mathcal{Y}|-1}\cdot K_n(\mathcal{R},\mathcal{C})$ uniformly for every $\mathcal{C}$ and $\mathcal{R}$. Now, obviously, $\hat{\mathcal{R}}$ minimizes $\hat{K}_n(\mathcal{C},\mathcal{R})$, and so, for every $\mathcal{R}$,

$$K_n(\mathcal{C},\hat{\mathcal{R}})\le\hat{K}_n(\mathcal{C},\hat{\mathcal{R}})\le\hat{K}_n(\mathcal{C},\mathcal{R})\le(n+1)^{|\mathcal{X}|\cdot|\mathcal{Y}|-1}\cdot K_n(\mathcal{C},\mathcal{R}), \qquad (21)$$

where the first and the third inequalites were just proved in the chain of inequalities (20), and the second inequality follows from the optimality of $\hat{\mathcal{R}}$ w.r.t. $\hat{K}_n(\mathcal{C},\mathcal{R})$. Since we have shown that

$$K_n(\mathcal{C},\hat{\mathcal{R}})\le(n+1)^{|\mathcal{X}|\cdot|\mathcal{Y}|-1}\cdot K_n(\mathcal{C},\mathcal{R})$$

for every $\mathcal{R}$, we can now minimize the r.h.s. w.r.t. $\mathcal{R}$ and the assertion of Theorem 1 is obtained. This completes the proof of Theorem 1.

5 Performance

In this section, we present an upper bound to $\bar{K}_n$ from which we derive a lower bound to ξ^*, the largest value of ξ for which $\bar{K}_n$ is sub–exponential in n.

We begin with a few definitions. The empirical disitribution $\hat{P}_{\boldsymbol{x}}$ of $\boldsymbol{x}$ is the vector of relative frequencies $\{\hat{P}_{\boldsymbol{x}}(a)=n_{\boldsymbol{x}}(a)/n,\ a\in\mathcal{X}\}$, $n_{\boldsymbol{x}}(a)$ being the number of occurrences of $a\in\mathcal{X}$ within $\boldsymbol{x}\in\mathcal{X}^n$. The type class $T_{\boldsymbol{x}}$ of $\boldsymbol{x}\in\mathcal{X}^n$ is the set of all $\boldsymbol{x}'\in\mathcal{X}^n$ such that $\hat{P}_{\boldsymbol{x}'}=\hat{P}_{\boldsymbol{x}}$. We next define the class $\mathcal{Q}$ of the sequences of random coding distributions $\{Q_n\}$ that we assume. For every positive integer n, let Q_n be a random coding distribution of the following form:

$$Q_n(\boldsymbol{x})=\frac{Q_n(T_{\boldsymbol{x}})}{|T_{\boldsymbol{x}}|}, \qquad (22)$$

where, of course, $\sum_{T_{\boldsymbol{x}}}Q_n(T_{\boldsymbol{x}})=1$. Let

$$\Delta_n(P_{\boldsymbol{x}})=-\frac{1}{n}\ln Q_n(T_{\boldsymbol{x}}),$$

and let $\Delta_n^*(P)$ be an extension of the function $\Delta_n^*(P_{\boldsymbol{x}})$ that is defined over the continuum of probability distributions over $\mathcal{X}$ (rather than just the set of rational probability distributions with denominator n). A sequence of random coding distributions $\{Q_n\}_{n\geq 1}$ is said to belong to the class $\mathcal{Q}$ if there exists such an extension $\Delta_n^*(P)$ that converges, as $n\to\infty$, to a certain non–negative functional $\Delta^*(P)$, uniformly over all probability distributions $\{P\}$ over $\mathcal{X}$.

It is easy to see that the class $\mathcal{Q}$ essentially covers all random coding distributions that are customarily used (and much more). In particular, to approximate a random coding distribution which is uniform within a small neighborhood of one type class – corresponding to a probability distribution P_0, and which vanishes elsewhere, we set $\Delta^*(P)=0$ for every P in that neighborhood of P_0, and $\Delta^*(P)=\infty$ elsewhere. For the case where Q is i.i.d., $\Delta^*(P)=D(P\|Q)=\sum_{a\in\mathcal{X}}P(a)\ln[P(a)/Q(a)]$, the Kullback–Leibler divergence between Q and P. In particular, if $Q(\boldsymbol{x})=1/|\mathcal{X}|^n$ for all $\boldsymbol{x}\in\mathcal{X}^n$, then $\Delta^*(P)=\ln|\mathcal{X}|-H(P)$, $H(P)$ being the entropy associated with the distribution P.

Given a distribution P_y on $\mathcal{Y}$, a positive real λ, and a value of θ, let

$$F(P_y,\lambda,\theta)\stackrel{\triangle}{=}\min_{P_{x|y}}\left[I(X;Y)+\Delta^*\left(\sum_{b\in\mathcal{Y}}P_y(b)P_{x|y}(\cdot|b)\right)-\lambda\boldsymbol{E}\ln P_\theta(Y|X)\right], \tag{23}$$

where $\boldsymbol{E}\{\cdot\}$ is the expectation and $I(X;Y)$ is the mutual information w.r.t. a generic joint distribution $P_{xy}(x,y)=P_y(y)P_{x|y}(x|y)$ of the RV's (X,Y). Next, for a pair $(\theta,\tilde{\theta})\in\Theta^2$, and for two real numbers s and ρ, $0\leq s\leq\rho\leq 1$, define:

$$E(\theta,\tilde{\theta},\rho,s)=\min_{P_y}[F(P_y,1-s,\theta)+\rho F(P_y,s/\rho,\tilde{\theta})-H(Y)], \tag{24}$$

where $H(Y)$ is the entropy of Y induced by P_y. Finally, let

$$\xi^*(R,T)\stackrel{\triangle}{=}\min_{\theta,\tilde{\theta}}\max_{0\leq s\leq\rho\leq 1}\frac{E(\theta,\tilde{\theta},s,\rho)-\rho R-sT}{(1-s)E_1(R,T,\theta)+sE_1(R,T,\tilde{\theta})}, \tag{25}$$

with the convention that if the denominator vanishes, then $\xi^*(R,T)\stackrel{\triangle}{=}1$. Our main result, in this section is the following:

Theorem 2 *Consider a sequence of ensemble of codes where each codeword is drawn independently, under a distribution Q_n, where the sequence $\{Q_n\}_{n\geq 1}$ is a member of the class $\mathcal{Q}$. Then,*

1. *For every $\xi\leq\xi^*(R,T)$,*

$$\limsup_{n\to\infty}\frac{1}{n}\log\bar{K}_n\leq 0.$$

2. There exists a sequence of encoders and decoders such that for every $\theta \in \Theta$,

$$\liminf_{n\to\infty}\left[-\frac{1}{n}\log\Gamma_\theta(\mathcal{C},\mathcal{R})\right] \geq \xi^*(R,T)\cdot E_1(R,T,\theta)+T.$$

The proof of the first part of Theorem 2 appears in the appendix. The second part follows immediately as discussed after eq. (14).

We now pause to discuss Theorem 2 and some of its aspects.

Theorem 2 suggests a conceptually simple strategy: Given R and T, first compute $\xi^*(R,T)$ using eq. (25). This may require some non-trivial optimization procedures, but it has to be done only once, and since this is a single–letter expression, it can be carried at least numerically, if closed–form analytic expressions are not apparent to be available (see the example of the BSC below). Once $\xi^*(R,T)$ has been computed, apply the decoding rule $\hat{\mathcal{R}}$ with $\xi = \xi^*(R,T)$, and the theorem guarantees that the resulting random coding error exponents of $\mathcal{E}_1$ and $\mathcal{E}_2$ are at least $\xi^*(R,T)\cdot E_1(R,T,\theta)$ and $\xi^*(R,T)\cdot E_1(R,T,\theta)+T$, respectively.

The theorem is interesting, of course, only when $\xi^*(R,T) > 0$, which is the case in many situations, at least as long as R and T are not too large. When $\xi^*(R,T) > 0$, the proposed universal decoder with $\xi = \xi^*(R,T)$ has the important property that whenever Forney's optimum decoder yields an exponential decay of $\overline{\Pr}\{\mathcal{E}_1\}$ ($E_1(R,T,\theta) > 0$), then so does the corresponding exponent of the proposed decoder, $\hat{\mathcal{R}}$. It should be pointed out that the exponential rates $\xi^*(R,T)\cdot E_1(R,T,\theta)$ and $\xi^*(R,T)\cdot E_1(R,T,\theta)+T$, guaranteed by Theorem 2, are only lower bounds to the real exponential rates, and that true exponential rate, at some points in Θ, might be larger.

Our last comment concerns the choice of the threshold T. Thus far, we assumed that T is a constant, independent of θ. However, in some situations, it makes sense to let T depend on the quality of the channel, and hence on the parameter θ. Intuitively, for fixed T, if the signal–to–noise ratio (SNR) becomes very high, the erasure option will be used so rarely, that it will effectively be non–existent. This means that we are actually no longer "enjoying" the benefits of the erasure option, and hence not the gain in the undetected error exponent that is associated with it. An alternative approach is to let $T = T_\theta$ depend on θ in a certain way. In this case, $K_n(\mathcal{C})$ would be redefined as follows:

$$K_n(\mathcal{C}) = \min_{\mathcal{R}}\max_{\theta\in\Theta}\frac{e^{-nT_\theta}\Pr\{\mathcal{E}_1\}+\Pr\{\mathcal{E}_2\}}{e^{-n[\xi E_1(R,T_\theta,\theta)+T_\theta]}}. \quad (26)$$

The corresponding generalized version of the competitive minimax decision rule $\hat{\mathcal{R}}$, would now be:

$$\begin{aligned}
\hat{\mathcal{R}}_m &= \left\{\boldsymbol{y}:\ g(\boldsymbol{x}_m,\boldsymbol{y}) \geq \sum_{m'\neq m} h(\boldsymbol{x}_{m'},\boldsymbol{y})\right\},\ m=1,\ldots,M \\
\hat{\mathcal{R}}_0 &= \bigcap_{m=1}^{M} \hat{\mathcal{R}}_m^c, \qquad (27)
\end{aligned}$$

where

$$g(\boldsymbol{x}_m,\boldsymbol{y}) \stackrel{\triangle}{=} \max_\theta [P_\theta(\boldsymbol{y}|\boldsymbol{x}_m)\cdot e^{n\xi E_1(R,T_\theta,\theta)}] \qquad (28)$$

and

$$h(\boldsymbol{x}_m,\boldsymbol{y}) \stackrel{\triangle}{=} \max_\theta [P_\theta(\boldsymbol{y}|\boldsymbol{x}_m)\cdot e^{n[\xi E_1(R,T_\theta,\theta)+T_\theta]}]. \qquad (29)$$

By extending the performance analysis carried out in the appendix, the resulting expression of ξ^* now becomes

$$\xi^*(R) \stackrel{\triangle}{=} \min_{\theta,\tilde{\theta}} \max_{0\leq s\leq\rho\leq 1} \frac{E(\theta,\tilde{\theta},s,\rho)-\rho R - sT_{\tilde{\theta}}}{(1-s)E_1(R,T_\theta,\theta)+sE_1(R,T_{\tilde{\theta}},\tilde{\theta})}. \qquad (30)$$

The main question that naturally arises, in this case, is: which function T_θ would be reasonable to choose? A plausible guideline could be based on the typical behavior of

$$\tau_\theta = \lim_{N\to\infty}\frac{1}{N}\boldsymbol{E}\ln\frac{P_\theta(\boldsymbol{Y}|\boldsymbol{x}_m)}{\sum_{m'\neq m}P_\theta(\boldsymbol{Y}|\boldsymbol{x}_{m'})}$$

which can be assessed, using standard bounding techniques, under the hypothesis that $\boldsymbol{x}_m$ is the correct message. For example, T_θ may be given by $\alpha\tau_\theta$ with some constant $\alpha\in[0,1]$, or $\tau_\theta-\beta$ for some $\beta>0$. This will make the probability of erasure (exponentially) small, but not *too* small, so that there would be some gain in the undetected error exponent for every θ.

6 Example – the Binary Symmetric Channel

Consider the BSC, where $\mathcal{X}=\mathcal{Y}=\{0,1\}$, and where θ designates the crossover probability, and let the sequence of random coding distributions be uniform, i.e., $Q_n(\boldsymbol{x})=1/|\mathcal{X}|^n$ for all $\boldsymbol{x}\in\mathcal{X}^n$, which as mentioned earlier, belongs to the class $\mathcal{Q}$ with $\Delta^*(P)=\ln|\mathcal{X}|-H(P)=\ln 2-H(P)$. We would like to examine, more closely, the expression of $\xi^*(R,T)$ and its behavior in this case. Let $h_2(u)$ denote the binary entropy function, $-u\ln u-(1-u)\ln(1-u)$,

$u \in [0,1]$. Denoting the modulo 2 sum of X and Y by $X \oplus Y$, we then have:

$$\begin{aligned}
F(P_y, \lambda, \theta) &= \min_{P_{x|y}} [I(X;Y) + (\ln 2 - H(X)) - \lambda \boldsymbol{E} \ln P(Y|X)] \\
&= \ln 2 - \max_{P_{x|y}} [H(X|Y) + \lambda \boldsymbol{E} \ln P(Y|X)] \\
&= \ln 2 - \max_{P_{x|y}} \left\{ H(X|Y) + \lambda \boldsymbol{E} \ln \left[(1-\theta) \left(\frac{\theta}{1-\theta} \right)^{X \oplus Y} \right] \right\} \\
&= \ln 2 - \lambda \ln(1-\theta) - \max_{P_{x|y}} \left[H(X|Y) + (\lambda \ln \frac{\theta}{1-\theta}) \cdot \boldsymbol{E}(X \oplus Y) \right] \\
&= \ln 2 - \lambda \ln(1-\theta) - \max_{P_{x|y}} \left[H(X \oplus Y|Y) + (\lambda \ln \frac{\theta}{1-\theta}) \cdot \boldsymbol{E}(X \oplus Y) \right] \\
&\geq \ln 2 - \lambda \ln(1-\theta) - \max_{P_{x|y}} \left[H(X \oplus Y) + (\lambda \ln \frac{\theta}{1-\theta}) \cdot \boldsymbol{E}(X \oplus Y) \right] \\
&= \ln 2 - \lambda \ln(1-\theta) - \max_{u} \left[h_2(u) + (\lambda \ln \frac{\theta}{1-\theta}) \cdot u \right] \\
&= \ln 2 - \lambda \ln(1-\theta) - \ln \left[1 + \left(\frac{\theta}{1-\theta} \right)^{\lambda} \right] \\
&= \ln 2 - \ln[\theta^{\lambda} + (1-\theta)^{\lambda}], \qquad (31)
\end{aligned}$$

where the inequality is, in fact, an equality achieved by a backward $P_{x|y}$ where $X \oplus Y$ is independent of Y. Since $F(P_y, \lambda, \theta)$ is independent of P_y, this easily yields

$$E(\theta, \tilde{\theta}, \rho, s) = \rho \ln 2 - \ln[\theta^{1-s} + (1-\theta)^{1-s}] - \rho \ln[\tilde{\theta}^{s/\rho} + (1-\tilde{\theta})^{s/\rho}] \qquad (32)$$

and so,

$$\xi^*(R,T) = \min_{\theta, \tilde{\theta}} \max_{0 \leq s \leq \rho \leq 1} \frac{\rho \ln 2 - \ln[\theta^{1-s} + (1-\theta)^{1-s}] - \rho \ln[\tilde{\theta}^{s/\rho} + (1-\tilde{\theta})^{s/\rho}] - \rho R - sT}{(1-s)E_1(R,T,\theta) + sE_1(R,T,\tilde{\theta})}, \qquad (33)$$

with

$$E_1(R,T,\theta) = \max_{0 \leq s \leq \rho \leq 1} \{\rho \ln 2 - \ln[\theta^{1-s} + (1-\theta)^{1-s}] - \rho \ln[\theta^{s/\rho} + (1-\theta)^{s/\rho}] - \rho R - sT\}. \qquad (34)$$

This expression, although still involves non–trivial optimizations, is much more explicit than the general one. We next offer a few observations regarding the function $\xi^*(R,T)$ for the example of the BSC.

First, observe that if Θ is a singleton, i.e., we are back to the case of a known channel, then $\theta = \tilde{\theta}$, and the numerator, after maximization over ρ and s, becomes $E_1(R,T,\theta)$, and so does the denominator, thus $\xi^*(R,T) = 1$, as expected. Secondly, we argue that there exists a region of R and T (both not too large) such that $\xi^*(R,T) > 0$. To see this, note that there are four possibilities regarding the minimizers θ and $\tilde{\theta}$ in the above minimax problem:

1. $\theta = \tilde{\theta} = 1/2$: In this case, the denominator vanishes too and so, $\xi^*(R,T) = 1$.

2. Both $\theta \neq 1/2$ and $\tilde{\theta} \neq 1/2$: Let $\hat{\theta}$ be the closer to 1/2 between θ and $\tilde{\theta}$. Then, the numertor is obviously lower bounded by

$$\rho \ln 2 - \ln[\hat{\theta}^{1-s} + (1-\hat{\theta})^{1-s}] - \rho \ln[\hat{\theta}^{s/\rho} + (1-\hat{\theta})^{s/\rho}] - \rho R - sT,$$

   which upon maximizing over ρ and s gives $E_1(R,T,\hat{\theta})$, which is positive as long as R and T are not too large.

3. $\theta = 1/2$ and $\tilde{\theta} \neq 1/2$: In this case, the numerator is given by

$$\rho \ln 2 - \rho \ln[\tilde{\theta}^{s/\rho} + (1-\tilde{\theta})^{s/\rho}] - \rho R - s(T + \ln 2).$$

   Choosing $\rho = 1$ and $s = 1/2$, we get

$$\frac{1}{2}\ln 2 - \ln\left[\sqrt{\tilde{\theta}} + \sqrt{1-\tilde{\theta}}\right] - \left(R + \frac{T}{2}\right),$$

   which is positive as long as R and T are not too large.

4. $\theta \neq 1/2$ and $\tilde{\theta} = 1/2$: In this case, the numerator is given by

$$s \ln 2 - \ln[\theta^{1-s} + (1-\theta)^{1-s}] - \rho R - sT,$$

   and once again, choosing $\rho = 1$ and $s = 1/2$ gives exactly the same expression as in item 3, except that θ replaces $\tilde{\theta}$, and hence the conclusion is identical.

We next demonstrate that $\xi^*(R,0) = 1$. This result is expected, as the case $T = 0$ is asymptotically equivalent (cf. [10]) to the case without erasures in the sense that $E_1(R,0,\theta) = E_2(R,0,\theta)$ coincide with Gallager's random coding exponent [11] (although erasures are still possible). This is in agreement with the aforementioned full universality result for ordinary universal decoding.

Referring to the definition of the Gallager function $E(\theta,\rho)$ for the BSC:

$$E(\theta,\rho) = \rho \ln 2 - (1+\rho)\ln[\theta^{1/(1+\rho)} + (1-\theta)^{1/(1+\rho)}] - \rho R, \qquad (35)$$

	$T=0.000$	$T=0.025$	$T=0.050$	$T=0.075$	$T=0.100$	$T=0.125$	$T=0.150$
$R=0.00$	1.000	0.364	0.523	0.418	0.396	0.422	0.298
$R=0.05$	1.000	0.756	0.713	0.656	0.535	0.562	0.495
$R=0.10$	1.000	0.858	0.774	0.648	0.655	0.585	0.518
$R=0.15$	1.000	0.877	0.809	0.720	0.713	0.662	0.622
$R=0.20$	1.000	0.905	0.815	0.729	0.729	0.684	0.647
$R=0.25$	1.000	0.912	0.832	0.763	0.706	0.661	0.627
$R=0.30$	1.000	0.896	0.850	0.788	0.738	0.644	0.613

Table 1: Numerical values of $\xi^*(R,T)$ for various values of R and T.

let us define $\rho' = 1/(1-s)-1$ and $\rho'' = \rho/s-1$, and rewrite the numerator of the expression for $\xi^*(R,0)$ as follows:

$$\begin{aligned}
& \rho\ln 2 - \ln[\theta^{1-s} + (1-\theta)^{1-s}] - \rho\ln[\tilde{\theta}^{s/\rho} + (1-\tilde{\theta})^{s/\rho}] - \rho R \\
= \;& \rho\ln 2 - \ln[\theta^{1/(1+\rho')} + (1-\theta)^{1/(1+\rho')}] - \rho\ln[\tilde{\theta}^{1/(1+\rho'')} + (1-\tilde{\theta})^{1/(1+\rho'')}] - \rho R \\
= \;& \frac{1}{1+\rho'}\{\rho'\ln 2 - (1+\rho')\ln[\theta^{1/(1+\rho')} + (1-\theta)^{1/(1+\rho')}] - \rho' R\} + \\
& + \frac{\rho}{1+\rho''}\{\rho''\ln 2 - (1+\rho'')\ln[\tilde{\theta}^{1/(1+\rho'')} + (1-\tilde{\theta})^{1/(1+\rho'')}] - \rho'' R\} \\
= \;& (1-s)E(\theta,\rho') + sE(\tilde{\theta},\rho'') \\
= \;& (1-s)E\left(\theta, \frac{1}{1-s}-1\right) + sE\left(\tilde{\theta}, \frac{\rho}{s}-1\right). \qquad (36)
\end{aligned}$$

Now, let us choose $s = \rho/(1+\tilde{\rho})$, where $\tilde{\rho}$ is the achiever of $E^*(\tilde{\theta}) = \max_{0\le\rho\le 1} E(\tilde{\theta},\rho)$, and $\rho = \rho^*(1+\tilde{\rho})/(1+\rho^*)$, where ρ^* is the achiever of $E^*(\theta) = \max_{0\le\rho\le 1} E(\theta,\rho)$ (observing that $\rho^*(1+\tilde{\rho})/(1+\rho^*) \le 1$, therefore this is choice is feasible). With this choice, the numerator of $\xi^*(R,0)$ becomes equal to the denominator, and so, $\xi^*(R,0) = 1$.

Finally, in Table 1, we provide some numerical results pertaining to the function $\xi^*(R,T)$, where all minimizations and maximizations were carried out by an exhaustive search with a step-size of 0.01 in each dimension. As can be seen, at the left–most column, corresponding to $T=0$, we indeed obtain $\xi^*(R,0) = 1$. As can also be seen, $\xi^*(R,T)$ is always strictly less than unity for $T>0$, and it in general decreases as T grows.

7 Conclusion

We have addressed the problem of universal decoding with erasures, using the competitive minimax methodology proposed in [9], which proved useful. This is in contrast to earlier approaches for deriving universal decoders, based on joint typicality considerations, for which we found no apparent extensions to accommodate Forney's erasure decoder. In order to guarantee the uniform achievability of a certain fraction of the exponent, the competitive minimax approach was applied to the Lagrangian, pertaining to a weighted sum of the two error probabilities.

The analysis of the minimax ratio, $\bar{K}_n$, resulted in a single–letter lower bound to the largest universally achievable fraction, $\xi^*(R,T)$ of Forney's exponent. An interesting problem for future work would be to derive a (hopefully compatible) lower bound. This requires the derivation of an exponentially tight lower bound to K_n, which is a challenge.

Our results cover performance analysis of competitive–minimax universal decoders with various types of random coding distributions in a considerable wide class $\mathcal{Q}$. This is in contrast to earlier works (see, e.g., [4], [22]), which were strongly based on the assumption that the random coding distribution is uniform within a set. A similar analysis technique can be applied also to universal decoding without erasures.

Finally, we analyzed the example of the BSC in full detail and demonstrated that $\xi^*(R,0)=1$. We have also provided some numerical results for this case.

Appendix – Proof of Theorem 2

For a given subset $\mathcal{E}\subseteq\mathcal{Y}^n$, let $1\{\boldsymbol{y}|\mathcal{E}\}$ denote the indicator function of $\mathcal{E}$, i.e., $1\{\boldsymbol{y}|\mathcal{E}\}=1$ if $\boldsymbol{y}\in\mathcal{E}$ and $1\{\boldsymbol{y}|\mathcal{E}\}=0$ otherwise. First, observe that

$$\begin{aligned}1\{\boldsymbol{y}|\hat{\mathcal{R}}_m\} &= 1\left\{\boldsymbol{y}|f(\boldsymbol{x}_m,\boldsymbol{y})\geq e^{nT}\sum_{m'\neq m}f(\boldsymbol{x}_{m'},\boldsymbol{y})\right\}\\ &\leq \min_{0\leq s\leq 1}\left[\frac{f(\boldsymbol{x}_m,\boldsymbol{y})}{e^{nT}\sum_{m'\neq m}f(\boldsymbol{x}_{m'},\boldsymbol{y})}\right]^{1-s} \qquad\text{(A.1)}\end{aligned}$$

and similarly,

$$1\{\boldsymbol{y}\in\hat{\mathcal{R}}_m^c\}\leq\min_{0\leq s\leq 1}\left[\frac{e^{nT}\sum_{m'\neq m}f(\boldsymbol{x}_{m'},\boldsymbol{y})}{f(\boldsymbol{x}_m,\boldsymbol{y})}\right]^s. \qquad\text{(A.2)}$$

Then, we have:

$$\begin{aligned}\bar{K}_n &\leq \boldsymbol{E}\{K_n(\hat{\mathcal{R}},\mathcal{C})\}\\ &= \boldsymbol{E}\left\{\max_\theta\left[\frac{e^{-nT}\mathrm{Pr}\{\mathcal{E}_1\}+\mathrm{Pr}\{\mathcal{E}_2\}}{e^{-n[\xi E_1(R,T,\theta)+T]}}\right]\right\}\\ &= \boldsymbol{E}\left\{\max_\theta\frac{1}{M}\sum_{m=1}^M\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\left[e^{-nT}\left(P_\theta(\boldsymbol{y}|\boldsymbol{X}_m)\cdot e^{n[\xi E_1(R,T,\theta)+T]}\right)\cdot 1\{\boldsymbol{y}|\hat{\mathcal{R}}_m^c\}+\right.\right.\end{aligned}$$

$$
\begin{aligned}
& \left(\sum_{m'\neq m} P_\theta(\boldsymbol{y}|\boldsymbol{X}_{m'})\cdot e^{n[\xi E_1(R,T,\theta)+T]}\right)\cdot 1\{\boldsymbol{y}|\hat{\mathcal{R}}_m\}\Bigg]\Bigg\} \\
\overset{\text{(a)}}{\leq}\; & \boldsymbol{E}\left\{\frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\left[e^{-nT}\max_\theta\left(P_\theta(\boldsymbol{y}|\boldsymbol{X}_m)\cdot e^{n[\xi E_1(R,T,\theta)+T]}\right)\cdot 1\{\boldsymbol{y}|\hat{\mathcal{R}}_m^c\}+\right.\right. \\
& \left.\left.\left(\sum_{m'\neq m}\max_\theta[P_\theta(\boldsymbol{y}|\boldsymbol{X}_{m'})\cdot e^{n[\xi E_1(R,T,\theta)+T]}]\right)\cdot 1\{\boldsymbol{y}|\hat{\mathcal{R}}_m\}\right]\right\} \\
=\; & \boldsymbol{E}\left\{\frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\left[e^{-nT}f(\boldsymbol{X}_m,\boldsymbol{y})\cdot 1\{\boldsymbol{y}|\hat{\mathcal{R}}_m^c\}+\left(\sum_{m'\neq m}f(\boldsymbol{X}_{m'},\boldsymbol{y})\right)\cdot 1\{\boldsymbol{y}|\hat{\mathcal{R}}_m\}\right]\right\} \\
\overset{\text{(b)}}{\leq}\; & \boldsymbol{E}\left\{\frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\left[e^{-nT}f(\boldsymbol{X}_m,\boldsymbol{y})\cdot\min_{0\leq s\leq 1}\left(\frac{e^{nT}\sum_{m'\neq m}f(\boldsymbol{X}_{m'},\boldsymbol{y})}{f(\boldsymbol{X}_m,\boldsymbol{y})}\right)^s+\right.\right. \\
& \left.\left.\left(\sum_{m'\neq m}f(\boldsymbol{X}_{m'},\boldsymbol{y})\right)\cdot\min_{0\leq s\leq 1}\left(\frac{f(\boldsymbol{X}_m,\boldsymbol{y})}{e^{nT}\sum_{m'\neq m}f(\boldsymbol{X}_{m'},\boldsymbol{y})}\right)^{1-s}\right]\right\} \\
=\; & \boldsymbol{E}\left\{\frac{2}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\min_{0\leq s\leq 1}\left\{e^{-nT(1-s)}f^{1-s}(\boldsymbol{X}_m,\boldsymbol{y})\left(\sum_{m'\neq m}f(\boldsymbol{X}_{m'},\boldsymbol{y})\right)^s\right\}\right\} \\
=\; & \boldsymbol{E}\left\{\frac{2}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\min_{0\leq s\leq 1}\left\{e^{-nT(1-s)}\left(\max_{\theta\in\Theta_n}P_\theta(\boldsymbol{y}|\boldsymbol{X}_m)e^{n[\xi E_1(R,T,\theta)+T]}\right)^{1-s}\cdot\right.\right. \\
& \left.\left.\left(\sum_{m'\neq m}\max_{\tilde\theta\in\Theta_n}P_{\tilde\theta}(\boldsymbol{y}|\boldsymbol{X}_{m'})e^{n[\xi E_1(R,T,\tilde\theta)+T]}\right)^s\right\}\right\} \\
\overset{\text{(c)}}{\leq}\; & \boldsymbol{E}\left\{\frac{2}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\min_{0\leq s\leq 1}\left\{e^{-nT(1-s)}\left[\max_{\theta\in\Theta_n}P_\theta(\boldsymbol{y}|\boldsymbol{X}_m)e^{n[\xi E_1(R,T,\theta)+T]}\right]^{1-s}\times\right.\right. \\
& \left.\left.\left(\sum_{m'\neq m}\sum_{\tilde\theta\in\Theta_n}P_{\tilde\theta}(\boldsymbol{y}|\boldsymbol{X}_{m'})e^{n[\xi E_1(R,T,\tilde\theta)+T]}\right)^s\right\}\right\} \\
=\; & \boldsymbol{E}\left\{\frac{2}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\min_{0\leq s\leq 1}\left\{e^{-nT(1-s)}\left[\max_{\theta\in\Theta_n}P_\theta(\boldsymbol{y}|\boldsymbol{X}_m)e^{n[\xi E_1(R,T,\theta)+T]}\right]^{1-s}\times\right.\right. \\
& \left.\left.\left(\sum_{\tilde\theta\in\Theta_n}\sum_{m'\neq m}P_{\tilde\theta}(\boldsymbol{y}|\boldsymbol{X}_{m'})e^{n[\xi E_1(R,T,\tilde\theta)+T]}\right)^s\right\}\right\} \\
\overset{\text{(d)}}{\leq}\; & \boldsymbol{E}\left\{\frac{2}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\min_{0\leq s\leq 1}\left\{e^{-nT(1-s)}\left[\max_{\theta\in\Theta_n}P_\theta(\boldsymbol{y}|\boldsymbol{X}_m)e^{n[\xi E_1(R,T,\theta)+T]}\right]^{1-s}\times\right.\right. \\
& \left.\left.\left(|\Theta_n|\cdot\max_{\tilde\theta\in\Theta_n}\sum_{m'\neq m}P_{\tilde\theta}(\boldsymbol{y}|\boldsymbol{X}_{m'})e^{n[\xi E_1(R,T,\tilde\theta)+T]}\right)^s\right\}\right\} \\
\leq\; & \boldsymbol{E}\left\{\frac{2|\Theta_n|}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\min_{0\leq s\leq 1}\left\{e^{-nT(1-s)}\left[\max_{\theta\in\Theta_n}P_\theta(\boldsymbol{y}|\boldsymbol{X}_m)e^{n[\xi E_1(R,T,\theta)+T]}\right]^{1-s}\times\right.\right.
\end{aligned}
$$

$$\left(\max_{\tilde{\theta}\in\Theta_n}\sum_{m'\neq m}P_{\tilde{\theta}}(\boldsymbol{y}|\boldsymbol{X}_{m'})e^{n[\xi E_1(R,T,\tilde{\theta})+T]}\right)^s\Bigg\}\Bigg\}, \tag{A.3}$$

where (a) follows from the fact that the maximum (over θ) of a summation is upper bounded by the summation of the maxima, (b) follows from (A.1) and (A.2), and (c) and (d) follow from the fact that if $g(\theta)$ is non–negative then

$$\max_{\theta\in\Theta_n} g(\theta) \leq \sum_{\theta\in\Theta_n} g(\theta) \leq |\Theta_n|\cdot\max_{\theta\in\Theta_n} g(\theta). \tag{A.4}$$

Now, for every given $\boldsymbol{y}$ and $\{\boldsymbol{x}_m\}$, let $\theta^*\in\Theta_n$ be the achiever of

$$\max_{\theta\in\Theta_n}(P_\theta(\boldsymbol{y}|\boldsymbol{x}_m)e^{n[\xi E_1(R,T,\theta)+T]}),$$

and let $\theta^{**}\in\Theta_n$ be the achiever of

$$\max_{\tilde{\theta}\in\Theta_n}\sum_{m'\neq m}P_{\tilde{\theta}}(\boldsymbol{y}|\boldsymbol{x}_{m'})e^{n[\xi E_1(R,T,\tilde{\theta})+T]}.$$

Note that θ^* and θ^{**} depend on $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_M$ and $\boldsymbol{y}$, but not on the parameter s. Let us denote

$$\begin{aligned} W(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_M,\boldsymbol{y},\theta,\tilde{\theta}) &= \min_{0\leq s\leq 1}\Bigg\{e^{-nT(1-s)}\left[P_\theta(\boldsymbol{y}|\boldsymbol{x}_m)e^{n[\xi E_1(R,T,\theta)+T]}\right]^{1-s}\times \\ &\qquad \left(\sum_{m'\neq m}P_{\tilde{\theta}}(\boldsymbol{y}|\boldsymbol{x}_{m'})e^{n[\xi E_1(R,T,\tilde{\theta})+T]}\right)^s\Bigg\}. \end{aligned} \tag{A.5}$$

Now, obviously,

$$\begin{aligned} \bar{K}_n &\leq \boldsymbol{E}\left\{\frac{2|\Theta_n|}{M}\sum_{m=1}^M\sum_{\boldsymbol{y}\in\mathcal{Y}^n}W(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_M,\boldsymbol{y},\theta^*,\theta^{**})\right\} \\ &\leq \boldsymbol{E}\left\{\frac{2|\Theta_n|}{M}\sum_{m=1}^M\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\sum_{\theta\in\Theta_n}\sum_{\tilde{\theta}\in\Theta_n}W(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_M,\boldsymbol{y},\theta,\tilde{\theta})\right\} \\ &= \boldsymbol{E}\Bigg\{\frac{2|\Theta_n|}{M}\sum_{m=1}^M\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\sum_{\theta\in\Theta_n}\sum_{\tilde{\theta}\in\Theta_n}\min_{0\leq s\leq 1}\Bigg\{e^{-nT(1-s)}\left[P_\theta(\boldsymbol{y}|\boldsymbol{X}_m)e^{n[\xi E_1(R,T,\theta)+T]}\right]^{1-s}\times \\ &\qquad \left(\sum_{m'\neq m}P_{\tilde{\theta}}(\boldsymbol{y}|\boldsymbol{X}_{m'})e^{n[\xi E_1(R,T,\tilde{\theta})+T]}\right)^s\Bigg\}\Bigg\} \\ &= \frac{2|\Theta_n|}{M}\sum_{\theta\in\Theta_n}\sum_{\tilde{\theta}\in\Theta_n}\sum_{m=1}^M\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\boldsymbol{E}\min_{0\leq s\leq 1}\Bigg\{e^{-nT(1-s)}\left[P_\theta(\boldsymbol{y}|\boldsymbol{X}_m)e^{n[\xi E_1(R,T,\theta)+T]}\right]^{1-s}\times \\ &\qquad \left(\sum_{m'\neq m}P_{\tilde{\theta}}(\boldsymbol{y}|\boldsymbol{X}_{m'})e^{n[\xi E_1(R,T,\tilde{\theta})+T]}\right)^s\Bigg\} \end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{\le} \frac{2|\Theta_n|^3}{M}\max_{\theta\in\Theta_n}\max_{\tilde{\theta}\in\Theta_n}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\boldsymbol{E}\min_{0\le s\le 1}\left\{e^{-nT(1-s)}\left[P_\theta(\boldsymbol{y}|\boldsymbol{X}_m)e^{n[\xi E_1(R,T,\theta)+T]}\right]^{1-s}\times\right.\\
&\left.\left(\sum_{m'\neq m}P_{\tilde{\theta}}(\boldsymbol{y}|\boldsymbol{X}_{m'})e^{n[\xi E_1(R,T,\tilde{\theta})+T]}\right)^{s}\right\}\\
&\le \frac{2|\Theta_n|^3}{M}\max_{\theta\in\Theta}\max_{\tilde{\theta}\in\Theta}\min_{0\le s\le 1}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\boldsymbol{E}\left\{e^{-nT(1-s)}\left[P_\theta(\boldsymbol{y}|\boldsymbol{X}_m)e^{n[\xi E_1(R,T,\theta)+T]}\right]^{1-s}\times\right.\\
&\left.\left(\sum_{m'\neq m}P_{\tilde{\theta}}(\boldsymbol{y}|\boldsymbol{X}_{m'})e^{n[\xi E_1(R,T,\tilde{\theta})+T]}\right)^{s}\right\}\\
&= \frac{2|\Theta_n|^3}{M}\max_{\theta\in\Theta}\max_{\tilde{\theta}\in\Theta}\min_{0\le s\le 1}e^{n[\xi\{(1-s)E_1(R,T,\theta)+sE_1(R,T,\tilde{\theta})\}+sT]}\times\\
&\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\boldsymbol{E}\left\{P_\theta^{1-s}(\boldsymbol{y}|\boldsymbol{X}_m)\cdot\left(\sum_{m'\neq m}P_{\tilde{\theta}}(\boldsymbol{y}|\boldsymbol{X}_{m'})\right)^{s}\right\}, \qquad \text{(A.6)}
\end{aligned}
$$

where in (a) we used again (A.4). Assuming that the codewords are drawn independently, we then have:

$$
\begin{aligned}
\bar{K}_n &\le \frac{2|\Theta_n|^3}{M}\max_{\theta\in\Theta}\max_{\tilde{\theta}\in\Theta}\min_{0\le s\le 1}e^{n[\xi\{(1-s)E_1(R,T,\theta)+sE_1(R,T,\tilde{\theta})\}+sT]}\times\\
&\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\boldsymbol{E}\{P_\theta^{1-s}(\boldsymbol{y}|\boldsymbol{X}_m)\}\cdot\boldsymbol{E}\left\{\left(\sum_{m'\neq m}P_{\tilde{\theta}}(\boldsymbol{y}|\boldsymbol{X}_{m'})\right)^{s}\right\}\\
&= \frac{2|\Theta_n|^3}{M}\max_{\theta\in\Theta}\max_{\tilde{\theta}\in\Theta}\min_{0\le s\le 1}e^{n[\xi\{(1-s)E_1(R,T,\theta)+sE_1(R,T,\tilde{\theta})\}+sT]}\times\\
&\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\boldsymbol{E}\{P_\theta^{1-s}(\boldsymbol{y}|\boldsymbol{X}_m)\}\cdot\min_{s\le\rho\le 1}\boldsymbol{E}\left\{\left(\left[\sum_{m'\neq m}P_{\tilde{\theta}}(\boldsymbol{y}|\boldsymbol{X}_{m'})\right]^{s/\rho}\right)^{\rho}\right\}\\
&\le \frac{2|\Theta_n|^3}{M}\max_{\theta\in\Theta}\max_{\tilde{\theta}\in\Theta}\min_{0\le s\le 1}e^{n[\xi\{(1-s)E_1(R,T,\theta)+sE_1(R,T,\tilde{\theta})\}+sT]}\times\\
&\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\boldsymbol{E}\{P_\theta^{1-s}(\boldsymbol{y}|\boldsymbol{X}_m)\}\cdot\min_{0\le s\le\rho\le 1}\boldsymbol{E}\left[\left(\sum_{m'\neq m}P_{\tilde{\theta}}^{s/\rho}(\boldsymbol{y}|\boldsymbol{X}_{m'})\right)^{\rho}\right\}\\
&\le \frac{2|\Theta_n|^3}{M}\max_{\theta\in\Theta}\max_{\tilde{\theta}\in\Theta}\min_{0\le s\le\rho\le 1}e^{n[\xi\{(1-s)E_1(R,T,\theta)+sE_1(R,T,\tilde{\theta})\}+sT]}\times\\
&\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\boldsymbol{E}\{P_\theta^{1-s}(\boldsymbol{y}|\boldsymbol{X}_m)\}\cdot\left(\sum_{m'\neq m}\boldsymbol{E}\{P_{\tilde{\theta}}^{s/\rho}(\boldsymbol{y}|\boldsymbol{X}_{m'})\}\right)^{\rho}, \qquad \text{(A.7)}
\end{aligned}
$$

where in the last step we have used Jensen's inequality. Now, observe that the summands do not depend on m, therefore, the effects of the summation over m and the factor of $1/M$ cancel each other. Also, the sum of $M-1$ contributions of identical expectations $\boldsymbol{E}\{P_{\tilde{\theta}}^{s/\rho}(\boldsymbol{y}|\boldsymbol{X}_{m'})\}$ create a factor of $M-1$ (upper bounded by M) raised to the power of ρ.

Denoting

$$U(\boldsymbol{y},\lambda,\theta)=\boldsymbol{E}\{P_\theta^\lambda(\boldsymbol{y}|\boldsymbol{X})\},$$

we have:

$$\begin{aligned}\bar{K}_n &\le 2|\Theta_n|^3\max_{\theta\in\Theta}\max_{\tilde{\theta}\in\Theta}\min_{0\le s\le\rho\le 1} M^\rho\cdot e^{n[\xi\{(1-s)E_1(R,T,\theta)+sE_1(R,T,\tilde{\theta})\}+sT]}\times\\ &\sum_{\boldsymbol{y}\in\mathcal{Y}^n}U(\boldsymbol{y},1-s,\theta)\cdot U^\rho(\boldsymbol{y},s/\rho,\tilde{\theta}). \qquad\text{(A.8)}\end{aligned}$$

To assess the exponential order of $U(\boldsymbol{y},\lambda,\theta)$, we use the method of types [4] as well as the assumption that the sequence of random coding distributions belongs to the class $\mathcal{Q}$. Let $\hat{P}_{\boldsymbol{y}}$ denote the empirical distribution of $\boldsymbol{y}$, and let $T_{\boldsymbol{y}}$ denote the type class of $\boldsymbol{y}$, i.e., the set of $\boldsymbol{y}'$ with $\hat{P}_{\boldsymbol{y}'}=\hat{P}_{\boldsymbol{y}}$. Let $\hat{H}_{\boldsymbol{y}}(Y)$ denote the corresponding empirical entropy of Y. Similarly, let $\hat{P}_{\boldsymbol{xy}}$ denote the empirical joint distribution of $(\boldsymbol{x},\boldsymbol{y})$ and let $\hat{\boldsymbol{E}}_{\boldsymbol{xy}}\{\cdot\}$ denote the corresponding empirical expectation, i.e., the expectation w.r.t. $\hat{P}_{\boldsymbol{xy}}$. Also, let $T_{\boldsymbol{x}|\boldsymbol{y}}$ denote the conditional type class of $\boldsymbol{x}$ given $\boldsymbol{y}$, i.e., the set of $\boldsymbol{x}'$ with $\hat{P}_{\boldsymbol{x}'\boldsymbol{y}}=\hat{P}_{\boldsymbol{xy}}$ and let $\hat{I}_{\boldsymbol{xy}}(X;Y)$ denote the corresponding empirical mutual information between X and Y. Then,

$$\begin{aligned}U(\boldsymbol{y},\lambda,\theta) &= \sum_{\boldsymbol{x}\in\mathcal{X}^n}Q_n(\boldsymbol{x})P_\theta^\lambda(\boldsymbol{y}|\boldsymbol{x})\\ &= \sum_{T_{\boldsymbol{x}|\boldsymbol{y}}\subset\mathcal{X}^n}|T_{\boldsymbol{x}|\boldsymbol{y}}|\cdot\frac{e^{-n\Delta_n^*(\hat{P}_{\boldsymbol{x}})}}{|T_{\boldsymbol{x}}|}\cdot e^{\lambda n\hat{\boldsymbol{E}}_{\boldsymbol{xy}}\ln P_\theta(Y|X)}\\ &\le (n+1)^{|\mathcal{X}|-1}\sum_{T_{\boldsymbol{x}|\boldsymbol{y}}\subset\mathcal{X}^n}e^{-n\hat{I}_{\boldsymbol{xy}}(X;Y)}\cdot e^{-n[\Delta^*(\hat{P}_{\boldsymbol{x}})-\epsilon_n]}\cdot e^{\lambda n\hat{\boldsymbol{E}}_{\boldsymbol{xy}}\ln P_\theta(Y|X)}\\ &\le (n+1)^{(|\mathcal{Y}|+1)\cdot(|\mathcal{X}|-1)}\cdot e^{-n[F(\hat{P}_{\boldsymbol{y}},\lambda,\theta)-\epsilon_n]}, \qquad\text{(A.9)}\end{aligned}$$

where $\epsilon_n\to 0$ independently of $\hat{P}_{\boldsymbol{x}}$ by the uniform convergence assumption that defines the class $\mathcal{Q}$, and where $F(P_{\boldsymbol{y}},\lambda,\theta)$ is defined as in eq. (23). On substituting this bound into the upper bound on $K_n(\hat{\mathcal{R}})$, we get:

$$\begin{aligned}\bar{K}_n &\le 2|\Theta_n|^3(n+1)^{2|\mathcal{Y}|\cdot(|\mathcal{X}|-1)}\max_{\theta\in\Theta}\max_{\tilde{\theta}\in\Theta}\min_{0\le s\le\rho\le 1}\\ &\quad M^\rho\cdot e^{n[\xi\{(1-s)E_1(R,T,\theta)+sE_1(R,T,\tilde{\theta})\}+sT]}\cdot\sum_{\boldsymbol{y}\in\mathcal{Y}^n}e^{-n[F(P_{\boldsymbol{y}},1-s,\theta)+\rho F(P_{\boldsymbol{y}},s/\rho,\tilde{\theta})]}\\ &\le 2|\Theta_n|^3(n+1)^{2|\mathcal{Y}|\cdot(|\mathcal{X}|-1)}\max_{\theta\in\Theta}\max_{\tilde{\theta}\in\Theta}\min_{0\le s\le\rho\le 1}\\ &\quad M^\rho\cdot e^{n[\xi\{(1-s)E_1(R,T,\theta)+sE_1(R,T,\tilde{\theta})\}+sT]}\cdot\sum_{T_{\boldsymbol{y}}\subset\mathcal{Y}^n}e^{n\hat{H}_{\boldsymbol{y}}(Y)}\cdot e^{-n[F(P_{\boldsymbol{y}},1-s,\theta)+\rho F(P_{\boldsymbol{y}},s/\rho,\tilde{\theta})]}\end{aligned}$$

$$\begin{aligned}&\leq 2|\Theta_n|^3(n+1)^{3|\mathcal{Y}|\cdot(|\mathcal{X}|-1)}\max_{\theta\in\Theta}\max_{\tilde{\theta}\in\Theta}\min_{0\leq s\leq\rho\leq 1}\\&\quad M^\rho\cdot e^{n[\xi\{(1-s)E_1(R,T,\theta)+sE_1(R,T,\tilde{\theta})\}+sT]}\cdot e^{-n\min_{P_y}[F(P_y,1-s,\theta)+\rho F(P_y,s/\rho,\tilde{\theta})-H(Y)]}\\&\leq 2|\Theta_n|^3(n+1)^{3|\mathcal{Y}|\cdot(|\mathcal{X}|-1)}\max_{\theta\in\Theta}\max_{\tilde{\theta}\in\Theta}\min_{0\leq s\leq\rho\leq 1}\\&\quad M^\rho\cdot e^{n[\xi\{(1-s)E_1(R,T,\theta)+sE_1(R,T,\tilde{\theta})\}+sT]}\cdot e^{-nE(\theta,\tilde{\theta},s,\rho)}.\end{aligned}\tag{A.10}$$

We would like to find the maximum value of ξ such that $\bar{K}_n$ would be guaranteed not to grow exponentially. To this end, we can now ignore the factor $2|\Theta_n|^3(n+1)^{3|\mathcal{Y}|\cdot(|\mathcal{X}|-1)}$, which is polynomial in n (cf. eq. (19)). Thus, the latter upper bound will be sub–exponential in n as long as

$$\min_{\theta,\tilde{\theta}}\max_{0\leq s\leq\rho\leq 1}[E(\theta,\tilde{\theta},s,\rho)-\xi\{(1-s)E_1(R,T,\theta)+sE_1(R,T,\tilde{\theta})\}-\rho R-sT]\geq 0,\tag{A.11}$$

or, equivalently, for every $(\theta,\tilde{\theta})$, there exist (ρ,s), $0\leq s\leq\rho\leq 1$, such that

$$E(\theta,\tilde{\theta},s,\rho)\geq\xi\{(1-s)E_1(R,T,\theta)+sE_1(R,T,\tilde{\theta})\}+\rho R+sT,\tag{A.12}$$

i.e.,

$$\xi\leq\frac{E(\theta,\tilde{\theta},s,\rho)-\rho R-sT}{(1-s)E_1(R,T,\theta)+sE_1(R,T,\tilde{\theta})}.\tag{A.13}$$

In other words, for every $\xi\leq\xi^*(R,T)$, where $\xi^*(R,T)$ is defined as in eq. (25)), $K_n(\hat{\mathcal{R}})$ is guaranteed not to grow exponentially with n. This completes the proof of Theorem 2.

Acknowledgement

We would like to thank the Associate Editor, Gerhard Kramer, and the three anonymous reviewers for their very useful comments, which certainly helped to improve the presentation of our results.

References

[1] R. Ahlswede, N. Cai, and Z. Zhang, "Erasure, list, and detection zero–error capacities for low noise and a relation to identification," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 55–62, January 1996.

[2] J. Byers, M. Luby, and M. Mitzenmacher, "A digital fountain approach to asynchronous reliable multicast," *IEEE Journal on Selected Areas in Communications*, vol. 20 no. 8, pp. 1528–1540, October 2002.

[3] J. Byers, M. Luby, M. Mitzenmacher and A. Rege, “A digital fountain approach to reliable distribution of bulk data”, *Proc. ACM SIGCOMM ‘98*, pp. 56–67, Vancouver, Canada, September 1998.

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

[5] S. Draper, B. J. Frey, and F. R. Kschischang, “Rateless coding for non–ergodic channels with decoder channel state information,” submitted to *IEEE Trans. Inform. Theory.*

[6] S. C. Draper, B. J. Frey, and F. R. Kschischang, “Efficient variable length channel coding for unknown DMCs,” *Proc. ISIT 2004*, p. 377, Chicago, U.S.A., June–July, 2004.

[7] U. Erez, G. W. Wornell, M. D. Trott, “Rateless space–time coding,” *Proc. ISIT 2005*, pp. 1937–1941, Adelaide, Australia, September 2005.

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

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

[10] G. D. Forney, Jr., “Exponential error bounds for erasure, list, and decision feedback schemes,” *IEEE Trans. Inform. Theory*, vol. IT–14, no. 2, pp. 206–220, March 1968.

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

[12] T. Hashimoto, “Composite scheme LT+Th for decoding with erasures and its effective equivalence to Forney’s rule,” *IEEE Trans. Inform. Theory*, vol. 45, no. 1, pp. 78–93, January 1999.

[13] T. Hashimoto and M. Taguchi, “Performance and explicit error detection and threshold decision in decoding with erasures,” *IEEE Trans. Inform. Theory*, vol. 43, no. 5, pp. 1650–1655, September 1997.

[14] J. Jiang and K. R. Narayanan, "Multilevel coding for channels with non-uniform inputs and rateless transmission over the BSC," arXiv:cs.IT/0601083, January 18, 2006.

[15] M. Luby, M. Mitzenmacher, A. Shokrollahi, and D. Spielman, "Efficient erasure correction codes," *IEEE Trans. Inform. Theory*, vol. 47, no. 2, pp. 569–584, February 2001.

[16] J. L. Massey, "Causality, feedback and directed information," *Proc. 1990 International Symposium on Information Theory and its Applications (ISITA '90)*, pp. 303–305, 1990.

[17] N. Shulman, "Communication over an unknown channel via common broadcasting," Ph.D. dissertation, Tel Aviv University, 2003.

[18] N. Shulman and M. Feder, "Static broadcasting," *Proc. ISIT 2000*, p. 23, Sorrento, Italy, June 2000.

[19] N. Shulman and M. Feder, "The uniform disitribution as a universal prior," *IEEE Trans. Inform. Theory*, vol. 50, no. 6, pp. 1356–1362, June 2004.

[20] A. Tchamkerten and E. I. Telatar, "Variable length codes over unknown channels," preprint 2004.

[21] A. J. Viterbi, "Error bounds for the white Gaussian and other very noisy memoryless channels with generalized decision regions," *IEEE Trans. Inform. Theory*, vol. IT–15, no. 2, pp. 279–287, March 1969.

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