Erasure/List Random Coding Error Exponents Are Not Universally Achievable*

Wasim Huleihel Nir Weinberger Neri Merhav
Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, ISRAEL
E-mail: {wh@tx, nirwein@tx, merhav@ee}.technion.ac.il

Abstract

We study the problem of universal decoding for unknown discrete memoryless channels in the presence of erasure/list option at the decoder, in the random coding regime. Specifically, we harness a universal version of Forney's classical erasure/list decoder developed in earlier studies, which is based on the competitive minimax methodology, and guarantees universal achievability of a certain fraction of the optimum random coding error exponents. In this paper, we derive an exact single-letter expression for the maximum achievable fraction. Examples are given in which the maximal achievable fraction is strictly less than unity, which imply that, in general, there is no universal erasure/list decoder which achieves the same random coding error exponents as the optimal decoder for a known channel. This is in contrast to the situation in ordinary decoding (without the erasure/list option), where optimum exponents are universally achievable, as is well known. It is also demonstrated that previous lower bounds derived for the maximal achievable fraction are not tight in general.

Index Terms

Universal decoding, error exponents, erasure/list decoding, maximum-likelihood decoding, random coding, generalized likelihood ratio test, channel uncertainty, competitive minimax.

I. Introduction

In many practical situations encountered in coded communication systems, the prevalent channel over which transmission takes place is unknown to the receiver. Typically, the optimal maximum likelihood

*This research was partially supported by The Israeli Science Foundation (ISF), grant no. 412/12.

(ML) decoder depends on the channel statistics, and therefore its usage is precluded. In such cases, universal decoders are sought which do not require knowledge of the actual channel present, but still preform well just as if the channel was known to the decoder. The design of such universal decoders was extensively addressed for ordinary decoding (without the erasure/list option), see, e.g., [1-7], and references therein. For example, for unknown discrete memoryless channels (DMCs), the maximum mutual information (MMI) decoder [1] is asymptotically optimal for ordinary decoding, in the sense that it achieves the same random coding error exponents as the ML decoder. However, for decoders with an erasure/list option, only partial results exist.

In this paper, we focus on universal erasure/list decoders proposed and analyzed by Forney for known channels [8]. Erasure/list decoding is especially attractive for unknown channels, since communicating 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 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. A possible approach to handle the problem described above is the *rateless coding* methodology, see, for example [9-14], in which at every time instant the decoder either makes a decision on one of the transmitted messages or decides to request an additional symbol via the feedback line. The latter case can be considered as an "erasure" event for the decoder, and so universal erasure decoders are required (see discussion in [15]).

In [4, Chapter 10, Theorem 10.11], Csiszár and Körner proposed a family of universal erasure decoders, parametrized by some real parameter, for DMCs, and analyzed the resulting error exponents. While this family is in the spirit of the MMI decoder, it does not achieve the same exponents as Forney's optimal erasure/list decoder. More recently, in [16], Moulin has generalized this family of decoders and proposed a family of decoders parametrized by a weighting function. Upon optimization of the weighting function within some class of possible functions, a few cases were identified in which the universal decoder achieves the same error exponents as if the channel was known.

In [15], Merhav and Feder studied the problem in a more systematic manner. Specifically, they considered the problem of universal decoding with an erasure/list option for the class of DMCs indexed by an unknown parameter θ. They invoked the competitive minimax methodology proposed in [17], in order to derive a universal version of Forney's classical erasure/list decoder. Recall that 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 total error 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, $\mathcal{E}_2$, for erasure decoder (or, average list size for list decoder), in the random coding regime. The universal erasure/list decoder of [15] guarantees achievability of an 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 for erasure decoder (or, average list size for list decoder) $\hat{E}_2(R,T,\theta) \geq \xi \cdot \hat{E}_1(R,T,\theta) + T$ for all θ. 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. It should be remarked, that the benchmark exponents in [15] were the classical lower bounds on $E_1(R,T,\theta)$ and $E_2(R,T,\theta)$ derived by Forney [8].

Clearly, to maximize the guaranteed exponents obtained by the universal decoder of [15], the maximal $0 \leq \xi \leq 1$ such that the above holds is of interest. This maximal fraction is the central quantity of this paper and will be denoted henceforth by $\xi^*(R,T)$. If, for example, $\xi^*(R,T)$ is strictly less than unity, then it means that there is a major difference between universal ordinary decoding and universal erasure/list decoding: while for the former, it is well known that optimum random coding error exponents are universally achievable (at least for some classes of channels and certain random coding distributions), in the latter, when the erasure/list options are available, this may no longer be the case. In [15], Merhav and Feder invoked Gallager's bounding techniques to analyze the exponential behavior of upper bounds on the probabilities $\mathcal{E}_1$ and $\mathcal{E}_2$. Accordingly, a single-letter expression for a lower bound to $\xi^*(R,T)$ was obtained, which we denote henceforth by $\xi_L(R,T)$. Since $\xi_L(R,T)$ was merely a lower bound, the question of achievability of Forney's erasure/list exponents was not fully settled in [15].

As was previously mentioned, even for a known channel, only lower bounds for the exponents were obtained by Forney [8]. More recently, inspired by a statistical-mechanical point of view on random code ensembles, Somekh-Baruch and Merhav [18] have found *exact* expressions for the exponents of the optimal erasure/list decoder, by assessing the moments of certain type class enumerators. In this paper, we tackle again the problem of erasure/list channel decoding using similar methods, and derive an *exact* expression for $\xi^*(R,T)$ with respect to the exact erasure/list exponents of a known channels found in [18]. This exact expression leads to the following conclusions:

1) In general, $\xi^*(R,T)$ is strictly less than 1. Therefore, the known channel exponents in erasure/list decoding cannot be achieved by any universal decoder. In this sense, channel knowledge is crucial for asymptotically optimum erasure/list decoding. This is in sharp contrast to the situation in ordinary decoding (without the erasure/list option), where, as said, optimum exponents are universally achievable, e.g., by the MMI decoder.

2) In general, $\xi_L(R,T)$ is strictly less than $\xi^*(R,T)$. Therefore, the Gallager-style analysis technique in [15] is not always powerful enough to obtain $\xi^*(R,T)$.

The outline of the rest of the paper is as follows. In Section II, we establish notation conventions, and in Section III we detail necessary background on erasure/list decoding, both for known and unknown channels. Then, in Section IV, we present our main result of an exact expression for $\xi^*(R,T)$, and discuss the special case of binary symmetric channel (BSC). Finally, we shed light on the differences between $\xi^*(R,T)$ and $\xi_L(R,T)$, along with some numerical results, which illustrate the main result of this paper. Finally, in Section V, we provide proofs for all our results.

II. Notation Conventions

Throughout this paper, scalar random variables (RVs) 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, e.g. X, x, and $\mathcal{X}$, respectively. A similar convention will apply to random vectors of dimension n and their sample values, which will be denoted with the same symbols in the boldface font. The set of all n-vectors with components taking values in a certain finite alphabet, will be denoted as the same alphabet superscripted by n, e.g., $\mathcal{X}^n$. Generic channels will be usually denoted by the letters P, Q, or W. We shall mainly consider joint distributions of two RVs (X,Y) over the Cartesian product of two finite alphabets $\mathcal{X}$ and $\mathcal{Y}$. For brevity, we will denote any joint distribution, e.g. Q_{XY}, simply by Q, the marginals will be denoted by Q_X and Q_Y, and the conditional distributions will be denoted by $Q_{X|Y}$ and $Q_{Y|X}$. The joint distribution induced by Q_X and $Q_{Y|X}$ will be denoted by $Q_X \times Q_{Y|X}$, and a similar notation will be used when the roles of X and Y are switched.

The expectation operator will be denoted by $\mathbb{E}\{\cdot\}$, and when we wish to make the dependence on the underlying distribution Q clear, we denote it by $\mathbb{E}_Q\{\cdot\}$. The entropy of X and the conditional entropy of X given Y, will be denoted $H_X(Q)$, $H_{X|Y}(Q)$, respectively, where Q is the underlying probability distribution. The mutual information of the joint distribution Q will be denoted by $I(Q)$. The divergence (or, Kullback-Liebler distance) between two probability measures Q and P will be denoted by $D(Q||P)$. For two numbers $0 \leq q,p \leq 1$, $D(q||p)$ will stand for the divergence between the binary measures $\{q, 1-q\}$ and $\{p, 1-p\}$.

For a given vector $\boldsymbol{x}$, let $\hat{Q}_{\boldsymbol{x}}$ denote the empirical distribution, that is, the vector $\{\hat{Q}_{\boldsymbol{x}}(x),\ x \in \mathcal{X}\}$, where $\hat{Q}_{\boldsymbol{x}}(x)$ is the relative frequency of the letter x in the vector $\boldsymbol{x}$. Let $\mathcal{T}_P$ denote the type class associated with P, that is, the set of all sequences $\boldsymbol{x}$ for which $\hat{Q}_{\boldsymbol{x}} = P$. Similarly, for a pair of vectors

$(\boldsymbol{x}, \boldsymbol{y})$, the empirical joint distribution will be denoted by $\hat{Q}_{\boldsymbol{xy}}$, or simply by $\hat{Q}$, for short. All the previously defined notations for regular distributions will also be used for empirical distributions.

The cardinality of a finite set $\mathcal{A}$ will be denoted by $|\mathcal{A}|$, its complement will be denoted by $\mathcal{A}^c$. The probability of an event $\mathcal{E}$ will be denoted by $\Pr\{\mathcal{E}\}$. The indicator function of an event $\mathcal{E}$ will be denoted by $\mathcal{I}\{\mathcal{E}\}$. For two sequences of positive numbers, $\{a_n\}$ and $\{b_n\}$, the notation $a_n \doteq b_n$ means that $\{a_n\}$ and $\{b_n\}$ are of the same exponential order, i.e., $n^{-1}\log a_n/b_n \to 0$ as $n \to \infty$, where in this paper, logarithms are defined with respect to (w.r.t.) the natural basis, that is, $\log(\cdot) \equiv \ln(\cdot)$. Finally, for a real number x, we let $|x|^+ \triangleq \max\{0, x\}$.

III. Model Formulation and Short Background

A. Known Channel

Consider a DMC with a finite input alphabet $\mathcal{X}$, finite output alphabet $\mathcal{Y}$, and a matrix of single-letter transition probabilities $\{W(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$. A rate-$R$ codebook consists of $M = \lceil e^{nR} \rceil$ length-n codewords $\boldsymbol{x}_m \in \mathcal{X}^n$, $m = 1, 2, \ldots, M$, representing the M messages. It will be assumed that all messages are a-priori equiprobable. We assume the ensemble of fixed composition random codes of blocklength n, where each codeword is selected at random, uniformly within a type class $\mathcal{T}(P_X)$ for some given random coding distribution P_X over the alphabet $\mathcal{X}$.

In the following, we give a short description on the operation of the erasure decoder and then the list decoder. A decoder with an erasure option is a partition of the observation space $\mathcal{Y}^n$ into $(M+1)$ regions, denoted by $\{\mathcal{R}_m\}_{m=0}^M$. An erasure decoder works as follows: If $\boldsymbol{y} \in \mathcal{Y}^n$ falls into the mth region, $\mathcal{R}_m$, for $m = 1, 2, \ldots, M$, then a decision is made in favor of message number m. If $\boldsymbol{y} \in \mathcal{R}_0$, then no decision is made and an erasure is declared. Accordingly, we shall refer to $\boldsymbol{y} \in \mathcal{R}_0$ as an *erasure event*. Given a code $\mathcal{C} \triangleq \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ and a decoder $\mathcal{R} \triangleq (\mathcal{R}_0, \ldots, \mathcal{R}_M)$, we define two error events. The event $\mathcal{E}_1$ is the event of deciding on erroneous codeword or making an erasure, and the event $\mathcal{E}_2$ which is the undetected error event, namely, the event of deciding on erroneous codeword. It is evident that $\mathcal{E}_1$ is the disjoint union of the erasure event and $\mathcal{E}_2$. The probabilities of all the aforementioned events are given by:

$$\Pr\{\mathcal{E}_1\} = \frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{R}_m^c} W(\boldsymbol{y}|\boldsymbol{x}_m), \tag{1}$$

$$\Pr\{\mathcal{E}_2\} = \frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{R}_m}\sum_{m'\neq m} W(\boldsymbol{y}|\boldsymbol{x}_{m'}), \tag{2}$$

and

$$\Pr\{\mathcal{R}_0\} = \Pr\{\mathcal{E}_1\} - \Pr\{\mathcal{E}_2\}. \tag{3}$$

A list decoder is a mapping from the space of received vectors $\mathcal{Y}^n$ into a collection of the subsets of $\{1,\ldots,M\}$. Alternatively, a list decoder is uniquely defined by a set of $M+1$ (not necessarily disjoint) decoding regions $\{\mathcal{R}_m\}_{m=0}^{M}$ such that $\mathcal{R}_m \subseteq \mathcal{Y}^n$ and $\mathcal{R}_0 = \mathcal{Y}^n \backslash \bigcup_{m=1}^{M} \mathcal{R}_m$. Given a received vector $\boldsymbol{y}$, the mth codeword belongs to the output list if $\boldsymbol{y} \in \mathcal{R}_m$, and if $\boldsymbol{y}$ does not belong to any of the regions $\mathcal{R}_m$ then $\boldsymbol{y} \in \mathcal{R}_0$, and an erasure is declared. The average error probability of a list decoder and a codebook $\mathcal{C}$ is the probability that the actual transmitted codeword does not belong to the output list, and it is defined similarly to (1). The average list size is the expected (w.r.t. the output of the channel) number of erroneous codewords in the output list, and defined similarly to (2).

Since the error events for the erasure and list decoders are defined in the same way, they can be treated on the same footing. Nonetheless, for descriptive purposes, we will refer to the erasure decoder, but we emphasize that all the following analysis and results are true also for the list decoder. When knowledge on the specific DMC is available at the decoder, Forney have shown [8], using Neyman-Pearson methodology, that the optimal tradeoff between $\Pr\{\mathcal{E}_1\}$ and $\Pr\{\mathcal{E}_2\}$ is attained by the decision regions $\mathcal{R}^* \triangleq (\mathcal{R}_0^*,\ldots,\mathcal{R}_M^*)$ given by:

$$\mathcal{R}_m^* \triangleq \left\{ \boldsymbol{y}:\ \frac{W(\boldsymbol{y}|\boldsymbol{x}_m)}{\sum_{m'\neq m} W(\boldsymbol{y}|\boldsymbol{x}_{m'})} \geq e^{nT} \right\},\ m = 1,2,\ldots M, \tag{4}$$

and

$$\mathcal{R}_0^* \triangleq \bigcap_{m=1}^{M} (\mathcal{R}_m^*)^c, \tag{5}$$

where T is a parameter, henceforth refereed as the *threshold*, which controls the balance between the probabilities of $\mathcal{E}_1$ and $\mathcal{E}_2$. When $T \geq 0$ the decoder operates in the erasure mode, and when it is in the list mode then $T < 0$. No other decision rule gives both a lower $\Pr\{\mathcal{E}_1\}$ and a lower $\Pr\{\mathcal{E}_2\}$ than the above choice. Finally, we define the error exponents $E_i(R,T)$, $i = 1,2$, as the exponents of the average probabilities of errors $\overline{\Pr}\{\mathcal{E}_i\}$ (associated with the optimal decoder $\mathcal{R}^*$), where the average is taken w.r.t. a given ensemble of the randomly selected codes, that is,

$$E_i(R,T) \triangleq -\liminf_{n\to\infty} \frac{1}{n} \log \overline{\Pr}\{\mathcal{E}_i\},\ i = 1,2. \tag{6}$$

An important observation is that Forney's decision rule for known DMCs can also be obtained by formulating the following optimization problem: Find a decoder $\mathcal{R}$ that minimizes $\Gamma(\mathcal{C},\mathcal{R})$ where

$$\Gamma(\mathcal{C},\mathcal{R}) \triangleq \Pr\{\mathcal{E}_2\} + e^{-nT} \Pr\{\mathcal{E}_1\} \tag{7}$$

$$= \frac{1}{M}\sum_{m=1}^{M}\left[\sum_{\boldsymbol{y}\in\mathcal{R}_m}\sum_{m'\neq m} W(\boldsymbol{y}|\boldsymbol{x}_{m'}) + \sum_{\boldsymbol{y}\in\mathcal{R}_m^c} e^{-nT}W(\boldsymbol{y}|\boldsymbol{x}_m)\right] \tag{8}$$

for a given codebook $\mathcal{C}$ and a given threshold T. Indeed, noting that (8) can be rewritten as

$$\Gamma(\mathcal{C},\mathcal{R}) = \sum_{\boldsymbol{y}\in\mathcal{Y}^n}\frac{1}{M}\sum_{m=1}^{M}\left[\sum_{m'\neq m} W(\boldsymbol{y}|\boldsymbol{x}_{m'})\mathcal{I}\{\boldsymbol{y}\in\mathcal{R}_m\} + e^{-nT}W(\boldsymbol{y}|\boldsymbol{x}_m)\mathcal{I}\{\boldsymbol{y}\in\mathcal{R}_m^c\}\right], \tag{9}$$

it is evident that for each m, the bracketed expression is minimized by $\mathcal{R}_m^*$ as defined above. By taking the ensemble average, we have

$$\mathbb{E}\{\Gamma(\mathcal{C},\mathcal{R}^*)\} \triangleq \overline{\Pr}\{\mathcal{E}_2\} + e^{-nT}\overline{\Pr}\{\mathcal{E}_1\}. \tag{10}$$

In [18], it was stated (without a proof) that, in the exponential scale, there is a balance between the two terms at the right hand side of (10), namely, the exponent of $\overline{\Pr}\{\mathcal{E}_2\}$ equals to the exponent of $e^{-nT}\overline{\Pr}\{\mathcal{E}_1\}$, for the optimal decoder $\mathcal{R}^*$. We rigorously assert this property in the following lemma. The proof appears in Appendix A.

Lemma 1 For all R and T, the optimal decoder $\mathcal{R}^*$ satisfies:

$$E_2(R,T) = T + E_1(R,T). \tag{11}$$

The significance of Lemma 1 is attributed to the fact that now we only need to assess the exponential behavior of either $\overline{\Pr}\{\mathcal{E}_1\}$, or, $\overline{\Pr}\{\mathcal{E}_2\}$, but not both. As was mentioned in the Introduction, in [18], Somekh-Baruch and Merhav have obtained exact single-letter formulas for the error exponents $E_1(R,T)$ and $E_2(R,T)$ associated with $\overline{\Pr}\{\mathcal{E}_1\}$ and $\overline{\Pr}\{\mathcal{E}_2\}$, respectively. Specifically, they show, that for the ensemble of fixed composition codes [18, Theorem 1][1,2]:

$$E_1(R,T) = \min\{E_a(R,T), E_b(R,T)\}, \tag{12}$$

where

$$E_a(R,T) \triangleq \min_{(Q,\tilde{Q})\in\hat{\mathcal{Q}}}\Big[D(\tilde{Q}||P_X\times W) + \ I(Q) - R\Big] \tag{13}$$

[1]In [18], each codeword in the codebook was drawn independently of all other codewords, and its symbols were drawn from an independent and identically (i.i.d.) distribution (identical for all the codewords). Nonetheless, the modification to the ensemble of fixed composition codes is straightforward.

[2]We note that there is an error at the end of the proof of Theorem 1 in [18], where it was claimed that $\min\{E_a(R,T), E_b(R,T)\} = E_a(R,T)$, which may not be true in general. The correct expression is as in (12).

and

$$E_b(R,T) \triangleq \min_{\tilde{Q}\in\hat{\mathcal{L}}} D(\tilde{Q}||P_X \times W) \tag{14}$$

where $\tilde{Q}$ is a probability distribution on $\mathcal{X}\times\mathcal{Y}$, and

$$\hat{\mathcal{Q}} \triangleq \left\{(Q,\tilde{Q}) \in \mathcal{D}:\ I(Q) \geq R,\ \hat{\Omega}(Q,\tilde{Q}) \leq 0\right\}, \tag{15}$$

$$\mathcal{D} \triangleq \left\{(Q,\tilde{Q}):\ Q_X = \tilde{Q}_X = P_X,\ Q_Y = \tilde{Q}_Y\right\}, \tag{16}$$

$$\hat{\Omega}(Q,\tilde{Q}) \triangleq \mathbb{E}_{\tilde{Q}} \log W(Y|X) - \mathbb{E}_Q \log W(Y|X) - T, \tag{17}$$

$$\tag{18}$$

and

$$\hat{\mathcal{L}} \triangleq \left\{\tilde{Q}:\ \mathbb{E}_{\tilde{Q}} \log W(Y|X) \leq R + T + \max_{Q:(Q,\tilde{Q})\in\mathcal{D}:\ I(Q)\leq R} \left[\mathbb{E}_Q \log W(Y|X) - I(Q)\right]\right\}. \tag{19}$$

As a special case, we shall consider in the sequel the problem of universal erasure/list decoding for the BSC, and to this end, we will use the exact expression of $E_1(R,T)$. Accordingly, for the BSC with crossover probability θ, it was shown that [18, Corollary 2]

$$E_{1,\text{BSC}}(R,T) = \min\left\{E_{a,\text{BSC}}(R,T), E_{b,\text{BSC}}(R,T)\right\}, \tag{20}$$

where

$$E_{a,\text{BSC}}(R,T) \triangleq \min_{\tilde{q}\in[\theta,\delta_{GV}(R)-T/\beta]} \left[D\left(\tilde{q}||\theta\right) - h\left(\tilde{q}+\frac{T}{\beta}\right) + \log 2 - R\right], \tag{21}$$

and

$$E_{b,\text{BSC}}(R,T) \triangleq \min_{\tilde{q}\in\hat{\mathcal{L}}_{\text{BSC}}} D\left(\tilde{q}||\theta\right), \tag{22}$$

where $\beta(\theta) \triangleq \log\left[(1-\theta)/\theta\right]$, and $\delta_{\text{GV}}(R)$ denote the normalized Gilbert-Varshamov (GV) distance, i.e., the smaller solution, δ, to the equation

$$h(\delta) = \log 2 - R, \tag{23}$$

where $h(\delta) \triangleq -\delta\log\delta - (1-\delta)\log(1-\delta)$ is the binary entropy function, and

$$\hat{\mathcal{L}}_{\text{BSC}} \triangleq \left\{\tilde{q}: -\tilde{q}\cdot\beta(\theta) \leq R + T + \max_{q:\ R\geq\log 2 - h(q)} \left[-q\cdot\beta(\theta) + h(q) - \log 2\right]\right\}. \tag{24}$$

B. Unknown Channel

We now move on to the case of an unknown channel. Consider a family of DMCs

$$\mathcal{W}_\Theta \triangleq \{W_\theta(y|x),\ x \in \mathcal{X}, y \in \mathcal{Y}, \theta \in \Theta\}, \tag{25}$$

with a finite input alphabet $\mathcal{X}$, a finite output alphabet $\mathcal{Y}$, and a matrix of single-letter transition probabilities $\{W_\theta(y|x)\}$, where θ is a parameter, or the index of the channel in the class, taking values in some set Θ, which may be countable or uncountable. For example, θ may be represent the set of all $|\mathcal{X}| \cdot (|\mathcal{Y}| - 1)$ single-letter transition probabilities that define the DMC with the given input and output alphabets. In our problem, the channel is unknown to the receiver designer, and the designer only knows that the channel belongs to the family of channels $\mathcal{W}_\Theta$, that is, θ itself is unknown.

When the channel is unknown, the competitive minimax methodology, proposed and developed in [15], proves useful. Specifically, let $\Gamma_\theta(\mathcal{C}, \mathcal{R})$ in (7) designate the above defined Lagrangian, where we now emphasize the dependence on the index of the channel, θ. Similarly, henceforth we shall denote the error exponents in (6) by $E_1(R, T, \theta)$ and $E_2(R, T, \theta)$. Also, let $\bar{\Gamma}_\theta^* \triangleq \mathbb{E}\{\min_{\mathcal{R}} \Gamma_\theta(\mathcal{C}, \mathcal{R})\}$, which is the ensemble average of the minimum of the above Lagrangian (achieved by Forney's optimum decision rule) w.r.t. the channel $W_\theta(y|x)$, for a given θ. Note that by Lemma 1, the exponential order of $\bar{\Gamma}_\theta^*$ is $e^{-n(E_1(R,T,\theta)+T)}$. A *competitive minimax decision rule* $\mathcal{R}$ is one that achieves

$$\min_{\mathcal{R}} \max_{\theta \in \Theta} \frac{\Gamma_\theta(\mathcal{C}, \mathcal{R})}{\bar{\Gamma}_\theta^*}, \tag{26}$$

which is asymptotically equivalent to

$$\min_{\mathcal{R}} \max_{\theta \in \Theta} \frac{\Gamma_\theta(\mathcal{C}, \mathcal{R})}{e^{-n[E_1(R,T,\theta)+T]}}. \tag{27}$$

However, as discussed in [15], such a minimax criterion, of competing with the optimum performance, may be too optimistic, and the value of the minimization problem in (27) may diverge to infinity for every R, as $n \to \infty$. A possible remedy for this situation is to compete with only a fraction $\xi \in [0, 1]$ of $E_1(R, T, \theta)$, which we would like to choose as large as possible. To wit, we are interested in the competitive minimax criterion

$$K_n(\mathcal{C}) = \min_{\mathcal{R}} K_n(\mathcal{R}, \mathcal{C}), \tag{28}$$

in which

$$K_n(\mathcal{R}, \mathcal{C}) = \max_{\theta \in \Theta} \frac{\Gamma_\theta(\mathcal{C}, \mathcal{R})}{e^{-n(\xi E_1(R,T,\theta)+T)}}. \tag{29}$$

Accordingly, we wish to find the largest value of ξ such that the ensemble average $\bar{K}_n \triangleq \mathbb{E}\{K_n(\mathcal{C})\}$ would *not* grow exponentially fast, i.e.,

$$\limsup_{n\to\infty} \frac{1}{n} \log \bar{K}_n \leq 0. \tag{30}$$

In [15], the following universal decoding metric was defined

$$f(\boldsymbol{x}_m, \boldsymbol{y}) \triangleq \max_{\theta\in\Theta} \left\{ e^{n[\xi E_1(R,T,\theta)+T]} W_\theta(\boldsymbol{y}|\boldsymbol{x}_m) \right\}, \tag{31}$$

and a universal erasure/list decoder was proposed which has the following decision regions

$$\hat{\mathcal{R}}_m \triangleq \left\{ \boldsymbol{y} : \frac{f(\boldsymbol{x}_m, \boldsymbol{y})}{\sum_{m'\neq m} f(\boldsymbol{x}_{m'}, \boldsymbol{y})} \geq e^{nT} \right\}, \ m = 1, 2, \ldots M, \tag{32}$$

and

$$\hat{\mathcal{R}}_0 \triangleq \bigcap_{m=1}^{M} \hat{\mathcal{R}}_m^c. \tag{33}$$

The property that makes $\hat{\mathcal{R}} \triangleq (\hat{\mathcal{R}}_0, \hat{\mathcal{R}}_1, \ldots, \hat{\mathcal{R}}_M)$ interesting is that it was shown in [15], that it is asymptotically optimal, i.e., for any given ξ, $K_n(\hat{\mathcal{R}}, \mathcal{C})$ may only be sub-exponentially larger than $K_n(\mathcal{C})$. Thus, the largest ξ such that $\bar{K}_n$ is sub-exponential is also attained by $\hat{\mathcal{R}}$. Hence, in order to find the largest achievable ξ, we would like to evaluate exactly the exponential order of $\mathbb{E}[K_n(\hat{\mathcal{R}}, \mathcal{C})]$.

Remark 1 Note that the results in this paper can be generalized to other random coding ensembles which assign equal probabilities within every type class (for more details see [15, Section V]). For conceptual simplicity, we confine attention to fixed-composition random coding.

IV. Results

In this section, our results are presented and discussed. Proofs are relegated to Section V.

A. Exact formula for the largest achievable fraction

We start with a few definitions. Let

$$G(R, T, \xi, \tilde{Q}) \triangleq \max_{\theta\in\Theta} \left\{ \xi E_1(R, T, \theta) + T + \mathbb{E}_{\tilde{Q}} \log W_\theta(Y|X) \right\}, \tag{34}$$

$$\Omega(R, T, \xi, Q, \tilde{Q}) \triangleq G(R, T, \xi, \tilde{Q}) - G(R, T, \xi, Q) - T, \tag{35}$$

where $E_1(R, T, \theta)$ is given in (12). Finally, let

$$\mathcal{Q} \triangleq \left\{ (Q, \tilde{Q}) \in \mathcal{D} : \ I(Q) \geq R, \ \Omega(R, T, \xi, Q, \tilde{Q}) \leq 0 \right\} \tag{36}$$

and

$$\mathcal{L} \triangleq \left\{ \tilde{Q} : G(R,T,\xi,\tilde{Q}) \leq R + T + \max_{Q:(Q,\tilde{Q})\in\mathcal{D},\, I(Q)\leq R} [G(R,T,\xi,Q) - I(Q)] \right\}. \tag{37}$$

where $\mathcal{D}$ is defined in (16).

Theorem 1 Consider the ensemble of fixed composition codes of type $\mathcal{T}(P_X)$. Then, $\xi^*(R,T)$ is equal to the largest number ξ that satisfies simultaneously:

$$\max_{\theta\in\Theta} \left\{ \xi E_1(R,T,\theta) - \min_{(Q,\tilde{Q})\in\mathcal{Q}} \left\{ D(\tilde{Q}||P_X \times W_\theta) + I(Q) - R \right\} \right\} \leq 0, \tag{38}$$

and

$$\max_{\theta\in\Theta} \left\{ \xi E_1(R,T,\theta) - \min_{\tilde{Q}\in\mathcal{L}} D(\tilde{Q}||P_X \times W_\theta) \right\} \leq 0. \tag{39}$$

Notice that in order to find $\xi^*(R,T)$, one can perform a simple line search over the interval $[0,1]$ using the conditions in Theorem 1. Also, note that one can readily find a single-letter formula for $\xi^*(R,T)$. The resulting formula is, however, complicated and does not provide much insight, therefore, it is not provided here. For the special case of the BSC, one can simplify the above minimization problems over the joint distributions $(Q,\tilde{Q})$, and obtain instead a one-dimensional minimization problem. Indeed, consider the family of BSCs where the unknown crossover probability θ belongs to $\Theta = [0,1]$. Recall that (c.f. end of Subsection III-A) $\beta(\theta) = \log[(1-\theta)/\theta]$. Define

$$\phi(\theta) = \frac{\xi E_1(R,T,\theta) + \log(1-\theta) + T - \max_{\theta'} \{\xi E_1(R,T,\theta') + \log(1-\theta') - \beta(\theta')\cdot\tilde{q}\}}{\beta(\theta)}, \tag{40}$$

and

$$q_1^* \triangleq \max_{\theta\leq 1/2} \phi(\theta), \tag{41}$$

$$q_2^* \triangleq \min_{\theta>1/2} \phi(\theta). \tag{42}$$

Finally, define

$$g(q_1^*, q_2^*) \triangleq \begin{cases} \log 2, & \text{if } q_1^* > 1/2, \text{ or, } q_2^* < 1/2, \\ \max\{h(q_1^*), h(q_2^*)\}, & \text{otherwise} \end{cases}, \tag{43}$$

and

$$\mathcal{L}_{\text{BSC}} \triangleq \left\{ \tilde{q} : \max_{0\leq\theta\leq 1} [\xi E_1(R,T,\theta) - \tilde{q}\cdot\beta(\theta) + \log\theta] \leq R + T \right.$$

$$+\max_{0\le\theta\le 1}\left[\xi E_1(R,T,\theta)-\max\left\{\theta,\delta_{\rm GV}(R)\right\}\cdot\beta(\theta)+\log\theta+h(\max\left\{\theta,\delta_{\rm GV}(R)\right\})-\log 2\right]\Bigg\}. \tag{44}$$

We have the following result.

Corollary 1 Consider a family of BSCs, where the unknown crossover probability θ belongs to $\Theta=[0,1]$, and with fixed composition codes of type $P_X=(1/2,1/2)$. Then, $\xi^*(R,T)$ equals to the largest number ξ that satisfies simultaneously:

$$\max_{0\le\theta\le 1}\left\{\xi\cdot E_{1,\rm BSC}(R,T,\theta)-\min_{\tilde{q}}\left[D\left(\tilde{q}||\theta\right)+\left|-g\left(q_1^*,q_2^*\right)+\log 2-R\right|^+\right]\right\}\le 0, \tag{45}$$

and

$$\max_{0\le\theta\le 1}\left\{\xi\cdot E_{1,\rm BSC}(R,T,\theta)-\min_{\tilde{q}\in\mathcal{L}_{\rm BSC}}D\left(\tilde{q}||\theta\right)\right\}\le 0, \tag{46}$$

where $E_{1,\rm BSC}(R,T,\theta)$ is given in (20).

Note that there is a major difference between ordinary and universal decoding in the context of the BSC: while for the former, the optimal detector depends only on whether $\theta\le 1/2$ or $\theta>1/2$ (i.e., minimum distance versus maximum distance decoders, respectively), for the latter, the dependence is on the exact value of θ.

B. Discussion and Comparison with Previous Results

While in this work we have derived the exact maximal achievable $\xi^*(R,T)$ for fixed composition coding of type P_X, in [15, Theorem 2], Merhav and Feder have obtained the following lower bound [15, Theorem 2]:

$$\xi^*(R,T)\ge\xi_L(R,T)\triangleq\min_{(\theta,\theta'')\in\Theta^2}\max_{0\le s\le\rho\le 1}\frac{E(\theta,\theta'',s,\rho)-\rho R-sT}{(1-s)E_1(R,T,\theta)+sE_1(R,T,\theta'')} \tag{47}$$

where

$$E(\theta,\tilde{\theta},s,\rho)\triangleq\min_{Q_Y}\left[F(Q_Y,1-s,\theta)+\rho F(Q_Y,s/\rho,\theta'')-H(Q_Y)\right] \tag{48}$$

and

$$F(Q_Y,1-s,\theta)\triangleq\min_{Q_{X|Y}:\,(Q_Y\times Q_{X|Y})_X=P_X}\left[I(Q)-\lambda\mathbb{E}_Q\log W_\theta(Y|X)\right]. \tag{49}$$

Before we continue, we remark that in [15], Forney's lower bound on $E_1(R,T,\theta)$ was used instead of its exact value as derived in [18], but for the sake of comparison any exponent can be used, and specifically,

the exact exponent. Now, note that an alternative (equivalent) representation of $\xi_L(R,T)$ in (47) is that it is given by the largest ξ such that for any pair $(\theta,\theta'')\in\Theta^2$

$$\min_{(\theta,\theta'')\in\Theta^2}\max_{0\le s\le\rho\le 1} E(\theta,\theta'',s,\rho)-\rho R-sT-\xi\left[(1-s)E_1(R,T,\theta)+sE_1(R,T,\theta'')\right]\ge 0. \tag{50}$$

Straightforward algebraic manipulations show that the last inequality can be rewritten as

$$\min_{(\theta,\theta'')\in\Theta^2}\max_{0\le s\le\rho\le 1}\min_{(Q,\tilde{Q})\in\mathcal{D}}\Psi(R,T,\theta,\theta,\theta'',Q,\tilde{Q},\rho,s)\ge 0 \tag{51}$$

where

$$\begin{aligned}\Psi(R,T,\theta,\theta',\theta'',Q,\tilde{Q},\rho,s,\xi)\triangleq\ & D(\tilde{Q}||P_X\times W_\theta)+\rho\left[I(Q)-R\right]\\ &+s\cdot\Big[\mathbb{E}_{\tilde{Q}}\log W_{\theta'}(Y|X)+\xi E_1(R,T,\theta')\\ &-\mathbb{E}_Q\log W_{\theta''}(Y|X)-\xi E_1(R,T,\theta'')-T\Big]-\xi E_1(R,T,\theta).\end{aligned} \tag{52}$$

For any given $(\theta,\theta'')\in\Theta^2$, and (s,ρ), $\Psi(R,T,\theta,\theta',\theta'',Q,\tilde{Q},\rho,s,\xi)$ is convex in $(Q,\tilde{Q})$, and for a given $(Q,\tilde{Q})$, it is linear (and hence concave) in (s,ρ) . Thus, the minimax theorem implies that (51) is equivalent to

$$\min_{(\theta,\theta'')\in\Theta^2}\min_{(Q,\tilde{Q})\in\mathcal{D}}\max_{0\le s\le\rho\le 1}\Psi(R,T,\theta,\theta,\theta'',Q,\tilde{Q},\rho,s,\xi)\ge 0. \tag{53}$$

On the other hand, the exact value of $\xi^*(R,T)$ in Theorem 1 is determined by two conditions (38)-(39). In what follows, we shall concentrate on the first condition in (38), as this condition can be compared to (53). Thus, assume, for a moment, that the condition in (39) is more lenient than the condition in (38). Then, according to (38), a fraction ξ is achievable if

$$\min_{\theta\in\Theta}\min_{(Q,\tilde{Q})\in\mathcal{D}} D(\tilde{Q}||P_X\times W_\theta)+I(Q)-R-\xi E_1(R,T,\theta)\ge 0 \tag{54}$$

where the minimum over $(Q,\tilde{Q})$ is such that $I(Q)\ge R$ and $\Omega(R,T,\xi,Q,\tilde{Q})\le 0$. Now, the optimization problem in (54) is equivalent to

$$\begin{aligned}\min_{\theta\in\Theta}\min_{(Q,\tilde{Q})\in\mathcal{D}}\max_{\rho'\ge 0}\max_{s\ge 0}\ &\Big[D(\tilde{Q}||P_X\times W_\theta)+(1-\rho')\left[I(Q)-R\right]\\ &+s\Omega(R,T,\xi,Q,\tilde{Q})-\xi E_1(R,T,\theta)\Big]\ge 0,\end{aligned} \tag{55}$$

or by letting $\rho=1-\rho'$ we get

$$\min_{\theta\in\Theta}\min_{(Q,\tilde{Q})\in\mathcal{D}}\max_{\rho\le 1}\max_{s\ge 0}\Big[D(\tilde{Q}||P_X\times W_\theta)+\rho\left[I(Q)-R\right]+s\Omega(R,T,\xi,Q,\tilde{Q})-\xi E_1(R,T,\theta)\Big]\ge 0, \tag{56}$$

which is equivalent to

$$\min_{\theta\in\Theta}\min_{(Q,\tilde{Q})\in\mathcal{D}}\max_{\rho\leq 1}\max_{s\geq 0}\max_{\theta'\in\Theta}\min_{\theta''\in\Theta}\Psi(R,T,\theta,\theta',\theta'',Q,\tilde{Q},\rho,s,\xi)\geq 0. \tag{57}$$

Moreover, for a given $(\theta, Q, \tilde{Q})$, we may write

$$\max_{\rho\leq 1}\max_{s\geq 0}\max_{\theta'\in\Theta}\min_{\theta''\in\Theta}\Psi(R,T,\theta,\theta',\theta'',Q,\tilde{Q},\rho,s,\xi)=\min_{\theta''\in\Theta}\max_{\theta'\in\Theta}\max_{0\leq\rho\leq 1}\max_{s\geq 0}\Psi(R,T,\theta,\theta',\theta'',Q,\tilde{Q},\rho,s,\xi), \tag{58}$$

because under the constraint $s \geq 0$, the inner minimization over $\theta'' \in \Theta$ does not depend on the value of (ρ, s, θ'): it is simply the $\theta'' \in \Theta$ which maximizes $\mathbb{E}_Q \log W_{\theta''}(Y|X) + \xi E_1(R, T, \theta'')$ [3]. Thus, the resulting condition is

$$\min_{(\theta,\theta'')\in\Theta^2}\min_{(Q,\tilde{Q})\in\mathcal{D}}\max_{0\leq\rho\leq 1}\max_{s\geq 0}\max_{\theta'\in\Theta}\Psi(R,T,\theta,\theta',\theta'',Q,\tilde{Q},\rho,s,\xi)\geq 0. \tag{59}$$

By comparing the condition in (59) to the condition of the lower bound of [15] in (53), the following differences are observed:

1) In (53) an additional constraint $s \leq \rho$ is imposed.
2) In (53) a sub-optimal choice of $\theta' = \theta$ is imposed.

Accordingly, these differences may cause the value of the minimax in (53) to be lower than the value of the optimization problem in (59), which results in a lower achievable ξ compared to $\xi^*(R, T)$, as one should expect. Next, we provide two examples where in one of which these differences are immaterial and in the other one they do. The former happens when the optimal solution in (59), denoted by $(\theta^*, \theta''^*, Q^*, \tilde{Q}^*, \rho^*, s^*)$, satisfies $s^* \leq \rho^*$, and the maximizer of $\mathbb{E}_{\tilde{Q}^*} \log W_{\theta'}(Y|X) + \xi_L(R, T) \cdot E_1(R, T, \theta')$ is given by θ^*. Accordingly, in this case, the value of (59) equals to (53), and therefore $\xi^*(R, T) = \xi_L(R, T)$, due the fact that the condition in (39) is more lenient than the condition in (38), as we have previously assumed. The conclusion that stems from this observation is that, in this case, the analysis in [15] is tight.

Example 1 In [15], a family of BSCs was considered where $\theta \in \Theta$ designates the cross-over probability of the BSC, and $\Theta = \{0, 1/100, 2/100, \ldots, 1\}$. The values of $\xi_L(R, T)$ were computed for various values of R and T. It was assumed that $T \geq 0$, which means that the decoder operates in the erasure mode. Numerical calculations of the bound derived in this work (and the exact formula), result in exactly the

[3] If for a given real function $f(u, v)$ the minimizer v^* w.r.t. v does not depend on u, then $\max_{u\in\mathcal{U}} \min_{v\in\mathcal{V}} f(u, v) = \max_{u\in\mathcal{U}} f(u, v^*) \geq \min_{v\in\mathcal{V}} \max_{u\in\mathcal{U}} f(u, v)$, and the minimax inequality results $\max_{u\in\mathcal{U}} \min_{v\in\mathcal{V}} f(u, v) = \min_{v\in\mathcal{V}} \max_{u\in\mathcal{U}} f(u, v)$, assuming that $\mathcal{U}$ and $\mathcal{V}$ are two independent sets (i.e., rectangular).

same values as given in [15, Table 1], and so in all these cases, the analysis of [15] was sufficient to provide tight results. For example, for $(R,T)=(0.05,0.15)$, and codebook type $P_X=(1/2,1/2)$, we obtain $\xi_L(R,T)=0.495$. Also, the two worst case channels (i.e., the solutions to (59)) are $\theta^*=0.18$ and $\theta''^*=0.22$ while $\theta'^*=\theta^*$ and $\rho^*=0.36>s^*=0.185$. So, since $s^*<\rho^*$ and $\theta'^*=\theta^*$, the discussion above implies that a tight result is obtained, that is, $\xi^*(R,T)=\xi_L(R,T)=0.495$.

Since $\xi^*(R,T)<1$ for some R and T, we arrive at the following conclusion: *In general, in the random coding regime of erasure/list decoding, there is no universal decoder which achieves the same error exponent as Forney's decoder for every channel in the class.* This fact is in contrast to ordinary decoding, in which the MMI decoder achieves the exact same error exponent as the ML decoder. In this sense, channel knowledge is crucial when erasure/list options are allowed.

Nonetheless, in general, we might have that $\xi_L(R,T)$ is strictly less than $\xi^*(R,T)$. Again, assume that the condition in (38) dominates $\xi^*(R,T)$. To provide intuition, notice that in (59) *triplets* $(\theta,\theta',\theta'')\in\Theta^3$ are optimized, in contrast to (53), where only *pairs* of channels $(\theta,\theta'')\in\Theta^2$ are optimized. Thus, for a family of only two channels, namely, $|\Theta|=2$, typically (but not necessarily) the second difference above, of imposing the constraint $\theta'=\theta$, is immaterial. Then, the only difference between the conditions in (53) and (59) is the constraint $s\le\rho$. Let us assume that this is indeed the case, and let us notice that s can be thought as a Lagrange multiplier for the constraint

$$\mathbb{E}_{\tilde{Q}}\log W_{\theta'}(Y|X)+\xi E_1(R,T,\theta')-\mathbb{E}_{Q}\log W_{\theta''}(Y|X)-\xi E_1(R,T,\theta'')-T\le 0. \tag{60}$$

Now, if the constraint, at the optimal solution, is slack, then the optimal Lagrange multiplier is $s^*=0$. In this case, the constraint $s\le\rho$ is immaterial and so (53) and (59) are exactly the same. However, as we shall see in the sequel, it is possible that $s^*>\rho^*$ in (59), and then the values of the objective in (53) and (59) are different. Observing (60), it is apparent that as T decreases, and especially in the list mode of $T<0$, the optimal s^* of (59) increases, perhaps beyond the optimal ρ^*. Thus, if both $s^*>\rho^*$ and the condition in (38) dominates $\xi^*(R,T)$, we get that $\xi_L(R,T)<\xi^*(R,T)$. The following example provides such a simple case. We remark, that such a phenomenon was already observed in a Slepian-Wolf erasure/list decoding scenario, for a known source [19]. There too, in the list regime of $T<0$, there is a gap between the Forney-style bound and the exact random binning error exponents.

Example 2 Consider a family of two BSC's, where $\Theta=\{0.1,0.15\}$, and a type $P_X=(1/2,1/2)$ for the random fixed composition codebook. We take $(R,T)=(0.4,-0.25)$, and since $T<0$, the decoder operates in the list mode. We obtain that $\xi_L(R,T)=0.716$ which is strictly less than $\xi^*(R,T)=0.727$.

In the optimization problem (53), the optimal values are $\rho^* = s^* = 0.231$, while if the constraint $s \leq \rho$ is relaxed, then the optimal values are $s = 0.231 > \rho = 0.217$. The resulting value of the optimization problem is exactly 0.727, just as $\xi^*(R,T)$. Moreover, for this example, the largest achievable ξ which satisfy condition (38) is the same for condition (39). While the difference between $\xi_L(R,T)$ and $\xi^*(R,T)$ is not very large, it is nevertheless existent and in more intricate scenarios, the differences might be more significant.

V. PROOFS

In the following, for simplicity of notations, we omit the dependency of the various quantities on R, T, and ξ, as they remain constants along the proofs.

Proof of Theorem 1: We analyze the total error term, following the steps of [18, Section V]. As was mentioned earlier, we want to assess the (exact) exponential behavior of $\mathbb{E}\left[K_n(\hat{\mathcal{R}},\mathcal{C})\right]$. In [15, Theorem 2], an upper bound was derived on this quantity, so here we seek a tight lower bound. Let Θ_n denote the set of values of θ that achieve $\{f(\boldsymbol{x},\boldsymbol{y}), \boldsymbol{x}\in\mathcal{X}^n, \boldsymbol{y}\in\mathcal{Y}^n\}$. Note that the elements of Θ_n depend on $\boldsymbol{x}$ and $\boldsymbol{y}$ only through their joint type, and whence, we have that $|\Theta_n| \leq (n+1)^{|\mathcal{X}|\cdot|\mathcal{Y}|-1}$, i.e. the size of Θ_n is a polynomial function of n. Now,

$$\mathbb{E}\left[K_n(\hat{\mathcal{R}},\mathcal{C})\right] = \mathbb{E}\left\{\max_{\theta\in\Theta}\frac{\Gamma_\theta(\mathcal{C},\hat{\mathcal{R}})}{e^{-n(\xi E_1(\theta)+T)}}\right\} \tag{61}$$

$$\geq \mathbb{E}\left\{\max_{\theta\in\Theta_n}\frac{\Gamma_\theta(\mathcal{C},\hat{\mathcal{R}})}{e^{-n(\xi E_1(\theta)+T)}}\right\} \tag{62}$$

$$\stackrel{(a)}{\doteq} \mathbb{E}\left\{\sum_{\theta\in\Theta_n}\frac{\Gamma_\theta(\mathcal{C},\hat{\mathcal{R}})}{e^{-n(\xi E_1(\theta)+T)}}\right\} \tag{63}$$

$$\stackrel{(b)}{\doteq} \mathbb{E}\left\{\sum_{\theta\in\Theta_n}\frac{\frac{1}{M}\sum_{m=1}^{M}\left[\sum_{\boldsymbol{y}\in\hat{\mathcal{R}}_m}\sum_{m'\neq m}W_\theta(\boldsymbol{y}|\boldsymbol{X}_{m'}) + \sum_{\boldsymbol{y}\in\hat{\mathcal{R}}_m^c}e^{-nT}W_\theta(\boldsymbol{y}|\boldsymbol{X}_m)\right]}{e^{-n(\xi E_1(\theta)+T)}}\right\} \tag{64}$$

$$= \mathbb{E}\left\{\frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\hat{\mathcal{R}}_m}\sum_{m'\neq m}\sum_{\theta\in\Theta_n}e^{n(\xi E_1(\theta)+T)}W_\theta(\boldsymbol{y}|\boldsymbol{X}_{m'})\right\} \tag{65}$$

$$+ \mathbb{E}\left\{\frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\hat{\mathcal{R}}_m^c}\sum_{\theta\in\Theta_n}e^{n\xi E_1(\theta)}W_\theta(\boldsymbol{y}|\boldsymbol{X}_m)\right\} \tag{66}$$

$$\stackrel{(a)}{\doteq} \mathbb{E}\left\{\frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\hat{\mathcal{R}}_m}\sum_{m'\neq m}\max_{\theta\in\Theta_n}e^{n(\xi E_1(\theta)+T)}W_\theta(\boldsymbol{y}|\boldsymbol{X}_{m'})\right\} \tag{67}$$

$$+\mathbb{E}\left\{\frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\hat{\mathcal{R}}_m^c}\max_{\theta\in\Theta_n}e^{n\xi E_1(\theta)}W_\theta(\boldsymbol{y}|\boldsymbol{X}_m)\right\} \tag{68}$$

$$=\mathbb{E}\left\{\frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\hat{\mathcal{R}}_m}\sum_{m'\neq m}f(\boldsymbol{X}_{m'},\boldsymbol{y})\right\}+\mathbb{E}\left\{\frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\hat{\mathcal{R}}_m^c}e^{-nT}f(\boldsymbol{X}_m,\boldsymbol{y})\right\} \tag{69}$$

where in (a) we have used the fact that the size of Θ_n is polynomial, and thus can be absorbed in the e^{nT} factor (see, [18, pp. 5, footnote 2]), and (b) follows from (8). As was shown in [15, eq. after (A.1)], the lower bound in (69) is, in fact, also an upper bound on $\mathbb{E}\left[K_n(\hat{\mathcal{R}},\mathcal{C})\right]$. Therefore, in the exponential scale, nothing was lost due to the above bounding, and we essentially have that

$$\mathbb{E}\left[K_n(\hat{\mathcal{R}},\mathcal{C})\right]\doteq\mathbb{E}\left\{\frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\hat{\mathcal{R}}_m}\sum_{m'\neq m}f(\boldsymbol{X}_{m'},\boldsymbol{y})\right\}+\mathbb{E}\left\{\frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\hat{\mathcal{R}}_m^c}e^{-nT}f(\boldsymbol{X}_m,\boldsymbol{y})\right\}. \tag{70}$$

Contrary to the proof technique used in [15] to assess the exponential behavior of (70), where Chernoff and Jensen bounds were invoked, here, we will evaluate the *exact* exponential scale of the two terms on the right hand side of (70). It can be noticed that the first expression is related to undetected errors (or average number of incorrect codewords on the list), and the second one is related to the total error (erasures and undetected errors). For brevity, we define

$$A_1\triangleq e^{-nT}\cdot\mathbb{E}\left\{\frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\hat{\mathcal{R}}_m^c}f(\boldsymbol{X}_m,\boldsymbol{y})\right\}, \tag{71}$$

and

$$A_2\triangleq\mathbb{E}\left\{\frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\hat{\mathcal{R}}_m}\sum_{m'\neq m}f(\boldsymbol{X}_{m'},\boldsymbol{y})\right\}, \tag{72}$$

and so

$$\mathbb{E}\left[K_n(\hat{\mathcal{R}},\mathcal{C})\right]\doteq A_1+A_2. \tag{73}$$

As was mentioned before, we would like to analyze the exponential rate of (70), or, equivalently, of (71) and (72). Now, note that

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}\left[K_n(\hat{\mathcal{R}},\mathcal{C})\right]=\max\left\{\lim_{n\to\infty}\frac{1}{n}\log A_1,\lim_{n\to\infty}\frac{1}{n}\log A_2\right\}. \tag{74}$$

Then, a fraction ξ is *achievable* if both $n^{-1}\log A_1$ and $n^{-1}\log A_2$ converge to a non-positive constant as $n\to\infty$. At this point, we would like to invoke Lemma 1, and conclude that it suffices to asses the exponential behavior of A_1 (or, A_2), and then the other one is immediately obtained. Note that while Lemma 1 was derived for the case of a known channel, it still remains true in the case of an unknown

channel due to the similar structure of our universal erasure decoder[4] (see the equivalence between (8) and (70)). Thus, while both A_1 and A_2 can be analyzed, it turns out that the analytical formula for the exponent of A_1 is more compact, and thus, in the following, we only present the analysis of A_1. Continuing from (71),

$$A_1 = e^{-nT}\mathbb{E}\left\{\frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}} f(\boldsymbol{X}_m,\boldsymbol{y})\cdot\mathcal{I}\{\boldsymbol{y}\in\hat{\mathcal{R}}_m^c\}\right\} \tag{75}$$

$$\overset{(a)}{=} e^{-nT}\mathbb{E}\left\{\sum_{\boldsymbol{y}} f(\boldsymbol{X}_m,\boldsymbol{y})\cdot\mathcal{I}\{\boldsymbol{y}\in\hat{\mathcal{R}}_m^c\}|m\text{th message transmitted}\right\} \tag{76}$$

$$= e^{-nT}\sum_{\boldsymbol{y}}\mathbb{E}\left\{f(\boldsymbol{X}_m,\boldsymbol{y})\cdot\mathcal{I}\{\boldsymbol{y}\in\hat{\mathcal{R}}_m^c\}|m\text{th message transmitted}\right\} \tag{77}$$

$$= e^{-nT}\sum_{\boldsymbol{x}_m} P_X(\boldsymbol{X}_m=\boldsymbol{x}_m)\sum_{\boldsymbol{y}}\mathbb{E}\left\{f(\boldsymbol{X}_m,\boldsymbol{y})\cdot\mathcal{I}\{\boldsymbol{y}\in\hat{\mathcal{R}}_m^c\}|\boldsymbol{X}_m=\boldsymbol{x}_m, m\text{th message transmitted}\right\} \tag{78}$$

$$= e^{-nT}\sum_{\boldsymbol{x}_m} P_X(\boldsymbol{X}_m=\boldsymbol{x}_m)\sum_{\boldsymbol{y}} f(\boldsymbol{x}_m,\boldsymbol{y})\cdot\Pr\left\{\boldsymbol{y}\in\hat{\mathcal{R}}_m^c|\boldsymbol{X}_m=\boldsymbol{x}_m, m\text{th message transmitted}\right\} \tag{79}$$

where (a) follows from the symmetry of the random coding mechanism. Next, let Q be the joint empirical probability distribution defined on $\mathcal{X}\times\mathcal{Y}$ of $\boldsymbol{x}_{m'}$ and $\boldsymbol{y}$. Then,

$$f(\boldsymbol{x}_{m'},\boldsymbol{y}) = \max_{\theta\in\Theta}\left\{e^{n(\xi E_1(\theta)+T)}W_\theta(\boldsymbol{y}|\boldsymbol{x}_m)\right\} \tag{80}$$

$$= \max_{\theta\in\Theta}\left\{e^{n(\xi E_1(\theta)+T)}e^{n\mathbb{E}_Q\log W_\theta(Y|X)}\right\} \tag{81}$$

$$= \exp\left[n\cdot\max_{\theta\in\Theta}\left\{(\xi E_1(\theta)+T)+\mathbb{E}_Q\log W_\theta(Y|X)\right\}\right] \tag{82}$$

$$= \exp\left[n\cdot G(Q)\right], \tag{83}$$

where we have defined $G(Q)$ in (34). Next, we shall focus on the latter probability in (79). For a given $\boldsymbol{x}_m$ and $\boldsymbol{y}$, let $\tilde{Q}=\hat{P}_{\boldsymbol{x}\boldsymbol{y}}$, let $N_{\boldsymbol{y}}(Q)$ denote the number of codewords (excluding $\boldsymbol{x}_m$) whose joint empirical probability distribution with a given $\boldsymbol{y}$ is Q. Accordingly, we have that

$$\Pr\left\{\boldsymbol{y}\in\hat{\mathcal{R}}_m^c|\boldsymbol{x}_m\right\} = \Pr\left\{\sum_{m'\neq m} f(\boldsymbol{x}_{m'},\boldsymbol{y})\geq f(\boldsymbol{x}_m,\boldsymbol{y})e^{-nT}\right\} \tag{84}$$

[4]Note that in the proof of Lemma 1, we have used the fact that $E_1(R,T,\theta)$ and $E_2(R,T,\theta)$ are both continuous functions of T. Accordingly, in order to apply Lemma 1 on the universal case, one should inspect that $\lim_{n\to\infty}\frac{1}{n}\log A_1$ and $\lim_{n\to\infty}\frac{1}{n}\log A_2$ are also continuous functions of T. As shall be seen in the sequel, $\lim_{n\to\infty}\frac{1}{n}\log A_1$ is indeed continuous in T, and similarly to the derivation of $E_2(R,T,\theta)$ in [18], it can be shown that $\lim_{n\to\infty}\frac{1}{n}\log A_2$ is continuous too.

$$= \Pr\left\{ \sum_{Q} N_{\boldsymbol{y}}(Q) \exp\left[n \cdot G(Q)\right] \geq \exp\left[n \cdot G(\tilde{Q})\right] e^{-nT} \right\} \tag{85}$$

$$\doteq \Pr\left\{ \max_{Q} N_{\boldsymbol{y}}(Q) \exp\left[n \cdot G(Q)\right] \geq \exp\left[n \cdot G(\tilde{Q})\right] e^{-nT} \right\} \tag{86}$$

$$= \Pr\left\{ \bigcup_{Q} \left\{ N_{\boldsymbol{y}}(Q) \exp\left[n \cdot G(Q)\right] \geq \exp\left[n \cdot G(\tilde{Q})\right] e^{-nT} \right\} \right\} \tag{87}$$

$$\doteq \sum_{Q} \Pr\left\{ N_{\boldsymbol{y}}(Q) \exp\left[n \cdot G(Q)\right] \geq \exp\left[n \cdot G(\tilde{Q})\right] e^{-nT} \right\} \tag{88}$$

$$\doteq \max_{Q} \Pr\left\{ N_{\boldsymbol{y}}(Q) \exp\left[n \cdot G(Q)\right] \geq \exp\left[n \cdot G(\tilde{Q})\right] e^{-nT} \right\} \tag{89}$$

$$= \max_{Q \in \mathcal{S}(\hat{P}_{\boldsymbol{y}})} \Pr\left\{ N_{\boldsymbol{y}}(Q) \geq \exp\left[n \cdot \Omega(Q, \tilde{Q})\right] \right\} \tag{90}$$

where for a given $\bar{Q}_Y$, $\mathcal{S}(\bar{Q}_Y) \triangleq \left\{Q : \, Q_Y = \bar{Q}_Y, \, Q_X = \bar{Q}_X\right\}$. The asymptotic analysis of the probability in (90) was carried out in [18, Section V] for any given Ω, and it is not different here. The result relies on the exponential decay of the probability that the joint type of a given $\boldsymbol{y}$ with a randomly chosen $\boldsymbol{x}_{m'}$ is Q, namely

$$p \triangleq \Pr\left\{ \hat{P}_{\boldsymbol{X}_{m'}, \boldsymbol{y}} = Q \right\}. \tag{91}$$

Under the assumed random coding ensemble, a simple application of the method of types reveals that [4]

$$p \doteq \exp\left\{-nI(Q)\right\}. \tag{92}$$

Next, standard large deviations arguments (cf. [18, Section V]) reveal that for $Q \in \mathcal{S}(\hat{P}_{\boldsymbol{y}})$

$$\Pr\left\{ N_{\boldsymbol{y}}(Q) \geq e^{n\Omega(Q,\tilde{Q})} \right\} \doteq \begin{cases} \exp\left\{-n\left|I(Q) - R\right|^{+}\right\} & \Omega(Q, \tilde{Q}) \leq 0 \\ 1 & 0 < \Omega(Q, \tilde{Q}) \leq R - I(Q) \\ 0 & \Omega(Q, \tilde{Q}) > R - I(Q) \end{cases}. \tag{93}$$

Let

$$U(\tilde{Q}) \triangleq \max_{Q \in \mathcal{S}(\tilde{Q}_Y)} \begin{cases} \exp\left[-n(I(Q) - R)\right] & \Omega(Q, \tilde{Q}) \leq 0, \; I(Q) > R \\ 1 & I(Q) \leq R, \; \Omega(Q, \tilde{Q}) \leq R - I(Q) \\ 0 & \text{otherwise} \end{cases}. \tag{94}$$

Thus, substituting (93) in (90) and then in (79), we obtain, using the method of types,

$$A_1 \doteq e^{-nT} \sum_{\boldsymbol{x}_m} P(\boldsymbol{X}_m = \boldsymbol{x}_m) \sum_{\boldsymbol{y}} f(\boldsymbol{x}_m, \boldsymbol{y}) \cdot U(\tilde{Q}) \tag{95}$$

$$\doteq e^{-nT} \max_{\tilde{Q}} \exp\left[nH_{Y|X}(\tilde{Q})\right] \exp\left[nG(\tilde{Q})\right] U(\tilde{Q}). \tag{96}$$

Note that the condition:

$$\Omega(Q, \tilde{Q}) \leq R - I(Q) \tag{97}$$

in (94) is equivalent to

$$G(\tilde{Q}) \leq G(Q) - I(Q) + R + T. \tag{98}$$

Thus, we obtain that the exponent of A_1 is given by

$$\lim_{n\to\infty} \frac{1}{n} \log A_1 = -T - \min\left\{\tilde{E}_a(R, T, \xi), \tilde{E}_b(R, T, \xi)\right\}, \tag{99}$$

in which

$$\tilde{E}_a(R, T, \xi) \triangleq \min_{(Q,\tilde{Q})\in\mathcal{Q}} \left[-H_{Y|X}(\tilde{Q}) - G(\tilde{Q}) + I(Q) - R\right] \tag{100}$$

where $\mathcal{Q}$ is defined in (36), and

$$\tilde{E}_b(R, T, \xi) \triangleq \min_{\tilde{Q}\in\mathcal{L}} \left[-H_{Y|X}(\tilde{Q}) - G(\tilde{Q})\right] \tag{101}$$

where $\mathcal{L}$ is defined in (37). Now, we want to find the maximal ξ for which

$$-T - \tilde{E}_a(R, T, \xi) \leq 0, \tag{102}$$

$$-T - \tilde{E}_b(R, T, \xi) \leq 0. \tag{103}$$

For $\tilde{E}_a(R, T, \xi)$, substituting $G(Q)$, given in (36), in (100), we obtain

$$-\tilde{E}_a(R, T, \xi) - T = \max_{(Q,\tilde{Q})\in\mathcal{Q}} \left[H_{Y|X}(\tilde{Q}) + G(\tilde{Q}) - I(Q) + R\right] - T \tag{104}$$

$$= \max_{(Q,\tilde{Q})\in\mathcal{Q}} \left[H_{Y|X}(\tilde{Q}) + \max_{\theta}\left\{\xi E_1(\theta) + \mathbb{E}_{\tilde{Q}} \log W_\theta(Y|X)\right\} - I(Q) + R\right] \tag{105}$$

$$= \max_{\theta}\left\{\xi E_1(\theta) + \max_{(Q,\tilde{Q})\in\mathcal{Q}} \left\{H_{Y|X}(\tilde{Q}) - I(Q) + R + \mathbb{E}_{\tilde{Q}} \log W_\theta(Y|X)\right\}\right\} \tag{106}$$

$$= \max_{\theta}\left\{\xi E_1(\theta) - \min_{(Q,\tilde{Q})\in\mathcal{Q}} \left\{D(\tilde{Q}||P_X \times W_\theta) + I(Q) - R\right\}\right\}, \tag{107}$$

which is exactly the condition in (38). In a similar manner, one obtains

$$-\tilde{E}_b(R, T, \xi) - T = \max_{\theta}\left\{\xi E_1(\theta) - \min_{\tilde{Q}\in\mathcal{L}} D(\tilde{Q}||P_X \times W_\theta)\right\}, \tag{108}$$

which is exactly the condition in (39). ■

Proof of Corollary 1: In the following, we analyze the objective in (38) for any θ. Starting with the left term, $E_1(\theta)$, note that this is just the expression that was considered in [18, pp. 6450-6451, eqs. (64)-(73)]. For completeness, we present here the main steps in the simplification of this term to the BSC. We start with the analysis of $E_a(R,T)$ given in (13). First, note that

$$\mathbb{E}_{\hat{Q}} \log W_\theta(Y|X) - \mathbb{E}_Q \log W_\theta(Y|X) = \Big[Q\left(X \neq Y\right) - \hat{Q}\left(X \neq Y\right)\Big]\beta \tag{109}$$

where $\beta = \log\left[\left(1-\theta\right)/\theta\right]$. Thus, recalling (12), $E_1(\theta)$ takes the form

$$\min_{\tilde{Q}} \left\{ D(\tilde{Q}||P_X \times W_\theta) + \left| \min_{Q \in \hat{\mathcal{Q}}_{\text{BSC}}(\tilde{Q})} \left(-H_{X|Y}(Q) + \log 2 - R\right) \right|^+ \right\} \tag{110}$$

where

$$\hat{\mathcal{Q}}_{\text{BSC}}(\tilde{Q}) \triangleq \left\{ Q:\ Q_Y = \tilde{Q}_Y,\ Q\left(X \neq Y\right) \leq \tilde{Q}\left(X \neq Y\right) + \frac{T}{\beta} \right\}. \tag{111}$$

Now, note that

$$H_{X|Y}(Q) = H_{\mathcal{I}\{X\neq Y\}|Y}\left(Q\right) \leq H_{\mathcal{I}\{X\neq Y\}}(Q), \tag{112}$$

and thus

$$\begin{aligned}
&\min_{\tilde{Q}} \left\{ D(\tilde{Q}||P_X \times W_\theta) + \left| \min_{Q \in \hat{\mathcal{Q}}_{\text{BSC}}(\tilde{Q})} \left(-H_{X|Y}(Q) + \log 2 - R\right) \right|^+ \right\} \\
&\geq \min_{\tilde{Q}} \left\{ D(\tilde{Q}||P_X \times W_\theta) + \left| \min_{Q \in \hat{\mathcal{Q}}_{\text{BSC}}(\tilde{Q})} \left(-H_{\mathcal{I}\{X\neq Y\}}(Q) + \log 2 - R\right) \right|^+ \right\} &(113)\\
&= \min_{\tilde{q}} \left\{ D\left(\tilde{q}||\theta\right) + \left| \min_{q \leq \tilde{q} + T/\beta} \left(-h\left(q\right) + \log 2 - R\right) \right|^+ \right\} &(114)
\end{aligned}$$

where the last step follows since the minimizing $\tilde{Q}$ is such that $\tilde{Q}_X = P_X$ to obtain minimal $D(\tilde{Q}||P_X \times W_\theta)$, and it is easy to verify using convexity arguments that given $\tilde{Q}(X \neq Y) = \tilde{q}$ the divergence $D(\tilde{Q}||P_X \times W_\theta)$ is minimized for a symmetric $\tilde{Q}_{Y|X}$, namely,

$$\tilde{Q}_{Y|X}\left(y|x\right) = \begin{cases} \tilde{q} & x = y \\ & \\ 1-\tilde{q} & x \neq y \end{cases}, \tag{115}$$

for which $D(\tilde{Q}||P_X \times W_\theta) = D\left(\tilde{q}||\theta\right)$. Finally, it is evident that we have equality in (113) if we choose

$$Q_{Y|X}\left(y|x\right) = \begin{cases} q & x = y \\ & \\ 1-q & x \neq y \end{cases}, \tag{116}$$

and thus it is the minimizer. Next, we observe that $-h\left(q\right)$ is a decreasing function of q for $q \in [0, 1/2]$ and increasing for $q \in [1/2, 1]$. Thus,

$$\begin{aligned}
&\min_{\tilde{q}} \left\{ D\left(\tilde{q}||\theta\right) + \left| \min_{q \leq \tilde{q}+T/\beta} \left(-h\left(q\right) + \log 2 - R\right) \right|^{+} \right\} \\
&= \min_{\tilde{q}} \left\{ D\left(\tilde{q}||\theta\right) + \left| -h\left(\min\left\{ \frac{1}{2}, \tilde{q} + \frac{T}{\beta} \right\} \right) + \log 2 - R \right|^{+} \right\} && (117) \\
&= \min_{\tilde{q}} \left\{ D\left(\tilde{q}||\theta\right) - h\left(\min\left\{ \delta_{GV}\left(R\right), \tilde{q} + \frac{T}{\beta} \right\} \right) + \log 2 - R \right\} && (118) \\
&= \min_{\tilde{q} \in [\theta, \delta_{GV}(R) - T/\beta]} \left[D\left(\tilde{q}||\theta\right) - h\left(\tilde{q} + \frac{T}{\beta} \right) \right] + \log 2 - R && (119)
\end{aligned}$$

where the last step can be easily verified using monotonicity properties of the binary entropy and divergence [18, p. 6451 after eq. (72)]. Now, we analyze $E_b(R, T)$ given in (14). Note that there is no conceptual difference between $E_a(R, T)$ and $E_b(R, T)$, and it can be verified that the latter can be written as

$$\min_{\tilde{q} \in \hat{\mathcal{L}}_{\text{BSC}}} D\left(\tilde{q}||\theta\right) \tag{120}$$

where

$$\hat{\mathcal{L}}_{\text{BSC}} \triangleq \left\{ \tilde{q} : -\tilde{q} \cdot \beta \leq R + T + \max_{q:\ R \geq \log 2 - h(q)} \left[-q \cdot \beta + h(q) - \log 2 \right] \right\}. \tag{121}$$

Next, for any θ, consider the right term in objective of (38). Note that the only difference between the left and the right terms in (38) is just the inner minimization region. Accordingly, the right term takes the form

$$\min_{\tilde{Q}} \left\{ D(\tilde{Q}||P_X \times W_\theta) + \left| \min_{Q \in \mathcal{Q}_{\text{BSC}}(\tilde{Q})} \left(-H_{X|Y}(Q) + \log 2 - R\right) \right|^{+} \right\} \tag{122}$$

where

$$\begin{aligned}
\mathcal{Q}_{\text{BSC}}(\tilde{Q}) \triangleq \Big\{ Q :\ & Q_Y = \tilde{Q}_Y,\ \max_{\theta'} \left\{ \xi E_1(\theta') - \beta(\theta') \tilde{Q}\left(X \neq Y\right) + \log\left(1 - \theta'\right) \right\} \\
& - \max_{\theta'} \left\{ \xi E_1(\theta') - \beta(\theta') Q\left(X \neq Y\right) + \log\left(1 - \theta'\right) \right\} - T \leq 0 \Big\}. && (123)
\end{aligned}$$

Let $\tilde{E}_1\left(\theta\right) \triangleq E_1\left(\theta\right) + \log(1-\theta)/\xi$. Then, using exactly the same steps as before, we get

$$\begin{aligned}
&\min_{\tilde{Q}} \left\{ D(\tilde{Q}||P_X \times W_\theta) + \left| \min_{Q \in \mathcal{Q}_{\text{BSC}}(\tilde{Q})} \left(-H_{X|Y}\left(Q\right) + \log 2 - R\right) \right|^{+} \right\} \\
&\geq \min_{\tilde{Q}} \left\{ D(\tilde{Q}||P_X \times W_\theta) + \left| \min_{Q \in \mathcal{Q}_{\text{BSC}}(\tilde{Q})} \left(-H_{\mathcal{I}\{X \neq Y\}}(Q) + \log 2 - R\right) \right|^{+} \right\} && (124)
\end{aligned}$$

$$= \min_{\tilde{q}} \left\{ D\left(\tilde{q}||\theta\right) + \left| \min_{q \in \tilde{\mathcal{Q}}_{\text{BSC}}(\tilde{q})} \left(-h\left(q\right) + \log 2 - R\right) \right|^{+} \right\}, \tag{125}$$

and equality can be achieved choosing Q to be symmetric, as before, and

$$\tilde{\mathcal{Q}}_{\text{BSC}}(\tilde{q}) \triangleq \left\{ q : \max_{\theta'} \left\{ \xi \tilde{E}_1(\theta') - \beta(\theta') \cdot \tilde{q} \right\} - \max_{\theta'} \left\{ \xi \tilde{E}_1(\theta') - \beta(\theta') \cdot q \right\} - T \leq 0 \right\} \tag{126}$$

$$= \left\{ q : -\max_{\theta'} \left\{ \xi \tilde{E}_1(\theta') - \beta(\theta') \cdot q \right\} \leq T - \max_{\theta'} \left\{ \xi \tilde{E}_1(\theta') - \beta(\theta') \cdot \tilde{q} \right\} \right\}. \tag{127}$$

Next, we simplify the set $\tilde{\mathcal{Q}}_{\text{BSC}}(\tilde{q})$. The constraint on q in the definition of $\tilde{\mathcal{Q}}_{\text{BSC}}(\tilde{q})$, is equivalent to demanding that there exist some $\theta' \in \Theta$ such that the following holds

$$\beta(\theta')q - \xi \tilde{E}_1(\theta') \leq T - \max_{\theta''} \left\{ \xi \tilde{E}_1(\theta'') - \beta(\theta'') \cdot \tilde{q} \right\}, \tag{128}$$

or equivalently

$$\beta(\theta')q \leq \xi \tilde{E}_1(\theta') + T - \max_{\theta''} \left\{ \xi \tilde{E}_1(\theta'') - \beta(\theta'') \cdot \tilde{q} \right\}. \tag{129}$$

Now, note that $\beta(\theta') \geq 0$ if and only if $\theta' \leq 1/2$. Accordingly, this means that, in terms of q, $\tilde{\mathcal{Q}}_{\text{BSC}}(\tilde{q})$ is equivalent to $q \leq q_1^*$ or $q \geq q_2^*$, where q_1^* and q_2^* are given in (41) and (42), respectively. Consequently,

$$\min_{\tilde{q}} \left\{ D\left(\tilde{q}||\theta\right) + \left| \min_{q \in \tilde{\mathcal{Q}}_{\text{BSC}}(\tilde{q})} \left(-h\left(q\right) + \log 2 - R\right) \right|^{+} \right\} = \min_{\tilde{q}} \left\{ D\left(\tilde{q}||\theta\right) + \left|\left(-g\left(q_1^*, q_2^*\right) + \log 2 - R\right)\right|^{+} \right\} \tag{130}$$

where $g\left(q_1^*, q_2^*\right)$ is defined in (43). Finally, we consider the right term in (39). Using the same steps as above we obtain that

$$\min_{\tilde{Q} \in \mathcal{L}} D(\tilde{Q}||P_X \times W_\theta) = \min_{\tilde{q} \in \mathcal{L}_{\text{BSC}}} D\left(\tilde{q}||\theta\right) \tag{131}$$

in which

$$\mathcal{L}_{\text{BSC}} \triangleq \left\{ \tilde{q} : \max_{\theta} \left[\xi E_1(\theta) - \tilde{q} \cdot \beta + \log \theta\right] \leq R + T \right.$$
$$\left. + \max_{q:\ R \geq \log 2 - h(q)} \left\{ \max_{\theta} \left[\xi E_1(\theta) - q \cdot \beta + \log \theta\right] + h(q) - \log 2 \right\} \right\} \tag{132}$$

$$= \left\{ \tilde{q} : \max_{\theta} \left[\xi E_1(R, T, \theta) - \tilde{q} \cdot \beta(\theta) + \log \theta\right] \leq R + T \right.$$
$$\left. + \max_{\theta} \left[\xi E_1(R, T, \theta) - \max\left\{\theta, \delta_{\text{GV}}(R)\right\} \cdot \beta(\theta) + \log \theta + h(\max\left\{\theta, \delta_{\text{GV}}(R)\right\}) - \log 2\right] \right\} \tag{133}$$

where the last step follows from the fact that the maximizer q in the optimization problem in (132) is given by $\max\left\{\theta, \delta_{\text{GV}}(R)\right\}$. ■

APPENDIX A
PROOF OF LEMMA 1

For the sake of this proof, we will explicitly designate the dependence on T, and denote the decoder in (4)-(5), with parameter T, by $\mathcal{R}^*(T)$. Similarly, we will denote the value of (7) as $\Gamma(\mathcal{C},\mathcal{R},T)$. As we have mentioned, the decoder minimizing $\Gamma(\mathcal{C},\mathcal{R},T)$ can be easily seen to be given by $\mathcal{R}^*(T)$. Now, assume conversely, that the exponents associated with $\mathbb{E}[\Gamma(\mathcal{C},\mathcal{R}^*(T),T)]$ satisfy

$$E_2(R,T) < T + E_1(R,T). \tag{A.1}$$

The opposite case, where the inequality in (A.1) is reversed, can be handled analogously. Accordingly, this means that in the exponential scale, we have

$$\mathbb{E}[\Gamma(\mathcal{C},\mathcal{R}^*(T),T)] \doteq e^{-nE_2(R,T)}. \tag{A.2}$$

Now, it is evident that $E_1(R,T)$ is a monotonically decreasing function of T (allowing more erasures increases $\overline{\Pr}\{\mathcal{E}_1\}$), and $E_2(R,T)$ is a monotonically increasing function of T (allowing more erasures decreases $\overline{\Pr}\{\mathcal{E}_2\}$) [18]. Now, due to the fact that $E_1(R,T)$ and $E_2(R,T)$ are continuous functions of T [18, eqs. (23) and (31)], without loss of essential generality, there exists $\epsilon > 0$ and $\delta_1 \geq 0, \delta_2 > 0$ such that

$$E_1(R,T+\epsilon) = E_1(R,T) - \delta_1 \tag{A.3}$$

and

$$E_2(R,T+\epsilon) = E_2(R,T) + \delta_2 \tag{A.4}$$

yet

$$E_2(R,T+\epsilon) < T + E_1(R,T+\epsilon). \tag{A.5}$$

Note that since it is not guaranteed that $E_1(R,T)$ or $E_2(R,T)$ are strictly monotonic, it might be the case that $\delta_2 = 0$ too, i.e., regions of plateau. Accordingly, there are several cases to consider. First, if just $E_1(R,T)$ is within a plateau region, then everything go along without any problem since $\delta_1 = 0$ but $\delta_2 > 0$. Secondly, if just $E_2(R,T)$ is within a plateau region, then we claim that this contradicts the optimality of Forney's decoder. Indeed, in this case, if we increase T by some small $\epsilon > 0$ (such that $E_2(R,T+\epsilon)$ is within the plateau), we obtain a decoder with exponents $E_2(R,T+\epsilon) = E_2(R,T)$ and $E_1(R,T+\epsilon) < E_1(R,T)$, and yet, due to continuity, $E_2(R,T) < E_1(R,T+\epsilon)$. Thus, we obtained that the optimal decoder $\mathcal{R}^*(T+\epsilon)$ has the same performance as $\mathcal{R}^*(T)$, in terms of $\mathbb{E}[\Gamma(\mathcal{C},\mathcal{R}^*(T),T)]$,

but with worse $\overline{\Pr}\{\mathcal{E}_1\}$, which means not the best tradeoff between $\overline{\Pr}\{\mathcal{E}_1\}$ and $\overline{\Pr}\{\mathcal{E}_2\}$, and thus contradicting the optimality of Forney's decoder at $T+\epsilon$. Finally, if both exponents are within a region of plateau, we can simply vary T until we leave this region, and thus we can assume that $\delta_2 > 0$. To conclude, we obtained that

$$\mathbb{E}[\Gamma(\mathcal{C}, \mathcal{R}(T+\epsilon), T)] \doteq e^{-nE_2(R,T+\epsilon)} \tag{A.6}$$

$$\dot{<} e^{-nE_2(R,T)} \doteq \mathbb{E}[\Gamma(\mathcal{C}, \mathcal{R}(T), T)] \tag{A.7}$$

which contradicts the property that $\mathcal{R}^*(T)$ is the minimizer of $\Gamma(\mathcal{C}, \mathcal{R}, T)$.

REFERENCES

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

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

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

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

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

[6] ——, "Universal decoding for arbitrary channels relative to a given class of decoding metrics," *IEEE Trans. on Inf. Theory*, vol. 59, no. 9, pp. 5566–576, Sep. 2013.

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

[8] G. D. Forney, Jr., "Exponential error bounds for erasure, list, and decision feedback schemes," *IEEE Trans. Inf. Theory*, vol. 14, no. 2, pp. 206–220, 1968.

[9] M. V. Burnashev, "Data transmission over a discrete channel with feedback," *Problems of Information Transmission*, pp. 250–265, 1976.

[10] N. Shulman, "*Communication over an unknown channel via common broadcasting*," Ph.D. dissertation, Tel-Aviv University, 2003, http://www.eng.tau.ac.il/ shulman/papers/Nadav_PhD.pdf.

[11] S. Draper, B. J. Frey, and F. R. Kschischang, "Rateless coding for non-ergodic channels with decoder channel state information," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 4119–4133, 2009.

[12] U. Erez, G. W. Wornell, and M. D. Trott, "Rateless space-time coding," in *Proc. ISIT 2005*, Sep. 2005, pp. 1937–1941.

[13] J. Jiang and K. R. Narayanan, "Multilevel coding for channels with non-uniform inputs and rateless transmission over the bsc," in *Proc. ISIT 2006*, 2006, pp. 518–522.

[14] A. Tchamkerten and E. I. Telatar, "Variable length codes over unknown channels," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 2126–2145, 2006.

[15] N. Merhav and M. Feder, "Minimax universal decoding with an erasure option," *IEEE Trans. Inf. Theory*, vol. 53, no. 5, pp. 1664–1675, May. 2007.

[16] P. Moulin, "A Neyman-Pearson approach to universal erasure and list decoding," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4462–4478, 2009.

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

[18] A. Somekh-Baruch and N. Merhav, "Exact random coding exponents for erasure decoding," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6444–6454, 2011.

[19] N. Merhav, "Erasure/list exponents for Slepian-Wolf decoding," *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4463–4471, Aug. 2014.