EE 80653 Information Theory Tutorial

Exponential Error Bounds

Lei Xiao

Department of Electrical Engineering,

University of Notre Dame, Notre Dame, Indiana, USA 46556

Email: lxiao@nd.edu

I. Introduction of the Topic

One objective of information theory is to study the range of the rates that achieve asymptotically error-free transmission. The the error probabilities, especially its dependence on parameters such as rate or block length, is an important theoretical subject. Moreover, the error probability expressions can help to optimize the choice of parameters in practical designs and reflect the possible trade-off.

Unfortunately, most of the coded communication systems are far too complicated to derive their error probabilities in closed-form. Tight error bounds can be used in these situations to provide insights while maintaining the analysis tractable. There are two different types of bounds. An *upper* bound is a conservative estimate of the error probability, and showing it approaching zero is sufficient to conclude the probability converges to zero. A *lower* bound gives the best possible performance, and thus shows what is impossible to reach. Upper bounds are usually useful in demonstrating achievability (e.g. the derivation of the coding theorem in [1] and the use of union bound in the analysis of typicality decoder in [2]), while lower bounds can be used in the converse (e.g. the result in [3] tells short Turbo codes cannot achieve capacity).

To make the study meaningful, we want to bound the performance of codes that are "reasonably good", instead of poorly designed codes whose performance does not have any practical insight. However, as one would soon find out, the characterization of a specific good code or the "best" code is in general hard to obtain. To circumvent this technical difficulty, the ensemble of codes, a.k.a. random codes, are considered in measuring performance. We consider exponential upper error bounds for the ensemble of codes in this tutorial.

The rest of this tutorial is organized as follows. Conventional ML decoder and decoders with erasure or list option are introduced in Section II. The exponential error bounds are summarized in Section III with only the proof of Gallager's bound. Some interesting properties and implications of these exponential error bounds are discussed in Section IV. We conclude this tutorial in Section V and present some related contributions not covered in this tutorial in Section VI.

II. Decoder with Erasure and List Options

A pictorial comparison among ordinary decoder, erasure decoder and list decoder is given in Fig. 1. The big circle represents the set of all possible value observed at the output of the channel, while the regions marked with numbers reflects the decision rule of the decoder.

1) In an ordinary decoder, one and only one hard decision is required from the observation of the channel input. The decision regions of all codewords therefore forms a *partition* of the space of the channel output, as shown in Fig. 1. Let x_m, y and R_m denotes the m-th codeword in the codebook, the channel output, and the decision region for the m-th codeword, respectively, the error probability of an ordinary decoder can be formulated as

$$\mathbb{P}(\mathcal{E}) = \sum_{m'} \sum_{m' \neq m} \sum_{\mathrm{y} \in R_{m'}} \mathbb{P}(\mathrm{y}, \mathrm{x}_m) \tag{1}$$

and the optimum decision rule the maximum a posteriori probability (APP) decoding, i.e. [4]

$$\mathrm{y} \in R_m \text{ iff } \mathbb{P}(\mathrm{x}_m|\mathrm{y}) > \mathbb{P}(\mathrm{x}_{m'}|\mathrm{y}) \text{ all } m' \neq m \tag{2}$$
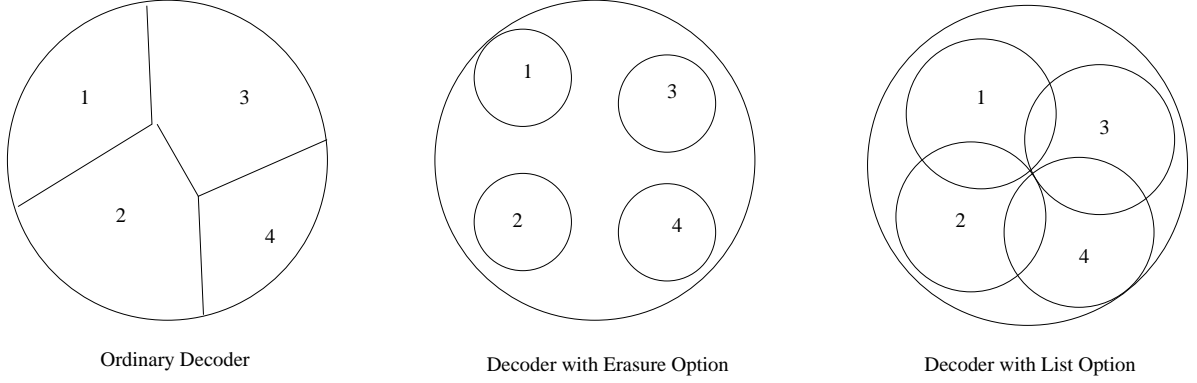
Fig. 1.   The decision regions for the ordinary decoder and decoders with erasure and list options.

2) In a decoder with erasure option, the decoder can reject all the hypothesis and refuse to make a decision. In such situations, we call the output of the decoder as erasure. The uncertainty introduced by the erasure can be resolved at a later stage by redundancy in the data, a repeat transmission (referred to as "decision feedback" by Forney), or by an outer code in concatenated coding setting [4]. Different from the ordinary decoders, the performance of erasure decoders is characterized by two probabilities, namely the probability of erasure and the probability of undetected errors. Define the event E_1 as "the received word y does not fall into the decision region of the transmitted codeword x_m", and the event E_2 as "y lies in R_m and $x_{m'}, m \neq m'$ was actually transmitted". Mathematically, these two quantities can be written as

$$\mathbb{P}(E_1) = \sum_m \sum_{y \notin R_m} \mathbb{P}(y, x_m) \tag{3a}$$

$$\mathbb{P}(E_2) = \sum_m \sum_{y \in R_m} \sum_{m' \neq m} \mathbb{P}(y, x_{m'}) \tag{3b}$$

Obviously, $\mathbb{P}(E_2)$ is the probability of undetected errors. When event E_1 happens, there is either an erasure or an undetected error, and the two possibilities are mutually exclusive. Hence the probability of erasure is $\mathbb{P}(E_1) - \mathbb{P}(E_2)$.

Due to the existence of an erasure option, there are certain tradeoff between the erasure rate and the probability of the undetected error. By allowing more erasure on unreliable receptions, the probability of undetected error will decrease. Forney formulated the decision criterion that is optimum in the Neyman-Pearson (N-P) sense.

Theorem 1 (Forney, 1968): Let the decision regions R_m be defined by

$$y \in R_m \text{ iff } \frac{\mathbb{P}(y, x_m)}{\sum_{m' \neq m} \mathbb{P}(y, x_{m'})} \geq \exp(NT) \tag{4}$$

where N is the block length and T is an arbitrary parameter. Then no other decision rules can gives both a lower $\mathbb{P}(E_1)$ and a lower $\mathbb{P}(E_2)$ than this set of decision regions do.

The proof of this N-P optimum decision rule follows directly from a generalization of the Neyman-Pearson theorem [4]. Note that to make sure that the decision regions are non-overlapping, T must be positive. Also, the increase of T makes fewer decision and more erasures, while decrease in

T provides trade-off in favor of decreasing erasure rate. The decision rule in (4) can be further simplified as [4]

$$\text{y} \in R_m \text{ iff } \mathbb{P}(\text{x}_m|\text{y}) \geq \eta \triangleq \frac{\exp(NT)}{1 + \exp(NT)} \tag{5}$$

3) In a decoder with list option, the decision regions are allowed to overlap with one and another. The decoder, therefore, can possibly output several candidates instead of just one most likely codeword. This option could be useful when the source is redundant, e.g. the transmission of English text, when there are decision feedback between transmitter and receiver (so that the transmitter can choose one of the candidates in the list in a subsequent transmission), or when an outer code is used to make trials on each of the candidates in the list and make the decision. The two parameters characterizing the performance of a list decoder is the probability of error and the average number of incorrect words on the list $\bar{N}$. The error event is defined as "x_m is transmitted but y in not in R_m", which is the same as E_1 defined in the case of erasure decoding. Moreover, if we relax the definition of $\mathbb{P}(E_2)$ and allow the sum in (3b) be performed in overlapped R_m's, then $\bar{N} = \mathbb{P}(E_2)$. Hence the N-P optimum decision criterion is again determined by (4), with the only difference being T should be a negative number to allow multiple codewords as output.

III. THE BOUNDS

A. Gallager's Bounds for ML Decoding

Theorem 2 (Gallager's bound): There exists $(\lceil e^{NR} \rceil, N)$ code over a discrete memoryless channel without feedback such that with APP decoding, the (block) error probability is bounded by

$$\mathbb{P}(\mathcal{E}) \leq \exp[-NE(R)], \qquad E(R) = \max_{q(x)} \max_{0 \leq \rho \leq 1} [E_0(\rho, q(x)) - \rho R] \tag{6}$$

where $E_0(\rho, q(x))$ is Gallager's error exponent defined as

$$E_0(\rho, q(x)) = -\ln \sum_j \left[\sum_k q(x_k) \mathbb{P}^{1/1+\rho}(y_j|x_k) \right]^{1+\rho} \tag{7}$$

Proof: Let the $(\lceil e^{NR} \rceil, N)$ code be randomly chosen over $\mathbb{Q}(\text{x})$ and the input message is uniformly distributed. Without loss of generality, we consider the error probability of a ML decoder when x_1 is transmitted. Condition on x_1 and y, the probability of error is

$$\begin{aligned}
\mathbb{P}(\mathcal{E}, 1|\text{x}_1, \text{y}) &= \mathbb{P}(\exists m \neq 1, \mathbb{P}(\text{y}|\text{x}_m) > \mathbb{P}(\text{y}|\text{x}_1)|\text{x}_1, \text{y}) \\
&\leq \min \left\{ 1, \sum_{m=2}^{\lceil e^{NR} \rceil} \mathbb{P}(\mathbb{P}(\text{y}|\text{x}_m) > \mathbb{P}(\text{y}|\text{x}_1)|\text{x}_1, \text{y}) \right\} \tag{8a} \\
&\leq \left[\sum_{m=2}^{\lceil e^{NR} \rceil} \mathbb{P}(\mathbb{P}(\text{y}|\text{x}_m) > \mathbb{P}(\text{y}|\text{x}_1)|\text{x}_1, \text{y}) \right]^\rho, \qquad 0 \leq \rho \leq 1 \tag{8b} \\
&= \left[(\lceil e^{NR} \rceil - 1) \mathbb{P}(\mathbb{P}(\text{y}|\text{x}_2) > \mathbb{P}(\text{y}|\text{x}_1)|\text{x}_1, \text{y}) \right]^\rho \tag{8c}
\end{aligned}$$

where (8a) is the union bound together with the fact that probability is smaller than one; (8b) is because $x^\rho > x$ if $x < 1$ and $x^\rho > 1$ if $x > 1$; while (8c) is because of the symmetry from random construction. The probability $\mathbb{P}(\mathbb{P}(\text{y}|\text{x}_2) > \mathbb{P}(\text{y}|\text{x}_1)|\text{x}_1, \text{y})$ is interpreted as the probability of a randomly chosen

codeword is more likely than the transmitted codeword x_1, given the transmitted codeword and the received vector y. This quantity can be further bounded as

$$
\mathbb{P}(\mathbb{P}(y|x_2) > \mathbb{P}(y|x_1)|x_1, y) = \sum_{x_2 : \mathbb{P}(y|x_2) > \mathbb{P}(y|x_1)} \mathbb{Q}(x_2)
$$

$$
= \sum_{x_2} \mathbb{Q}(x_2) \times \mathbf{1}_{\mathbb{P}(y|x_2) > \mathbb{P}(y|x_1)} \tag{9a}
$$

$$
\leq \sum_{x_2} \mathbb{Q}(x_2) \left[\frac{\mathbb{P}(y|x_2)}{\mathbb{P}(y|x_1)} \right]^s, \qquad s \geq 0 \tag{9b}
$$

where $\mathbf{1}_{\mathbb{P}(y|x_2) > \mathbb{P}(y|x_1)}$ in (9a) is the indicator function, i.e. $\mathbf{1}_{a>b} = 1$ if $a > b$ and $\mathbf{1}_{a>b} = 0$ otherwise. The inequality in (9b) is explained by the fact that $[\frac{a}{b}]^s \geq 1$ if $a > b$ and $1 \geq [\frac{a}{b}]^s \geq 0$ if $a < b$ $(a, b, s \geq 0)$. Substitute (9b) in (8c) and average over the distribution $\mathbb{Q}(x_1)\mathbb{P}(y|x_1)$, the probability of error when the first codeword is transmitted is bounded by

$$
\mathbb{P}(\mathcal{E}, 1) = \sum_{y, x_1} \mathbb{Q}(x_1)\mathbb{P}(y|x_1)\mathbb{P}(\mathcal{E}, 1|x_1, y)
$$

$$
\leq \sum_y \sum_{x_1} \mathbb{Q}(x_1)\mathbb{P}(y|x_1) \left\{ (\lceil e^{NR} \rceil - 1) \sum_{x_2} \mathbb{Q}(x_2) \left[\frac{\mathbb{P}(y|x_2)}{\mathbb{P}(y|x_1)} \right]^s \right\}^\rho
$$

$$
\leq e^{N\rho R} \sum_y \left[\sum_{x_1} \mathbb{Q}(x_1)\mathbb{P}^{1-s\rho}(y|x_1) \right] \left[\sum_{x_2} \mathbb{Q}(x_2)\mathbb{P}^s(y|x_2) \right]^\rho \tag{10}
$$

Now note that due to the random generation and uniform input $\mathbb{P}(\mathcal{E}, 1) = \mathbb{P}(\mathcal{E}, 2) = \cdots = \mathbb{P}(\mathcal{E}, \lceil e^{NR} \rceil) = \mathbb{P}(\mathcal{E})$ and $x_m, m = 1, 2, \cdots, \lceil e^{NR} \rceil$ are identically distributed with probability mass function $\mathbb{Q}(x)$. Choose $s = 1/(1 + \rho)$ (Gallager showed this choice tighten the bound [5]) we obtain the bound on the error probability

$$
\mathbb{P}(\mathcal{E}) \leq e^{N\rho R} \sum_y \left[\sum_x \mathbb{Q}(x)\mathbb{P}^{1/1+\rho}(y|x) \right]^{1+\rho} \tag{11}
$$

Use the property of discrete memoryless channel without feedback and choose the input distribution to be independent and identically distributed $\mathbb{Q}(x) = \Pi_{i=1}^N q(x(i))$, the above bound can be further simplified as

$$
\mathbb{P}(\mathcal{E}) \leq e^{N\rho R} \left\{ \sum_y \left[\sum_x q(x)\mathbb{P}^{1/1+\rho}(y|x) \right]^{1+\rho} \right\}^n \tag{12}
$$

The Gallager's bound is obtained by optimize ρ and input distribution $q(x)$. Q.E.D.

It worth noting that the error exponent can be further tighten at low rates by the expurgated error bound [5, p. 153]. However, the above error bound tight in high rate and is sufficient to prove important conclusions such as the achievability of the channel coding theorem.

B. Forney's Bound for Erasure/List Decoding

Theorem 3 (Forney, 1968): There exists $(\lceil e^{NR}, N \rceil)$ code such that when (4) is used with threshold T, one can obtain at the same time

$$
\mathbb{P}(E_1) \leq \exp[-NE_1(R, T)] \tag{13a}
$$

$$
\mathbb{P}(E_2) \leq \exp[-NE_2(R, T)] \tag{13b}
$$

$E_1(R, T)$ is

$$E_1(R, T) = \max_{0 \geq s \geq \rho \geq 1, q(x)} E_0(s, \rho, q(x)) - \rho R - sT \tag{14a}$$

$$E_0(s, \rho, q(x)) = -\ln \sum_j \left[\sum_k q(x_k) \mathbb{P}^{1-s}(y_j | x_k) \right] \left[\sum_{k'} q(x_{k'}) \mathbb{P}^{s/\rho}(y_j | x_{k'}) \right]^\rho \tag{14b}$$

whereas $E_2(R, T)$ is given by

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

The proof is very similar to that of Gallager's bound and the details can be found in [4, Appendix]. As in the ML decoding case, the bound can be tighten at low rates by the expurgation argument.

IV. SOME PROPERTIES

A. Gallager's Bound and Channel Capacity

Consider the limit [5]

$$\lim_{\rho \to 0} \frac{E_0(\rho, q(x))}{\rho} = \lim_{\rho \to 0} \frac{\mathrm{d}}{\mathrm{d}\rho} E_0(\rho, q(x))$$
$$= I(X; Y) \tag{16}$$

Hence, for any $R < I(X; Y)$, we can find a ρ close to zero enough such that $\frac{E_0(\rho, q(x))}{\rho} > R \implies E_0(\rho, q(x)) - \rho R > 0$. Maximize $I(X; Y)$ over the distribution $q(x)$, we reach the conclusion that for every $R < C = \max_{q(x)} I(X; Y)$ there exists at least one sequence of $(\lceil e^{NR} \rceil, N)$ code whose (block) error rate using ML decoding decrease exponentially with the block length.

B. Implication to Decoder with Erasure/List Options

We first consider a system with erasure decoder and request for repeat transmission. Denote the probability of an erasure as $\mathbb{P}(X)$, then the average number of transmission per codeword will be

$$1 + \mathbb{P}(X) + [\mathbb{P}(X)]^2 + \cdots = \frac{1}{1 - \mathbb{P}(X)} \tag{17}$$

The rate of the information transfer will be reduced to $R[1 - \mathbb{P}(X)]$ due to retransmission of the same codeword. Since

$$\mathbb{P}(X) < \exp[-NE_1(R, T)] \tag{18}$$
$$\mathbb{P}(\mathcal{E}) < \exp[-NE_2(R, T)] \tag{19}$$

the effective rate $R[1 - \mathbb{P}(X)]$ can be pushed as close to R as one might wanted by increasing N. By some mathematical manipulations, Forney showed that "near capacity the achievement of low error probabilities is dramatically simplified by use of erasure decoder and request for repeat transmission", although the capacity cannot be increased by feedback for DMC.

For list decoders, as $\bar{N}$ is bounded by the error exponent, the average number of candidates in the output is close to one as N approaches infinity. Hence the average complexity in processing the list is rather low, although the list could be as large as $\exp[N(-T)]$. Forney showed the benefit of list decoder is most visible in low rate region [4]. The derivation of such results made use of expurgated error bounds,

and we just quote the conclusions drawn by Forney as "at rates above R_{comp}, not much improvement is obtained by going to lists over ordinary decoding; at rates below R_{comp}, however, significant improvement can be achieved, approaching that attainable with feedback at very low rates."

V. SUMMARY

We first compare ordinary decoder and decoders with erasure/list options in this tutorial. Application of Neyman-Pearson theorem provides us with optimum decision rule for erasure and list decoders in a uniform framework as threshold testing problem. We derived the Gallager's exponential error bound and summarized corresponding bounds for erasure/list decoders derived by Forney. The connection between Gallager's bound and channel capacity is demonstrate, and some implication for erasure/list decoders revealed by Forney are outlined.

VI. LITERATURE AT A GLANCE

The detailed derivation of exponential error bounds for ML decoders can be found in [1] and [5, Chapter 5]. Forney presented the details of the mathematical derivation of bounds for list and erasure decoders in [4, Appendix]. Also discussed in Forney's paper [4] is the zero error capacity with request for repeat transmission or list decoder. Improved bounds for Forney's list/erasure decoder setting were proved in [6] for the family of spherical codes and binary linear codes. Finally, a summary of different variations of Gallager bounds and their connections can be found in [7].

REFERENCES

[1] R. G. Gallager, "A simple derivation of the coding theorem and some applications," *IEEE Transactions on Information Theory*, vol. 11, pp. 3–18, Jan. 1965.

[2] T.M.Cover and J.A.Thomas, *Elements of Information Theory*. New York: Wiley-Interscience, 1991.

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

[4] G. D. Forney, Jr., "Exponential error bounds for erasure, list, and decision feedback schemes," *IEEE Transactions on Information Theory*, vol. 14, pp. 206–220, Mar. 1968.

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

[6] A. Barg, "Improved error bounds for the erasure/list scheme: The binary and spherical cases," *IEEE Transactions on Information Theory*, vol. 50, pp. 2503–2511, Oct. 2004.

[7] S. Shamai and I. Sason, "Variations on the Gallager bounds, connections, and applications," *IEEE Transactions on Information Theory*, vol. 48, pp. 3029–3051, Dec. 2002.