List Decoding – Random Coding Exponents and Expurgated Exponents*

Neri Merhav

Department of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, ISRAEL
E–mail: merhav@ee.technion.ac.il

Abstract

Some new results are derived concerning random coding error exponents and expurgated exponents for list decoding with a deterministic list size L. Two asymptotic regimes are considered, the *fixed list–size* regime, where L is fixed independently of the block length n, and the *exponential list–size*, where L grows exponentially with n. We first derive a general upper bound on the list–decoding average error probability, which is suitable for both regimes. This bound leads to more specific bounds in the two regimes. In the fixed list–size regime, the bound is related to known bounds and we establish its exponential tightness. In the exponential list–size regime, we establish the achievability of the well known sphere packing lower bound. Relations to guessing exponents are also provided. An immediate byproduct of our analysis in both regimes is the universality of the maximum mutual information (MMI) list decoder in the error exponent sense. Finally, we consider expurgated bounds at low rates, both using Gallager's approach and the Csiszár–Körner–Marton approach, which is, in general better (at least for $L = 1$). The latter expurgated bound, which involves the notion of *multi–information*, is also modified to apply to continuous alphabet channels, and in particular, to the Gaussian memoryless channel, where the expression of the expurgated bound becomes quite explicit.

Index Terms: List decoding, error exponent, random coding, sphere packing, expurgated exponent.

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

1 Introduction

The concept of list decoding was first introduced independently by Elias [8] and Wozencraft [27] in the late fifties of the previous century. A list decoder, rather than outputting a single estimate of the transmitted message, produces a list of L candidates, among which the final 'winner' will be eventually selected upon receiving additional information. This is naturally applicable in concatenated coded communication systems, most notably, in wireless systems (see, e.g., [20], [21]), where the list decoder corresponds to the inner code, and the above–mentioned additional information is available to the outer decoder, for example, information concerning the structure of the outer code or contextual information that is present when not all the source redundancy has been removed. Accordingly, the error event associated with a list decoder is the event where the actual message that has been sent is not in the list.

While considerable work has been done on list decoding for codes with certain algebraic/combinatorial structures (most notably, linear codes along with some subclasses of linear codes – see, e.g., [3], [12] and many references therein), with emphasis on algorithmic issues, the main focus of this work is on Shannon–theoretic issues, like achievable error exponents pertaining to list decoding.

A few words of background on list decoding error exponents are therefore in order. First, a clear distinction should be made between two different classes of list decoders, according to whether the list size is a deterministic number or a random variable, which depends on the random channel output vector (see, e.g., [9], [24], [25] as well as many other later further developments). In this paper, we confine ourselves to the former class, which in turn, is also subdivided into two sub–classes with two different asymptotic regimes: (i) the *fixed list–size regime*, where the deterministic list size L is fixed, independently of the block length n, and (ii) the *exponential list–size regime*, where L is an exponential function of the block length n, namely, $L = e^{\lambda n}$, for some fixed $\lambda > 0$, that is independent of n. Both regimes will be considered in this paper.

First, regarding the fixed list–size regime, the extension of the ordinary random coding error exponent analysis to account for list decoding with a fixed list size L, was considered by both Gallager [10, p. 538, Exercise 5.20] and Viterbi and Omura [26, p. 215, Exercise 3.16] simple and straightforward enough to be left as an exercise for the reader. In particular, for the ensemble of independent random selection of codewords according to an i.i.d. distribution, in the above exercises

it is shown that the average probability of list–decoding error is upper bounded by an exponential function, whose exponential decay rate is given by the same expression as that of Gallager's random coding exponent, $E_{\rm r}(R)$ [10, eq. (5.6.16)] (R being the coding rate), except that the interval of the maximization over the auxiliary parameter ρ is expanded from $[0,1]$ to $[0,L]$ (thus $L=1$ recovers ordinary decoding as a special case). Obviously, when L exceeds the global maximizer of the same expression over $[0,\infty)$, the resulting exponent coincides with the sphere–packing exponent $E_{\rm sp}(R)$ [10, Theorem 5.8.1]. Similarly as in ordinary decoding, while the upper bound and the lower bound agree at high rates, there is some gap at low rates and the random coding achievability result can be improved at low rates by expurgation. Blinovsky [4] has shown that, in analogy to ordinary decoding, the expurgated exponent of the fixed list–size regime at $R=0$ is tight in the sense that there is also a lower bound of the same exponential rate.

The exponential list–size regime is interesting at least as much. In 1967, Shannon, Gallager and Berlekamp [22, Theorem 2] have established (among other results), a lower bound on the probability of list decoding, which is meaningful also for the exponential list–size regime. According to this lower bound, the probability of list error cannot decay exponentially more rapidly than $e^{-nE_{\rm sp}(R-\lambda)}$, where again, λ is the list–size exponent (see also [6, p. 196, Problem 27, part (b)], [26, p. 179, Lemma 3.8.1]). On the other hand, in [6, p. 196, Problem 27, part (a)], the reader is asked to show that an exponential decay at the rate $e^{-nE_{\rm r}(R-\lambda)}$ is achievable (see also [13, p. 3767, eq. (46)]). One of our results is that a decay rate of $e^{-nE_{\rm sp}(R-\lambda)}$ is achievable, thus closing the gap between the upper bound and the lower bound and fully characterizing the best achievable error exponent function (reliability function for list decoding) in the exponential list–size regime according to $E_{\rm sp}(R-\lambda)$.

In this paper, we contribute several additional results, both on the fixed list–size regime and the exponential list–size regime. We first derive a general upper bound on the average list–decoding error probability, pertaining to the ordinary ensemble of fixed composition codes, i.e., the ensemble defined by an independent, random selection of each codeword from a single type class. This general bound is suitable for both regimes. The general upper bound leads to more specific upper bounds in the two regimes. In the fixed list–size regime, our bound is intimately related to the above mentioned bounds with the extended interval of optimization – $[0,L]$, except that it corresponds to the ensemble of fixed composition codes (rather than an i.i.d. codeword distribution). We also show that this random coding bound is exponentially tight for the average code by deriving a compatible

lower bound with the same exponent, thereby extending the result in [11] to list decoding. In the exponential list size–regime, we establish, as mentioned earlier, the achievability of $E_{\rm sp}(R-\lambda)$ and thereby close the gap between the lower bound and the upper bound. An immediate byproduct of these derivations is the universality of the maximum mutual information (MMI) list decoder in the error exponent sense. The derivations involve a random variable that is defined as the number of incorrect codewords whose likelihood score exceed the score of the correct codeword. Accordingly, we also provide bounds on the moments of this random variable, which are intimately related to guessing exponents [1], [2]. Finally, we consider expurgated bounds at low rates, using both Gallager's approach [10, Section 5.7], [26, Section 3.3] and the Csiszár–Körner–Marton approach [6, p. 185, problem 17], [7]. which is in general better at least for $L = 1$ [19]. The latter expurgated bound, which happens to involve the notion of *multi–information*,[1] is also modified to apply to continuous alphabet channels, and in particular, to the Gaussian memoryless channel, where the expression of the expurgated bound becomes quite explicit.

The outline of the remaining part of this paper is as follows. In Section 2, we establish notation conventions, formalize the problem and provide some background and preliminaries. In Section 3, we first derive the general upper bound on the average probability of list error for the ordinary ensemble of fixed composition code (Subsection 3.1), and then particularize its analysis to both the fixed list–size regime (Subsection 3.2) and the exponential list–size regime (Subsection 3.3), and finally, we relate our findings to the problem of guessing (Subsection 3.4). Section 4 is devoted to expurgated exponents. Finally, in Section 5, we present a problem for future work.

2 Notation Conventions, Problem Formulation and Preliminaries

2.1 Notation Conventions

Throughout the paper, random variables will be denoted by capital letters, specific values they may take will be denoted by the corresponding lower case letters, and their alphabets will be denoted by calligraphic letters. Random vectors and their realizations will be denoted, respectively, by capital

[1] Multi–information [23] is a natural generalization of the notion of mutual information that accommodates the statistical dependence of more than two random variables. It is defined as the sum of the marginal entropies of these random variables minus their joint entropy, or equivalently, as the Kullback–Leibler divergence between the joint distribution and the product of marginals. To the best of our knowledge, in this paper, it is the first occasion that the multi–information plays a natural role in a concrete information–theoretic formula.

letters and the corresponding lower case letters, both in the bold face font. Their alphabets will be superscripted by their dimensions. For example, the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$, ($n$ – positive integer) may take a specific vector value $\boldsymbol{x} = (x_1, \ldots, x_n)$ in $\mathcal{X}^n$, the n–th order Cartesian power of $\mathcal{X}$, which is the alphabet of each component of this vector. Sources and channels will be denoted by the letter P or Q, superscripted by the names of the relevant random variables/vectors and their conditionings, if applicable, following the standard notation conventions, e.g., Q_X, $P_{Y|X}$, and so on. When there is no room for ambiguity, these subscripts will be omitted. The probability of an event $\mathcal{E}$ will be denoted by $\Pr\{\mathcal{E}\}$, and the expectation operator with respect to (w.r.t.) a probability distribution P will be denoted by $\boldsymbol{E}_P\{\cdot\}$. Again, the subscript will be omitted if the underlying probability distribution is clear from the context. The entropy of a generic distribution Q on $\mathcal{X}$ will be denoted by $H(Q)$. For two positive sequences a_n and b_n, the notation $a_n \doteq b_n$ will stand for equality in the exponential scale, that is, $\lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0$. Similarly, $a_n \dot{\leq} b_n$ means that $\limsup_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} \leq 0$, and so on. The indicator function of an event $\mathcal{E}$ will be denoted by $\mathcal{I}\{E\}$. The notation $[x]_+$ will stand for $\max\{0, x\}$.

The empirical distribution of a sequence $\boldsymbol{x} \in \mathcal{X}^n$, which will be denoted by $\hat{P}_{\boldsymbol{x}}$, is the vector of relative frequencies $\hat{P}_{\boldsymbol{x}}(x)$ of each symbol $x \in \mathcal{X}$ in $\boldsymbol{x}$. The type class of $\boldsymbol{x} \in \mathcal{X}^n$, denoted $\mathcal{T}(\boldsymbol{x})$, is the set of all vectors $\boldsymbol{x}'$ with $\hat{P}_{\boldsymbol{x}'} = \hat{P}_{\boldsymbol{x}}$. When we wish to emphasize the dependence of the type class on the empirical distribution $\hat{P}$, we will denote it by $\mathcal{T}(\hat{P})$. Information measures associated with empirical distributions will be denoted with 'hats' and will be subscripted by the sequences from which they are induced. For example, the entropy associated with $\hat{P}_{\boldsymbol{x}}$, which is the empirical entropy of $\boldsymbol{x}$, will be denoted by $\hat{H}_{\boldsymbol{x}}(X)$. An alternative notation, following the conventions described in the previous paragraph, is $H(\hat{P}_{\boldsymbol{x}})$. Similar conventions will apply to the joint empirical distribution, the joint type class, the conditional empirical distributions and the conditional type classes associated with pairs (and multiples) of sequences of length n. Accordingly, $\hat{P}_{\boldsymbol{xy}}$ would be the joint empirical distribution of $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_i, y_i)\}_{i=1}^n$, $\mathcal{T}(\boldsymbol{x}, \boldsymbol{y})$ or $\mathcal{T}(\hat{P}_{\boldsymbol{xy}})$ will denote the joint type class of $(\boldsymbol{x}, \boldsymbol{y})$, $\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})$ will stand for the conditional type class of $\boldsymbol{x}$ given $\boldsymbol{y}$, $\hat{H}_{\boldsymbol{xy}}(X, Y)$ will designate the empirical joint entropy of $\boldsymbol{x}$ and $\boldsymbol{y}$, $\hat{H}_{\boldsymbol{xy}}(X|Y)$ will be the empirical conditional entropy, $\hat{I}_{\boldsymbol{xy}}(X; Y)$ will denote empirical mutual information, and so on.

2.2 Problem Formulation

Let $\{P(y|x)\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$ be a matrix single–letter transition probabilities of a discrete memoryless channel (DMC) with finite input and output alphabets, $\mathcal{X}$ and $\mathcal{Y}$, respectively. For a given $R > 0$, and a block length n, let $\mathcal{C} = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{M-1}\}$, $M = e^{nR} + 1$, $\boldsymbol{x}_m \in \mathcal{X}^n$, $m = 0, 1, \ldots, M-1$, be a given codebook, known to both the encoder and decoder.[2] We consider a list decoder that outputs a list of L candidate estimates of the transmitted message, based on the channel output $\boldsymbol{y} \in \mathcal{Y}^n$. We will be interested in two asymptotic regimes. In the first regime, which will be referred to as the *fixed list–size regime*, L is fixed and independent of n and in the second regime, which will be referred to as the *exponential list–size regime*, $L = e^{\lambda n}$, where $0 < \lambda < R$ is a given constant, independent of n. The figure of merit, in both cases, is the probability of list error, namely, the probability that the correct message is not in the list. Our focus will be on achievable error exponents associated with the probability of list error as functions of (R, L) in the fixed list–size regime, or as functions of (R, λ) in the exponential list–size regime.

As usual, the main mechanism for deriving achievability results will be random coding. The random coding ensemble will be defined by independent random selection of each codeword according to a probability distribution $P(\boldsymbol{x})$, which, unless specified otherwise, will be the uniform distribution across a given type class $\mathcal{T}(Q)$ (except for a few places, where it will be the product measure $\prod_{i=1}^{n} Q(x_i)$). Once selected, the codebook is revealed to both the encoder and the decoder.

The discussion in the subsections 2.3 and 3.1 is non–asymptotic – it applies to any two positive integers n and $L \leq e^{nR}$. This means, of course, that in these subsections, it is immaterial if we consider the fixed list–size regime or the exponential list–size regime.

2.3 Preliminaries

We begin from the fundamental fact that, as intuition suggests, the optimal list decoder generates the list according to the L largest likelihood values $\{P(\boldsymbol{y}|\boldsymbol{x}_m)\}$. This follows from the following simple consideration, which is a fairly simple extension of the one used to prove the optimality of the maximum likelihood (ML) decoder in ordinary decoding ($L = 1$). For a given unordered list of L distinct integers $m_1, \ldots, m_L$, all in the range $\{0, 1, \ldots, M-1\}$, let $\Omega(m_1, \ldots, m_L) \subseteq \mathcal{Y}^n$ denote the

[2] To avoid the need for rounding functions, we assume throughout that R is such that e^{nR} is integer. Also, M is defined as $e^{nR} + 1$, rather than e^{nR}, simply for reasons of convenience.

region where the list decoder outputs the messages $m_1, \ldots, m_L$. Obviously, the $\left(\begin{array}{c} M \\ L \end{array} \right)$ different regions $\{\Omega(m_1, \ldots, m_L)\}$ form a partition of $\mathcal{Y}^n$. For a given $\boldsymbol{y}$, let $m_1^*(\boldsymbol{y}), \ldots m_L^*(\boldsymbol{y})$ achieve the L highest rankings of $P(\boldsymbol{y}|\boldsymbol{x}_m)$. Then, the probability of correct list decoding (i.e., the probability that the correct message is in the list) is given by

$$\begin{aligned}
P_c &= \frac{1}{M} \sum_{m=1}^{M} \sum_{m_1,\ldots,m_L} \left[\sum_{i=1}^{L} \delta(m_i - m) \right] \sum_{\boldsymbol{y} \in \Omega(m_1,\ldots,m_L)} P(\boldsymbol{y}|\boldsymbol{x}_m) \\
&= \frac{1}{M} \sum_{m_1,\ldots,m_L} \sum_{\boldsymbol{y} \in \Omega(m_1,\ldots,m_L)} \sum_{i=1}^{L} \sum_{m=1}^{M} \delta(m_i - m) P(\boldsymbol{y}|\boldsymbol{x}_m) \\
&= \frac{1}{M} \sum_{m_1,\ldots,m_L} \sum_{\boldsymbol{y} \in \Omega(m_1,\ldots,m_L)} \sum_{i=1}^{L} P(\boldsymbol{y}|\boldsymbol{x}_{m_i}) \\
&\leq \frac{1}{M} \sum_{m_1,\ldots,m_L} \sum_{\boldsymbol{y} \in \Omega(m_1,\ldots,m_L)} \sum_{i=1}^{L} P(\boldsymbol{y}|\boldsymbol{x}_{m_i^*(\boldsymbol{y})}) \\
&= \frac{1}{M} \sum_{\boldsymbol{y} \in \mathcal{Y}^n} \sum_{i=1}^{L} P(\boldsymbol{y}|\boldsymbol{x}_{m_i^*(\boldsymbol{y})}) \qquad (1)
\end{aligned}$$

and the inequality becomes an equality for the list decoder that outputs $m_1^*(\boldsymbol{y}), \ldots, m_L^*(\boldsymbol{y})$.

The following results, on random coding exponents for list decoding, are well known. First, it is shown both in Gallager [10, p. 538, Exercise 5.20] and Viterbi and Omura [26, p. 215, Exercise 3.16], that the average probability of list error is upper bounded by

$$\overline{P_e} \leq \min_{0 \leq \rho \leq L} M^{\rho} \sum_{\boldsymbol{y} \in \mathcal{Y}^n} \left[\sum_{\boldsymbol{x} \in \mathcal{X}^n} P(\boldsymbol{x}) P(\boldsymbol{y}|\boldsymbol{x})^{1/(1+\rho)} \right]^{1+\rho}, \qquad (2)$$

which is very similar to Gallager's well–known random coding bound for ordinary decoding, except that here, the range of optimization of the parameter ρ is stretched from $[0, 1]$ to $[0, L]$. When the random coding distribution $P(\boldsymbol{x})$ is i.i.d. with a single–letter marginal Q, and the fixed list–size regime is considered, this is, of course, an exponential function whose exponential rate is

$$E_{\mathrm{r}}(R, L) = \sup_{0 \leq \rho \leq L} \sup_{Q} [E_0(\rho, Q) - \rho R], \qquad (3)$$

where $E_0(\rho)$ is the well–known Gallager function

$$E_0(\rho, Q) = -\ln \left(\sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} Q(x) P(y|x)^{1/(1+\rho)} \right]^{1+\rho} \right). \qquad (4)$$

Thus, $E_{\rm r}(R,1)$ is the ordinary random coding error exponent, which will also be denoted by $E_{\rm r}(R)$, following traditional notation conventions.

Concerning the exponential list–size regime, where $L=e^{\lambda n}$, Shannon, Gallager and Berlekamp [22] have proved that the exponential rate of the list error probability cannot be faster than $E_{\rm sp}(R-\lambda)$, where $E_{\rm sp}(R)$ is the sphere–packing exponent, defined as

$$E_{\rm sp}(R)=\sup_{\rho\geq 0}\sup_{Q}[E_0(\rho,Q)-\rho R] \tag{5}$$

or, equivalently,[3] as

$$E_{\rm sp}(R)=\sup_{Q}\inf_{\{\tilde{P}_{Y|X}:\ \tilde{I}(X;Y)\leq R\}} D(\tilde{P}_{Y|X}\|P_{Y|X}|Q), \tag{6}$$

where $\tilde{I}(X;Y)$ is the mutual information induced by $X\sim Q$ and $\tilde{P}_{Y|X}$ and

$$D(\tilde{P}_{Y|X}\|P_{Y|X}|Q)\stackrel{\Delta}{=}\sum_{x\in\mathcal{X}}Q(x)\sum_{y\in\mathcal{Y}}\tilde{P}_{Y|X}(y|x)\ln\frac{\tilde{P}_{Y|X}(y|x)}{P_{Y|X}(y|x)}. \tag{7}$$

In [6, Problem 27], the reader is asked to prove that $E_{\rm r}(R-\lambda)$ is achievable.

3 Upper Bounds for an Ordinary Ensemble of Fixed Composition Codes

In this section, we first derive the general upper bound on the average probability of list error for the ordinary ensemble of fixed composition codes defined above (Subsection 3.1), and then particularize its analysis to both the fixed list–size regime (Subsection 3.2) and the exponential list–size regime (Subsection 3.3). Finally, we relate our findings to the problem of guessing (Subsection 3.4).

3.1 A General Upper Bound

The following theorem holds for every positive integer n and every $L\leq e^{nR}$.

Theorem 1 *Consider the random coding ensemble of rate R codes for a DMC $\{P(y|x),\ x\in\mathcal{X},\ y\in\mathcal{Y}\}$, as described in Subsection 2.2. The average probability of list error, $\overline{P_e}$, associated with the optimal list decoder, is upper bounded by*

$$\overline{P_e}\leq\sum_{\boldsymbol{x},\boldsymbol{y}}P(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x})\exp\left\{-nL\left[\hat{I}_{\boldsymbol{xy}}(X;Y)+\frac{\ln L}{n}-R-\delta_n-\frac{1}{n}\right]_+\right\}, \tag{8}$$

where $P(\boldsymbol{x})$ is the uniform distribution over $\mathcal{T}(Q)$ and $\delta_n=O((\log n)/n)$.

[3] In [6], see eq. (5.19) on page 166 and compare with Problem 23 on page 192. See also the next footnote below.

Eq. (8) serves as our general upper bound. It can be further analyzed, using the method of types, both in the fixed list–size regime and in the in the exponential list–size regime. These analyses will be carried out in the following two subsections, respectively. Note that the parameter L appears twice in the right–hand side (r.h.s.) of eq. (8). In the fixed list–size regime, the first occurrence of L (outside the square brackets) will be the important one, whereas in the exponential list–size regime, it will be the second occurrence of L that will play the more important role. The remaining part of this subsection is devoted to the proof of Theorem 1.

Proof. For the purpose of deriving an upper bound, it is legitimate to analyze a sub–optimal list decoder that outputs the L messages with the L highest values of the empirical mutual information $\hat{I}_{\boldsymbol{x}_m\boldsymbol{y}}(X;Y)$, that is, the *maximum mutual information (MMI) list decoder*. Without loss of generality, assume that $\boldsymbol{x}_0$ was transmitted and $\boldsymbol{y}$ was received. The average probability of list error, for a given $(\boldsymbol{x}_0, \boldsymbol{y})$, denoted $\overline{P_e(\boldsymbol{x}_0, \boldsymbol{y})}$, is the probability of list error which accounts for the randomness of all other $M-1$ codewords. The overall average probability of list error, $\overline{P_e}$, is the expectation of $\overline{P_e(\boldsymbol{X}_0, \boldsymbol{Y})}$ w.r.t. $P(\boldsymbol{x}_0)P(\boldsymbol{y}|\boldsymbol{x}_0)$, where $P(\boldsymbol{x}_0)$ is the uniform distribution across $\mathcal{T}(Q)$. For the given $\boldsymbol{x}_0$ and $\boldsymbol{y}$, we then have

$$\overline{P_e(\boldsymbol{x}_0, \boldsymbol{y})} = \sum_{\ell=L}^{M} \left(\begin{array}{c} M \\ \ell \end{array} \right) [q(\boldsymbol{x}_0, \boldsymbol{y})]^{\ell} [1 - q(\boldsymbol{x}_0, \boldsymbol{y})]^{M-\ell}, \tag{9}$$

where

$$q(\boldsymbol{x}_0, \boldsymbol{y}) = \frac{|\mathcal{T}(Q) \cap \{\boldsymbol{x} : \ \hat{I}_{\boldsymbol{x}\boldsymbol{y}}(X;Y) \geq \hat{I}_{\boldsymbol{x}_0\boldsymbol{y}}(X;Y)\}|}{|\mathcal{T}(Q)|}. \tag{10}$$

An alternative representation is the following: For a given $(\boldsymbol{x}_0, \boldsymbol{y})$, let the random variable $N(\boldsymbol{x}_0, \boldsymbol{y})$ be defined as

$$N(\boldsymbol{x}_0, \boldsymbol{y}) = \sum_{m=1}^{M-1} \mathcal{I}\{\hat{I}_{\boldsymbol{x}_m\boldsymbol{y}}(X;Y) \geq \hat{I}_{\boldsymbol{x}_0\boldsymbol{y}}(X;Y)\}. \tag{11}$$

Then,

$$\overline{P_e(\boldsymbol{x}_0, \boldsymbol{y})} = \Pr\{N(\boldsymbol{x}_0, \boldsymbol{y}) \geq L\}, \tag{12}$$

and

$$\overline{P_e} = \Pr\{N(\boldsymbol{X}_0, \boldsymbol{Y}) \geq L\}, \tag{13}$$

where it should be understood that the random variable $N(\boldsymbol{X}_0, \boldsymbol{Y})$ incorporates also the randomness of $\boldsymbol{X}_0$ and $\boldsymbol{Y}$, in addition to the randomness of $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{M-1}\}$, unlike $N(\boldsymbol{x}_0, \boldsymbol{y})$, for which

$\boldsymbol{x}_0$ and $\boldsymbol{y}$ are fixed and only $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{M-1}\}$ are random. By the method of types [6],

$$\begin{aligned} q(\boldsymbol{x}_0, \boldsymbol{y}) &= \sum_{\mathcal{T}(\boldsymbol{x}|\boldsymbol{y}) \subseteq \mathcal{T}(Q):\ \hat{I}_{\boldsymbol{x}\boldsymbol{y}}(X;Y) \geq \hat{I}_{\boldsymbol{x}_0\boldsymbol{y}}(X;Y)} \frac{|\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})|}{|\mathcal{T}(\boldsymbol{x})|} \\ &\leq \sum_{\mathcal{T}(\boldsymbol{x}|\boldsymbol{y}) \subseteq \mathcal{T}(Q):\ \hat{I}_{\boldsymbol{x}\boldsymbol{y}}(X;Y) \geq \hat{I}_{\boldsymbol{x}_0\boldsymbol{y}}(X;Y)} e^{-n[\hat{I}_{\boldsymbol{x}\boldsymbol{y}}(X;Y) - \delta_n/2]} \\ &\leq e^{-n[\hat{I}_{\boldsymbol{x}_0\boldsymbol{y}}(X;Y) - \delta_n]} \stackrel{\Delta}{=} q'(\boldsymbol{x}_0, \boldsymbol{y}), \end{aligned} \tag{14}$$

where $\delta_n = O((\log n)/n)$ is a term that accounts for the normalized logarithm of the number of different types classes of sequences of length n. Obviously, $\Pr\{N(\boldsymbol{x}_0, \boldsymbol{y}) \geq L\}$ is a monotonically non–decreasing function of $q(\boldsymbol{x}_0, \boldsymbol{y})$, and so

$$\overline{P_e(\boldsymbol{x}_0, \boldsymbol{y})} \leq \sum_{\ell=L}^{M} \binom{M}{\ell} [q'(\boldsymbol{x}_0, \boldsymbol{y})]^{\ell} [1 - q'(\boldsymbol{x}_0, \boldsymbol{y})]^{M-\ell} \stackrel{\Delta}{=} \Pr\{N'(\boldsymbol{x}_0, \boldsymbol{y}) \geq L\} \tag{15}$$

where $N'(\boldsymbol{x}_0, \boldsymbol{y})$ is defined as a Bernoulli random variable of $M-1$ independent trials and a probability of success $q'(\boldsymbol{x}_0, \boldsymbol{y})$. Now, for a given $(\boldsymbol{x}_0, \boldsymbol{y})$, if $q'(\boldsymbol{x}_0, \boldsymbol{y}) < L/(M-1)$, the event $\{N'(\boldsymbol{x}_0, \boldsymbol{y}) \geq L\}$ is a large deviations event, otherwise it occurs with high probability. Accordingly, the Chernoff bound on $\Pr\{N(\boldsymbol{x}_0, \boldsymbol{y}) \geq L\}$ is as follows.

$$\begin{aligned} \overline{P_e(\boldsymbol{x}_0, \boldsymbol{y})} &\leq \begin{cases} \exp\{-MD(\frac{L}{M} \| q'(\boldsymbol{x}_0, \boldsymbol{y})\} & q'(\boldsymbol{x}_0, \boldsymbol{y}) < L/(M-1) \\ 1 & q'(\boldsymbol{x}_0, \boldsymbol{y}) \geq L/(M-1) \end{cases} \\ &= \begin{cases} \exp\{-MD(\frac{L}{M} \| e^{-n[\hat{I}_{\boldsymbol{x}_0\boldsymbol{y}}(X;Y) - \delta_n]})\} & \hat{I}_{\boldsymbol{x}_0,\boldsymbol{y}}(X;Y)) > R - \frac{\ln L}{n} + \delta_n \\ 1 & \hat{I}_{\boldsymbol{x}_0,\boldsymbol{y}}(X;Y)) \leq R - \frac{\ln L}{n} + \delta_n \end{cases} \end{aligned} \tag{16}$$

where $D(a\|b)$, for $a, b \in [0,1]$, is the binary divergence function, that is

$$D(a\|b) = a \ln \frac{a}{b} + (1-a) \ln \frac{1-a}{1-b}. \tag{17}$$

Now, for $a \geq b$, the following inequality is proved in [16, pp. 167–168]:

$$D(a\|b) \geq a \left[\ln \frac{a}{b} - 1 \right]_+. \tag{18}$$

Thus, the first line of (16) is further upper bounded by

$$\begin{aligned} & \exp\left\{ -e^{nR} L e^{-nR} \left[\ln \left(\frac{L e^{-nR}}{\exp\{-n[\hat{I}_{\boldsymbol{x}_0\boldsymbol{y}}(X;Y) - \delta_n]\}} \right) - 1 \right]_+ \right\} \\ = & \exp\left\{ -nL \left[\hat{I}_{\boldsymbol{x}_0\boldsymbol{y}}(X;Y) + \frac{\ln L}{n} - R - \delta_n - \frac{1}{n} \right]_+ \right\}, \end{aligned} \tag{19}$$

and so, the upper bound on $\overline{P_e(\boldsymbol{x}_0, \boldsymbol{y})}$ can eventually be presented, for every $(\boldsymbol{x}_0, \boldsymbol{y})$, as

$$\overline{P_e(\boldsymbol{x}_0, \boldsymbol{y})} \leq \exp\left\{-nL\left[\hat{I}_{\boldsymbol{x}_0\boldsymbol{y}}(X;Y) + \frac{\ln L}{n} - R - \delta_n - \frac{1}{n}\right]_+\right\}. \tag{20}$$

The overall average list–error probability is obtained, of course, by averaging over the randomness of $\boldsymbol{X}_0$ and $\boldsymbol{Y}$. This completes the proof of Theorem 1.

3.2 The Fixed List–Size Regime

In the fixed list–size regime, the term $\frac{\ln L}{n}$, in the exponent of eq. (8), vanishes as n grows without bound. Since $P(\boldsymbol{x}_0)$ is the uniform distribution within $\mathcal{T}(Q)$, it is easy to see, by using the method of types, that eq. (8) leads to an exponential upper bound $\overline{P_e} \doteq e^{-nE(R,L,Q)}$, where $E(R,L,Q)$ is given by

$$E(R,L,Q) = \min_{\tilde{P}_{Y|X}}\{D(\tilde{P}_{Y|X}\|P_{Y|X}|Q) + L\cdot[\tilde{I}(X;Y) - R]_+\}, \tag{21}$$

where $\tilde{I}(X;Y)$ and $D(\tilde{P}_{Y|X}\|P_{Y|X}|Q)$ are defined as in Subsection 2.3. Of course, the best bound is obtained by selecting Q so as to maximize $E(R,L,Q)$. The resulting exponent is then obviously $E(R,L) = \max_Q E(R,L,Q)$.

We would like to compare eq. (21) with eq. (2). As mentioned in Subsection 2.3, when $P(\boldsymbol{x}) = \prod_{i=1}^n Q(x_i)$, for the optimal Q, the resulting exponent is $E_{\rm r}(R,L)$, as defined in eq. (3). Let us now evaluate the exponential rate of the r.h.s. of eq. (2) for the case where $P(\boldsymbol{x})$ is uniform within $\mathcal{T}(Q)$, as defined in Subsection 2.2. Using the method of types, this gives the following:[4]

$$\min_{0\leq\rho\leq L} M^\rho \sum_{\boldsymbol{y}}\left[\sum_{\boldsymbol{x}\in\mathcal{T}(Q)}\frac{1}{|\mathcal{T}(Q)|}\cdot P(\boldsymbol{y}|\boldsymbol{x})^{1/(1+\rho)}\right]^{1+\rho} \tag{22}$$

$$\doteq \min_{0\leq\rho\leq L} M^\rho \sum_{\boldsymbol{y}}\left[\sum_{\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})\subseteq\mathcal{T}(Q)}|\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})|e^{-n\hat{H}(Q)}\cdot\exp\left\{\frac{n}{1+\rho}\hat{\boldsymbol{E}}_{\boldsymbol{x}\boldsymbol{y}}\ln P(Y|X)\right\}\right]^{1+\rho} \tag{23}$$

$$\doteq \min_{0\leq\rho\leq L} M^\rho \sum_{\boldsymbol{y}}\left[\max_{\mathcal{T}(\boldsymbol{x}|\boldsymbol{y})\subseteq\mathcal{T}(Q)}\exp\left\{-n\hat{I}_{\boldsymbol{x}\boldsymbol{y}}(X;Y) + \frac{n}{1+\rho}\hat{\boldsymbol{E}}_{\boldsymbol{x}\boldsymbol{y}}\ln P(Y|X)\right\}\right]^{1+\rho} \tag{24}$$

[4]This chain of exponential equalities can serve as the basis for the proof that the two expressions of $E_{\rm sp}(R)$ (in Subsection 2.3) are equivalent. By taking $L\to\infty$, the second to the last line becomes (6). On the other hand, when $P(\boldsymbol{x})$ is the n–fold product of Q, the resulting exponent would be (5). For a given Q, (6) cannot be smaller than (5) since $\mathcal{I}\{\boldsymbol{x}\in\mathcal{T}(Q)\}/|\mathcal{T}(Q)| \doteq \prod_{i=1}^n Q(x_i)$ for all $\boldsymbol{x}$. On the other hand, the optimum input distribution that minimizes the r.h.s. of (2) for a DMC is a product distribution (see, e.g., [15, Theorem 3]), and therefore, for the optimum Q, (5) cannot be smaller than (6). Consequently, for the optimum Q, the two expressions must be equal.

$$\dot{=} \min_{0\le\rho\le L} M^{\rho} \sum_{\boldsymbol{y}} \exp\left\{-n \min_{\hat{P}_{X|Y}:\ \hat{P}_X=Q} [(1+\rho)\hat{I}_{\boldsymbol{xy}}(X;Y) - \hat{\boldsymbol{E}}_{\boldsymbol{xy}} \ln P(Y|X)]\right\} \tag{25}$$

$$\dot{=} \exp\left\{-n \max_{0\le\rho\le L} \min_{\tilde{P}_{XY}:\ \tilde{P}_X=Q} [(1+\rho)\tilde{I}(X;Y) - \tilde{H}(Y) - \tilde{\boldsymbol{E}} \ln P(Y|X) - \rho R]\right\} \tag{26}$$

$$= \exp\left\{-n \max_{0\le\rho\le L} \min_{\tilde{P}_{Y|X}} [\rho\tilde{I}(X;Y) - \tilde{H}(Y|X) - \tilde{\boldsymbol{E}} \ln P(Y|X) - \rho R]\right\} \tag{27}$$

$$= \exp\left\{-n \max_{0\le\rho\le L} \min_{\tilde{P}_{Y|X}} (D(\tilde{P}_{Y|X}\|P_{Y|X}|Q) + \rho[\tilde{I}(X;Y) - R])\right\} \tag{28}$$

$$= \exp\left\{-n \min_{\tilde{P}_{Y|X}} \max_{0\le\rho\le L} (D(\tilde{P}_{Y|X}\|P_{Y|X}|Q) + \rho \cdot [\tilde{I}(X;Y) - R])\right\} \tag{29}$$

$$= \exp\left\{-n \min_{\tilde{P}_{Y|X}} (D(\tilde{P}_{Y|X}\|P_{Y|X}|Q) + L \cdot [\tilde{I}(X;Y) - R]_+)\right\} \tag{30}$$

$$= e^{-nE(R,L,Q)}, \tag{31}$$

where $\hat{\boldsymbol{E}}_{\boldsymbol{xy}}\{\cdot\}$ and $\tilde{\boldsymbol{E}}\{\cdot\}$ denote the expectation operators w.r.t. $\hat{P}_{\boldsymbol{xy}}$ and $\tilde{P}_{XY}$, respectively, and $\tilde{H}(Y)$ and $\tilde{H}(Y|X)$ denote the unconditional and the conditional output entropy induced by $\tilde{P}_{XY}$. Here, we have used the fact that the minimization over $\tilde{P}_{Y|X}$ and the maximization over ρ are commutative since the objective function is convex–concave. We see that the two upper bounds are exponentially equivalent in the fixed list–size regime, for the ensemble of independent random selection within a given type class. For $L \to \infty$, this is equivalent $\min\{D(\tilde{P}_{Y|X}\|P_{Y|X}|Q) :\ \tilde{I}(X;Y) \le R\}$, which is the sphere–packing exponent $E_{sp}(R)$. In fact, it is achieved for all $L \ge L_c$, where L_c is the smallest value of L for which the minimizing $\tilde{P}_{Y|X}$ of $D(\tilde{P}_{Y|X}\|P_{Y|X}|Q) + L \cdot [\tilde{I}(X;Y) - R]_+$ achieves $\tilde{I}(X;Y) \le R$ (think of L as a Lagrange multiplier).

We next show that this bound is also exponentially tight for the average code (thus extending the main result of [11] from ordinary decoding to list decoding), by deriving a compatible lower bound on the average error probability. Consider now the optimum list decoder, that outputs the L most likely messages. First, the probability that a single incorrect codeword will receive a likelihood score higher than that of the correct message, for a given $\boldsymbol{x}_0$ and $\boldsymbol{y}$, is lower bounded as follows:

$$q_0(\boldsymbol{x}_0, \boldsymbol{y}) \triangleq \sum_{\mathcal{T}(\boldsymbol{x}|\boldsymbol{y}) \subset \mathcal{T}(Q):\ P(\boldsymbol{y}|\boldsymbol{x}) \ge P(\boldsymbol{y}|\boldsymbol{x}_0)} \frac{|T(\boldsymbol{x}|\boldsymbol{y})|}{|\mathcal{T}(Q)|} \tag{32}$$

$$\ge \frac{|T(\boldsymbol{x}_0|\boldsymbol{y})|}{|\mathcal{T}(Q)|} \tag{33}$$

$$\geq \quad \exp\{-n[\hat{I}_{\boldsymbol{x}_0\boldsymbol{y}}(X;Y)+\delta_n]\} \stackrel{\Delta}{=} q_0'(\boldsymbol{x}_0,\boldsymbol{y}). \tag{34}$$

Let $N''(\boldsymbol{x}_0,\boldsymbol{y})$ be a Bernoulli random variable with $M-1$ independent trials and probability of success $q_0'(\boldsymbol{x}_0,\boldsymbol{y})$. Then,

$$\begin{aligned}
\overline{P_e} &\geq \Pr\{N''(\boldsymbol{X}_0,\boldsymbol{Y})\geq L\} && (35)\\
&= \Pr\{N''(\boldsymbol{X}_0,\boldsymbol{Y})\geq L,\ \hat{I}_{\boldsymbol{x}_0\boldsymbol{y}}(X;Y)>R-\delta_n\}+ \\
&\quad \Pr\{N''(\boldsymbol{X}_0,\boldsymbol{Y})\geq L|\hat{I}_{\boldsymbol{x}_0\boldsymbol{y}}(X;Y)\leq R-\delta_n\}\cdot\Pr\{\hat{I}_{\boldsymbol{x}_0\boldsymbol{y}}(X;Y)\leq R-\delta_n\} && (36)\\
&\dot{=} \Pr\{N''(\boldsymbol{X}_0,\boldsymbol{Y})\geq L,\ \hat{I}_{\boldsymbol{x}_0\boldsymbol{y}}(X;Y)>R-\delta_n\}+\Pr\{\hat{I}_{\boldsymbol{x}_0\boldsymbol{y}}(X;Y)\leq R-\delta_n\} && (37)\\
&\dot{=} \Pr\{N''(\boldsymbol{X}_0,\boldsymbol{Y})\geq L,\ \hat{I}_{\boldsymbol{x}_0\boldsymbol{y}}(X;Y)>R-\delta_n\}+ \\
&\quad \exp\left\{-n\min_{\tilde{P}_{Y|X}:\ \tilde{I}(X;Y)\leq R} D(\tilde{P}_{Y|X}\|P_{Y|X}|Q)\right\}, && (38)
\end{aligned}$$

where we have used fact that given $\hat{I}_{\boldsymbol{x}_0\boldsymbol{y}}(X;Y)\leq R-\delta_n$, the event $\{N''(\boldsymbol{X}_0,\boldsymbol{Y})\geq L\}$ occurs with high probability. As for the first term on the right–most side of (38), consider first a given $(\boldsymbol{x}_0,\boldsymbol{y})$ with $\hat{I}_{\boldsymbol{x}_0\boldsymbol{y}}(X;Y)>R-\delta_n$ and let $\hat{N}(\boldsymbol{x}_0,\boldsymbol{y})$ be a Bernoulli random variable with $e^{n(R-\delta_n)}$ independent trials and probability of success $q_0'(\boldsymbol{x}_0,\boldsymbol{y})$. Then,

$$\begin{aligned}
\Pr\{N''(\boldsymbol{x}_0,\boldsymbol{y})\geq L\} &\geq \Pr\{\hat{N}(\boldsymbol{x}_0,\boldsymbol{y})\geq L\} && (39)\\
&\geq \Pr\{\hat{N}(\boldsymbol{x}_0,\boldsymbol{y})= L\} && (40)\\
&= \left(\begin{array}{c} e^{n(R-\delta_n)} \\ L \end{array}\right)\cdot q_0'(\boldsymbol{x}_0,\boldsymbol{y})^L[1-q_0'(\boldsymbol{x}_0,\boldsymbol{y})]^{e^{n(R-\delta_n)}-L} && (41)\\
&\geq \left(\begin{array}{c} e^{n(R-\delta_n)} \\ L \end{array}\right)\cdot q_0'(\boldsymbol{x}_0,\boldsymbol{y})^L[1-e^{-n[R-\delta_n]}]^{e^{n(R-\delta_n)}} && (42)\\
&\dot{=} e^{nRL}\cdot\exp\{-nL\hat{I}_{\boldsymbol{x}_0\boldsymbol{y}}(X;Y)\} && (43)\\
&= \exp\{-nL[\hat{I}_{\boldsymbol{x}_0\boldsymbol{y}}(X;Y)-R]\}. && (44)
\end{aligned}$$

Thus,

$$\begin{aligned}
&\Pr\{N''(\boldsymbol{X}_0,\boldsymbol{Y})\geq L,\ \hat{I}_{\boldsymbol{X}_0\boldsymbol{Y}}(X;Y)>R-\delta_n\} && (45)\\
\dot{\geq}\ &\sum_{(\boldsymbol{x}_0,\boldsymbol{y}):\ \hat{I}_{\boldsymbol{x}_0\boldsymbol{y}}(X;Y)>R-\delta_n} P(\boldsymbol{x}_0,\boldsymbol{y})\cdot\exp\{-nL[\hat{I}_{\boldsymbol{x}_0\boldsymbol{y}}(X;Y)-R]\} && (46)\\
\dot{=}\ &\exp\left\{-n\min_{\tilde{P}_{Y|X}:\ \tilde{I}(X;Y)>R}(D(\tilde{P}_{Y|X}\|P_{Y|X}\|Q)+L[\tilde{I}(X;Y)-R])\right\}. && (47)
\end{aligned}$$

Combining this with the second term on the right–most side of (38), the overall exponent becomes the minimum between

$$\min_{\tilde{P}_{Y|X}:\ \tilde{I}(X;Y)\le R} D(\tilde{P}_{Y|X}\|P_{Y|X}|Q)$$

and

$$\min_{\tilde{P}_{Y|X}:\ \tilde{I}(X;Y)> R} (D(\tilde{P}_{Y|X}\|P_{Y|X}\|Q) + L[\tilde{I}(X;Y) - R]),$$

namely,

$$\min_{\tilde{P}_{Y|X}} \{D(\tilde{P}_{Y|X}\|P_{Y|X}\|Q) + L\cdot[\tilde{I}(X;Y) - R]_+\},$$

which is again exactly $E(R, L, Q)$ of eq. (21). Thus, the upper bound (2) and our fixed list–size bound (21) are equivalent under the random coding ensemble of fixed composition codes, and they both give the exact random coding exponent for the average code. Also, since the upper bound was obtained by the MMI list decoder and the compatible lower bound applies to the optimal, ML list decoder, we have also proved, as a byproduct, the universal optimality of the MMI list decoder in the error exponent sense.

3.3 The Exponential List–Size Regime

In the exponential list–size regime, $L = e^{\lambda n}$. Now, let $\epsilon > 0$ be arbitrarily small, and define $\mathcal{E}$ to be the set of all $\{(\boldsymbol{x}, \boldsymbol{y})\}$ for which

$$\hat{I}_{\boldsymbol{xy}}(X;Y) + \lambda - R - \delta_n - \frac{1}{n} \ge \epsilon. \tag{48}$$

Then, eq. (8) is averaged as follows:

$$\begin{aligned}
\overline{P_e} &\doteq \sum_{\boldsymbol{x},\boldsymbol{y}} P(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x}) \exp\left\{-ne^{n\lambda}\left[\hat{I}_{\boldsymbol{xy}}(X;Y) + \lambda - R - \delta_n - \frac{1}{n}\right]_+\right\} && (49)\\
&= \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{E}} P(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x}) \exp\left\{-ne^{n\lambda}\left[\hat{I}_{\boldsymbol{xy}}(X;Y) + \lambda - R - \delta_n - \frac{1}{n}\right]_+\right\} + \\
&\quad + \sum_{(\boldsymbol{x}_0,\boldsymbol{y})\in\mathcal{E}^c} P(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x}) \exp\left\{-ne^{n\lambda}\left[\hat{I}_{\boldsymbol{xy}}(X;Y) + \lambda - R - \delta_n - \frac{1}{n}\right]_+\right\} && (50)\\
&\le \exp\{-\epsilon e^{\lambda n}\} + \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{E}^c} P(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x}). && (51)
\end{aligned}$$

Now, the first term decays double–exponentially and hence is negligible. Therefore, $\overline{P_e}$ is dominated by the second term. Using the method of types, the arbitrariness of ϵ, and the fact that δ_n and

$1/n$ vanish as $n \to \infty$, the exponent of the second term is easily found to be

$$\min_{\{\tilde{P}_{Y|X}:\ \tilde{I}(X;Y) \le R-\lambda\}} D(\tilde{P}_{Y|X} \| P_{Y|X} | Q),$$

whose maximum over Q is $E_{\rm sp}(R-\lambda)$, as mentioned in Subsection 2.3. Thus, we have shown that the Shannon–Gallager–Berlekamp bound [22, Theorem 2] is actually achievable across the whole relevant range of rates, $(\lambda, C+\lambda)$, where C is the channel capacity. Once again, we see that in the exponential list–size regime too, the MMI list decoder is universal in the error exponent in sense.

3.4 Relation to the Guessing Problem

It is interesting to look at moments of the random variable $N(\boldsymbol{X}_0, \boldsymbol{Y})$ (or any of the other Bernoulli random variables that we have define earlier). Consider the following lower bound. Let $\epsilon > 0$ be arbitrarily small. Then,

$$\begin{aligned}
\boldsymbol{E}\{N(\boldsymbol{X}_0, \boldsymbol{Y})^\rho\} &= \sum_{L=1}^{M} L^\rho \cdot \Pr\{N(\boldsymbol{X}_0, \boldsymbol{Y}) = L\} && (52)\\
&\ge \sum_{i=0}^{R/\epsilon-1} e^{in\epsilon\rho} \cdot \Pr\{e^{ni\epsilon} \le N(\boldsymbol{X}_0, \boldsymbol{Y}) < e^{n(i+1)\epsilon}\} && (53)\\
&\dot{\ge} \sum_{i=0}^{R/\epsilon-1} e^{in\epsilon\rho} \cdot e^{-nE_{\rm sp}(R-i\epsilon)} && (54)\\
&\dot{=} \exp\left\{n \max_i [i\epsilon\rho - E_{\rm sp}(R-i\epsilon)]\right\}. && (55)
\end{aligned}$$

Since $\epsilon > 0$ is arbitrarily small, one can take the limit $\epsilon \to 0$, and obtain

$$\liminf_{n\to\infty} \frac{\ln \boldsymbol{E}\{N(\boldsymbol{X}_0, \boldsymbol{Y})^\rho\}}{n} \ge \sup_{0<\lambda<R} [\rho\lambda - E_{\rm sp}(R-\lambda)]. \quad (56)$$

Let $\varrho(R)$ denote the achiever of $E_{\rm sp}(R)$ in eq. (5). Then obviously, $\dot{E}_{\rm sp}(R) = -\varrho(R)$, where $\dot{E}_{\rm sp}(R)$ denotes the derivative of $E_{\rm sp}(R)$. Thus, the supremum is achieved by $\lambda^* = R - \varrho^{-1}(\rho)$, provided that $\varrho^{-1}(\rho) \le R$, namely, $\rho \ge \varrho(R)$. In this case, the exponential lower bound becomes

$$\begin{aligned}
\rho[R - \varrho^{-1}(\rho)] - E_{\rm sp}(\varrho^{-1}(\rho)) &= \rho R - \rho\varrho^{-1}(\rho) - E_0(\rho) + \rho\varrho^{-1}(\rho) && (57)\\
&= \rho R - E_0(\rho). && (58)
\end{aligned}$$

When $\rho \ge \varrho(R)$, the supremum is achieved at $\lambda = 0$ and the result is $-E_{\rm sp}(R)$. In summary, then

$$\liminf_{n\to\infty} \frac{\ln \boldsymbol{E}\{N(\boldsymbol{X}, \boldsymbol{Y})^\rho\}}{n} \ge \begin{cases} -E_{\rm sp}(R) & \rho \le \varrho(R) \\ \rho R - E_0(\rho) & \rho > \varrho(R) \end{cases} \quad (59)$$

This result is intimately related to those on guessing exponents [1], [2]. In the guessing problem, the decoder of a coded communication system submits, upon receiving the channel output $\boldsymbol{y}$, a sequence of guesses, $\boldsymbol{x}_{m_1}, \boldsymbol{x}_{m_2}, \ldots$ concerning the transmitted message until it hits the correct message. Let $G(\boldsymbol{x}_0|\boldsymbol{y})$ denote the number of guesses when the transmitted codeword is $\boldsymbol{x}_0$ and the received vector is $\boldsymbol{y}$. The aim is to minimize $\boldsymbol{E}\{[G(\boldsymbol{X}_0|\boldsymbol{Y})]^\rho\}$ for a given $\rho > 0$, and the best guessing strategy is, of course, according to decreasing likelihood scores, hence the close relation to list decoding. In fact, by definition, the relation between $N(\boldsymbol{x}_0, \boldsymbol{y})$ and $G(\boldsymbol{x}_0|\boldsymbol{y})$ is extremely simple: $G(\boldsymbol{x}_0|\boldsymbol{y}) = N(\boldsymbol{x}_0, \boldsymbol{y}) + 1$. While this seems like a very minor difference, it should be noted $\boldsymbol{E}\{G(\boldsymbol{X}_0|\boldsymbol{Y})^\rho\}$ can only have a non–negative exponent since $G(\boldsymbol{x}_0|\boldsymbol{y}) \geq 1$, whereas $\boldsymbol{E}\{N(\boldsymbol{X}_0, \boldsymbol{Y})^\rho\}$ may also have a negative exponent since there is a high probability that $N(\boldsymbol{X}_0, \boldsymbol{Y}) = 0$ (the event of correct decoding in the ordinary sense). Since $G(\boldsymbol{X}_0|\boldsymbol{Y}) \geq \max\{N(\boldsymbol{X}_0, \boldsymbol{Y}), 1\}$, the exponent of $\boldsymbol{E}\{G(\boldsymbol{X}_0|\boldsymbol{Y})^\rho\}$ is lower bounded by an exponential function whose exponent is given by the positive part of the exponent of $\boldsymbol{E}\{N(\boldsymbol{X}_0, \boldsymbol{Y})^\rho\}$, and so, some information is lost when one considers the guessing exponent rather than the exponent of $\boldsymbol{E}\{N(\boldsymbol{X}_0, \boldsymbol{Y})^\rho\}$. In other words, the latter is a more informative measure. Indeed, the second line of eq. (59) agrees with the results on guessing exponents in [2].[5] Our results thus far seem to indicate that the following performance is achievable (provided that the input distribution that attains $E_{\rm sp}(R)$ is independent of R):

$$\liminf_{n\to\infty} \frac{\ln \boldsymbol{E}\{N(\boldsymbol{X}_0, \boldsymbol{Y})^\rho\}}{n} \leq \begin{cases} -E_{\rm r}(R) & \rho \leq \varrho'(R) \\ \rho R - E_0(\rho) & \rho > \varrho'(R) \end{cases} \tag{60}$$

where $\varrho'(R)$ is the solution to the equation $E_{\rm r}(R) = E_0(\rho) - \rho R$. Thus, at least for large values of ρ, where the dominant value of λ is large enough, the bound is tight (just like in the guessing problem).

4 Upper Bounds for an Expurgated Ensemble

While in the exponential list–size regime, we have an exact characterization of the list–decoding reliability function, as $E_{\rm sp}(R-\lambda)$, in the fixed list–size regime, there is some gap between the upper bound and the lower bound, $E_{\rm sp}(R)$, at low rates, just like in ordinary decoding. Although this problematic range of low rates shrinks as L increases, it is nevertheless still existent for every finite

[5] The results in [2] are for lossy joint source–channel coding, however, it is easy to particularize them to pure channel coding by considering the binary symmetric source and allowing zero distortion.

L. Like in ordinary decoding, one the way to reduce this gap is by improving the upper bound at low rates using expurgation. In this section, we discuss expurgated bounds for list decoding.

For a given code $\mathcal{C}$, consider first the following union bound on $P_{e|m}(\mathcal{C})$, the probability of list error, given that message no. m was transmitted.

$$\begin{aligned} P_{e|m}(\mathcal{C}) &= \sum_{m_1,\ldots,m_L \neq m} \sum_{\boldsymbol{y}\in\mathcal{Y}^n} P(\boldsymbol{y}|\boldsymbol{x}_m) \prod_{i=1}^{L} \mathcal{I}\{P(\boldsymbol{y}|\boldsymbol{x}_{m_i}) \geq P(\boldsymbol{y}|\boldsymbol{x}_m)\} \\ &\leq \sum_{m_1,\ldots,m_L \neq m} \sum_{\boldsymbol{y}\in\mathcal{Y}^n} P(\boldsymbol{y}|\boldsymbol{x}_m) \prod_{i=1}^{L} \left[\frac{P(\boldsymbol{y}|\boldsymbol{x}_{m_i})}{P(\boldsymbol{y}|\boldsymbol{x}_m)}\right]^{1/(L+1)} \\ &= \sum_{m_1,\ldots,m_L \neq m} \sum_{\boldsymbol{y}\in\mathcal{Y}^n} \left[P(\boldsymbol{y}|\boldsymbol{x}_m) \prod_{i=1}^{L} P(\boldsymbol{y}|\boldsymbol{x}_{m_i})\right]^{1/(L+1)}, \end{aligned} \tag{61}$$

where summation over $\{m_1,\ldots,m_L\}$ is over all L–tuples of distinct integers in $\{0,1,\ldots,M-1\} \setminus \{m\}$.

The first approach to the derivation of expurgated bounds is a straightforward generalization of Gallager's approach [10, Section 5.7] (see also [26, Section 3.3]). Using the same argumentation as in the derivation of the classical expurgated bound in these references, it is apparent that there exists a code for which the *maximal* probability of error satisfies, for every $s \geq 0$:

$$\max_m P_{e|m}(\mathcal{C}) \leq \left[\boldsymbol{E}\left(\sum_{m_1,\ldots,m_L \neq m} \sum_{\boldsymbol{y}\in\mathcal{Y}^n} \left[P(\boldsymbol{y}|\boldsymbol{X}_m) \prod_{i=1}^{L} P(\boldsymbol{y}|\boldsymbol{X}_{m_i})\right]^{1/(L+1)}\right)^s\right]^{1/s}, \tag{62}$$

where now the summation over $\{m_1,\ldots,m_L\}$ is across all L–tuples of distinct integers in $\{0,1,\ldots,2M-1\} \setminus \{m\}$. Now, assuming $s \in [0,1]$, the Jensen inequality can be used to obtain

$$\begin{aligned} \max_m P_{e|m}(\mathcal{C}) &\leq \left[\boldsymbol{E}\left(\sum_{m_1,\ldots,m_L \neq m} \sum_{\boldsymbol{y}\in\mathcal{Y}^n} \left[P(\boldsymbol{y}|\boldsymbol{X}_m) \prod_{i=1}^{L} P(\boldsymbol{y}|\boldsymbol{X}_{m_i})\right]^{1/(L+1)}\right)^s\right]^{1/s} && (63) \\ &\leq \left[\boldsymbol{E} \sum_{m_1,\ldots,m_L \neq m} \left(\sum_{\boldsymbol{y}\in\mathcal{Y}^n} \left[P(\boldsymbol{y}|\boldsymbol{X}_m) \prod_{i=1}^{L} P(\boldsymbol{y}|\boldsymbol{X}_{m_i})\right]^{1/(L+1)}\right)^s\right]^{1/s} && (64) \\ &\leq \left[(2M)^L \sum_{\boldsymbol{x}_0,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_L\in\mathcal{X}^n} \prod_{i=0}^{L} Q(\boldsymbol{x}_i) \left(\sum_{\boldsymbol{y}\in\mathcal{Y}^n} \prod_{0=1}^{L} P(\boldsymbol{y}|\boldsymbol{x}_i)^{1/(L+1)}\right)^s\right]^{1/s}, && (65) \end{aligned}$$

which can be further developed using the method of types, for the ensemble of fixed composition codes, or using Gallager's technique, for the ensemble where the distribution of each codeword is the

n–fold product measure Q. In the latter case, the resulting single–letter formula of the expurgated exponent is

$$E_{\rm ex}^{\rm G}(R,L) = \sup_{\rho\geq 1}\sup_{Q}\left\{-\rho\ln\left[\sum_{x_0,x_1,\ldots,x_L\in\mathcal{X}}\prod_{i=0}^{L}Q(x_i)\left(\sum_{y\in\mathcal{Y}}\prod_{i=0}^{L}[P(y|x_i)^{1/(L+1)}\right)^{1/\rho}\right]-\rho LR\right\}, \quad (66)$$

where $\rho = 1/s$ and where the superscript "G" stands for "Gallager". At zero rate,

$$E_{\rm ex}^{\rm G}(0,L) = -\sum_{x_0,x_1,\ldots,x_L}\prod_{i=0}^{L}Q(x_i)\ln\left(\sum_{y}\prod_{i=0}^{L}P(y|x_i)^{1/(L+1)}\right) \quad (67)$$

has been shown by Blinovsky [4] to be tight in the sense that there is also a compatible lower bound with the same exponential rate.

Another approach is in the spirit of the methodology of Csiszár, Körner and Marton [6, page 185, Problem 17], [7], which for $L = 1$ (ordinary decoding) gives an exponent at least as large as that of Gallager, with equality for the optimal choice of Q [6, p. 193, Problem 23(ii)], [19]. To the best of our knowledge, the extension of the Csiszár–Körner–Marton (CKM) expurgated exponent to list decoding is new.

Define the following function, which is an extension of the Bhattacharyya distance from two to $L+1$ variables:

$$d(x_0,x_1,\ldots,x_L) = -\ln\left[\sum_{y\in\mathcal{Y}}\prod_{i=0}^{L}P(y|x_i)^{1/(L+1)}\right]. \quad (68)$$

For a random vector $(X_0,X_1,\ldots,X_L)$ with a given joint distribution $P_{X_0X_1\ldots X_L}$, let us define also the *multi–information* as

$$I(X_0;X_1;\ldots;X_L) = \sum_{i=0}^{L}H(X_i) - H(X_0,X_1,\ldots,X_L) = D(P_{X_0X_1\ldots X_L}\|P_{X_0}\times P_{X_1}\times\ldots P_{X_L}), \quad (69)$$

which is a natural extension of the mutual information as a measure of joint dependence among several random variables [23]. Our main result in this section is the following.theorem, the proof of which appears in the Appendix.

Theorem 2 *There exists a sequence of rate–R codes for which*

$$\lim_{n\to\infty}\left[-\frac{\ln\max_m P_{e|m}(\mathcal{C})}{n}\right] \geq E_{ex}^{CKM}(R,L), \quad (70)$$

where

$$E_{ex}^{CKM}(R,L) \stackrel{\Delta}{=} \sup_{Q} \inf_{\{P_{X_0X_1\ldots X_L}\in\mathcal{A}(R,Q)\}} [\boldsymbol{E}d(X_0,X_1,\ldots,X_L) + I(X_0;X_1;\ldots;X_L)] - LR, \tag{71}$$

and

$$\mathcal{A}(R,Q) \stackrel{\Delta}{=} \{P_{X_0X_1\ldots X_L}:\ I(X_0;X_1;\ldots;X_L) \le LR,\ P_{X_0} = P_{X_1} = \ldots = P_{X_L} = Q\}. \tag{72}$$

It is easy to see that, similarly as in the case $L = 1$, here too, $E_{\rm ex}^{\rm CKM}(R,L)$ is a monotonically decreasing, convex function for some range of small rates, and then at a certain critical rate, it becomes an affine function (with slope $-L$) in the range of large R. In particular, the convex function in the low–rate range is the "distortion–rate function"

$$D(R) = \min_{P_{X_0X_1\ldots X_L}\in\mathcal{A}(R,Q)} \boldsymbol{E}\{d(X_0,X_1,\ldots,X_L)\}, \tag{73}$$

and the critical rate is $R_{\rm crit} = I^*(X_0;X_1;\ldots;X_L)/L$, where $I^*(X_0;X_1;\ldots;X_L)$ is the multi–information pertaining to the minimizer $P^*_{X_0X_1\ldots X_L}$ of $\boldsymbol{E}d(X_0,X_1,\ldots,X_L) + I(X_0;X_1;\ldots;X_L)$ over $\mathcal{A}(\infty,Q)$. The affine part is given by the straight line that is tangential to the curve $D(R)$ at $R = R_{\rm crit}$. The special case $L = 1$ obviously recovers the CKM expurgated exponent for ordinary decoding.

A few comments concerning the continuous alphabet case are in order. The proof technique (see Appendix) can be extended, in principle, to the continuous alphabet case (with input constraints), provided that we can assess exponential growth rates of volumes[6] of groups of $(L+1)$ n–vectors with every given value of $d(\boldsymbol{x}_0,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_L) = \sum_{i=1}^n d(x_{0,i},x_{1,i},\ldots,x_{L,i})$ (within an arbitrarily tolerance ϵ). For concreteness, consider the memoryless Gaussian channel, $Y_i = X_i + Z_i$, where $Z_i \sim \mathcal{N}(0,\sigma^2)$ are i.i.d., independent of $\boldsymbol{X} = (X_1,\ldots,X_n)$, and the random codewords $\{\boldsymbol{X}_m\}$ are independently selected under the uniform distribution over the surface of the n–dimensional hypersphere of radius $\sqrt{nS}$ ($S > 0$ being the power). Then,

$$\begin{aligned} d(x_0,x_1,\ldots,x_L) &= -\ln\left[\int_{-\infty}^{+\infty} {\rm d}y\cdot(2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2(L+1)}\sum_{i=0}^{L}(y-x_i)^2\right\}\right] &(74)\\ &= \frac{L}{2\sigma^2(L+1)^2}\left[\sum_{i=0}^{L}x_i^2 - \frac{2}{L}\sum_{i<j}x_ix_j\right] &(75)\end{aligned}$$

[6]See, e.g., [1, Proof of Theorem 5], [14], where an extension of the method of types to continuous alphabet situations was used extensively.

and for vectors of length n, we have

$$d(\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) = \frac{L}{2\sigma^2(L+1)^2} \sum_{t=1}^{n} \left[\sum_{i=0}^{L} x_{i,t}^2 - \frac{2}{L} \sum_{i<j} x_{i,t} x_{j,t} \right] \tag{76}$$

$$= \frac{L}{2\sigma^2(L+1)^2} \left[n(L+1)S - \frac{2}{L} \sum_{i<j} \sum_{t=1}^{n} x_{i,t} x_{j,t} \right] \tag{77}$$

$$= \frac{nLS}{2\sigma^2(L+1)^2} \left(L + 1 - \frac{2}{L} \sum_{i<j} \rho_{ij} \right), \tag{78}$$

where $\rho_{ij} = \sum_{t=1}^{n} x_{t,i} x_{t,j}/(nS)$. Thus, $d(\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)$ is completely determined by the empirical correlations $\{\rho_{ij}\}$. Therefore, in order to modify the proof of Theorem 2 to the continuous case considered here, we have to assess the probability that $L+1$ independently selected vectors across the hyper-surface of the sphere would happen to have prescribed values of empirical correlation coefficients (within a small tolerance $\pm\epsilon$), because the desired probability is equal to the ratio between the volume pertaining these correlation coefficients and the total available volume. Using the techniques of [1, Proof of Theorem 5] and [14], it is not difficult to show that the former volume is of the exponential order of $\exp\{nh_G(X_0, X_1, \ldots, X_L)\}$, where $h_G(X_0, X_1, \ldots, X_L)$ is the differential entropy of a Gaussian vector $(X_0, X_1, \ldots, X_L)$ with covariance matrix $\{\rho_{ij}S\}$ ($\rho_{ii} \stackrel{\Delta}{=} 1$, $i = 0, 1, \ldots, L$), and the latter volume is of the exponential order of $\exp\{\frac{n(L+1)}{2} \ln[2\pi e S]\}$. Thus, the probability under discussion is of the exponential order of $\exp\{-nI_G(X_0; X_1; \ldots; X_L)\}$, where $I_G(X_0; X_1; \ldots; X_L)$ is the multi–information associated with the above defined Gaussian vector $(X_0, X_1, \ldots, X_L)$, in analogy to the finite–alphabet case. In particular, let Λ denote the $(L+1) \times (L+1)$ matrix of correlation coefficients $\{\rho_{ij}\}$. Then,

$$I_G(X_0; X_1; \ldots; X_L) = -\frac{1}{2} \ln |\Lambda|, \tag{79}$$

and so,

$$E_{\rm ex}^{\rm CKM}(R, L) = \inf_{\Lambda \in \mathcal{A}(R)} \left[\frac{LS}{2\sigma^2(L+1)^2} \left(L + 1 - \frac{2}{L} \sum_{i<j} \rho_{ij} \right) - \frac{1}{2} \ln |\Lambda| - LR \right]. \tag{80}$$

where $\mathcal{A}(R)$ is the set of all positive definite matrices with $-\frac{1}{2} \ln |\Lambda| \le LR$. For $L = 1$, we get

$$E_{\rm ex}^{\rm CKM}(R, 1) = \inf_{|\rho|<1:\ -\frac{1}{2}\ln(1-\rho^2) \le R} \left[\frac{S}{4\sigma^2} \cdot (1 - \rho) - \frac{1}{2} \ln(1 - \rho^2) - R \right] \tag{81}$$

which in the curvy part is attained with $\rho = \sqrt{1 - e^{-2R}}$ to yield

$$E_{\rm ex}^{\rm CKM}(R,1) = \frac{S(1-\sqrt{1-e^{-2R}})}{4\sigma^2} \tag{82}$$

and the affine part is the tangential straight line with slope -1. For a general L, we argue that $E_{\rm ex}^{\rm CKM}(R,L)$ is always attained by a totally symmetric correlation matrix, i.e., $\rho_{ij} = \rho$ for all $i \neq j$, for some ρ which keeps this matrix positive definite. This facilitates considerably the minimization problem associated with $E_{\rm ex}^{\rm CKM}(R,L)$, since it reduces to a minimization over the single variable ρ. This total symmetry follows from the concavity of the function $\ln|\Lambda|$ [5, p. 679, Theorem 17.9.1]. In particular, if Λ_0 attains $E_{\rm ex}^{\rm CKM}(R,L)$, then so does $\Pi\Lambda_0\Pi^T$ for every permutation matrix Π since neither the objective function nor the constraint is sensitive to permutations. Now, let

$$\Lambda^* = \frac{1}{(L+1)!}\sum_{\Pi}\Pi\Lambda_0\Pi^T, \tag{83}$$

which is obviously has a totally symmetric structure where all diagonal elements are 1 and all off–diagonal elements are the same. By the concavity property,

$$\ln|\Lambda^*| \geq \frac{1}{(L+1)!}\sum_{\Pi}\ln|\Pi\Lambda_0\Pi^T| = \frac{1}{(L+1)!}\sum_{\Pi}\ln|\Lambda_0| = \ln|\Lambda_0|. \tag{84}$$

It follows then that if Λ_0 is replaced by Λ^*, the value of the objective function is reduced without violating the constraint, and so, Λ^* cannot be worse than Λ_0. Now, for an $(L+1)\times(L+1)$ matrix Λ whose entries are $\rho_{ij} = \rho + (1-\rho)\delta_{i-j}$ (δ_k being the Kronecker delta function), the eigenvalues are easily found to be $\lambda_1 = 1+\rho L$ (with an eigenvector being the all–one vector) and $\lambda_2 = \ldots = \lambda_{L+1} = 1-\rho$ (with L independent eigenvectors, all orthogonal to the all–one vector). Therefore,

$$\ln|\Lambda| = \ln(1+\rho L) + L\ln(1-\rho). \tag{85}$$

and so, the expurgated exponent function becomes

$$\begin{aligned} E_{\rm ex}^{\rm CKM}(R,L) \;=\; & \inf_{\{-1/L<\rho<1:\ -\frac{1}{2}\ln(1+\rho L)-\frac{L}{2}\ln(1-\rho)\leq LR\}}\left[\frac{SL(1-\rho)}{2\sigma^2(L+1)} - \right. \\ & \left. \frac{1}{2}\ln(1+\rho L) - \frac{L}{2}\ln(1-\rho) - LR\right]. \end{aligned} \tag{86}$$

The behavior of this function can be characterized as follows. Let ρ_0 be the unconstrained minimizer of the function

$$\frac{SL(1-\rho)}{2\sigma^2(L+1)} - \frac{1}{2}\ln(1+\rho L) - \frac{L}{2}\ln(1-\rho)$$

over $[0,1]$ and let $\varrho(R)$ be the solution to the equation

$$-\frac{1}{2}\ln(1+\rho L)-\frac{L}{2}\ln(1-\rho)=LR.$$

Then,

$$E_{\rm ex}^{\rm CKM}(R,L)=\begin{cases}\frac{SL[1-\varrho(R)]}{2\sigma^2(L+1)} & R<\varrho^{-1}(\rho_0)\\ \frac{SL(1-\rho_0)}{2\sigma^2(L+1)}-\frac{1}{2}\ln(1+\rho_0 L)-\frac{L}{2}\ln(1-\rho_0)-LR & R\geq\varrho^{-1}(\rho_0)\end{cases}\tag{87}$$

The critical rate, $R_0\stackrel{\Delta}{=}\varrho^{-1}(\rho_0)$, is the point at which the derivative of the function $SL[1-\varrho(R)]/[2\sigma^2(L+1)]$ w.r.t. R is equal to $-L$, so that the affine function in the second line of the last equation represents a straight line that is tangential to the curve of the first line at $R=R_0$. Such a point always exists because it can easily be shown that the derivative of the curvy function begins from $-\infty$ at $R=0$ and then increases. This is related to the fact that the multi–information, as a function of ρ, has a zero derivative at $\rho=0$.

For $L=1$, we have already seen that $\varrho(R)=\sqrt{1-e^{-2R}}$. For $L=2$, the derivation of $\varrho(R)$ is associated with the relevant solution of a cubic equation in ρ, which is given by

$$\varrho(R)=\frac{\sqrt{1-e^{-4R}}}{2\cos\left[\frac{1}{3}\cos^{-1}(-\sqrt{1-e^{-4R}})\right]},\tag{88}$$

with $\cos^{-1}(t)$ being defined as the unique solution x to the equation $\cos x=t$ in the range $[0,\pi]$.

5 Future Work

In Subsection 2.3, we have proved the well–known fact the optimal list decoder provides the L messages with highest likelihoods. This proof suggests an extension to a recent work [17] (see also [18]) concerning a decoder/detector that first has to decide whether the received channel output $\boldsymbol{y}$ really contains a transmitted message or it is simply pure noise (that is received when the transmitter is silent), and in the former case to decode the message in the ordinary way ($L=1$). Three figures of merit should therefore be traded off here: the false–alarm probability (deciding that a message has been sent while $\boldsymbol{y}$ is actually pure noise), the mis–detection probability (deciding that $\boldsymbol{y}$ is pure noise while it actually contains a codeword) and the probability of decoding error (detecting rightfully that $\boldsymbol{y}$ contains a codeword by decoding it erroneously). It was shown in [17] that the optimum decision rule in the sense of minimizing the probability of decoding error subject

to given constraints on the false alarm and the mis–detection probabilities, is as follows: Accept $\boldsymbol{y}$ as containing a message if and only if

$$e^{n\alpha}\sum_{m=1}^{M}P(\boldsymbol{y}|\boldsymbol{x}_m)+\max_m P(\boldsymbol{y}|\boldsymbol{x}_m)\geq e^{n\beta}P(\boldsymbol{y}|\text{no transmission}), \tag{89}$$

where α and β are constants that are chosen to meet the false–alarm and the mis-detection constraints and $P(\boldsymbol{y}|\text{no transmission})$ is the probability of $\boldsymbol{y}$ when no transmission takes place (e.g., the channel input is the zero signal), in other words, the probability distribution of pure noise. The reason for the term $\max_m P(\boldsymbol{y}|\boldsymbol{x}_m)$ in this test turns out[7] to be associated with the fact that the probability of correct decoding is proportional to $\sum_{\boldsymbol{y}}\max_m P(\boldsymbol{y}|\boldsymbol{x}_m)$. Now, when list decoding is considered, then in view of the last line of (1), this decision rule is generalized to the form:

$$e^{n\alpha}\sum_{m=1}^{M}P(\boldsymbol{y}|\boldsymbol{x}_m)+\sum_{i=1}^{L}P(\boldsymbol{y}|\boldsymbol{x}_{m_i^*(\boldsymbol{y})})\geq e^{n\beta}P(\boldsymbol{y}|\text{no transmission}). \tag{90}$$

The analysis of the error exponents associated with this detector/decoder can be carried out using the same methods as in [17] for the fixed list–size regime, but it is considerably more challenging in the exponential list–size regime.

Appendix

Proof of Theorem 2. For a given code $\mathcal{C}$ and a given message m, let $N_m(\hat{P},\mathcal{C})$ be the number of L–tuples of distinct integers, $\{m_1,\ldots,m_L\}$, all different from m, for which $(\boldsymbol{x}_m,\boldsymbol{x}_{m_1},\ldots,\boldsymbol{x}_{m_L})$ have a given joint empirical distribution $\hat{P}$ of an $(L+1)$–tuple of random variables all taking on values in $\mathcal{X}$, whose single–letter marginals (which are the individual empirical distributions of the various codewords) all coincide with Q. Now, for the given code,

$$\begin{aligned}P_{e|m}(\mathcal{C}) &\leq \sum_{m_1,\ldots,m_L\neq m}\sum_{\boldsymbol{y}}\left[P(\boldsymbol{y}|\boldsymbol{X}_m)\prod_{i=1}^{L}P(\boldsymbol{y}|\boldsymbol{x}_{m_i})\right]^{1/(L+1)}\\ &= \sum_{\hat{P}}N_m(\hat{P},\mathcal{C})\exp\{-n\hat{\boldsymbol{E}}d(X_0,X_1,\ldots,X_L)\}.\end{aligned} \tag{A.1}$$

Our key argument here is that for every $\epsilon>0$ and sufficiently large n, there exists a code $\mathcal{C}$ of rate (essentially) R, that satisfies, for every message m and every $\hat{P}$,

$$N_m(\hat{P},\mathcal{C}) \leq N^*(\hat{P})$$

[7] See [17] for details.

$$\triangleq \begin{cases} \exp\{n[LR - \hat{I}(X_0; X_1; \ldots; X_L) + \epsilon]\} & LR \geq \hat{I}(X_0; X_1; \ldots; X_L) - \epsilon \\ 0 & LR < \hat{I}(X_0; X_1; \ldots; X_L) - \epsilon \end{cases} \tag{A.2}$$

where $\hat{I}(X_0; X_1; \ldots; X_L)$ is the empirical multi–information pertaining to $\hat{P}$. To see why this is true, consider a random selection of the code $\mathcal{C}$. Then, obviously,

$$\begin{aligned}
\overline{N(\hat{P})} &\triangleq \frac{1}{M} \sum_{m=0}^{M-1} \boldsymbol{E}\{N_m(\hat{P}, \mathcal{C})\} && \text{(A.3)} \\
&= \boldsymbol{E}\{N_0(\hat{P}, \mathcal{C})\} && \text{(A.4)} \\
&\leq M^L \cdot \Pr\{(\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_L) \in \mathcal{T}(\hat{P})\} && \text{(A.5)} \\
&= M^L \cdot \frac{|\mathcal{T}(\hat{P})|}{|\mathcal{T}(Q)|^{L+1}} && \text{(A.6)} \\
&\doteq M^L \cdot \frac{\exp\{n\hat{H}(X_0, X_1, \ldots, X_L)\}}{e^{n(L+1)H_Q}} && \text{(A.7)} \\
&= \exp\{n[LR - \hat{I}(X_0; X_1; \ldots; X_L)]\}. && \text{(A.8)}
\end{aligned}$$

It follows then that in the ensemble of the randomly selected codes

$$\begin{aligned}
& \Pr \bigcup_{\hat{P}} \left\{ \mathcal{C} : \ \frac{1}{M} \sum_{m=0}^{M-1} N_m(\hat{P}, \mathcal{C}) > \exp\{n[LR - \hat{I}(X_0; X_1; \ldots; X_L) + \epsilon/2]\} \right\} \\
\leq & \sum_{\hat{P}} \Pr \left\{ \mathcal{C} : \ \frac{1}{M} \sum_{m=0}^{M-1} N_m(\hat{P}, , \mathcal{C}) > \exp\{n[LR - \hat{I}(X_0; X_1; \ldots; X_L) + \epsilon/2]\} \right\} \\
\leq & \sum_{\hat{P}} \frac{\overline{N(\hat{P})}}{\exp\{n[LR - \hat{I}(X_0; X_1; \ldots; X_L) + \epsilon/2]\}} \\
\dot{\leq} & \sum_{\hat{P}} e^{-n\epsilon/2} \\
\leq & (n+1)^{|\mathcal{X}|^{L+1}} \cdot e^{-n\epsilon/2} \to 0, && \text{(A.9)}
\end{aligned}$$

which means that there exists a code (and in fact, for almost every code),

$$\frac{1}{M} \sum_{m=0}^{M-1} N_m(\hat{P}, \mathcal{C}) \leq \exp\{n[LR - \hat{I}(X_0; X_1; \ldots; X_L) + \epsilon/2]\} \quad \forall \hat{P}. \tag{A.10}$$

For a given such code and every given $\hat{P}$, there must then exist at least $(1 - e^{-n\epsilon/2}) \cdot M$ values of m such that

$$N_m(\hat{P}, \mathcal{C}) \leq \exp\{n[LR - \hat{I}(X_0; X_1; \ldots; X_L) + \epsilon]\}. \tag{A.11}$$

Upon eliminating the exceptional codewords from the code, for all $\hat{P}$, one ends up with at least $[1-(n+1)^{|\mathcal{X}|^{L+1}}e^{-n\epsilon/2}]\cdot M$ for which

$$N_m(\hat{P},\mathcal{C}) \leq \exp\{n[LR-\hat{I}(X_0;X_1;\ldots;X_L)+\epsilon]\} \quad \forall \hat{P}. \tag{A.12}$$

Let $\mathcal{C}'$ denote the sub–code formed by these $[1-(n+1)^{|\mathcal{X}|^{L+1}}e^{-n\epsilon/2}]\cdot M$ remaining codewords. Since $N_m(\hat{P},\mathcal{C}') \leq N_m(\hat{P},\mathcal{C})$, then the sub–code $\mathcal{C}'$ certainly satisfies

$$N_m(\hat{P},\mathcal{C}') \leq \exp\{n[LR-\hat{I}(X_0;X_1;\ldots;X_L)+\epsilon]\} \quad \forall \hat{P}. \tag{A.13}$$

Finally, observe that since $N_m(\hat{P},\mathcal{C}')$ is a non–negative integer, then for $\hat{P}$ with $LR-\hat{I}(X_0;X_1;\ldots;X_L)+\epsilon<0$, the last inequality means $N_m(\hat{P},\mathcal{C}')=0$, in which case the r.h.s. of the last equation becomes $N^*(\hat{P})$. Thus, we have shown that there exists a code $\mathcal{C}'$ of size $M'=[1-(n+1)^{|\mathcal{X}|^{L+1}}e^{-n\epsilon/2}]\cdot e^{nR}$ for which all codewords satisfy $N_m(\hat{P},\mathcal{C}') \leq N^*(\hat{P})$ for all $\hat{P}$.

As a consequence of the above observation, we have seen the existence of a code $\mathcal{C}'$ for which

$$\begin{aligned}
& \max_m P_{e|m}(\mathcal{C}') \\
\leq\ & \sum_{\hat{P}} N^*(\hat{P})\exp\{-n\hat{\boldsymbol{E}}d(X_0,X_1,\ldots,X_L)\} && \text{(A.14)}\\
=\ & \sum_{\hat{P}:N^*(\hat{P})>0} \exp\{n[LR-\hat{I}(X_0;X_1;\ldots;X_L)-\hat{\boldsymbol{E}}d(X_0,X_1,\ldots,X_L)+\epsilon]\} && \text{(A.15)}
\end{aligned}$$

and due to the arbitrariness of $\epsilon>0$, this means (upon maximization over Q) that there exists a sequence of rate–R codes for which

$$\lim_{n\to\infty}\left[-\frac{\ln \max_m P_{e|m}(\mathcal{C})}{n}\right] \geq E_{\rm ex}^{\rm CKM}(R,L), \tag{A.16}$$

where $E_{\rm ex}^{\rm CKM}(R,L)$ is defined as in Theorem 2. This completes the proof of Theorem 2.

References

[1] E. Arikan and N. Merhav, "Guessing subject to distortion," *IEEE Trans. Inform. Theory*, vol. 44, no. 3, pp. 1041–1056, May 1998.

[2] E. Arikan and N. Merhav, "Joint source–channel coding and guessing with application to sequential decoding," *IEEE Trans. Inform. Theory*, vol. 44, no. 5, pp. 1756–1769, September 1998.

[3] C. Bai, B. Mielczarek, W. A. Krzymień, and I. J. Fair, "Improved analysis of list decoding and its application to convolutional codes and Turbo codes," *IEEE Trans. Inform. Theory*, vol. 53, no. 2, pp. 615–627, February 2007.

[4] V. M. Blinovsky, "Error probability exponent of list decoding at low rates," *Problems of Information Transmission*, vol. 27, no. 4, pp. 277–287, 2001 (translated from *Problemy Peredachi Informatsii*, no. 4, pp. 3–14, 2001).

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Second Edition, Hoboken, New Jersey, U.S.A., 2006.

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

[7] I. Csiszár, J. Körner, and K. Marton, "A new look at the error exponent of a discrete memoryless channel," *Proc. ISIT '77*, p. 107 (abstract), Cornell University, Itacha, New York, U.S.A., 1977.

[8] P. Elias, "List decoding for noisy channels," *Proc. IRE WESCON Conf. Rec.*, vol. 2, pp. 94–104, 1957.

[9] 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.

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

[11] R. G. Gallager, "The random coding bound is tight for the average code," *IEEE Trans. Inform. Theory*, pp. 244–246, March 1973.

[12] V. Guruswami, *List Decoding of Error–Correcting Codes*, Ph.D. dissertation, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA, September 2001. Available on–line at: http://www.cs.cmu.edu/∼venkatg/pubs/papers/frozen.pdf

[13] E. Hof, I. Sason, and S. Shamai (Shitz), "Performance bounds for erasure, list and decision feedback schemes with linear codes," *IEEE Trans. Inform. Theory*, vol. 56, no. 8, pp. 3754–3778, August 2010.

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

[15] N. Merhav, "On random coding error exponents of watermarking systems," *IEEE Trans. Inform. Theory*, vol. 46, no. 2, pp. 420–430, March 2000.

[16] N. Merhav, "Statistical physics and information theory," *Foundations and Trends in Communications and Information Theory*, vol. 6, nos. 1–2, pp. 1–212, 2009.

[17] N. Merhav, "Codeword or noise? Exact random coding exponents for slotted asynchronism," submitted to *IEEE Trans. Inform. Theory*, August 2013. http://arxiv.org/pdf/1308.4572.pdf

[18] N. Merhav, "Asymptotically optimal decision rules for joint detection and source coding," submitted to *IEEE Trans. Inform. Theory*, October 2013. http://arxiv.org/pdf/1310.4939.pdf

[19] J. Scarlett, L. Peng, N. Merhav, A. Martinéz, and A. G. i Fàbregas, "Expurgated random–coding ensembles: exponents, refinements and connections," submitted to *IEEE Trans. Inform. Theory*, July 2013.

[20] N. Seshadri and C.-W. E. Sundberg, "Generalized Viterbi algorithms for error detection and convolutional codes," *Proc. IEEE GLOBECOM89*, Dallas, TX, pp. 1534–1538, November 1989.

[21] N. Seshadri and C.-W. E. Sundberg, "List Viterbi decoding algorithms with applications," *IEEE Trans. Commun.*, vol. 42, pp. 313–323, Feb./Mar./Apr. 1994.

[22] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels. I", *Inform. Contr.*, vol. 10, pp. 65–103, January 1967.

[23] M. Studený and J. Vejnarová, "The multiinformation function as a tool for measuring stochastic dependence," *Learning in Graphical Models*, NATO ASI Series, vol. 89, pp. 261–297, 1998.

[24] E. Telatar, "Zero–error list capacities of discrete memoryless channels," *IEEE Trans. Inform. Theory*, vol. 43, no. 6, pp. 1977–1982, November 1997.

[25] E. Telatar, "Exponential bounds for list size moments and error probability," *Proc. 1998 IEEE Information Theory Workshop (ITW 1998)*, p. 60, Killarney, Ireland, June 1998.

[26] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding*, McGraw–Hill, New York, 1979.

[27] J. M. Wozencraft, "List decoding," MIT Res. Lab. Electron. Cambridge, MA, January 15, 1958.