CCIT Report #647 November 2007

Error Exponents of Erasure/List Decoding Revisited via Moments of Distance Enumerators

Neri Merhav

November 15, 2007

Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, ISRAEL

Abstract

The analysis of random coding error exponents pertaining to erasure/list decoding, due to Forney, is revisited. Instead of using Jensen's inequality as well as some other inequalities in the derivation, we demonstrate that an exponentially tight analysis can be carried out by assessing the relevant moments of a certain distance enumerator. The resulting bound has the following advantages: (i) it is at least as tight as Forney's bound, (ii) under certain symmetry conditions associated with the channel and the random coding distribution, it is simpler than Forney's bound in the sense that it involves an optimization over one parameter only (rather than two), and (iii) in certain special cases, like the binary symmetric channel (BSC), the optimum value of this parameter can be found in closed form, and so, there is no need to conduct a numerical search. We have not found yet, however, a numerical example where this new bound is strictly better than Forney's bound. This may provide an additional evidence to support Forney's conjecture that his bound is tight for the average code. We believe that the technique we suggest in this paper can be useful in simplifying, and hopefully also improving, exponential error bounds in other problem settings as well.

Index Terms: random coding, erasure, list, error exponent, distance enumerator.

1 Introduction

In his celebrated paper [3], Forney extended Gallager's bounding techniques [2] and found exponential error bounds for the ensemble performance of optimum generalized decoding rules that include the options of erasure, variable size lists, and decision feedback (see also later studies, e.g., [1], [4],[5], [6], [8], and [10]).

Stated informally, Forney [3] considered a communication system where a code of block length n and size $M = e^{nR}$ (R being the coding rate), drawn independently at random under a distribution $\{P(x)\}$, is used for a discrete memoryless channel (DMC) $\{P(y|x)\}$ and decoded with an erasure/list option. For the erasure case, in which we focus hereafter, an optimum tradeoff was sought between the probability of erasure (no decoding) and the probability of undetected decoding error. 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 [3] 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. This lower bound is given by

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

where

$$E_0(s,\rho) = -\ln\left[\sum_y\left(\sum_x P(x)P^{1-s}(y|x)\right)\cdot\left(\sum_{x'}P(x')P^{s/\rho}(y|x')\right)^{\rho}\right]. \tag{2}$$

The probability of the undetected error event $\mathcal{E}_2$ (i.e., the event of not erasing but making a wrong estimate of the transmitted message) is given by $E_2(R,T) = E_1(R,T) + T$.[1] As can be seen, the computation of $E_1(R,T)$ involves an optimization over two auxiliary parameters, ρ and s, which in general requires a two–dimensional search over these two parameters by some method. This is different from Gallager's random coding error exponent function for ordinary decoding (without

[1]Forney also provides improved (expurgated) exponents at low rates, but we will focus here solely on (1).

erasures), which is given by:

$$E_r(R) = \max_{0 \le \rho \le 1} [E_0(\rho) - \rho R], \tag{3}$$

with $E_0(\rho)$ being defined as

$$E_0(\rho) = -\ln \left[\sum_y \left(\sum_x P(x) P^{1/(1+\rho)}(y|x) \right)^{1+\rho} \right], \tag{4}$$

where there is only one parameter to be optimized. In [3], one of the steps in the derivation involves the inequality $(\sum_i a_i)^r \le \sum_i a_i^r$, which holds for $r \le 1$ and non–negative $\{a_i\}$ (cf. eq. (90) in [3]), and another step (eq. (91e) therein) applies Jensen's inequality. The former inequality introduces an additional parameter, denoted ρ, to be optimized together with the original parameter, s.[2]

In this paper, we offer a different technique for deriving a lower bound to the exponent of the probability of $\mathcal{E}_1$, which avoids the use of these inequalities. Instead, an exponentially tight evaluation of the relevant expression is derived by assessing the moments of a certain distance enumerator, and so, the resulting bound is at least as tight as Forney's bound. Since the first above–mentioned inequality is bypassed, there is no need for the additional parameter ρ, and so, under certain symmetry conditions (that often hold) on the random coding distribution and the channel, the resulting bound is not only at least as tight as Forney's bound, but it 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). We have not found yet, however, a convincing[3] numerical example where the new bound is *strictly* better than Forney's bound. This may serve as an additional evidence to support Forney's conjecture that his bound is tight for the average code. Nevertheless, the question whether there

[2]The parameter s is introduced, as in many other derivations, as the power of the likelihood ratio that bounds the indicator function of the error event. This point will be elaborated on in Section 4.

[3]In a few cases, small differences were found, but these could attributed to insufficient resolution of the two–dimensional search for the optimum ρ and s in Forney's bound.

exist situations where the new bound is strictly better, remains open.

We wish to emphasize that the main message of this contribution, is not merely in the simplification of the error exponent bound in this specific problem of decoding with erasures, but more importantly, in the analysis technique we offer here, which, we believe, is applicable to quite many other problem settings as well. It is conceivable that in some of these problems, the proposed technique could not only simply, but perhaps also improve on currently known bounds. The underlying ideas behind this technique are inspired from the statistical mechanical point of view on random code ensembles, offered in [9] and further elaborated on in [7].

The outline of this paper is as follows. In Section 2, we establish notation conventions and give some necessary background in more detail. In Section 3, we present the main result along with a short discussion. Finally, in Section 4, we provide the detailed derivation of the new bound, first for the special case of the BSC, and then more generally.

2 Notation and Preliminaries

Throughout this paper, scalar random variables (RV's) will be denoted by capital letters, their sample values will be denoted by the respective lower case letters, and their alphabets will be denoted by the respective calligraphic letters. A similar convention will apply to random vectors of dimension n and their sample values, which will be denoted with same symbols in the bold face font. The set of all n–vectors with components taking values in a certain finite alphabet, will be denoted as the same alphabet superscripted by n. Thus, for example, a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ may assume a specific vector value $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ as each component takes values in $\mathcal{X}$. Sources and channels will be denoted generically by the letter P or Q. Information theoretic

quantities 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 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 $\mathcal{A}$ will be denoted by $|\mathcal{A}|$. 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, $a_n \dot{\leq} b_n$ means that $\limsup_{n\to\infty}\frac{1}{n}\ln\frac{a_n}{b_n} \leq 0$, and so on.

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

$$P(y_i|x_1,\ldots,x_i,y_1,\ldots,y_{i-1}) = P(y_i|x_i), \quad i = 1,2,\ldots,n \tag{5}$$

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

A decoder with an erasure option is a partition of $\mathcal{Y}^n$ into $(M+1)$ regions, $\mathcal{R}_0,\mathcal{R}_1,\ldots,\mathcal{R}_M$.

Such a decoder works as follows: If $\boldsymbol{y}$ falls into $\mathcal{R}_m$, $m = 1, 2, \ldots, M$, then a decision is made in favor of message number m. If $\boldsymbol{y} \in \mathcal{R}_0$, no decision is made and an erasure is declared. We will refer to $\mathcal{R}_0$ as the *erasure event*. Given a code $\mathcal{C} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ and a decoder $\mathcal{R} = (\mathcal{R}_0, \mathcal{R}_1, \ldots, \mathcal{R}_m)$, let us now define two additional undesired events. The event $\mathcal{E}_1$ is the event of not making the right decision. This event is the disjoint union of the erasure event and the event $\mathcal{E}_2$, which is the *undetected error* event, namely, the event of making the wrong decision. The probabilities of all three events are defined as follows:

$$\begin{aligned}
\Pr\{\mathcal{E}_1\} &= \sum_{m=1}^{M} \sum_{\boldsymbol{y} \in \mathcal{R}_m^c} P(\boldsymbol{x}_m, \boldsymbol{y}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{\boldsymbol{y} \in \mathcal{R}_m^c} P(\boldsymbol{y}|\boldsymbol{x}_m) && (6)\\
\Pr\{\mathcal{E}_2\} &= \sum_{m=1}^{M} \sum_{\boldsymbol{y} \in \mathcal{R}_m} \sum_{m' \neq m} P(\boldsymbol{x}_{m'}, \boldsymbol{y}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{\boldsymbol{y} \in \mathcal{R}_m} \sum_{m' \neq m} P(\boldsymbol{y}|\boldsymbol{x}_{m'}) && (7)\\
\Pr\{\mathcal{R}_0\} &= \Pr\{\mathcal{E}_1\} - \Pr\{\mathcal{E}_2\}. && (8)
\end{aligned}$$

Forney [3] shows, using the Neyman–Pearson theorem, that the best tradeoff between $\Pr\{\mathcal{E}_1\}$ and $\Pr\{\mathcal{E}_2\}$ (or, equivalently, between $\Pr\{\mathcal{R}_0\}$ and $\Pr\{\mathcal{E}_2\}$) is attained by the decoder $\mathcal{R}^* = (\mathcal{R}_0^*, \mathcal{R}_1^*, \ldots, \mathcal{R}_M^*)$ defined by

$$\begin{aligned}
\mathcal{R}_m^* &= \left\{ \boldsymbol{y} : \frac{P(\boldsymbol{y}|\boldsymbol{x}_m)}{\sum_{m' \neq m} P(\boldsymbol{y}|\boldsymbol{x}_{m'})} \geq e^{nT} \right\}, \quad m = 1, 2, \ldots, M\\
\mathcal{R}_0^* &= \bigcap_{m=1}^{M} (\mathcal{R}_m^*)^c, && (9)
\end{aligned}$$

where $(\mathcal{R}_m^*)^c$ is the complement of $\mathcal{R}_m^*$, and where $T \geq 0$ is a parameter, henceforth referred to as the *threshold*, which controls the balance between the probabilities of $\mathcal{E}_1$ and $\mathcal{E}_2$. Forney devotes the remaining part of his paper [3] to derive lower bounds, as well as to investigate properties, of the random coding exponents (associated with $\mathcal{R}^*$), $E_1(R, T)$ and $E_2(R, T)$, of $\overline{\Pr}\{\mathcal{E}_1\}$ and $\overline{\Pr}\{\mathcal{E}_2\}$, the average probabilities of $\mathcal{E}_1$ and $\mathcal{E}_2$, respectively, (w.r.t.) the ensemble of randomly selected codes, drawn independently according to an i.i.d. distribution $P(\boldsymbol{x}) = \prod_{i=1}^{n} P(x_i)$. As mentioned in the

Introduction, $E_1(R,T)$ is given by (1) and $E_2(R,T) = E_1(R,T) + T$.

3 Main Result

Our main result in this paper is the following:

Theorem 1 *Assume that 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,

$$\gamma_y(s) \stackrel{\Delta}{=} -\ln\left[\sum_{x\in\mathcal{X}} P(x)P^s(y|x)\right] \tag{10}$$

is independent of y, *in which case, it will be denoted by* $\gamma(s)$. *Let* s_R *be the solution to the equation*

$$\gamma(s) - s\gamma'(s) = R, \tag{11}$$

where $\gamma'(s)$ *is the derivative of* $\gamma(s)$. *Finally, let*

$$E_1^*(R,T,s) = \Lambda(R,s) + \gamma(1-s) - sT - \ln|\mathcal{Y}| \tag{12}$$

where

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

Then,

$$\overline{Pr}\{\mathcal{E}_1\} \stackrel{\cdot}{\leq} e^{-nE_1^*(R,T)} \tag{14}$$

where $E_1^*(R,T) = \sup_{s\geq 0} E_1^*(R,T,s)$ *and*

$$\overline{Pr}\{\mathcal{E}_2\} \stackrel{\cdot}{\leq} e^{-nE_2^*(R,T)} \tag{15}$$

where $E_2^*(R,T) = E_1^*(R,T) + T$. *Also,* $E_1^*(R,T) \geq E_1(R,T)$, *where* $E_1(R,T)$ *is given in (1).*

Three comments are in order regarding the condition that $\gamma_y(s)$ of eq. (10) is independent of y.

The first comment is that this condition is obviously satisfied when $\{P(x)\}$ is uniform and the columns of the matrix $\{a_{xy}\} = \{P(y|x)\}$ are permutations of each other, because then the summations $\sum_x P(x)P^\beta(y|x)$, for the various y's, consist of exactly the same terms, just in a different order. This is the case, for example, when $\mathcal{X} = \mathcal{Y}$ is a group endowed with an addition/subtraction operation (e.g., addition/subtraction modulo the alphabet size), and the channel is additive in the sense that the 'noise' $(Y - X)$ is statistically independent of X. Somewhat more generally, the condition $\gamma_y(s) = \gamma(s)$ for all y holds when the different columns of the matrix $\{P(y|x)\}$ are formed by permutations of each other subject to the following rule: $P(y|x)$ can be permuted with $P(y|x')$ if $P(x) = P(x')$. For example, let $\mathcal{X} = \{A, B, C\}$ and $\mathcal{Y} = \{1, 2\}$, let the random coding distribution be given by $P(A) = a$, $P(B) = P(C) = (1 - a)/2$, and let the channel be given by $P(0|A) = P(1|A) = 1/2$, $P(0|B) = P(1|C) = 1 - P(0|C) = 1 - P(1|B) = b$. In this case, the condition is satisfied and $\gamma(s) = -\ln[(1 - a)/2)(b^s + (1 - b)^s) + a \cdot 2^{-s}]$.

The second comment is that the derivation of the bound, using the proposed technique, can be carried out, in principle, even without this condition on $\gamma_y(s)$. In the absence of this condition, one ends up with an exponential expression that depends, for each $\boldsymbol{y}$, on the empirical distribution $\hat{P}_{\boldsymbol{y}}$, and its summation over $\boldsymbol{y}$ can then be handled using the method of types, which involves optimization over $\{\hat{P}_{\boldsymbol{y}}\}$, or in the limit of large n, optimization over the continuum of probability distributions on $\mathcal{Y}$. But then we are loosing the simplicity of the bound relative to Forney's bound, since this optimization is at least as complicated as the optimization over the additional parameter ρ in [3].

Our last comment, in the context of this condition on $\gamma_y(s)$, is that even when it holds, it is not apparent that the expression of Forney's bound $E_1(R,T)$ can be simplified directly in a trivial manner, nor can we see how the optimum parameters ρ and s can be found analytically in closed form. This is true even in the simplest case of the BSC.

4 Derivation of the New Bound

4.1 Background

The first few steps of the derivation are similar to those in [3]: For a given code and for every $s \geq 0$,

$$\begin{aligned}
\Pr\{\mathcal{E}_1\} &= \frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in(\mathcal{R}_m^*)^c} P(\boldsymbol{y}|\boldsymbol{x}_m) \\
&= \frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n} P(\boldsymbol{y}|\boldsymbol{x}_m)\cdot 1\left\{\frac{e^{nT}\sum_{m'\neq m}P(\boldsymbol{y}|\boldsymbol{x}_{m'})}{P(\boldsymbol{y}|\boldsymbol{x}_m)}\geq 1\right\} \\
&\leq \frac{1}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n} P(\boldsymbol{y}|\boldsymbol{x}_m)\left(\frac{e^{nT}\sum_{m'\neq m}P(\boldsymbol{y}|\boldsymbol{x}_{m'})}{P(\boldsymbol{y}|\boldsymbol{x}_m)}\right)^s \\
&= \frac{e^{nsT}}{M}\sum_{m=1}^{M}\sum_{\boldsymbol{y}\in\mathcal{Y}^n} P^{1-s}(\boldsymbol{y}|\boldsymbol{x}_m)\left(\sum_{m'\neq m}P(\boldsymbol{y}|\boldsymbol{x}_{m'})\right)^s. \qquad (16)
\end{aligned}$$

As for $\mathcal{E}_2$, we have similarly,

$$\Pr\{\mathcal{E}_2\} \leq e^{-n(1-s)T}\sum_{\boldsymbol{y}\in\mathcal{Y}^n} P^{1-s}(\boldsymbol{y}|\boldsymbol{X}_m)\left(\sum_{m'\neq m}P(\boldsymbol{y}|\boldsymbol{X}_{m'})\right)^s. \qquad (17)$$

Since this differs from the bound on $\Pr\{\mathcal{E}_1\}$ only by the constant factor e^{-nT}, it will be sufficient to focus on $\mathcal{E}_1$ only. Taking now the expectation w.r.t. the ensemble of codes, and using the fact that $\boldsymbol{X}_m$ is independent of all other codewords, we get:

$$\overline{\Pr}\{\mathcal{E}_1\} \leq e^{nsT}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\boldsymbol{E}\{P^{1-s}(\boldsymbol{y}|\boldsymbol{X}_m)\}\cdot\boldsymbol{E}\left\{\left(\sum_{m'\neq m}P(\boldsymbol{y}|\boldsymbol{X}_{m'})\right)^s\right\}. \qquad (18)$$

The first factor of the summand is easy to handle:

$$\boldsymbol{E}\{P^{1-s}(\boldsymbol{y}|\boldsymbol{X}_m)\} = \sum_{\boldsymbol{x}\in\mathcal{X}^n}P(\boldsymbol{x})P^{1-s}(\boldsymbol{y}|\boldsymbol{x})$$

$$= \prod_{i=1}^{n} [\sum_{x \in \mathcal{X}} P(x) P^{1-s}(y_i|x)]$$
$$= e^{-n\gamma(1-s)}. \tag{19}$$

Concerning the second factor of the summand, Forney's approach is to use the inequality $(\sum_i a_i)^r \leq \sum_i a_i^r$, which holds when $\{a_i\}$ are positive and $r \leq 1$, in order to upper bound

$$\boldsymbol{E}\left\{\left(\sum_{m' \neq m} P(\boldsymbol{y}|\boldsymbol{X}_{m'})\right)^s\right\}$$

by

$$\boldsymbol{E}\left\{\left(\sum_{m' \neq m} P(\boldsymbol{y}|\boldsymbol{X}_{m'})^{s/\rho}\right)^\rho\right\}, \quad \rho \geq s,$$

and then (similarly to Gallager) use Jensen's inequality to insert the expectation into the bracketed expression, which is allowed by limiting ρ to be less than unity. In other words, the above expression is further upper bounded in [3] by

$$\left(\sum_{m' \neq m} \boldsymbol{E}\left\{P(\boldsymbol{y}|\boldsymbol{X}_{m'})^{s/\rho}\right\}\right)^\rho, \quad \rho \leq 1.$$

The idea here, as we understand it, is that the parameter ρ controls the tradeoff between the gap pertaining to the first inequality and the one associated with the second inequality. The first gap is maximum when $\rho = 1$ and non–existent when $\rho = s$ ($s \leq 1$), whereas for the second gap it is vice versa.

We, on the other hand, will use a different route, where all steps of the derivation will be clearly exponentially tight, and without introducing the additional parameter ρ. To simplify the exposition and make it easier to gain some geometrical insight, it will be instructive to begin with the special case of the BSC and the uniform random coding distribution. The extension to more general DMC's and random coding distributions will be given in Subsection 4.3. Readers who are interested in the more general case only may skip to Subsection 4.3 without loss of continuity.

4.2 The BSC with the uniform random coding distribution

Consider the special case where $\mathcal{X} = \mathcal{Y} = \{0,1\}$, the channel is a BSC with a crossover probability p, and the random coding distribution is uniform over the Hamming space $\{0,1\}^n$, namely, $P(\boldsymbol{x}) = 2^{-n}$ for all $\boldsymbol{x} \in \{0,1\}^n$. First, concerning the first factor in the summand of (18), we have, in this special case:

$$\gamma(1-s) = -\ln\left[\frac{1}{2}p^{1-s} + \frac{1}{2}(1-p)^{1-s}\right] = \ln 2 - \ln[p^{1-s} + (1-p)^{1-s}]. \tag{20}$$

As for the second factor, we proceed as follows. Define $\beta = \ln\frac{1-p}{p}$ and for a given $\boldsymbol{y}$, let $N_{\boldsymbol{y}}(d)$ denote distance enumerator relative to $\boldsymbol{y}$, that is, the number of incorrect codewords $\{\boldsymbol{x}_{m'},\ m' \neq m\}$ at Hamming distance d from $\boldsymbol{y}$. We then have:

$$\begin{aligned} \boldsymbol{E}\left\{\left(\sum_{m'\neq m} P(\boldsymbol{y}|\boldsymbol{X}_{m'})\right)^s\right\} &= \boldsymbol{E}\left\{\left[(1-p)^n \sum_{d=0}^{n} N_{\boldsymbol{y}}(d) e^{-\beta d}\right]^s\right\} \\ &\doteq (1-p)^{ns} \sum_{d=0}^{n} \boldsymbol{E}\{N_{\boldsymbol{y}}^s(d)\} e^{-\beta s d}. \end{aligned} \tag{21}$$

The second (exponential) equality is the *first main point* in our approach: It holds, even before taking the expectations, because the summation over d consists of a *subexponential* number of terms, and so, both $[\sum_d N_{\boldsymbol{y}}(d)e^{-\beta d}]^s$ and $\sum_d N_{\boldsymbol{y}}^s(d)e^{-\beta s d}$ are of the same exponential order as $\max_d N_{\boldsymbol{y}}^s(d)e^{-\beta s d} = [\max_d N_{\boldsymbol{y}}(d)e^{-\beta d}]^s$. This is different from the original summation over m', which contains an *exponential* number of terms. Thus, the key issue here is how to assess the power–s moments of the distance enumerator $N_{\boldsymbol{y}}(d)$. To this end, we have to distinguish between two ranges of d, or equivalently, $\delta = d/n$. Let $\delta_{GV}(R)$ denote the normalized Gilbert–Varshamov (GV) distance, $\delta_{GV} = d_{GV}/n$, i.e., the smaller solution, δ, to the equation

$$h(\delta) = \ln 2 - R,$$

where

$$h(\delta) = -\delta \ln \delta - (1-\delta)\ln(1-\delta), \quad \delta \in [0,1].$$

Now, the *second main point* of the proposed analysis approach is that $\boldsymbol{E}\{N_{\boldsymbol{y}}^s(d)\}$ behaves differently[4] for the case $\delta_{GV}(R) \le \delta \le 1-\delta_{GV}(R)$ and for the case $\delta < \delta_{GV}(R)$ or $\delta > 1-\delta_{GV}(R)$. Let us define then $\mathcal{G}_R = \{\delta:\ \delta_{GV}(R) \le \delta \le 1-\delta_{GV}(R)\}$. In particular, using the large deviations behavior of $N_{\boldsymbol{y}}(n\delta)$, $\delta \in [0,1]$, as the sum of $e^{nR}-1$ binary i.i.d. RV's, it is easily seen (see Appendix) that

$$\boldsymbol{E}\{N_{\boldsymbol{y}}^s(n\delta)\} \doteq \begin{cases} e^{ns[R+h(\delta)-\ln 2]} & \delta \in \mathcal{G}_R \\ e^{n[R+h(\delta)-\ln 2]} & \delta \in \mathcal{G}_R^c. \end{cases} \tag{22}$$

Thus,

$$\begin{aligned} & \boldsymbol{E}\left\{\left(\sum_{m' \neq m} P(\boldsymbol{y}|\boldsymbol{X}_{m'})\right)^s\right\} \\ \doteq\ & (1-p)^{ns}\left[\sum_{\delta\in\mathcal{G}_R} e^{ns[R+h(\delta)-\ln 2]}\cdot e^{-\beta sn\delta} + \sum_{\delta\in\mathcal{G}_R^c} e^{n[R+h(\delta)-\ln 2]}\cdot e^{-\beta sn\delta}\right] \\ \doteq\ & (1-p)^{ns}\left[e^{ns(R-\ln 2)}\cdot \exp\{ns\max_{\delta\in\mathcal{G}_R}[h(\delta)-\beta\delta]\} + e^{n(R-\ln 2)}\cdot\exp\{n\max_{\delta\in\mathcal{G}_R^c}[h(\delta)-\beta s\delta]\}\right] \end{aligned} \tag{23}$$

We are assuming, of course, $R < C = \ln 2 - h(p)$, which is equivalent to $p < \delta_{GV}(R)$ or $p > 1-\delta_{GV}(R)$. We also assume, for the sake of simplicity and without essential loss of generality, that $p < 1/2$, which will leave us only with the first possibility of $p < \delta_{GV}(R)$. Therefore, the global (unconstrained) maximum of $h(\delta)-\beta\delta$, which is attained at $\delta = p$, falls outside $\mathcal{G}_R$, and so, $\max_{\delta\in\mathcal{G}_R}[h(\delta)-\beta\delta]$ is attained at $\delta = \delta_{GV}(R)$ which yields

$$\max_{\delta\in\mathcal{G}_R}[h(\delta)-\beta\delta] = h(\delta_{GV}(R)) - \beta\delta_{GV}(R) = \ln 2 - R - \beta\delta_{GV}(R).$$

[4]The intuition behind this different behavior is that when $h(\delta)+R-\ln 2 > 0$, the RV $N_{\boldsymbol{y}}(d)$, which is the sum of $e^{nR}-1$ many i.i.d. binary RV's, $1\{d(\boldsymbol{X}_{m'},\boldsymbol{y})=d\}$, concentrates extremely (double–exponentially) rapidly around its expectation $e^{n[R+h(\delta)-\ln 2]}$, whereas for $h(\delta)+R-\ln 2 < 0$, $N_{\boldsymbol{y}}(d)$ is typically zero, and so, the dominant term of $\boldsymbol{E}\{N_{\boldsymbol{y}}^s(d)\}$ is $1^s\cdot\Pr\{N_{\boldsymbol{y}}(d)=1\} \approx e^{n[R+h(\delta)-\ln 2]}$.

Thus, the first term in the large square brackets of the r.h.s. of (23) is of the exponential order of $e^{-ns\beta\delta_{GV}(R)}$. As for the second term, the unconstrained maximum of $h(\delta) - \beta s\delta$ is obtained at $\delta = p_s \stackrel{\Delta}{=} \frac{p^s}{p^s+(1-p)^s}$, which can be either larger or smaller than $\delta_{GV}(R)$, depending on s. Specifically,

$$\max_{\delta\in\mathcal{G}_R^c}[h(\delta) - \beta s\delta] = \begin{cases} h(p_s) - \beta s p_s & p_s \le \delta_{GV}(R) \\ \ln 2 - R - \beta s\delta_{GV}(R) & p_s > \delta_{GV}(R) \end{cases} \tag{24}$$

The condition $p_s \le \delta_{GV}(R)$ is equivalent to

$$s \ge s_R \stackrel{\Delta}{=} \frac{\ln[(1-\delta_{GV}(R))/\delta_{GV}(R)]}{\beta}.$$

Thus, the second term in the square brackets of the r.h.s. of eq. (23) is of the order of $e^{-n\mu(s,R)}$, where

$$\mu(s,R) = \begin{cases} \mu_0(s,R) & s \ge s_R \\ \beta s\delta_{GV}(R) & s < s_R \end{cases} \tag{25}$$

and where

$$\begin{aligned} \mu_0(s,R) &= \beta s p_s - h(p_s) + \ln 2 - R \\ &= s\ln(1-p) - \ln[p^s + (1-p)^s] + \ln 2 - R. \end{aligned} \tag{26}$$

Next, observe that the second term, $e^{-n\mu(s,R)}$, is always the dominant term: For $s < s_R$, this is trivial as both terms behave like $e^{-n\beta s\delta_{GV}(R)}$. For $s \ge s_R$ (namely, $p_s \le \delta_{GV}(R)$), as $\delta = p_s$ achieves the *global* minimum of the function $f(\delta) \stackrel{\Delta}{=} \beta s\delta - h(\delta) + \ln 2 - R$, we have

$$\mu_0(s,R) = f(p_s) \le f(\delta_{GV}(R)) = \beta s\delta_{GV}(R).$$

Therefore, we have established that

$$\boldsymbol{E}\left\{\left(\sum_{m'\neq m} P(\boldsymbol{y}|\boldsymbol{X}_{m'})\right)^s\right\} \doteq \exp\left\{-n\left[s\ln\frac{1}{1-p} + \mu(s,R)\right]\right\} \tag{27}$$

independently of $\boldsymbol{y}$. Finally, we get:

$$\overline{\Pr}\{\mathcal{E}_1\} \dot{\leq} e^{nsT} \cdot 2^n \cdot e^{-n[\ln 2 - \ln(p^{1-s}+(1-p)^{1-s})]} \cdot \exp\left\{-n\left[s\ln\frac{1}{1-p} + \mu(s,R)\right]\right\} = e^{-nE_1(R,T,s)} \quad (28)$$

where

$$E_1(R,T,s) \stackrel{\Delta}{=} \mu(s,R) + s\ln\frac{1}{1-p} - \ln[p^{1-s}+(1-p)^{1-s}] - sT.$$

It is immediate to verify that this coincides with the expression of Theorem 1 when specialized to the case of the BSC with a uniform random coding distribution.

We next derive closed form expressions for the optimum value of s, denoted $s_{\rm opt}$, using the following consideration: We have seen that $E_1^*(R,T,s)$ is given by

$$F(s) \stackrel{\Delta}{=} \mu_0(s,R) + s\ln\frac{1}{1-p} - \ln[p^{1-s}+(1-p)^{1-s}] - sT$$

for $s \geq s_R$, and by

$$G(s) \stackrel{\Delta}{=} \beta s\delta_{GV}(R) + s\ln\frac{1}{1-p} - \ln[p^{1-s}+(1-p)^{1-s}] - sT$$

for $s < s_R$. Both $F(s)$ and $G(s)$ are easily seen to be concave functions and hence have a unique maximum each, which can be found by equating the corresponding derivative to zero. We have also seen that $F(s) \leq G(s)$ for all s, with equality at $s = s_R$ and only at that point. This means that $F(s)$ and $G(s)$ are tangential to each other at $s = s_R$, in other words, $F(s_R) = G(s_R)$ and $F'(s_R) = G'(s_R)$, where F' and G' are the derivatives of F and G, respectively. Now, there are three possibilities: If $F'(s_R) = G'(s_R) = 0$, then $s_{\rm opt} = s_R$. If $F'(s_R) = G'(s_R) < 0$, then $s_{\rm opt} < s_R$ is found by solving the equation $G'(s) = 0$. If $F'(s_R) = G'(s_R) > 0$, then $s_{\rm opt} > s_R$ is found by solving the equation $F'(s) = 0$.

Let us assume first that $s_{\rm opt} < s_R$. Then, the equation $G'(s) = 0$ is equivalent to:

$$\beta\delta_{GV}(R) + \ln\frac{1}{1-p} + p_{1-s}\ln p + (1-p_{1-s})\ln(1-p) - T = 0$$

or

$$\beta p_{1-s} = \beta\delta_{GV}(R) - T$$

whose solution is:

$$s^* = 1 - \frac{1}{\beta}\ln\frac{\beta(1-\delta_{GV}(R))+T}{\beta\delta_{GV}(R)-T}. \tag{29}$$

Of course, if the r.h.s. of (29) turns out to be negative, then $s_{\rm opt} = 0$. Thus, overall

$$s_{\rm opt} = s_1(p,R,T) \stackrel{\Delta}{=} \left[1 - \frac{1}{\beta}\ln\frac{\beta(1-\delta_{GV}(R))+T}{\beta\delta_{GV}(R)-T}\right]_+, \tag{30}$$

where $[x]_+ \stackrel{\Delta}{=} \max\{x,0\}$. Of course, when $s_{\rm opt} = 0$, the new bound $E_1^*(R,T)$ vanishes.

Next, assume that $s_{\rm opt} > s_R$. In this case,

$$\begin{aligned} E_1(R,T,s) &= F(s) \\ &= \ln 2 - \ln[p^s + (1-p)^s] - \ln[p^{1-s} + (1-p)^{1-s}] - R - sT. \end{aligned} \tag{31}$$

Thus, the optimum s minimizes the convex function

$$\begin{aligned} f(s) &= \ln[p^s + (1-p)^s] + \ln[p^{1-s} + (1-p)^{1-s}] + sT \\ &= \ln\left[1 + (1-p)\left(\frac{p}{1-p}\right)^s + p\left(\frac{1-p}{p}\right)^s\right] + sT. \end{aligned} \tag{32}$$

Equating the derivative to zero, we get:

$$f'(s) \equiv \frac{-\left(\frac{p}{1-p}\right)^s \cdot (1-p)\beta + \left(\frac{1-p}{p}\right)^s \cdot p\beta}{1 + (1-p)\left(\frac{p}{1-p}\right)^s + p\left(\frac{1-p}{p}\right)^s} + T = 0 \tag{33}$$

or equivalently, defining $z = e^{\beta s}$ as the unknown, we get:

$$\frac{-(1-p)/z + pz}{1 + (1-p)/z + pz} = -\frac{T}{\beta},$$

which is a quadratic equation whose relevant (positive) solution is:

$$z = z_0 \stackrel{\Delta}{=} \frac{\sqrt{T^2 + 4p(1-p)(\beta^2 - T^2)} - T}{2p(T+\beta)}$$

provided[5] that $T < \beta$, and so the derivative vanishes at

$$s_{\rm opt} = s_2(p,T) \stackrel{\Delta}{=} \frac{1}{\beta}\ln\left[\frac{\sqrt{T^2+4p(1-p)(\beta^2-T^2)}-T}{2p(T+\beta)}\right].$$

It is not difficult to verify that $s_{\rm opt}$ never exceeds unity. Also, $s_{\rm opt}$ is always positive ($z_0 \geq 1$) since the condition $F'(s_R) > 0$, which is equivalent to the condition $T < \beta(p_{s_R} - p_{1-s_R})$, implies $T < \beta/2$, which in turn is the condition for $s_{\rm opt} > 0$. Note that for $T = 0$, we obtain $s_2(p,0) = 1/2$, in agreement with the Bhattacharyya bound.

In summary, the behavior of the solution can be described as follows: As R increases from 0 to $C = \ln 2 - h(p)$, s_R increases correspondingly from 0 to 1, and so, the expression $\beta(p_{s_R} - p_{1-s_R})$ (which is positive as long as $R < \ln 2 - h(p_{1/2})$) decreases. As long as this expression is still larger than T, we have $F'(s_R) > 0$ and the relevant expression of $E_1^*(R,T,s)$ is $F(s)$, which is maximized at $s = s_2(p,T)$ independently of R. At this range, the slope of $E_1^*(R,T)$, as a function of R, is -1. As R continues to increase, we cross the point where $F'(s_R) = 0$ (a point which can be thought of as an analogue to the critical rate of ordinary decoding) and enter into the region where $F'(s_R) < 0$, for which $E_1^*(R,T) = G(s_1(p,R,T))$.

4.3 More General DMC's and Random Coding Distributions

Assume now a general DMC $\{P(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$ and a general i.i.d. random coding distribution $P(\boldsymbol{x}) = \prod_{i=1}^n P(x_i)$ that satisfy the condition of Theorem 1. As for the second factor of the summand of (18), we have the following:

$$\begin{aligned}\boldsymbol{E}\left\{\left(\sum_{m'\neq m} P(\boldsymbol{y}|\boldsymbol{X}_{m'})\right)^s\right\} &= \boldsymbol{E}\left\{\left(\sum_{Q_{x|y}} N_{\boldsymbol{y}}(Q_{x|y})\cdot\exp\{n\boldsymbol{E}_Q \ln P(Y|X)\}\right)^s\right\} \\ &\doteq \sum_{Q_{x|y}} \boldsymbol{E}\{N_{\boldsymbol{y}}^s(Q_{x|y})\}\cdot\exp\{ns\boldsymbol{E}_Q \ln P(Y|X)\}, \qquad (34)\end{aligned}$$

[5]Note that if $T > \beta$, the decoder will always erase (even for $R = 0$) since for $p < 1/2$, we have $P(\boldsymbol{y}|\boldsymbol{x}_m)/[\sum_{m'\neq m} P(\boldsymbol{y}|\boldsymbol{x}_{m'})] \leq (1-p)^n/p^n = e^{\beta n} < e^{nT}$.

where $N_{\boldsymbol{y}}(Q_{x|y})$ is the number of incorrect codewords whose conditional empirical distribution with $\boldsymbol{y}$ is $Q_{x|y}$ and $\boldsymbol{E}_Q$ is the expectation operator associated with $\hat{P}_{\boldsymbol{y}} \times Q_{x|y}$. Define

$$\mathcal{G}_R = \{Q_{x|y} : \ R + H_Q(X|Y) + \boldsymbol{E}_Q \ln P(x) \geq 0\},$$

where $H_Q(X|Y)$ is the conditional entropy induced by $\hat{P}_{\boldsymbol{y}} \times Q_{x|y}$. Analogously to the case of the BSC (see also Appendix), we have:

$$\boldsymbol{E}\{N^s_{\boldsymbol{y}}(Q_{x|y})\} \doteq \begin{cases} \exp\{ns[R + H_Q(X|Y) + \boldsymbol{E}_Q \ln P(x)]\} & Q_{x|y} \in \mathcal{G}_R \\ \exp\{n[R + H_Q(X|Y) + \boldsymbol{E}_Q \ln P(x)]\} & Q_{x|y} \in \mathcal{G}^c_R \end{cases} \tag{35}$$

Thus,

$$\begin{aligned} \boldsymbol{E}\left\{\left(\sum_{m' \neq m} P(\boldsymbol{y}|\boldsymbol{X}_{m'})\right)^s\right\} \ &\doteq \ \sum_{Q_{x|y} \in \mathcal{G}_R} \exp\{ns[R + H_Q(X|Y) + \boldsymbol{E}_Q \ln P(x)]\} \times \\ &\quad \exp\{ns\boldsymbol{E}_Q \ln P(Y|X)\} + \\ &\quad \sum_{Q_{x|y} \in \mathcal{G}^c_R} \exp\{n[R + H_Q(X|Y) + \boldsymbol{E}_Q \ln P(x)]\} \times \\ &\quad \exp\{ns\boldsymbol{E}_Q \ln P(Y|X)\} \\ &\triangleq \ A + B. \end{aligned} \tag{36}$$

As for A, we obtain:

$$A \doteq \exp\{ns[R + \max_{Q_{x|y} \in \mathcal{G}_R} (H_Q(X|Y) + \boldsymbol{E}_Q \ln[P(X)P(Y|X)])]\} \tag{37}$$

Note that without the constraint $Q_{x|y} \in \mathcal{G}_R$, the maximum of $(H_Q(X|Y) + \boldsymbol{E}_Q \ln[P(X)P(Y|X)])$ is attained at

$$Q_{x|y}(x|y) = P_{x|y}(x|y) \triangleq \frac{P(x)P(y|x)}{\sum_{x \in \mathcal{X}} P(x')P(y|x')}.$$

But since $R < I(X;Y)$, then $P_{x|y}$ is in $\mathcal{G}^c_R$. We argue then that the optimum $Q_{x|y}$ in $\mathcal{G}_R$ is on the boundary of $\mathcal{G}_R$, i.e., it satisfies $R + H_Q(X|Y) + \boldsymbol{E}_Q \ln P(X) = 0$. To see why this is true, consider

the following argument: Let $Q^0_{x|y}$ be any internal point in $\mathcal{G}_R$ and consider the conditional pmf $Q^t = (1-t)Q^0_{x|y} + tP_{x|y}$, $t \in [0,1]$. Define $f(t) = H_{Q^t}(X|Y) + \boldsymbol{E}_{Q^t} \ln[P(X)P(Y|X)]$. Obviously, f is concave and $f(0) \le f(1)$. Now, since $Q^0 \in \mathcal{G}_R$ and $Q^1 = P_{x|y} \in \mathcal{G}_R^c$, then by the continuity of the function $R + H_{Q^t}(X|Y) + \boldsymbol{E}_{Q^t} \ln P(X)$, there must be some $t = t_0$ for which Q^{t_0} is on the boundary of $\mathcal{G}_R$. By the concavity of f, $f(t_0) \ge (1-t_0)f(0) + t_0 f(1) \ge f(0)$. Thus, any internal point of $\mathcal{G}_R$ can be improved by a point on the boundary between $\mathcal{G}_R$ and $\mathcal{G}_R^c$. Therefore, we have

$$\begin{aligned}
& \max_{Q_{x|y} \in \mathcal{G}_R} (H_Q(X|Y) + \boldsymbol{E}_Q \ln[P(X)P(Y|X)])] \\
= & \max_{\{Q_{x|y}:\ H_Q(X|Y) + \boldsymbol{E}_Q \ln P(X) = -R\}} [H_Q(X|Y) + \boldsymbol{E}_Q \ln P(X) + \boldsymbol{E}_Q \ln P(Y|X)] \\
= & \max_{\{Q_{x|y}:\ H_Q(X|Y) + \boldsymbol{E}_Q \ln P(X) = -R\}} [-R + \boldsymbol{E}_Q \ln P(Y|X)] \\
= & -R + \max_{\{Q_{x|y}:\ H_Q(X|Y) + \boldsymbol{E}_Q \ln P(X) = -R\}} \boldsymbol{E}_Q \ln P(Y|X) \\
= & -R + \max_{Q_{x|y} \in \mathcal{G}_R} \boldsymbol{E}_Q \ln P(Y|X) \qquad (38)
\end{aligned}$$

which means that $A \doteq e^{-ns\Delta(R)}$, where

$$\Delta(R) = \min_{Q_{x|y} \in \mathcal{G}_R} \boldsymbol{E}_Q \ln[1/P(Y|X)].$$

The achiever of $\Delta(R)$ is of the form

$$Q(x|y) = \frac{P(x)P^{s_R}(y|x)}{\sum_{x' \in \mathcal{X}} P(x')P^{s_R}(y|x')},$$

where s_R is such that $H_Q(X|Y) + \boldsymbol{E}_Q \ln P(X) = -R$, or equivalently, s_R is the solution [6] to the equation $s\gamma'(s) - \gamma(s) = R$. In other words,

$$\Delta(R) = \frac{\sum_{x \in \mathcal{X}} P(x)P^{s_R}(y|x) \ln[1/P(y|x)]}{\sum_{x \in \mathcal{X}} P(x)P^{s_R}(y|x)} = \gamma'(s_R).$$

[6] Observe that for $s = 0$, $H_Q(X|Y) + \boldsymbol{E}_Q \ln P(X) = 0$ and for $s = 1$, $H_Q(X|Y) + \boldsymbol{E}_Q \ln P(X) = -I(X;Y) < -R$. Thus for $R < I(X;Y)$, $s_R \in [0,1)$.

Considering next the expression of B, we have:

$$B \doteq \exp\{n[R + \max_{Q_{x|y} \in \mathcal{G}_R^c} (H_Q(X|Y) + \boldsymbol{E}_Q \ln P(X) + s\boldsymbol{E}_Q \ln P(Y|X))]\}.$$

The unconstrained maximizer of $(H_Q(X|Y) + \boldsymbol{E}_Q \ln P(X) + s\boldsymbol{E}_Q \ln P(Y|X))$ is

$$Q_{x|y}^{(s)}(x|y) = \frac{P(x)P^s(y|x)}{\sum_{x' \in \mathcal{X}} P(x')P^s(y|x')}.$$

Now, there are two cases, depending on the value of s: If s is such that $Q_{x|y}^{(s)} \in \mathcal{G}_R^c$, or equivalently, $s > s_R$, then $B \doteq e^{-n[\gamma(s)-R]}$. If $Q_{x|y}^{(s)} \in \mathcal{G}_R$, namely, $s \leq s_R$, then once again, the optimum is attained at the boundary between $\mathcal{G}_R$ and $\mathcal{G}_R^c$, and then $B \doteq e^{-ns\gamma'(s_R)}$. In summary, $B \doteq e^{-n\Lambda(R,s)}$, where

$$\Lambda(R,s) = \begin{cases} \gamma(s) - R & s > s_R \\ s\gamma'(s_R) & s \leq s_R \end{cases}$$

The dominant term between A and B is obviously always B because it is either of the same exponential order of A, in the case $s \leq s_R$, or has a slower exponential decay, when $s > s_R$, as then the global (unconstrained) maximum of $[H_Q(X|Y) + \boldsymbol{E}_Q \ln P(X) + s\boldsymbol{E}_Q \ln P(Y|X)]$ is achieved. Thus, putting it all together, we get:

$$\begin{aligned} \overline{\Pr}\{\mathcal{E}_1\} &\doteq e^{nsT} \cdot |\mathcal{Y}|^n \cdot e^{-n\gamma(1-s)} \cdot e^{-n\Lambda(R,s)} \\ &= e^{-nE_1^*(R,T,s)} \end{aligned} \tag{39}$$

and the optimum $s \geq 0$ gives $E_1^*(R,T)$.

Appendix

We begin with a simple large deviations bound regarding the distance enumerator. In fact, this bound (in a slightly different form) was given already in [7], but we present here too for the sake of completeness. For $a, b \in [0,1]$, consider the binary divergence

$$\begin{aligned} D(a\|b) &\stackrel{\Delta}{=} a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b} \\ &= a\ln\frac{a}{b} + (1-a)\ln\left[1+\frac{b-a}{1-b}\right]. \end{aligned} \tag{A.1}$$

To derive a lower bound to $D(a\|b)$, let us use the inequality

$$\ln(1+x) = -\ln\frac{1}{1+x} = -\ln\left(1-\frac{x}{1+x}\right) \geq \frac{x}{1+x}, \tag{A.2}$$

and then

$$\begin{aligned} D(a\|b) &\geq a\ln\frac{a}{b} + (1-a)\cdot\frac{(b-a)/(1-b)}{1+(b-a)/(1-b)} \\ &= a\ln\frac{a}{b} + b - a \\ &> a\left(\ln\frac{a}{b} - 1\right). \end{aligned} \tag{A.3}$$

Consider first the binary case (the extension to the general case is straightforward as will be explained below). For every given $\boldsymbol{y}$, $N_{\boldsymbol{y}}(d)$ is the sum of the $e^{nR}-1$ independent binary random variables, $\{1\{d(\boldsymbol{X}_{m'},\boldsymbol{y})=d\}\}_{m'\neq m}$, where the probability that $d(\boldsymbol{X}_{m'},\boldsymbol{y}) = n\delta$ is exponentially $b = e^{-n[\ln 2 - h(\delta)]}$. The event $N_{\boldsymbol{y}}(n\delta) \geq e^{nA}$, for $A \in [0,R)$, means that the relative frequency of the event $1\{d(\boldsymbol{X}_{m'},\boldsymbol{y}) = n\delta\}$ is at least $a = e^{-n(R-A)}$. Thus, by the Chernoff bound:

$$\begin{aligned} \Pr\{N_{\boldsymbol{y}}(n\delta) \geq e^{nA}\} &\stackrel{\cdot}{\leq} \exp\left\{-(e^{nR}-1)D(e^{-n(R-A)}\|e^{-n[\ln 2-h(\delta)]})\right\} \\ &\stackrel{\cdot}{\leq} \exp\left\{-e^{nR}\cdot e^{-n(R-A)}(n[(\ln 2 - R - h(\delta) + A] - 1)\right\} \\ &\leq \exp\left\{-e^{nA}(n[\ln 2 - R - h(\delta) + A] - 1)\right\}. \end{aligned} \tag{A.4}$$

Therefore, for $\delta \in \mathcal{G}_R^c$, we have:

$$\begin{aligned}
\boldsymbol{E}\{N_{\boldsymbol{y}}^s(n\delta)\} &\leq e^{n\epsilon s} \cdot \Pr\{1 \leq N_{\boldsymbol{y}}(n\delta) \leq e^{n\epsilon}\} + e^{nRs} \cdot \Pr\{N_{\boldsymbol{y}}(n\delta) \geq e^{n\epsilon}\} \\
&\leq e^{n\epsilon s} \cdot \Pr\{N_{\boldsymbol{y}}(n\delta) \geq 1\} + e^{nRs} \cdot \Pr\{N_{\boldsymbol{y}}(n\delta) \geq e^{n\epsilon}\} \\
&\leq e^{n\epsilon s} \cdot \boldsymbol{E}\{N_{\boldsymbol{y}}(n\delta)\} + e^{nRs} \cdot e^{-(n\epsilon-1)e^{n\epsilon}} \\
&\leq e^{n\epsilon s} \cdot e^{n[R+h(\delta)-\ln 2]} + e^{nRs} \cdot e^{-(n\epsilon-1)e^{n\epsilon}}. \qquad \text{(A.5)}
\end{aligned}$$

One can let ϵ vanish with n sufficiently slowly that the second term is still superexponentially small, e.g., $\epsilon = 1/\sqrt{n}$. Thus, for $\delta \in \mathcal{G}_R^c$, $\boldsymbol{E}\{N_{\boldsymbol{y}}^s(n\delta)\}$ is exponentially bounded by $e^{n[R+h(\delta)-\ln 2]}$ independently of s. For $\delta \in \mathcal{G}_R$, we have:

$$\begin{aligned}
\boldsymbol{E}\{N_{\boldsymbol{y}}^s(n\delta)\} &\leq e^{ns[R+h(\delta)-\ln 2+\epsilon]} \cdot \Pr\{N_{\boldsymbol{y}}(n\delta) \leq e^{n[R+h(\delta)-\ln 2+\epsilon]}\} + \\
&\qquad e^{nRs} \cdot \Pr\{N_{\boldsymbol{y}}(n\delta) \geq e^{n[R+h(\delta)-\ln 2+\epsilon]}\} \\
&\leq e^{ns[R+h(\delta)-\ln 2+\epsilon]} + e^{nRs} \cdot e^{-(n\epsilon-1)e^{n\epsilon}} \qquad \text{(A.6)}
\end{aligned}$$

where again, the second term is exponentially negligible.

To see that both bounds are exponentially tight, consider the following lower bounds. For $\delta \in \mathcal{G}_R^c$,

$$\begin{aligned}
\boldsymbol{E}\{N_{\boldsymbol{y}}^s(n\delta)\} &\geq 1^s \cdot \Pr\{N_{\boldsymbol{y}}(n\delta) = 1\} \\
&= e^{nR} \cdot \Pr\{d_H(\boldsymbol{X}, \boldsymbol{y}) = n\delta\} \cdot [1 - \Pr\{d_H(\boldsymbol{X}, \boldsymbol{y}) = n\delta\}]^{e^{nR}-1} \\
&\doteq e^{nR} e^{-n[\ln 2 - h(\delta)]} \cdot \left[1 - e^{-n[\ln 2 - h(\delta)]}\right]^{e^{nR}} \\
&= e^{n[R+h(\delta)-\ln 2]} \cdot \exp\{e^{nR} \ln[1 - e^{-n[\ln 2 - h(\delta)]}]\}. \qquad \text{(A.7)}
\end{aligned}$$

Using again the inequality in (A.2), the second factor is lower bounded by

$$\exp\{-e^{nR} e^{-n[\ln 2 - h(\delta)]} / (1 - e^{-n[\ln 2 - h(\delta)]})\} = \exp\{-e^{-n[\ln 2 - R - h(\delta)]} / (1 - e^{-n[\ln 2 - h(\delta)]})\}$$

which clearly tends to unity as $\ln 2 - R - h(\delta) > 0$ for $\delta \in \mathcal{G}_R^c$. Thus, $\boldsymbol{E}\{N_{\boldsymbol{y}}^s(n\delta)\}$ is exponentially lower bounded by $e^{n[R+h(\delta)-\ln 2]}$. For $\delta \in \mathcal{G}_R$, and an arbitrarily small $\epsilon > 0$, we have:

$$\begin{aligned} \boldsymbol{E}\{N_{\boldsymbol{y}}^s(n\delta)\} &\geq e^{ns[R+h(\delta)-\ln 2-\epsilon]} \cdot \Pr\{N_{\boldsymbol{y}}(n\delta) \geq e^{n[R+h(\delta)-\ln 2-\epsilon]}\} \\ &= e^{ns[R+h(\delta)-\ln 2-\epsilon]} \cdot \Big(1 - \Pr\{N_{\boldsymbol{y}}(n\delta) < e^{n[R+h(\delta)-\ln 2-\epsilon]}\}\Big) \end{aligned} \tag{A.8}$$

where $\Pr\{N_{\boldsymbol{y}}(n\delta) < e^{n[R+h(\delta)-\ln 2-\epsilon]}\}$ is again upper bounded, for an internal point in $\mathcal{G}_R$, by a double exponentially small quantity as above. For δ near the boundary of $\mathcal{G}_R$, namely, when $R + h(\delta) - \ln 2 \approx 0$, we can lower bound $\boldsymbol{E}\{N_{\boldsymbol{y}}^s(n\delta)\}$ by slightly reducing R to $R' = R - \epsilon$ (where $\epsilon > 0$ is very small). This will make δ an internal point of $\mathcal{G}_{R'}^c$ for which the previous bound applies, and this bound is of the exponential order of $e^{n[R'+h(\delta)-\ln 2]}$. Since $R' + h(\delta) - \ln 2$ is still very close to zero, then $e^{n[R'+h(\delta)-\ln 2]}$ is of the same exponential order as $e^{ns[R+h(\delta)-\ln 2]}$ since both are about $e^{0 \cdot n}$.

The above proof extends straightforwardly from the binary case to the more general case. The only difference is that in the general case, for a given $\boldsymbol{y}$, the probability that a random codeword, drawn under $\{P(x)\}$, would have a given conditional empirical distribution $Q_{x|y}$ with $\boldsymbol{y}$, is exponentially $e^{n[H_Q(X|Y)+\boldsymbol{E}_Q \ln P(X)]}$. Thus, $h(\delta) - \ln 2$ of the binary case has to be replaced by $H_Q(X|Y) + \boldsymbol{E}_Q \ln P(X)$ in all places.

References

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

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

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

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

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

[6] 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, November 2007.

[7] N. Merhav, "Relations between random coding exponents and the statistical physics of random codes," submitted to *IEEE Trans. Inform. Theory*, August 2007. Also, available on–line at: [http://www.ee.technion.ac.il/people/merhav/papers/p117.pdf].

[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] M. Mézard and A. Montanari, *Constraint satisfaction networks in physics and computation*, draft, February 27, 2006. Available on–line at: [http://www.stanford.edu/∼montanar/BOOK/book.html].

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