IRWIN AND JOAN JACOBS
CENTER FOR COMMUNICATION AND INFORMATION TECHNOLOGIES

Exact Random Coding Exponents for Erasure Decoding

Anelia Somekh-Baruch and Neri Merhav

CCIT Report #771
August 2010

Electronics
Computers
Communications

DEPARTMENT OF ELECTRICAL ENGINEERING
TECHNION - ISRAEL INSTITUTE OF TECHNOLOGY, HAIFA 32000, ISRAEL

CCIT REPORT #771 August 2010

Exact Random Coding Exponents for Erasure Decoding

Anelia Somekh-Baruch and Neri Merhav

***Abstract*—Random coding of channel decoding with an erasure option is studied. By analyzing the large deviations behavior of the code ensemble, we obtain *exact* single-letter formulas for the error exponents in lieu of Forney's lower bounds. The analysis technique we use is based on an enhancement and specialization of tools for assessing the moments of certain distance enumerators. We specialize our results to the setup of the binary symmetric channel case with uniform random coding distribution and derive an *explicit* expression for the error exponent which, unlike Forney's bounds, does not involve optimization over two parameters. We also establish the fact that for this setup, the difference between the exact error exponent corresponding to the probability of undetected decoding error and the exponent corresponding to the erasure event is equal to the threshold parameter. Numerical calculations indicate that for this setup, as well as for a Z-channel, Forney's bound coincides with the exact random coding exponent.**

***Index Terms*—random coding, erasure, list, error exponent, distance enumerator**

I. Introduction

In [1], Forney derived lower bounds on the random coding exponents associated with decoding rules that allow for erasure and list decoding (see also later related studies[4] -[9]). The channel model he considered was a single user discrete memoryless channel (DMC), where a codebook of block length n is randomly drawn with i.i.d. codewords having i.i.d. symbols. When erasure is concerned, the decoder may fully decode the message, or, decide to declare that an erasure has occurred. An optimum tradeoff between the probability of erasure and the probability of undetected decoding error was investigated. This tradeoff is optimally controlled by a threshold parameter T of the function e^{nT} to which one compares the ratio between the likelihood of each hypothesized message and the sum of likelihoods of all other messages. If this ratio exceeds e^{nT} for some message, a decision is made in favor of that message, otherwise, an erasure is declared. Forney's main result in [1] is a single-letter lower bound, $E_1(R,T)$, to the exponent of the probability of the event $\mathcal{E}_1$ of not making the correct decision, namely, either erasing or making the wrong decision, and a single-letter lower bound, $E_2(R,T)$, to the exponent of the probability of the event $\mathcal{E}_2$ of undetected error.

In [2, Th. 5.11], Csiszár and Körner derived universally achievable error exponents for a decoder with an erasure option for DMCs. These error exponents were obtained by analyzing a decoder which generalizes the MMI decoder for constant composition (CC) codes. Unlike Forneys decoder, no optimality claims were made for this decoder, but, in [3, Sec. 4.4.3] Telatar stated that these bounds are essentially the same as those in [1].

Inspired by a statistical–mechanical point of view on random code ensembles (offered in [10] and further elaborated on in [12]), Merhav [11] applied a different technique to derive a lower bound to the exponents of the probabilities of $\mathcal{E}_1, \mathcal{E}_2$ by assessing the moments of certain distance enumerators. This approach, which also proved fruitful in several other applications (see [13], [14], [15]), resulted in a bound that is at least as tight as Forney's bound. It is shown in [11] that under certain symmetry conditions (that often hold) on the random coding distribution and the channel, the resulting bound is also simpler in the sense that there is only one parameter to optimize rather than two. Moreover, this optimization can be carried out in closed form at least in some special cases like the binary symmetric channel (BSC). It is not clear though, whether the bounds of [11] are strictly tighter than those of Forney.

In this paper, we use the approach of distance enumerators to tackle again the problem of random of channel decoding with an erasure option. Unlike the approach of [1] and [11], our starting point is not a Gallager-type bound [16] on the probability of error, but rather the *exact* expression. This approach results in single-letter expressions for the exact exponential behavior of the probabilities of the events $\mathcal{E}_1, \mathcal{E}_2$ when random coding is used. So far, we have not been able to determine analytically whether our results coincide with Forney's bounds, i.e., we cannot say whether Forney's bounds are tight or not, but the tightness of the our expressions is guaranteed. While our analysis pertains to the ensemble of codes where each symbol of each codeword is drawn i.i.d. (mainly in order to enable a fair comparison to [1]), our technique can easily be used for other ensembles, like the ensemble where each codeword is drawn independently according to the uniform distribution within a given type class. For the case of the BSC with uniform random coding distribution we have conducted several numerical calculations, which indicate that Forney's bound coincides with the exact random coding exponent.

The outline of this paper is as follows. In Section II, we present notation conventions and in Section III, we give some necessary background in more detail. Section IV is devoted to a description of the main results. In the following sections, V-IX, we provide detailed derivations of the main results: in Section V, we derive the exact expression for the error exponent corresponding to the probability of $\mathcal{E}_1$, and in Sections VI and VII, we study two special cases of

A. Somekh-Baruch is with the School of Engineering, Bar-Ilan University, Ramat-Gan, Israel 52900. E-mail: anelia.somekhbaruch@gmail.com

N. Merhav is with the Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa, 32000, Israel. E-mail: merhav@ee.technion.ac.il.

channels. Section VIII is dedicated to the derivation of the exact expression for the error exponent corresponding to the probability of $\mathcal{E}_2$, and in Section IX, we specialize the proof to the case of the BSC with the uniform random coding distribution.

II. Notation

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., $\mathcal{X}^n$.

Sources and channels will be denoted generically by the letter P or Q. Information theoretic quantities, such as 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. When we wish to emphasize the dependence of the entropy on a certain underlying probability distribution, say Q, we subscript it by Q, i.e., use notations like $H_Q(X)$, $H_Q(X|Y)$, etc. The divergence (or, Kullback - Liebler distance) between two probability measures Q and P will be denoted by $D(Q\|P)$, and when there is a need to make a distinction between P and Q as joint distributions of (X,Y) as opposed to the corresponding marginal distributions of, say, X, we will use subscripts to avoid ambiguity, that is, we shall use the notations $D(Q_{XY}\|P_{XY})$ and $D(Q_X\|P_X)$. For two number $0 \le q,p \le 1$, $D(q\|p)$ will stand for the divergence between the binary measures $\{q, 1-q\}$ and $\{p, 1-p\}$. The expectation operator will be denoted by $\boldsymbol{E}\{\cdot\}$, and once again, when we wish to make the dependence on the underlying distribution Q clear, we denote it by $\boldsymbol{E}_Q\{\cdot\}$.

The cardinality of a finite set A will be denoted by $|A|$. The indicator function of an event $\mathcal{E}$ will be denoted by $1\{\mathcal{E}\}$. For a given sequence $\boldsymbol{y} \in \mathcal{Y}^n$, $\mathcal{Y}$ being a finite alphabet, $\hat{P}_{\boldsymbol{y}}$ will denote the empirical distribution on $\mathcal{Y}$ extracted from $\boldsymbol{y}$, in other words, $\hat{P}_{\boldsymbol{y}}$ is the vector $\{\hat{P}_{\boldsymbol{y}}(y), y \in \mathcal{Y}\}$, where $\hat{P}_{\boldsymbol{y}}(y)$ is the relative frequency of the letter y in the vector $\boldsymbol{y}$. 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., $\frac{1}{n}\ln\frac{a_n}{b_n} \to 0$ as $n \to \infty$. Similarly, $a_n \dot{\le} b_n$ means that $\limsup_n \frac{1}{n}\ln\frac{a_n}{b_n} \le 0$, and so on. Another notation that we shall use is that for a real number x, $|x|^+ = \max\{0, x\}$.

III. Preliminaries

Consider a 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 of conditional probability distributions

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

where for $i=1$, $(y_1,...,y_{i-1})$ is understood as the null string. A rate-R block code of length n consists of $M = e^{nR}$ n-vectors $\boldsymbol{x}_m$, $m = 1,2,...,M$, which represent M different messages. We will assume that all possible messages are a-priori equiprobable, i.e., $P(m) = 1/M$ for all $m = 1,2,...,M$. A decoder with an erasure option is a partition of $\mathcal{Y}^n$ into $(M+1)$ regions, $\mathcal{R}_0, \mathcal{R}_1, ..., \mathcal{R}_M$. Such a decoder works as follows: If $\boldsymbol{y}$ falls into $R_m, m = 1,2,...,M$, 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 $C = \{\boldsymbol{x}_1,...,\boldsymbol{x}_M\}$ and a decoder $\mathcal{R} = (\mathcal{R}_0, \mathcal{R}_1, ..., R_m)$, let us now define two 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:

$$\Pr\{\mathcal{E}_1\} = \sum_{m=1}^{M} \sum_{\boldsymbol{y}\in\mathcal{R}_m^c} P(\boldsymbol{x}_m, \boldsymbol{y}) = \frac{1}{M}\sum_{m=1}^{M} \sum_{\boldsymbol{y}\in\mathcal{R}_m^c} P(\boldsymbol{y}|\boldsymbol{x}_m) \tag{2}$$

$$\begin{aligned}\Pr\{\mathcal{E}_2\} &= \sum_{m=1}^{M} \sum_{\boldsymbol{y}\in\mathcal{R}_m} \sum_{m'\ne m} P(\boldsymbol{x}_{m'}, \boldsymbol{y}) \\ &= \frac{1}{M}\sum_{m=1}^{M} \sum_{\boldsymbol{y}\in\mathcal{R}_m} \sum_{m'\ne m} P(\boldsymbol{y}|\boldsymbol{x}_{m'})\end{aligned} \tag{3}$$

$$\Pr\{\mathcal{R}_0\} = \Pr\{\mathcal{E}_1\} - \Pr\{\mathcal{E}_2\}. \tag{4}$$

Forney [1] shows, using the Neyman-Pearson Theorem, that the best tradeoff between $\Pr\{\mathcal{E}_1\}$ and $\Pr\{\mathcal{E}_2\}$ is attained by the decoder $\mathcal{R}^* = (\mathcal{R}_0^*, \mathcal{R}_1^*, ..., \mathcal{R}_M^*)$ defined by

$$\mathcal{R}_m^* = \left\{\boldsymbol{y}:\ \frac{P(\boldsymbol{y}|\boldsymbol{x}_m)}{\sum_{m'\ne m} P(\boldsymbol{y}|\boldsymbol{x}_{m'})} \ge e^{nT}\right\}, m = 1,2,...,M \tag{5}$$

$$\mathcal{R}_0^* = \bigcup_{m=1}^{M} (\mathcal{R}_m^*)^c, \tag{6}$$

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

Define the error exponents $e_i(R,T), i = 1,2$, as the exponents associated with the average probabilities of error $\overline{\Pr}\{\mathcal{E}_i\}, i = 1,2$, where the average is taken with respect to (w.r.t.) the ensemble of randomly selected codes, drawn independently according to an i.i.d. distribution $P(\boldsymbol{x}) = \prod_{i=1}^n P(x_i)$, that is,

$$e_i(R,T) = \limsup_{n\to\infty}\left[-\frac{1}{n}\ln\overline{\Pr}\{\mathcal{E}_i\}\right], \quad i = 1,2. \tag{7}$$

Forney derives lower bounds, $E_1(R,T)$ and $E_2(R,T)$, to $e_1(R,T)$ and $e_2(R,T)$, (or, upper bounds to the average probabilities of error), respectively, given by

$$E_1(R,T) = \max_{0\le s\le\rho\le 1}\left[E_0(s,\rho) - \rho R - sT\right], \tag{8}$$

where

$$E_0(s,\rho) = \tag{9}$$
$$-\ln\left[\sum_y\left(\sum_x P(x)P^{1-s}(y|x)\right)\left(\sum_{x'}P(x')P^{s/\rho}(y|x')\right)^{\rho}\right]$$

and

$$E_2(R,T) = E_1(R,T) + T. \tag{10}$$

Merhav [11] established tighter (though not necessarily strictly tighter) upper bounds to $e_1(R,T)$ and $e_2(R,T)$ denoted $E_1^*(R,T)$ and $E_2^*(R,T)$. These bounds take on a very simple form under the following condition:

Condition 1: The random coding distribution $\{P(x), x \in \mathcal{X}\}$ and the channel transition matrix $\{P(y|x), x \in \mathcal{X}, y \in \mathcal{Y}\}$ are such that for every real s,

$$\gamma_y(s) \stackrel{\Delta}{=} -\ln\left[\sum_{x\in\mathcal{X}} P(x)P^s(y|x)\right] \tag{11}$$

is independent of y.

When the condition holds, $\gamma_y(s)$ will be denoted by $\gamma(s)$.

Under Condition 1, the bounds to the exponents are

$$E_1^*(R,T) = \sup_{s\geq 0}\left[\Lambda(R,s) + \gamma(1-s) - sT - \ln|\mathcal{Y}|\right] \tag{12}$$

where

$$\Lambda(R,s) = \begin{cases} \gamma(s) - R, & s \geq s_R \\ s\gamma'(s_R), & s < s_R, \end{cases} \tag{13}$$

$\gamma'(s) = \frac{d\gamma(s)}{ds}$ and where s_R is the solution to the equation

$$\gamma(s) - s\gamma'(s) = R. \tag{14}$$

Also, similarly as in (10),

$$E_2^*(R,T) = E_1^*(R,T) + T. \tag{15}$$

Let $\delta_{GV}(R)$ denote the normalized Gilbert-Varshamov (GV) distance, i.e., the smaller solution, δ, to the equation

$$h(\delta) = \ln 2 - R, \tag{16}$$

where $h(\delta) = -\delta\ln(\delta) - (1-\delta)\ln(1-\delta)$ is the binary entropy function.

For the BSC with uniform[1] random coding distribution, the upper bound is given by

$$E_1^*(R,T) \stackrel{\Delta}{=} \tag{17}$$
$$\sup_{s\geq 0}\left(\mu(s,R) + s\ln\frac{1}{1-p} - \ln[p^{1-s} + (1-p)^{1-s}] - sT\right)$$

where

$$\mu(s,R) = \begin{cases} \mu_0(s,R) & s \geq s_R \\ \beta s\delta_{GV}(R) & s < s_R \end{cases}, \tag{18}$$

$$\mu_0(s,R) = s\ln(1-p) - \ln[p^s + (1-p)^s] + \ln 2 - R, \tag{19}$$

and

$$\beta \;=\; \ln\frac{1-p}{p}. \tag{20}$$

It is noted that the optimal s in (17) has an explicit expression given in [11].

[1]By the term *uniform random coding distribution, we mean uniform over* $\{0,1\}^n$.

IV. Main Results

The main results in this paper are stated in Theorems 1 and 2, establishing exact expressions for the random coding error exponents $e_1(R,T)$ and $e_2(R,T)$ for the general DMC.

For a given probability distribution Q on $\mathcal{X}\times\mathcal{Y}$ define

$$\begin{aligned} K(Q,R) &= \boldsymbol{E}_Q \ln\frac{1}{P(X)} - H_Q(X|Y) - R \\ &= D(Q_X\|P_X) + I_Q(X;Y) - R, \end{aligned} \tag{21}$$

and for a given probability distribution Q_Y on $\mathcal{Y}$, define

$$\mathcal{G}_R(Q_Y) \stackrel{\Delta}{=} \left\{Q_{X|Y} : K(Q,R) \leq 0\right\}, \tag{22}$$

where Q is the probability distribution $Q_Y \times Q_{X|Y}$.

Theorem 1: The error exponent $e_1(R,T)$ is given by

$$e_1(R,T) = \min_{\tilde{Q}}\left[D(\tilde{Q}_{XY}||P_{XY}) + \min_{Q_{X|Y}} K(Q,R)\right], \tag{23}$$

where $\tilde{Q}$ is a probability distribution on $\mathcal{X}\times\mathcal{Y}$, Q and $\tilde{Q}$ share the same marginal pmf of Y, that is, $Q = \tilde{Q}_Y \times Q_{X|Y}$, and the inner minimization is over $Q_{X|Y} \in \mathcal{G}_R^c(\tilde{Q}_Y)$ such that

$$\Omega(\tilde{Q},Q,T) \stackrel{\Delta}{=} \boldsymbol{E}_{\tilde{Q}}\ln P(Y|X) + \boldsymbol{E}_Q\ln\frac{1}{P(Y|X)} - T \leq 0. \tag{24}$$

Corollary 1: Under Condition 1 (see (11)) the error exponent $e_1(R,T)$ is given by

$$\begin{aligned} &e_1(R,T) \\ &= \min_{\tilde{Q}:\ \xi(\tilde{Q},T)\leq\gamma'(s_R)}[D(\tilde{Q}_{XY}\|P_{XY}) + \psi(s(\xi(\tilde{Q},T)))] - R, \end{aligned} \tag{25}$$

where

$$\xi(\tilde{Q},T) = T + \boldsymbol{E}_{\tilde{Q}}\ln\frac{1}{P(Y|X)}, \tag{26}$$

$s(\xi)$ is the solution of the equation $\gamma'(s) = \xi$, s_R is defined as in (14), and

$$\psi(s) \stackrel{\Delta}{=} \gamma(s) - s\gamma'(s). \tag{27}$$

Corollary 2: For the BSC with uniform random coding distribution, if $R \geq \ln 2 - h(p + T/\beta)$, $e_1(R,T) = 0$ and otherwise

$$\begin{aligned} &e_1(R,T) = \\ &\min_{q\in[p,\delta_{GV}(R)-T/\beta]}[D(q\|p) - h(q+T/\beta)] + \ln 2 - R, \end{aligned} \tag{28}$$

where β is defined in (20).

It is easy to verify, by equating the derivative of $D(q||p) - h(q+T/\beta)$ to zero, that the minimizing q is either a boundary point of the interval $[p, \delta_{GV}(R) - T/\beta]$, or q that satisfies the quadratic equation

$$q^2(1-e^{-\beta}) + q\left(\frac{T}{\beta}(1-e^{-\beta}) + 2e^{-\beta}\right) - e^{-\beta} = 0, \tag{29}$$

that is, denoting $\tau \triangleq e^{-\beta}$

$$q = \frac{-\left(\frac{T}{\beta}(1-\tau)+2\tau\right) + \sqrt{\left(\frac{T}{\beta}(1-\tau)+2\tau\right)^2 + 4(1-\tau)\tau}}{2(1-\tau)}. \tag{30}$$

We note that in the BSC case, the exact exponent $e_1(R,T)$ has a surprisingly simple explicit expression (28) in the sense that there is an optimization over one parameter only and that its optimum value is found in closed form.

Theorem 2: The error exponent $e_2(R,T)$ is given by

$$e_2(R,T) = \min_{Q_Y}\Bigg[D(Q_Y\|P_Y) + \min_{\Theta\le\Theta_0(Q_Y)}\{E_{\mathcal{A}}(Q_Y,\Theta) + E_{\mathcal{B}}(Q_Y,\Theta)\} - R\Bigg], \tag{31}$$

where

$$\Theta_0(Q_Y) = \min_{Q_{X|Y}\in\mathcal{G}_R(Q_Y)}\left[\boldsymbol{E}_Q \ln\frac{1}{P(X,Y)} - H_Q(X|Y)\right] - R, \tag{32}$$

with $Q = Q_Y \times Q_{X|Y}$,

$$E_{\mathcal{A}}(Q_Y,\Theta) = \min_{Q_{X|Y}:\ \boldsymbol{E}_Q \ln P(Y|X) = T-\Theta}\left[\boldsymbol{E}_Q \ln\frac{1}{P(X)} - H_Q(X|Y)\right], \tag{33}$$

$$E_{\mathcal{B}}(Q_Y,\Theta) = \min_{Q_{X|Y}:\ \boldsymbol{E}_Q \ln P(Y|X) \le -\Theta}\left[\boldsymbol{E}_Q \ln\frac{1}{P(X|Y)} - H_Q(X|Y)\right]. \tag{34}$$

Corollary 3: For the BSC with uniform random coding distribution,

$$e_1(R,T) = e_2(R,T) + T. \tag{35}$$

Discussion: The relation $e_1(R,T) = e_2(R,T)+T$, which is proved to hold in the case of the BSC with uniform random coding distribution, is not surprising. The intuition behind it can be explained as follows: Recall that the decision rule (5) was chosen to minimize the tradeoff between $\Pr\{\mathcal{E}_1\}$ and $\Pr\{\mathcal{E}_2\}$, and consider an equivalent problem of minimizing a Lagrangian which is a linear combination of $\Pr\{\mathcal{E}_1\}$ and $\Pr\{\mathcal{E}_2\}$, where the Lagrange multiplier is e^{nT}. Had $e_1(R,T)$ been different from $e_2(R,T)+T$, the exponents of these two terms would differ, and hence one could improve the overall exponent by changing their balance.

While our results imply that $e_i(R,T) \ge E_i(R,T)$ and $e_i(R,T) \ge E_i^*(R,T)$, $i = 1,2$, we have not been able to determine analytically whether or not there are cases in which at least one of these inequalities is a strong one. If such cases will be found, then the conclusion will be that we have strictly improved on the results of [1] or [11] or both. If not, then the conclusion would be that the exponents of [1] and [11] are tight, a fact which was not determined unequivocally before. In either case, the tools proposed in this paper provide us with a yardstick to determine the tightness of the results in [1] and [11]. As mentioned earlier, we have conducted a numerical study for the case of the BSC with uniform random coding distribution which indicates that in this case, $e_i(R,T)$ appearing in (28) is equal to $E_i^*(R,T)$ (see (12)) and to $E_i(R,T)$. We have shown analytically for this case that the lowest rate for which $e_1(R,T)=0$ is equal to that of $E_1^*(R,T)$. A numerical study of the Z-channel (i.e., a binary channel for which $P(Y=0|X=1)=0$) indicates that $e_i(R,T) = E_i(R,T)$ also for this case. In light of these two examples, we conjecture that Forney's exponent is tight in general.

V. Proof of Theorem 1

The probability of error given that the message that was sent is m, averaged over the codebooks is given by

$$\overline{\Pr}(\mathcal{E}_1|m) = \sum_{\boldsymbol{x}_m} P(\boldsymbol{x}_m)\sum_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x}_m) \cdot \Pr\left\{\sum_{m'\ne m} P(\boldsymbol{y}|\boldsymbol{x}_{m'}) > P(\boldsymbol{y}|\boldsymbol{x}_m)e^{-nT}\,\middle|\,\boldsymbol{x}_m,\boldsymbol{y}\right\}. \tag{36}$$

Next, let Q be an empirical probability distribution defined on $\mathcal{X}\times\mathcal{Y}$ and let $N_{\boldsymbol{y}}(Q)$ denote the number of codewords (excluding $\boldsymbol{x}_m$) whose joint empirical probability distribution with $\boldsymbol{y}$ is Q, and denote $f(\boldsymbol{y},Q) \triangleq N_{\boldsymbol{y}}(Q)e^{n\boldsymbol{E}_Q \ln P(Y|X)}$ then

$$\begin{aligned}
&\Pr\left\{\sum_{m'\ne m} P(\boldsymbol{y}|\boldsymbol{x}_{m'}) > P(\boldsymbol{y}|\boldsymbol{x}_m)e^{-nT}\,\middle|\,\boldsymbol{x}_m,\boldsymbol{y}\right\}\\
&= \Pr\left\{\sum_{Q} N_{\boldsymbol{y}}(Q)e^{n\boldsymbol{E}_Q \ln P(Y|X)} > P(\boldsymbol{y}|\boldsymbol{x}_m)e^{-nT}\,\middle|\,\boldsymbol{x}_m,\boldsymbol{y}\right\}\\
&= \Pr\left\{\sum_{Q} f(\boldsymbol{y},Q) > P(\boldsymbol{y}|\boldsymbol{x}_m)e^{-nT}\,\middle|\,\boldsymbol{x}_m,\boldsymbol{y}\right\}\\
&\overset{(a)}{\doteq} \Pr\left\{\max_{Q} f(\boldsymbol{y},Q) > P(\boldsymbol{y}|\boldsymbol{x}_m)e^{-nT}\,\middle|\,\boldsymbol{x}_m,\boldsymbol{y}\right\}\\
&= \Pr\left\{\bigcup_{Q}\{f(\boldsymbol{y},Q) > P(\boldsymbol{y}|\boldsymbol{x}_m)e^{-nT}\}\,\middle|\,\boldsymbol{x}_m,\boldsymbol{y}\right\}\\
&\doteq \sum_{Q}\Pr\left\{f(\boldsymbol{y},Q) > P(\boldsymbol{y}|\boldsymbol{x}_m)e^{-nT}\,\middle|\,\boldsymbol{x}_m,\boldsymbol{y}\right\}\\
&\doteq \max_{Q}\Pr\left\{N_{\boldsymbol{y}}(Q) > P(\boldsymbol{y}|\boldsymbol{x}_m)e^{-n[\boldsymbol{E}_Q \ln P(Y|X)+T]}\,\middle|\,\boldsymbol{x}_m,\boldsymbol{y}\right\},
\end{aligned} \tag{37}$$

where (a) follows from monotonicity and continuity of the

exponent [2] in T. Now, we calculate

$$\Pr\left\{N_{\boldsymbol{y}}(Q) > P(\boldsymbol{y}|\boldsymbol{x}_m)e^{-n[\boldsymbol{E}_Q \ln P(Y|X)+T]}\right\}.$$

Recall the definition of $\Omega(\tilde{Q}, Q, T)$ (24), and note that

$$\begin{aligned} & \frac{1}{n}\ln\left(P(\boldsymbol{y}|\boldsymbol{x}_m)e^{-n[\boldsymbol{E}_Q \ln P(Y|X)+T]}\right) \\ = \; & \boldsymbol{E}_{\hat{P}_{\boldsymbol{x}_m,\boldsymbol{y}}} \ln P(Y|X) - \boldsymbol{E}_Q \ln P(Y|X) - T \\ = \; & \Omega(\hat{P}_{\boldsymbol{x}_m,\boldsymbol{y}}, Q, T), \end{aligned} \tag{38}$$

so we will be interested in evaluating $\Pr\left\{N_{\boldsymbol{y}}(Q) > e^{n\Omega(\hat{P}_{\boldsymbol{x}_m,\boldsymbol{y}},Q,T)}\right\}$. There are two cases to consider depending on the sign of $\Omega = \Omega(\hat{P}_{\boldsymbol{x}_m,\boldsymbol{y}}, Q, T)$.

The case $\Omega \leq 0$: Here $e^{n\Omega} \leq 1$ and since $N_{\boldsymbol{y}}(Q)$ takes on integer values,

$$\begin{aligned} & \Pr\left\{N_{\boldsymbol{y}}(Q) > e^{n\Omega}\right\} \\ & = \Pr\left\{N_{\boldsymbol{y}}(Q) \geq 1\right\} \\ & \doteq 1 - \left[1 - e^{n[H_Q(X|Y)+\boldsymbol{E}_Q \ln P(X)]}\right]^{M-1} \\ & \doteq \exp\left\{-n\left|-H_Q(X|Y) - \boldsymbol{E}_Q \ln P(X) - R\right|^+\right\} \\ & = \exp\left\{-n\left|K(Q,R)\right|^+\right\}, \end{aligned} \tag{39}$$

where $K(Q,R)$ is defined in (21), the first exponential equality is because the random variable $N_{\boldsymbol{y}}(Q)$ is equal to the sum of the i.i.d. binary-p random variables $\mathbf{1}\left\{\hat{P}_{\boldsymbol{X}_i,\boldsymbol{y}} = Q\right\}, i = 1, ..., e^{nR} - 1$, with

$$p \doteq e^{n\left[H_Q(X|Y)+\boldsymbol{E}_Q \ln P(X)\right]}, \tag{40}$$

and the second exponential equality is because if $a \in [0,1]$, then

$$\frac{1}{2}\min\{1, aM\} \leq 1 - (1-a)^M \leq \min\{1, aM\} \tag{41}$$

(see Lemma 1 in [17]).

The case $\Omega > 0$: There are two sub-cases to consider:

• If $\Omega > 0$ and $\Omega \geq -K(Q,R)$ we can use the Chernoff bound, similarly to [12] Appendix B

$$\begin{aligned} & \Pr\left\{N_{\boldsymbol{y}}(Q) > e^{n\Omega}\right\} \\ & \dot{\leq} \min\left\{1, \exp\left\{-e^{n\Omega}\left[n\left\{K(Q,R)+\Omega\right\} - 1\right]\right\}\right\}. \end{aligned} \tag{42}$$

This term decays at least double–exponentially rapidly and hence is negligible in the exponential scale.

• If $0 < \Omega \leq -K(Q,R)$ we prove in the following lemma (see Appendix A) that $\Pr\left\{N_{\boldsymbol{y}}(Q) > e^{n\Omega}\right\}$ is very close to one.

Lemma 1: If $Q \in \mathcal{G}_R$ and $\Omega < R + H_Q(X|Y) + E_Q \ln P(X)$, then $\Pr\left\{N_{\boldsymbol{y}}(Q) \leq e^{n\Omega}\right\}$ vanishes superexponentially.

[2]Formally, the difference between $\sum_Q f(\boldsymbol{y}, Q)$ and $\max_Q f(\boldsymbol{y}, Q)$, which is $O(\log n/n)$ in the exponential scale, can be absorbed in the parameter T. Thus, one can derive upper and lower bounds in terms of $e_1(R, T + O(\log n/n))$ and $e_1(R, T - O(\log n/n))$, with $e_1(\cdot,\cdot)$ defined as in (23). These are asymptotically the same wherever $e_1(R,T)$ is continuous in T. And, in fact, since $e_1(R,T)$ a monotonically non-increasing function of T for a given R, it is continuous in T almost everywhere.

Therefore, for $\Omega > 0$, we have

$$\Pr\left\{N_{\boldsymbol{y}}(Q) > e^{n\Omega}\right\} \doteq \mathbf{1}\left\{0 < \Omega \leq -K(Q,R)\right\} \tag{43}$$

Combining this with the expression for the range where $\Omega \leq 0$ (39), we get.

$$\begin{aligned} \Pr\left\{N_{\boldsymbol{y}}(Q) > e^{n\Omega}\right\} \doteq \; & \mathbf{1}\left\{\Omega \leq 0\right\}\exp\left\{-n\left|K(Q,R)\right|^+\right\} \\ & + \mathbf{1}\left\{0 < \Omega \leq -K(Q,R)\right\}. \end{aligned} \tag{44}$$

This can be rewritten as

$$\begin{aligned} & \Pr\left\{N_{\boldsymbol{y}}(Q) > e^{n\Omega}\right\} \\ & \doteq \mathbf{1}\left\{Q \in \mathcal{G}_R^c(\hat{P}_{\boldsymbol{y}}), \Omega \leq 0\right\}\exp\left\{-n\left(K(Q,R)\right)\right\} \\ & \quad + \mathbf{1}\left\{Q \in \mathcal{G}_R(\hat{P}_{\boldsymbol{y}}), \Omega \leq -K(Q,R)\right\}. \end{aligned} \tag{45}$$

Recall that, in fact, $\Omega = \Omega(\tilde{Q}, Q, T)$ (with $\tilde{Q} = \hat{P}_{\boldsymbol{x}_m,\boldsymbol{y}}$, see (38)), so the condition in the second term,

$$\Omega \leq -K(Q,R) = H_Q(X|Y) + \boldsymbol{E}_Q \ln P(X) + R \tag{46}$$

is equivalent to

$$E_{\tilde{Q}} \ln P(Y|X) - T \leq R + \boldsymbol{E}_Q \ln P(X,Y) + H_Q(X|Y). \tag{47}$$

Thus, after maximizing over Q and taking the expectation w.r.t. $(\boldsymbol{x}_m, \boldsymbol{y})$, the resulting exponent is

$$e_1(R,T) \;=\; \min\left\{E_a(R,T), E_b(R,T)\right\} \tag{48}$$

where

$$\begin{aligned} E_a(R,T) = \min_{\tilde{Q}} \Bigg[& D(\tilde{Q}_{XY}||P_{XY}) \\ & + \min_{Q_{X|Y}:\; Q\in\mathcal{G}_R^c(\tilde{Q}_Y),\; \Omega(\tilde{Q},Q,T)\leq 0} K(Q,R)\Bigg], \end{aligned} \tag{49}$$

$$E_b(R,T) = \min_{\tilde{Q}\in\mathcal{L}_{R,T}} D(\tilde{Q}_{XY}||P_{XY}), \tag{50}$$

with $Q = \tilde{Q}_Y \times Q_{X|Y}$, and where

$$\mathcal{L}_{R,T} = \left\{ \begin{array}{l} \tilde{Q}(X,Y) : E_{\tilde{Q}} \ln P(Y|X) \leq R + T + \\ \max_{Q\in\mathcal{G}_R(\tilde{Q}_Y)}\left[\boldsymbol{E}_Q \ln P(X,Y) + H_Q(X|Y)\right] \end{array} \right\}, \tag{51}$$

and the probability distributions $\tilde{Q}$ are defined on the set $\mathcal{X} \times \mathcal{Y}$. Next, we show that $E_a(R,T)$ is the dominant term in the minimization (48):

$$\begin{aligned} & E_a(R,T) = \\ & \min_{\tilde{Q}}\left[D(\tilde{Q}_{XY}||P_{XY})) + \min_{Q\in\mathcal{G}_R^c(\tilde{Q}_Y):\; \Omega(\tilde{Q},Q,T)\leq 0} K(Q,R)\right] \\ & \leq \min_{\tilde{Q}\in\mathcal{L}_{R,T}}\left[D(\tilde{Q}_{XY}||P_{XY}) + \min_{Q\in\mathcal{G}_R^c(\tilde{Q}_Y):\; \Omega(\tilde{Q},Q,T)\leq 0} K(Q,R)\right] \\ & \overset{(a)}{=} \min_{\tilde{Q}\in\mathcal{L}_{R,T}}\left[D(\tilde{Q}_{XY}||P_{XY}) + \min_{Q\in\mathcal{G}_R^c(\tilde{Q}_Y)} K(Q,R)\right] \\ & \overset{(b)}{=} \min_{\tilde{Q}\in\mathcal{L}_{R,T}} D(\tilde{Q}_{XY}||P_{XY}) \\ & = E_b(R,T), \end{aligned} \tag{52}$$

where (a) follows since when $\tilde{Q} \in \mathcal{L}_{R,T}$, the constraint $\Omega(\tilde{Q}, Q, T) \leq 0$ becomes inactive, and (b) is by definition of $\mathcal{G}_R(\tilde{Q}_Y)$, and by the fact that when the rate R is below capacity, the boundary of $\mathcal{G}_R^c(\tilde{Q}_Y)$ is non empty[3].

This results in the exponent $e_1(R,T) = E_a(R,T)$.

VI. Proof of Corollary 1

To prove the corollary we use the general expression (23) of Theorem 1. For a measure $\tilde{Q}$ on $\mathcal{X} \times \mathcal{Y}$ define

$$Z(\tilde{Q}, R, T) \stackrel{\Delta}{=} \min_{Q_{X|Y}} \{\boldsymbol{E}_Q \ln[1/P(X)] - H_Q(X|Y)\} \quad (53)$$

where $Q = \tilde{Q}_Y \times Q_{X|Y}$ and the minimum over $Q_{X|Y}$ being across $\mathcal{G}_R^c(\tilde{Q}_Y) \bigcap \{\Omega(\tilde{Q}, Q, T) \leq 0\}$. Assuming that the condition of Theorem 1 holds, it is easy to see (using Lagrange multipliers) that the minimizing $Q_{X|Y}$ of Z is always of the form

$$\mu_s(x|y) = \frac{P(x)P^s(y|x)}{\sum_{x'} P(x')P^s(y|x')}, \quad (54)$$

and it is easy to check that for $Q_s = \tilde{Q}_Y \times \mu_s$

$$\boldsymbol{E}_{Q_s} \ln[1/P(X)] - H_{Q_s}(X|Y) = \gamma(s) - s\gamma'(s), \quad (55)$$

and that $\boldsymbol{E}_{Q_s} \ln[1/P(Y|X)] = \gamma'(s)$. Therefore, the computation of $Z(\tilde{Q}, R, T)$ boils down to a problem of minimizing over one parameter only, that is s. Specifically,

$$Z(\tilde{Q}, R, T) = \min\{\psi(s) : \ \psi(s) \geq R, \ \gamma'(s) \leq \xi(\tilde{Q}, T\} \quad (56)$$

where

$$\xi(\tilde{Q}, T) = T + \boldsymbol{E}_{\tilde{Q}} \ln[1/P(Y|X)]. \quad (57)$$

Now, $\gamma(s)$ is a monotonically non–decreasing concave function (as $\gamma'(s) \geq 0$, $\gamma''(s) \leq 0$), and so, $\psi'(s) = -s\gamma''(s) \geq 0$, which means that ψ is a non–decreasing function. Therefore, minimizing ψ is equivalent to minimizing s subject to the constraints. Now, the constraint $\psi(s) \geq R$ is equivalent to a constraint $s \geq s_R$ (see the definition of s_R preceding (14)), and the constraint $\gamma'(s) \leq \xi$ is equivalent to $s \geq s(\xi(\tilde{Q}, T))$, $s(\xi)$ being the solution to the equation $\gamma'(s) = \xi$. Thus, to satisfy both constraints s must be larger than $\max\{s_R, s(\xi(\tilde{Q}, T))\}$. So, the optimum s is $s^* = \max\{s_R, s(\xi(\tilde{Q}, T))\}$. We therefore obtain from (23) that the exponent is equal to

$$\min_{\tilde{Q}}\{D(\tilde{Q}_{XY}\|P_{XY}) + \psi(\max\{s_R, s(\xi(\tilde{Q}, T))\})\} - R. \quad (58)$$

We now argue that the achiever $\tilde{Q}$ must satisfy $s(\xi(\tilde{Q}, T)) \geq s_R$, and so, and alternative expression of the above is

$$\min_{\tilde{Q}:\ s(\xi(\tilde{Q},T)) \geq s_R} [D(\tilde{Q}_{XY}\|P_{XY}) + \psi(s(\xi(\tilde{Q}, T)))] - R. \quad (59)$$

[3]To realize this, note that for rates below the capacity $P \in \mathcal{G}_R^c(\tilde{Q}_Y)$ since $[-H_Q(X|Y) - \boldsymbol{E}_Q \ln P(X)]|_{Q=P} - R = I_P(X;Y) - R \leq 0$, and obviously $\mathcal{G}_R(\tilde{Q}_Y)$ is not empty, e.g., it contains $Q_{XY} = P(X)Q_Y$ which yields $[-H_Q(X|Y) - \boldsymbol{E}_Q \ln P(X)] - R = -R$ hence, and by continuity there exists a linear combination of these two probability distributions that lies on the boundary of $\mathcal{G}_R^c(\tilde{Q}_Y)$.

or, equivalently,

$$\min_{\tilde{Q}:\ \xi(\tilde{Q},T)) \leq \gamma'(s_R)} [D(\tilde{Q}_{XY}\|P_{XY}) + \psi(s(\xi(\tilde{Q}, T)))] - R. \quad (60)$$

To see why this is true, assume conversely, that the achiever $\tilde{Q}^*$ satisfies $s(\xi(\tilde{Q}^*, T)) < s_R$, or equivalently, $\boldsymbol{E}_{\tilde{Q}^*} \ln[1/P(Y|X)] > \gamma'(s_R) - T$. In this case we get,

$$e_1(R,T) = D(\tilde{Q}^*_{XY}\|P_{XY}) + \psi(s_R) - R = D(\tilde{Q}^*_{XY}\|P_{XY}). \quad (61)$$

But we know already that $\boldsymbol{E}_{P_{XY}} \ln[1/P(Y|X)] = H(Y|X) < \gamma'(s_R) - T$. Thus, if we look at the convex combination $Q^{(t)} = (1-t)\tilde{Q}^* + tP_{XY}$, and choose t such that $\boldsymbol{E}_{Q^{(t)}} \ln[1/P(Y|X)] = \gamma'(s_R) - T$ (namely, $s(\xi(Q^{(t)}, T)) = s_R$), we have, by convexity of the divergence, $D(Q^{(t)}_{XY}\|P_{XY}) \leq (1-t)D(\tilde{Q}^*_{XY}\|P_{XY}) < D(\tilde{Q}^*_{XY}\|P_{XY})$, which contradicts the optimality of $\tilde{Q}^*$ and hence proves the claim.

VII. Proof of Corollary 2

Consider the general expression (23) of Theorem 1. For the BSC, $P(Y|X)$, we have

$$\begin{aligned} &E_{\tilde{Q}} \ln P(Y|X) - E_Q \ln P(Y|X) \\ &= \Big[\tilde{Q}(X \neq Y) - Q(X \neq Y)\Big] \cdot \beta. \end{aligned} \quad (62)$$

We also note that since $P(X) = \frac{1}{2}$, for all Q, $E_Q \ln P(X) = -\ln 2$, thus the expression we get from (23) is

$$\min_{\tilde{Q}} \left[D(\tilde{Q}_{XY}||P_{XY}) + \min_{Q_{X|Y} \in \mathcal{S}} (-H_Q(X|Y) + \ln 2 - R)\right] \quad (63)$$

where $Q = \tilde{Q}_T \times Q_{X|Y}$ and

$$\mathcal{S}_1 \stackrel{\Delta}{=} \left\{ Q_{X|Y} : \begin{array}{c} -H_Q(X|Y) + \ln 2 - R \geq 0, \\ Q(X \neq Y) \leq \tilde{Q}(X \neq Y) + T/\beta \end{array} \right\}. \quad (64)$$

Consequently, (63) is equal to

$$\begin{aligned} \min_{\tilde{Q}} \Bigg[& D(\tilde{Q}_{XY}||P_{XY}) \\ & + \left| \min_{Q_{X|Y} \in \mathcal{S}_2} (-H_Q(X|Y) + \ln 2 - R) \right|^+ \Bigg]. \end{aligned} \quad (65)$$

where $\mathcal{S}_2 \stackrel{\Delta}{=} \{Q_{X|Y} : \ Q(X \neq Y) \leq \tilde{Q}(X \neq Y) + T/\beta\}$.

Now, on one hand we have

$$H_Q(X|Y) = H_Q\left(1\{X \neq Y\}|Y\right) \leq H_Q\left(1_{\{X \neq Y\}}\right), \quad (66)$$

thus,

$$\begin{aligned}
&\min_{\tilde{Q}} \Bigg[D(\tilde{Q}_{XY}||P_{XY}) \\
&+ \left| \min_{Q_{X|Y}\in\mathcal{S}_2} \left(-H_Q(X|Y) + \ln 2 - R\right) \right|^+ \Bigg] \\
&\geq \min_{\tilde{Q}} \Bigg[D(\tilde{Q}_{XY}||P_{XY}) \\
&+ \left| \min_{Q_{X|Y}\in\mathcal{S}_2} \left(-H_Q(1_{\{X\neq Y\}}) + \ln 2 - R\right) \right|^+ \Bigg] \\
&= \min_{\tilde{q}} \left[D(\tilde{q}||p) + \left| \min_{q\leq\tilde{q}+T/\beta} (-h(q) + \ln 2 - R) \right|^+ \right], \quad (67)
\end{aligned}$$

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

$$\tilde{Q}(y|x) = \begin{cases} \tilde{q} & x = y \\ 1-\tilde{q} & x \neq y \end{cases}, \quad (68)$$

for which we have $D(\tilde{Q}_{XY}\|P_{XY}) = D(\tilde{q}\|p)$.

On the other hand, one can choose

$$Q(y|x) = \begin{cases} q & x = y \\ 1-q & x \neq y \end{cases}, \quad (69)$$

which obtains the inequality in (67) with equality and thus is the minimizer.

Next, we observe that $-h(q)$ is decreasing in q for $q \in [0, \frac{1}{2}]$ and increasing for $q \in [\frac{1}{2}, 1]$ so,

$$\begin{aligned}
&\min_{\tilde{q}} \left[D(\tilde{q}||p) + \left| \min_{q\leq\tilde{q}+T/\beta} (-h(q) + \ln 2 - R) \right|^+ \right] \\
&= \min_{\tilde{q}} \left[D(\tilde{q}||p) + \left| -h\left(\min\left\{ \frac{1}{2}, \tilde{q} + T/\beta \right\} \right) + \ln 2 - R \right|^+ \right] \\
&= \min_{\tilde{q}} \left[D(\tilde{q}||p) - h\left(\min\left\{ \tilde{q} + T/\beta, \delta_{GV}(R) \right\} \right) + \ln 2 - R \right]. \quad (70)
\end{aligned}$$

Finally, we see that the minimum over $\tilde{q}$ cannot be attained at $\tilde{q} > \delta_{GV}(R) - T/\beta$, because beyond $\delta_{GV}(R) - T/\beta$, $D(\tilde{q}\|p)$ grows while $h(\min\{\tilde{q} + T/\beta, \delta_{GV}(R)\}) = h(\delta_{GV}(R))$ remains constant. Thus, it is enough to limit the range of the minimization to $\tilde{q} \in [p, \delta_{GV}(R) - T/\beta]$ in which case the minimization within the argument of $h(\cdot)$ becomes redundant. In summary,

$$\begin{aligned}
&e_1(R,T) \\
&= \min_{\tilde{q}\in[p,\delta_{GV}(R)-T/\beta]} [D(\tilde{q}\|p) - h(\tilde{q} + T/\beta)] + \ln 2 - R. \quad (71)
\end{aligned}$$

VIII. Proof of Theorem 2

The exponent $e_2(R,T)$ is associated with the probability that $\boldsymbol{y}$ falls in $\mathcal{R}_k$ for some k while the true message sent was m, $m \neq k$. Since $\mathcal{R}_k, k = 1, ..., e^{nR}$ are disjoint sets

$$\begin{aligned}
&\Pr\{\mathcal{E}_2|m, \boldsymbol{y}\} \\
&= \sum_{k\neq m} \Pr\left\{ e^{-nT} P(\boldsymbol{y}|\boldsymbol{x}_k) \geq \sum_{l\neq k} P(\boldsymbol{y}|\boldsymbol{x}_l) \,\middle|\, \boldsymbol{x}_m, \boldsymbol{y} \right\}. \quad (72)
\end{aligned}$$

Now, fix k and consider the following chain of equalities (on the exponential scale):

$$\begin{aligned}
&\Pr\left\{ e^{-nT} P(\boldsymbol{y}|\boldsymbol{x}_k) \geq \sum_{l\neq k} P(\boldsymbol{y}|\boldsymbol{x}_l) \,\middle|\, \boldsymbol{x}_m, \boldsymbol{y} \right\} \\
\dot{=} &\Pr\left\{ \frac{P(\boldsymbol{y}|\boldsymbol{x}_k)}{e^{nT}} \geq \max\left\{ P(\boldsymbol{y}|\boldsymbol{x}_m), \sum_{l\neq k,m} P(\boldsymbol{y}|\boldsymbol{x}_l) \right\} \,\middle|\, \boldsymbol{x}_m, \boldsymbol{y} \right\} \\
= &\Pr\left\{ \begin{array}{c} P(\boldsymbol{y}|\boldsymbol{x}_m) \leq e^{-nT} P(\boldsymbol{y}|\boldsymbol{x}_k), \\ \sum_{l\neq k,m} P(\boldsymbol{y}|\boldsymbol{x}_l) \leq e^{-nT} P(\boldsymbol{y}|\boldsymbol{x}_k) \end{array} \,\middle|\, \boldsymbol{x}_m, \boldsymbol{y} \right\}, \quad (73)
\end{aligned}$$

where, as mentioned, $\boldsymbol{x}_m$ is the true message transmitted. Consider now the random selection of the codebook, where the random codewords will be denoted by capital letters. Then,

$$\begin{aligned}
&\overline{\Pr}\{\mathcal{E}_2\} \\
&= \sum_{k\neq m} \sum_{\boldsymbol{y}\in\mathcal{Y}^n} \sum_{\Theta} P(\boldsymbol{y}) \cdot \Pr\{e^{-nT} P(\boldsymbol{y}|\boldsymbol{X}_k) = e^{-n\Theta}|\boldsymbol{y}\} \\
&\quad \times \Pr\{P(\boldsymbol{y}|\boldsymbol{X}_m) \leq e^{-n\Theta}|\boldsymbol{y}\} \\
&\quad \times \Pr\left\{ \sum_{l\neq k,m} P(\boldsymbol{y}|\boldsymbol{X}_l) \leq e^{-n\Theta}|\boldsymbol{y} \right\} \\
&\triangleq \sum_{k\neq m} \sum_{\boldsymbol{y}\in\mathcal{Y}^n} \sum_{\Theta} P(\boldsymbol{y}) \cdot \Pr(\mathcal{A}|\boldsymbol{y}) \cdot \Pr(\mathcal{B}|\boldsymbol{y}) \cdot \Pr(\mathcal{C}|\boldsymbol{y}). \quad (74)
\end{aligned}$$

Now,

$$\begin{aligned}
&\Pr(\mathcal{A}|\boldsymbol{y}) \\
&= \sum_{\boldsymbol{x}:\ \ln P(\boldsymbol{y}|\boldsymbol{x}) = n(T-\Theta)} P(\boldsymbol{x}) \\
&\dot{=} \exp\Bigg\{ n\times \\
&\quad \max_{Q_{X|Y}:\ \boldsymbol{E}_Q \ln P(Y|X) = T-\Theta} [H_Q(X|Y) + \boldsymbol{E}_Q \ln P(X)] \Bigg\} \\
&= \exp\Bigg\{ -n\times \\
&\times \min_{Q_{X|Y}:\ \boldsymbol{E}_Q \ln P(Y|X) = T-\Theta} [\boldsymbol{E}_Q \ln 1/P(X) - H_Q(X|Y)] \Bigg\} \\
&\triangleq e^{-nE_{\mathcal{A}}(\hat{P}_{\boldsymbol{y}}, \Theta)}. \quad (75)
\end{aligned}$$

Similarly, letting $P(\boldsymbol{x}|\boldsymbol{y}) = \prod_{i=1}^n P(x_i|y_i)$ denote the posterior induced by $P(\boldsymbol{x},\boldsymbol{y}) = \prod_{i=1}^n P(x_i)P(y_i|x_i)$, we have:

$$\begin{aligned}
&\Pr(\mathcal{B}|\boldsymbol{y}) \\
&= \sum_{\boldsymbol{x}:\ \ln P(\boldsymbol{y}|\boldsymbol{x}) \le -n\Theta} P(\boldsymbol{x}|\boldsymbol{y}) \\
&\doteq \exp\Bigg\{ n\times \\
&\quad \times \max_{Q_{X|Y}:\ \boldsymbol{E}_Q \ln P(Y|X) \le -\Theta} [H_Q(X|Y) + \boldsymbol{E}_Q \ln P(X|Y)] \Bigg\} \\
&= \exp\Bigg\{ -n\times \\
&\quad \min_{Q_{X|Y}:\ \boldsymbol{E}_Q \ln P(Y|X) \le -\Theta} [\boldsymbol{E}_Q \ln 1/P(X|Y) - H_Q(X|Y)] \Bigg\} \\
&\stackrel{\Delta}{=} e^{-nE_{\mathcal{B}}(\hat{P}\boldsymbol{y},\Theta)}. \qquad (76)
\end{aligned}$$

As for $\Pr(\mathcal{C}|\boldsymbol{y})$, we proceed as follows.

$$\begin{aligned}
&\Pr(\mathcal{C}|\boldsymbol{y}) \\
&= \Pr\left\{ \sum_Q N_{\boldsymbol{y}}(Q)\cdot e^{n\boldsymbol{E}_Q \ln P(Y|X)} \le e^{-n\Theta} \,\middle|\, \boldsymbol{y} \right\} \\
&\doteq \Pr\left\{ \max_Q \left[N_{\boldsymbol{y}}(Q)\cdot e^{n\boldsymbol{E}_Q \ln P(Y|X)} \right] \le e^{-n\Theta} \,\middle|\, \boldsymbol{y} \right\} \\
&= \Pr\left\{ \bigcap_Q \left\{ N_{\boldsymbol{y}}(Q) \le e^{n[\boldsymbol{E}_Q \ln 1/P(Y|X) - \Theta]} \right\} \,\middle|\, \boldsymbol{y} \right\} \qquad (77)
\end{aligned}$$

Now, if Θ is such that there exists $Q_{X|Y} \in \mathcal{G}_R(\hat{P}\boldsymbol{y})$ such that $Q = \hat{P}\boldsymbol{y} \times Q_{X|Y}$ satisfies $\boldsymbol{E}_Q \ln 1/P(Y|X) - \Theta < R + H_Q(X|Y) + \boldsymbol{E}_Q \ln P(X)$ (or equivalently, $\Theta > \min_{Q_{X|Y}\in\mathcal{G}_R(\hat{P}\boldsymbol{y})}[\boldsymbol{E}_Q \ln 1/P(X,Y) - H_Q(X|Y)] - R$), then this Q alone is responsible for a double–exponential decay of $\Pr(\mathcal{C}|\boldsymbol{y})$, let alone the intersection over all Q. Thus, in this range of large Θ, the contribution is double-exponential. Conversely, in the complementary event, namely, if Θ is such that for all $Q_{X|Y} \in \mathcal{G}_R(\hat{P}\boldsymbol{y})$, the measure $Q = \hat{P}\boldsymbol{y} \times Q_{X|Y}$ satisfies $\boldsymbol{E}_Q \ln 1/P(Y|X) - \Theta > R + H_Q(X|Y) + \boldsymbol{E}_Q \ln P(X)$ (or equivalently, $\Theta < \min_{Q_{X|Y}\in\mathcal{G}_R(\hat{P}\boldsymbol{y})}[\boldsymbol{E}_Q \ln 1/P(X,Y) - H_Q(X|Y)] - R$), then $\Pr(\mathcal{C}|\boldsymbol{y})$ is close to unity because it is lower bounded by $1 - \sum_Q \Pr\{N_{\boldsymbol{y}}(Q) > e^{n[\boldsymbol{E}_Q \ln 1/P(Y|X) - \Theta]}\}$, where the summation is over polynomially many terms that decay at least exponentially rapidly (at least exponentially in $\mathcal{G}_R^c(\hat{P}\boldsymbol{y})$ and at least double–exponentially in $\mathcal{G}_R(\hat{P}\boldsymbol{y})$). Thus, $\Pr(\mathcal{C}|\boldsymbol{y})$ can be approximated exponentially tightly by an indicator for the event $\Theta < \Theta_0(\hat{P}\boldsymbol{y}) \stackrel{\Delta}{=} \min_{Q_{X|Y}\in\mathcal{G}_R(\hat{P}\boldsymbol{y})}[\boldsymbol{E}_Q \ln 1/P(X,Y) - H_Q(X|Y)] - R$. Putting it all together, we then have:

$$\begin{aligned}
&\overline{\Pr}\{\mathcal{E}_2\} \\
&\doteq \sum_{k\neq m} \sum_{\boldsymbol{y}} \sum_{\Theta \le \Theta_0(\hat{P}\boldsymbol{y})} P(\boldsymbol{y}) \times \\
&\qquad \times \exp\Big\{ -n[E_{\mathcal{A}}(\hat{P}\boldsymbol{y},\Theta) + E_{\mathcal{B}}(\hat{P}\boldsymbol{y},\Theta)] \Big\} \\
&= \exp\Bigg\{ -n \min_{Q_Y} \big[D(Q_Y\|P_Y) \\
&\qquad + \min_{\Theta} \{E_{\mathcal{A}}(Q_Y,\Theta) + E_{\mathcal{B}}(Q_Y,\Theta)\} - R \big] \Bigg\}, \qquad (78)
\end{aligned}$$

where the minimization is over $\Theta \le \Theta_0(Q_Y)$

IX. Proof of Corollary 3

First, we compute $E_{\mathcal{A}}(Q_Y,\Theta)$ and $E_{\mathcal{B}}(Q_Y,\Theta)$ for this case. Clearly we have $E_Q \ln 1/P(X) = \ln 2$ and $E_Q \ln P(Y|X) = -Q(X\neq Y)\cdot\beta + \ln(1-p)$, therefore,

$$\begin{aligned}
&E_{\mathcal{A}}(Q_Y,\Theta \\
&= \min_{Q_{X|Y}:\ \boldsymbol{E}_Q \ln P(Y|X) = T - \Theta} [\boldsymbol{E}_Q \ln 1/P(X) - H_Q(X|Y)] \\
&= \min_{Q_{X|Y}:\ -Q(X\neq Y)\cdot\beta + \ln(1-p) = T-\Theta} [\ln 2 - H_Q(X|Y)] \\
&= \ln 2 - h\left([\ln(1-p) + \Theta - T]/\beta\right), \qquad (79)
\end{aligned}$$

where the last step is because we have (66) so the expression is minimized by choosing $Q_{X|Y}$ as in (69) with $q = Q(X\neq Y) = \frac{\ln(1-p)+\Theta-T}{\beta}$.

As for $E_{\mathcal{B}}(Q_Y,\Theta)$, we have similarly

$$\begin{aligned}
&E_{\mathcal{B}}(Q_Y,\Theta) \\
&= \min_{Q_{X|Y}:\ \boldsymbol{E}_Q \ln P(Y|X) \le -\Theta} [\boldsymbol{E}_Q \ln 1/P(X|Y) - H_Q(X|Y)] \\
&= \min_{Q_{X|Y}:\ -Q(X\neq Y)\cdot\beta + \ln(1-p) \le -\Theta} [Q(X\neq Y)\cdot\beta \\
&\qquad\qquad - \ln(1-p) - H_Q(X|Y)] \\
&= \min_{q \ge \frac{\Theta + \ln(1-p)}{\beta}} [q\cdot\beta - \ln(1-p) - h(q)] \\
&= \max\{p, q_m\}\,\beta - \ln(1-p) - h\left(\max\{p,q_m\}\right), \qquad (80)
\end{aligned}$$

where the last step of (80) is because the unconstrained minimization is achieved at $q = p$ and the function $q\beta - h(q)$ is convex for $q \in [0,1]$.

Next, we calculate $\Theta_0(Q_Y)$ similarly

$$\begin{aligned}
&\Theta_0(Q_Y) \\
&= \min_{Q_{X|Y}\in\mathcal{G}_R(Q_Y)} [\boldsymbol{E}_Q \ln 1/P(X,Y) - H_Q(X|Y)] - R \\
&\quad = \min_{q:\ \ln 2 - h(q) - R \le 0} q\beta - \ln(1-p) + \ln 2 - h(q) - R \\
&\quad = \max\{p, \delta_{GV}(R)\}\,\beta - \ln(1-p) + \ln 2 \\
&\qquad - h\left(\max\{p,\delta_{GV}(R)\}\right) - R. \qquad (81)
\end{aligned}$$

To conclude we note that since all of the relevant expressions $E_{\mathcal{A}}(Q_Y,\Theta), E_{\mathcal{B}}(Q_Y,\Theta), \Theta_0(Q_Y)$ actually do not depend on Q_Y, the optimal Q_Y is P_Y to minimize the divergence, and therefore we get (35).

APPENDIX

To prove Lemma 1, we shall use the lower bound on the divergence given in [12, eq. (27)]

$$D(a\|b) \geq a\left(\ln\frac{a}{b}-1\right)+b. \tag{82}$$

Using the Chernoff bound for $\Omega < R + H_Q(X|Y) + E_Q \ln P(X)$ we get

$$\begin{aligned}
&\Pr\left\{N_{\boldsymbol{y}}(Q) \leq e^{n\Omega}\right\} \\
&\leq \exp\left\{(1-e^{nR})D\left(e^{-n(R-\Omega)}\|e^{-n(-H_Q(X|Y)-E_Q\ln P(X))}\right)\right\} \\
&\leq \exp\left\{-e^{nR}D\left(e^{-n(R-\Omega)}\|e^{-n(-H_Q(X|Y)-E_Q\ln P(X))}\right)\right\} \\
&\dot{\leq} \exp\Big\{-e^{n\Omega}n\left(-H_Q(X|Y)-E_Q\ln P(X)-R+\Omega-1\right) \\
&\qquad -e^{n(H_Q(X|Y)+E_Q\ln P(X)+R)}\Big\}.
\end{aligned} \tag{83}$$

Now, it is evident that for $Q \in \mathcal{G}_R$, the term $e^{n(H_Q(X|Y)+E_Q\ln P(X)+R)}$ increases exponentially and the term $e^{n\Omega}n\left(-H_Q(X|Y)-E_Q\ln P(X)-R+\Omega-1\right)$ is negative but with exponent Ω which is smaller than $H_Q(X|Y)+E_Q\ln P(X)+R$ and therefore negligible, hence this probability vanishes superexponentially.

REFERENCES

[1] 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, Mar. 1968.

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

[3] E. Telatar, "Multi-access communications with decision feedback decoding," Ph.D. dissertation, M.I.T., Cambridge, Massachusetts, May 1992.

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

[5] T. Hashimoto, "Composite scheme LR+Th for decoding with era- sures and its effective equivalence to Forneys rule," *IEEE Trans. Inform. Theory*, vol. 45, no. 1, pp. 78-93, Jan. 1999.

[6] T. Hashimoto and M. Taguchi, "Performance and explicit error detec- tion and threshold decision in decoding with erasures," *IEEE Trans. Inform. Theory*, vol. 43, no. 5, pp. 1650-1655, Sep. 1997.

[7] P. Kumar, Y.-H. Nam, and H. El Gamal, "On the error exponents of arq channels with deadlines," *IEEE Trans. Inform. Theory*, vol. 53, no. 11, pp. 4265-4273, Nov. 2007.

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

[9] 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, Mar. 1969.

[10] M. Mézard and A. Montanari, *Information, Physics, and Computation*, Oxford University Press, 2009.

[11] N. Merhav, "Error exponents of erasure/list decoding revisited via moments of distance enumerators," *IEEE Trans. Inform. Theory*, vol. 54, no. 10, pp. 4439-4447, Oct. 2008.

[12] N. Merhav, "Relations between random coding exponents and the statistical physics of random codes," *IEEE Trans. Inform. Theory*, vol. 55, no. 1, pp. 83-92, Jan. 2009.

[13] N. Merhav, "The generalized random energy model and its application to the statistical physics of ensembles of hierarchical codes," *IEEE Trans. Inform. Theory*, vol. 55, no. 3, pp. 1250-1268, Mar. 2009.

[14] R. Etkin, N. Merhav and E. Ordentlich, "Error exponents of optimum decoding for the interference channel," *IEEE Trans. Inform. Theory*, vol. 56 no. 1 pp. 40-56, Jan. 2010.

[15] Y. Kaspi and N. Merhav, "Error exponents for broadcast channels with degraded message sets," accepted to *IEEE Trans. Inform. Theory*, June 2010.

[16] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[17] A. Somekh-Baruch and N. Merhav, "Achievable error exponents for the private fingerprinting game," vol. 53, no. 5, pp. 1827-1838, May 2007.