Unequal Error Protection: Some Fundamental Limits*

Shashi Borade   Barış Nakiboğlu   Lizhong Zheng
EECS, Massachusetts Institute of Technology
{ spb, nakib, lizhong } @mit.edu

Abstract

Various scenarios are considered where some information is more important than other and needs better protection. A general theoretical framework for unequal error protection is developed in terms of exponential error bounds. It provides some fundamental limits and optimal strategies for such problems. New class of *message-wise* unequal error protection problems are also introduced.

Even for data-rates approaching the channel capacity, it is shown that a crucial part of information can be protected with exponential reliability. Channels without feedback are analyzed first, which is useful later while analyzing channels with feedback.

1 Introduction

Classical theoretical framework for communication [1] assumes that all information is equally important. In this framework, the communication system aims to provide a uniform error protection to all messages: any particular message being mistaken as any other is viewed to be equally costly. With such uniformity assumptions, reliability of a communication scheme is measured by either the average or the worst case probability of error, over all possible messages to be transmitted. In information theory literature, a communication scheme is said to be *reliable* if this error probability can be made small. Communication schemes designed with this framework turn out to be optimal in sending any source over any channel, provided that long enough codes can be employed. This homogeneous view of information motivates the universal interface of "bits" between any source and any channel [1], and is often viewed as Shannon's most significant contribution.

In many communication scenarios, such as wireless networks, interactive systems, and control applications, where uniformly good error protection becomes a luxury, providing such a protection to the entire information might be wasteful, if not infeasible. Instead, it is more efficient here to protect a crucial part of information better than the rest. For example,

- In a wireless network, control signals like channel state, power control, and scheduling information are often more important than the payload data, and should be protected more carefully. Thus even though the final objective is delivering the payload data, the physical layer should provide a better protection to such protocol information.

*This research is supported by DARPA ITMANET project and an AFOSR grant FA9550-06-0156. Initial part of this paper was submitted to IEEE International Symposium on Information Theory, 2008.

Similarly for the Internet, packet headers are more important for delivering the packet and need better protection to ensure that the actual data gets through.

- Another example is transmission of a multiple resolution source code. The coarse resolution needs a better protection than the fine resolution so that the user at least obtains some crude reconstruction after bad noise realizations.
- Controlling unstable plants over noisy communication link [23] and compressing unstable sources [24] provide more examples where different parts of information need different reliability.

These examples demonstrate the heterogeneous nature of information in contrast with the classical homogeneous view. For these situations, unequal error protection (UEP) is a natural generalization to the conventional content-blind information processing.

The simplest method of unequal error protection is to allocate different channels for different types of data. For example, many wireless systems allocate a separate "control channel", often with short codes and low spectral efficiency, to transmit control signals with high reliability. The well known Gray code, assigning similar bit strings to close by constellation points, can be viewed as UEP: even if there is some error in identifying the transmitted symbol, there is a good chance that some of the bits are correctly received. More systematic designs for UEP can be found in [13, 14, 15] and references therein. For erasure channels, this problem is known as "priority encoded transmission" (PET) [12]. For wireless channels, [17] analyzes this problem in terms of diversity-multiplexing tradeoffs. Most of these approaches focus on designing good codes for specific channel models. The optimality of these designs was established in only limited cases. This paper aims to provide a general information theoretic framework for understanding fundamental limits in UEP.

Consider a channel encoder which takes the input of k information bits, $\mathbf{b} = [b_1, b_2, \dots b_k]$, which is equivalent to a random variable M taking values from the set $\{1, 2, 3, \dots, 2^k\}$. Each message in this set corresponds to a particular value of the bit-sequence $\mathbf{b}$. This set of possible values of M are referred to as "messages". After a message is encoded and transmitted over the channel, a decoding error is defined as the event that the receiver decodes to a different message than the transmitted one. In most information theory texts, when a decoding error occurs, the entire bit sequence $\mathbf{b}$ is rejected. That is, errors in decoding the message and in decoding the information bits are treated similarly.

In the existing formulations of unequal error protection codes, the information bits are divided into subsets, and the decoding errors in different subsets of bits are viewed as different kinds of errors. For example, one might want to provide a better protection to one subset of bits by ensuring that errors in these bits are less probable than the other bits. We call such problems as "bit-wise UEP". Previous examples of packet headers, multiple resolution codes, etc. belong to this category of UEP.

However, in some situations, instead of *bits* one might want to provide a better protection to a subset of *messages*. For example, one might consider embedding a special message in a normal k-bit code, i.e., transmitting one of $2^k + 1$ messages, where the extra message has a special meaning and requires a smaller error probability. Note that the error event for the special message is not associated to error in any particular bit. Instead, it corresponds to a particular bit-sequence (*i.e.*, message) being decoded as some other bit-sequence. Borrowing from hypothesis testing, we can define two kinds of errors corresponding to a special message.

- We say that *missed-detection* of a message i occurred when that message was transmitted, but the receiver missed it by decoding to some other message $j \neq i$. Consider a special message indicating some system emergency, which is too costly to be missed. Clearly, such special messages demand a small missed detection probability. Missed detection probability of a message is simply the conditional error probability after its transmission.

- We say that *false-alarm* of a message i occurred when some other message $j \neq i$ was transmitted, but the receiver decoded it to message i. Consider the reboot message for a remote-controlled system such as a robot or satellite or the "disconnect" message to a cell-phone. Its false-alarm could cause unnecessary shutdowns and other system troubles. Such special messages demand small false alarm probability.

We call such problems as "message-wise UEP". In conventional framework, every bit is as important as every other bit and every message is as important as every other message. In short in conventional framework it is assumed that all the information is "created equal". In such a framework there is no reason to distinguishing between bit-wise or message wise error probabilities because message wise error probability larger than bit wise error probability by an insignificant factor. However, in the UEP setting, it is necessary to differentiate between message-errors and bit-errors. We will see that in many situations, error probability of special bits and messages have very different behavior.

The main contribution of this paper is a set of results, identifying the performance limits and optimal coding strategies, for a variety of UEP scenarios. We will focus on a few simplified notions of UEP, most with immediate practical applications, and try to illustrate the main insights for them. One can imagine using these UEP strategies for embedding protocol information within the actual data. By eliminating a separate control channel, this can enhance the overall bandwidth and/or energy efficiency.

For conceptual clarity, this article focuses on situations where the data-rate is essentially equal to the channel capacity[1]. By this analysis, we will be addressing UEP issues for scenarios where data rate is a crucial system resource that can not be compromised. In these situations, no positive error exponent in the conventional sense can be achieved. That is, if we aim to protect the entire information uniformly well, neither bit-wise nor message-wise error probabilities can decay exponentially fast with increasing code length. We ask the question then "can we make the error probability of a particular bit, or a particular message, decay exponentially fast with block length?"

The question of fundamental limits of UEP was clearly of interest in previous works on code designs for UEP. To the best of our knowledge, however, there was no general characterization of these limits in terms of error exponents; partially due to the difficulty in proving converses. In this paper and [10], we develop such converses as well as optimal strategies. More importantly, the notion of message-wise UEP was essentially never addressed in the past (except in the paper "*Joint Source-Channel Error Exponent*" by Csiszár, [16]).

When we break away from the conventional framework and start to provide better protection to selected parts of information, these parts of information need not be only *bits*. A general formulation of UEP could be an arbitrary combination of protection demands from messages, where each message demands better protection against some specific kinds of errors. In this general definition of UEP, bit-wise UEP and message-wise UEP are simply two particular ways of specifying which kinds of errors are too costly compared to others.

[1]In another write-up [10], we will analyze similar problems in a more general framework to allow data-rates below capacity.

In the following, we start by defining the channel model and some basic definitions in Section 2. Then Section 3 discusses bit-wise UEP and message-wise UEP for the block codes without feedback. Its Theorem 1 shows that for data-rates approaching capacity, even a single bit cannot achieve any positive error exponent. Thus in bit-wise UEP, the data-rate must back-off from capacity for achieving any error exponent even for a single bit. On the contrary, in message-wise UEP, positive error exponents can be achieved even at capacity. If only one message in a capacity achieving code was special and demanded an error exponent, Theorem 2 shows its optimal value is equal to a new fundamental channel parameter called the *red-alert exponent*. We then consider situations where an exponentially large subset of messages is special and each message in it demands a positive error exponent. Theorem 3 shows a surprising result that these special messages can achieve the same exponent as if all the other (non-special) messages were absent. In other words, a capacity achieving code and an error exponent-optimal code below capacity can coexist without hurting each other. These results also shed some new light on the structure of capacity achieving codes.

Insights from the block codes without feedback becomes useful in Section 4 where we investigate similar problems for variable length block codes with feedback. Feedback together with variable decoding time creates some fundamental connections between bit-wise UEP and message-wise UEP. Now even for bit-wise UEP, positive error exponent can be achieved at capacity. Theorem 5 shows that a single special bit can achieve the same exponent as a single special message—the red-alert exponent. As the number of special bits increases, the achievable exponent for them decays linearly with their rate as shown in Theorem 6. Then Theorem 7 generalizes this result to the case when there are multiple levels of specialty—most special, second-most special and so on. It uses a strategy similar to onion-peeling and achieves error exponents which are successively refinable over multiple layers. For a single special message however, Theorem 8 shows that feedback does not improve the achievable exponent. The case of exponentially many messages is resolved in Theorem 9. Of course, many special messages cannot achieve a better exponent compared to a single special message. We will see that the special messages can achieve the same error exponent with feedback as if all other messages were absent, at rates beyond certain threshold.

Section 5 then addresses message-wise UEP situations where special messages demand small probability of false-alarms instead of missed-detections. It considers the case of no-feedback as well as full feedback. This discussion for false-alarms was postponed from earlier sections to avoid confusion with the missed-detection results in earlier sections. Later, some future directions are discussed briefly in Section 6.

After discussing each theorem, we have provided a brief description of the optimal strategy. More proof details can be found in later sections. Section 7 discusses proofs of the results on for block codes without feedback in Section 3. Section 8 discusses proofs of the results on variable length block codes with feedback in Section 4 and lastly, Section 9 discusses proofs for the false-alarm results in Section 5.

2 Channel Model and Notation

2.1 Channel Model and Block Codes

We will consider a discrete memoryless channel $W_{Y|X}$, with input alphabet $\mathcal{X} = \{1, 2, \ldots, |\mathcal{X}|\}$ and output alphabet $\mathcal{Y} = \{1, 2, \ldots, |\mathcal{Y}|\}$. The conditional distribution of output letter

Y when the channel input letter X equals $i \in \mathcal{X}$ is denoted by $W_{Y|X}(\cdot|i)$.

$$\Pr\left[Y = j|\, X = i\right] = W_{Y|X}(j|i) \qquad \forall i \in \mathcal{X},\ \forall j \in \mathcal{Y}. \tag{1}$$

We assume that all the entries of the channel transition matrix are non-zero, that is, every output letter is reachable from every input letter. This assumption is indeed a crucial one. Many of the results we will present in this paper will change when there are zero-probability transitions.

A length n block code without feedback with message set $\mathcal{M} = \{1, 2, \ldots, |\mathcal{M}|\}$ is composed of two mappings, encoder mapping and decoder mapping. Encoder mapping assigns a length n codeword[2], $\bar{x}^n(k) \triangleq (\bar{x}_1(k), \bar{x}_2(k) \cdots, \bar{x}_n(k))$, for each element k of the message set, where $\bar{x}_t(k)$ denotes the input at time t. Decoder mapping will assign a message, $\hat{M}(Y^n) \in \mathcal{M}$ for each possible channel output sequence Y^n.

At time zero transmitter will be given the message M, which is chosen from $\mathcal{M}$ according to a uniform distribution. In the following n time units, it will send the corresponding codeword. After observing Y^n, receiver will decode a message. The average error probability P_e and rate R of the code is given by

$$P_e \triangleq \Pr\left[\hat{M} \neq M\right] \qquad \text{and} \qquad R \triangleq \tfrac{\log|\mathcal{M}|}{n}. \tag{2}$$

2.2 Different Kinds of Errors

While discussing message-wise UEP, we will consider the conditional error probability for a particular message $i \in \mathcal{M}$:

$$\Pr\left[\hat{M} \neq i \middle| M = i\right]. \tag{3}$$

Recall that this is the same as the missed detection probability for message i.

On the other hand when we are talking about bit-wise UEP, the overall message is composed of two components, $M = (M_1, M_2)$, where M_i is chosen uniformly from message set $\mathcal{M}_i$. For example, M_1 may correspond to the high-priority bits while M_2 corresponds to the low-priority bits. Note that now the message set $\mathcal{M}$ is equal to the Cartesian product $\mathcal{M}_1 \times \mathcal{M}_2$. Error probability of decoding M_j is given by

$$\Pr\left[\hat{M}_j \neq M_j\right] \qquad j = 1, 2 \tag{4}$$

Note that the overall message M is decoded incorrectly when either M_1 or M_2 or both are decoded incorrectly. The goal of bit-wise UEP is to achieve best possible $\Pr\left[\hat{M}_1 \neq M_1\right]$ while ensuring reasonably small $\Pr\left[\hat{M} \neq M\right] = P_e$.

2.3 Reliable Code Sequences

This investigation focuses on systems where reliable communication is achieved and aims to find exponentially tight bounds for error probabilities of special parts of information. We use the notion of code-sequences to simplify our discussion.

[2]Unless mentioned otherwise, small letters (e.g. x) will denote a particular value of the corresponding random variable denoted in capital letters (e.g. X).

An infinite sequence of codes indexed by their block length is called *reliable* if

$$\lim_{n\to\infty} {P_e}^{(n)} = 0 \tag{5}$$

For any reliable code-sequence $\mathcal{Q}$, its rate $R_{\mathcal{Q}}$ is given by

$$R_{\mathcal{Q}} \triangleq \liminf_{n\to\infty} \tfrac{\log|\mathcal{M}^{(n)}|}{n} \tag{6}$$

The (conventional) error exponent of a reliable sequence is then

$$E_{\mathcal{Q}} \triangleq \liminf_{n\to\infty} \tfrac{-\log {P_e}^{(n)}}{n} \tag{7}$$

Thus the number of messages in $\mathcal{Q}$ is[3] $\doteq e^{nR_{\mathcal{Q}}}$ and its average error probability equals ${P_e}^{(n)} \doteq e^{n\mathcal{E}_{\mathcal{Q}}}$. Now we can define error exponent $E(R)$ in the conventional sense, which is equivalent to the ones given in [2],[3],[4],[5],[7].

Definition 1 *For any $R \leq \mathrm{C}$ the error exponent $E(R)$ is defined as*

$$E(R) \triangleq \sup_{\mathcal{Q}:R_{\mathcal{Q}}\geq R} E_{\mathcal{Q}} \tag{9}$$

As mentioned previously, we are interested in UEP when operating at capacity. We already know that $E(C) = 0$, [3], i.e. the overall error probability cannot decay exponentially at capacity. In the following sections, we will show how certain parts of information can still achieve a positive exponent at capacity. In doing that, we will be solely focusing on the reliable $\mathcal{Q}$'s whose rates are equal to C. We will call such reliable code sequences as *capacity-achieving sequences.*

Through out the text we will denote Kullback-Leibler (KL) divergence between two distributions $\alpha_X(\cdot)$ and $\beta_X(\cdot)$ as $D\left(\alpha_X(\cdot)\|\,\beta_X(\cdot)\right)$.

$$D\left(\alpha_X(\cdot)\|\,\beta_X(\cdot)\right) = \sum_{i\in\mathcal{X}} \alpha_X(i) \log \tfrac{\alpha_X(i)}{\beta_X(i)}$$

Similar conditional KL divergence between $W_{Y|X}(\cdot|\cdot)$ and $\Psi_{Y|X}(\cdot|\cdot)$ under $P_X(\cdot)$ will be denoted by $D\left(W_{Y|X}(\cdot|X)\middle\|\,\Psi_{Y|X}(\cdot|X)\middle|\,P_X\right)$.

$$D\left(W_{Y|X}(\cdot|X)\middle\|\,\Psi_{Y|X}(\cdot|X)\middle|\,P_X\right) = \sum_{i\in\mathcal{X}} P_X(i) \sum_{j\in\mathcal{Y}} W_{Y|X}(j|i) \log \tfrac{W_{Y|X}(j|i)}{\Psi_{Y|X}(j|i)}$$

The input distribution that achieves the capacity will be denoted by P_X^*. Corresponding output distribution will be denoted by P_Y^*.

[3]The $\doteq$ sign denotes equality in the exponential sense. For a sequence $a^{(n)}$,

$$a^{(n)} \doteq e^{nF} \Leftrightarrow F = \liminf_{n\to\infty} \frac{\log a^{(n)}}{n} \tag{8}$$

3 UEP at Capacity: Block Codes without Feedback

3.1 Special bit

We first address the situation where one particular (say the first) information bit out of the total $\lfloor \log_2 |\mathcal{M}| \rfloor$ information bits is a special bit—it needs a much better error protection than the overall information. If this first bit is denoted as b_1 and its decoded value is denoted by $\hat{b}_1$, we require the error probability for b_1 to decay exponentially while ensuring reliable communication at capacity for the remaining bits.

In the Cartesian-product terminology, the single special bit is equivalent to defining $\mathcal{M}_1 = \{0, 1\}$, from which $M_1 \equiv b_1$ is chosen uniformly. The overall message equals $M = (M_1, M_2)$, where M_2 is independent of M_1 and chosen uniformly from $\mathcal{M}_2$. The optimal error exponent E_{b} for the special bit is be defined as follows[4].

Definition 2 *For a capacity-achieving sequence $\mathcal{Q}$ with message sets $\mathcal{M}^{(n)} = \mathcal{M}_1 \times \mathcal{M}_2^{(n)}$ where $\mathcal{M}_1 = \{0, 1\}$, the special bit error exponent is defined as*

$$E_{b,\mathcal{Q}} \triangleq \liminf_{n\to\infty} \frac{-\log \Pr{}^{(n)}\left[\hat{M}_1 \neq M_1\right]}{n} \tag{10}$$

Then E_b is defined as $E_b \triangleq \sup_{\mathcal{Q}} E_{b,\mathcal{Q}}$.

Thus if $\Pr{}^{(n)}\left[\hat{b}_1 \neq b_1\right] \doteq \exp(-nE_{\mathrm{b},\mathcal{Q}})$ for a reliable sequence $\mathcal{Q}$, then E_{b} is the supremum of $E_{\mathrm{b},\mathcal{Q}}$ over all capacity-achieving $\mathcal{Q}$.

Since $E(C) = 0$, it was clear that the entire information cannot achieve any positive error exponent at capacity. However, it is not clear whether a single special bit can steal a positive error exponent E_{b} at capacity.

Theorem 1 $E_b = 0$

This implies that even if we are aiming to protect a single bit with exponential reliability, the data-rate should back-off from capacity.

Intuitive Interpretation: Let the shaded balls in Fig. 3.1 denote the minimal decoding regions of the $\doteq e^{nC}$ messages. These decoding regions to ensure reliable communication, they essentially denote the typical noise-balls [9] around codewords.

The decoding regions on the left of the thick line corresponds to $\hat{b}_1 = 1$ and those on the right correspond to the same when $\hat{b}_1 = 0$. Each of these halves includes half of the decoding regions.

For achieving a positive error exponent for the special bit, the codewords in the two halves should be sufficiently separated from each other as seen in Fig. 3.1. Such separation is necessary to ensure exponentially small probability of landing in the wrong half. However, above theorem indicates that such a thick patch takes too much volume, and is impossible when we have to fill $\doteq e^{nC}$ typical noise balls in this output space.

3.2 Special message

Now consider situations where one particular message (say $M = 1$) out of the $\doteq e^{nC}$ total messages is a special message—it needs a superior error protection. The missed

[4]Appendix A discusses a different but equivalent type of definition and shows why its equivalence to this one. These two types of definitions are equivalent for all the UEP exponents discussed in this paper.

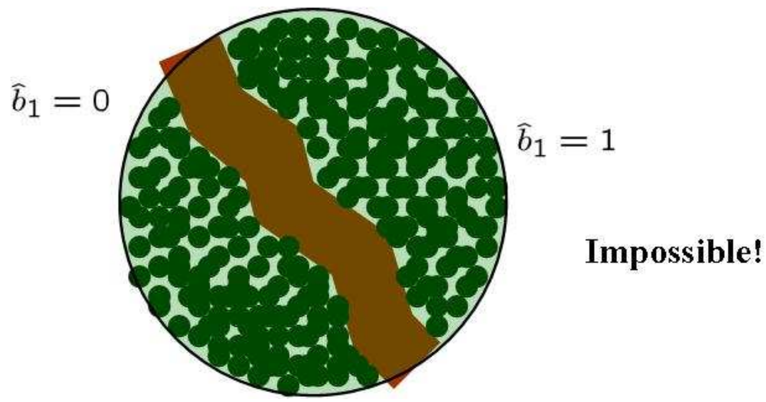


Figure 1: Splitting the output space into 2 distant enough clusters.

detection probability for this 'emergency' message needs to be minimized. The best missed detection exponent $E_{\rm md}$ is defined on similar lines of $E_{\rm b}$.

Definition 3 *For a capacity-achieving sequence $\mathcal{Q}$, missed detection exponent is defined as*

$$E_{md,\mathcal{Q}} \triangleq \liminf_{n\to\infty} \frac{-\log \Pr{}^{(n)}\left[\hat{M}\neq 1 \middle| M=1\right]}{n}. \tag{11}$$

Then define $E_{md} = \sup_{\mathcal{Q}} E_{md,\mathcal{Q}}$.

Compare this with the situation where we aim to protect all the messages uniformly well. If all the messages demand equally good missed detection exponent, then no positive exponent is achievable at capacity. This follows from the earlier discussion about $E(C) = 0$. Below theorem shows the improvement in this exponent if we only demand it for a single message instead of all.

Definition 4 *The parameter $\tilde{\rm C}$ is defined*[5] *as the* red-alert exponent *of a channel.*

$$\tilde{\rm C} \triangleq \max_{i\in\mathcal{X}} D\left(P_Y^*(\cdot)\| \, W_{Y|X}(\cdot|i)\right) \tag{12}$$

We will denote the input letter achieving above maximum by x_r.

Theorem 2 $E_{md} = \tilde{\rm C}$.

Notice the relation between $\tilde{C}$ and C: the arguments to KL divergence are flipped. It is because Karush-Kuhn-Tucker (KKT) conditions for achieving capacity imply the following expression for C [4].

$$C = \max_{i\in\mathcal{X}} D\left(W_{Y|X}(\cdot|i)\middle\| \, P_Y^*(\cdot)\right) \tag{13}$$

If capacity C represents the best possible data-rate over a channel, then red-alert exponent $\tilde{C}$ represents the best possible protection for a message achievable at capacity.

It is worth mentioning here the "very noisy" channel in [2]. In this formulation [11], the KL divergence is symmetric, which implies $D\left(P_Y^*(\cdot)\| \, W_{Y|X}(\cdot|i)\right) \approx D\left(W_{Y|X}(\cdot|i)\middle\| \, P_Y^*(\cdot)\right)$. Hence the red-alert exponent and capacity become essentially equal. For a symmetric

[5]Thanks to Krishnan Eswaran of UC Berkeley for suggesting this name.

channel like BSC, all inputs can be used as x_r. Since the P_Y^* is the uniform distribution for these channels, $\tilde{C} = D\left(P_Y^*(\cdot) \| W_{Y|X}(\cdot|i)\right)$ for any input letter i. This also happens to be the sphere-packing exponent $E_{\mathrm{sp}}(0)$ of this channel [3] at rate 0.

Optimal strategy: Codeword for the special message is a repetition sequence of the input x_r. Its decoding region $\mathcal{G}(1)$ contains every output sequence with empirical distribution (type) is not approximately equal to P_Y^*. For ordinary messages, codewords of a capacity achieving code will be used. Receiver will use a maximum-likelihood (ML) decoding over them for output sequences outside $\mathcal{G}(1)$.

Intuitive Interpretation: Missed detection exponent for the special message corresponds to having a large decoding region $\mathcal{G}(1)$ for the special message. This ensures that when the special message is transmitted, probability of landing outside $\mathcal{G}(1)$ is exponentially small. In a sense, E_{md} indicates how large $\mathcal{G}(1)$ could be made, while still filling $\doteq e^{nC}$ typical noise balls in the remaining space. The red region in Fig. 2 denotes such a large region. Note that the actual decoding region $\mathcal{G}(1)$ is much larger than this illustration, because it consists of all output types except P_Y^*, whereas the ordinary decoding regions only contain the output type P_Y^*.

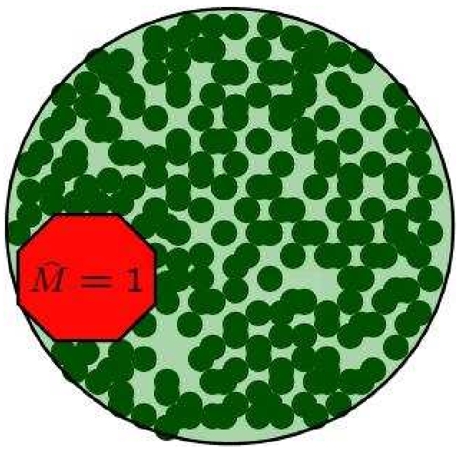


Figure 2: Avoiding missed-detection

Utility of this result is two folds: first, the optimally of such a simple scheme was not obvious before; second, protecting a single special message can be a key building block for many other problems when some feedback is available.

3.3 Many special messages

Now consider that instead of a single special message, exponentially many of the total $\doteq e^{nC}$ messages are special. Let $\mathcal{M}_s^{(n)} \subseteq \mathcal{M}^{(n)}$ denote this set of special messages[6]:

$$\mathcal{M}_s^{(n)} = \{1, 2, \cdots, \lceil e^{nr} \rceil\}$$

The best missed detection exponent, achievable simultaneously for all these special messages, is denoted by $E_{\mathrm{md}}(r)$.

[6]We are assuming that total number of messages $|\mathcal{M}^{(n)}|$ is larger than $\lceil e^{nr} \rceil$ because $\mathcal{M}_s^{(n)}$ cannot be larger than $\mathcal{M}^{(n)}$ itself.

Definition 5 *For a capacity-achieving sequence $\mathcal{Q}$, the missed detection exponent achieved on sequence of subsets $\mathcal{M}_s$ is defined as*

$$E_{md,\mathcal{Q},\mathcal{M}_s} \triangleq \liminf_{n\to\infty} \frac{-\log \max_{i\in\mathcal{M}_s^{(n)}} \Pr{}^{(n)}[\hat{M}\neq i|M=i]}{n}.$$

Then for a given $r < C$, we define $E_{md}(r) \triangleq \sup_{\mathcal{Q},\mathcal{M}_s} E_{md,\mathcal{Q},\mathcal{M}_s}$ where maximization is over $\mathcal{M}_s$ such that $\liminf_{n\to\infty} \frac{\log|\mathcal{M}_s^{(n)}|}{n} \geq r$.

Essentially, $E_{\mathrm{md}}(r)$ is the best value for which missed detection probability of every special message is $\doteq \exp(-nE_{\mathrm{md}}(r))$ or smaller. Note that if the only messages in the code are these $\lceil e^{nr}\rceil$ special messages (instead of $|\mathcal{M}^{(n)}| \doteq e^{nC}$ total messages), their best missed detection exponent equals the classical exponent $E(r)$ discussed earlier.

Theorem 3 $\quad E_{md}(r) = E(r) \quad \forall\ r \in [0, \mathrm{C})$.

Thus we can communicate reliably at capacity and still protect the special messages as if we are only communicating the special messages. Note that the classical error exponent $E(r)$ is yet unknown for the rates below critical rate (except zero rate). Nonetheless, this theorem says that whatever $E(r)$ can be achieved for only $\lceil e^{nr}\rceil$ messages, can still be achieved when there are $\doteq e^{nC}$ additional ordinary messages requiring reliable communication [7]

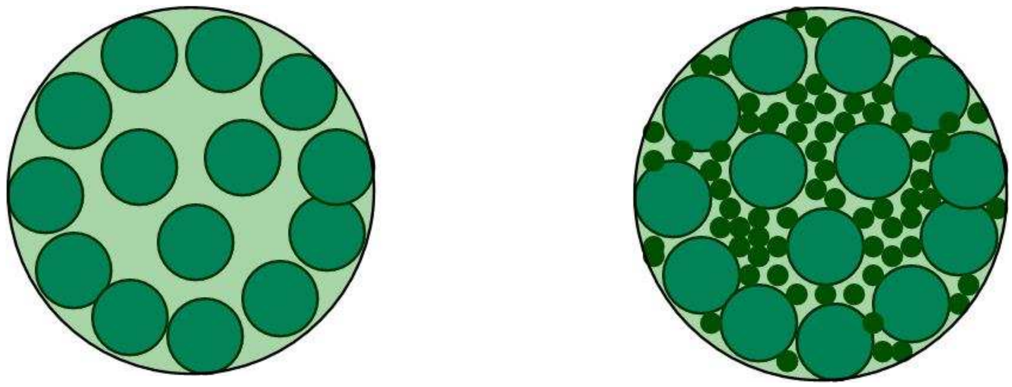
Optimal strategy: Start with an optimal code-book for $\lceil e^{nr}\rceil$ messages which achieves error exponent $E(r)$. These codewords are used for the special messages. Now the ordinary codewords are added using random coding. The ordinary codewords which land close to a special codeword may be discarded without essentially any effect on the rate of communication. At the decoder, a two-stage decoding rule is employed. The first stage decides that some special codeword was sent if at least one of the special codewords is 'close enough' to the received sequence. Otherwise, the first stage decides that an ordinary codeword was sent. Depending on the first stage decision, the second stage ignores all codewords of one kind and applies ML decoding to the rest.

The overall missed detection exponent $E_{\mathrm{md}}(r)$ is bottle-necked by the second stage errors. It is because the first stage error exponent is essentially the sphere-packing exponent $E_{\mathrm{sp}}(r)$, which is never smaller than the second stage error exponent $E(r)$.

Intuitive Interpretation: This means that we can start with a code of $\lceil e^{nr}\rceil$ messages, where the decoding regions are large enough to provide a missed detection exponent of $E(r)$. Consider the balls around each codeword with sphere-packing radius (see Fig. 3(a)). For each message, the probability of going outside its ball decays exponentially with the sphere-packing exponent.

Although, these $\lceil e^{nr}\rceil$ balls fill up most of the output space, there are still some cavities left between them. These small cavities can still accommodate $\doteq e^{nC}$ typical noise balls for the ordinary messages (see Fig. 3(b)), which are much smaller than the original $\lceil e^{nr}\rceil$ balls. This is analogous to filling sand particles in a box full of large boulders. This theorem is like saying that the number of sand particles remains unaffected (exponentially) in spite of the large boulders.

[7]Recently, a closely related result to this theorem, in a paper with somewhat unrelated name,[16], was kindly pointed to us by Pulkit Grover of UC Berkeley.



(a) Exponent optimal code (b) Achieving capacity

Figure 3: "There is always room for capacity"

3.4 Allowing erasures

In some situations, A decoder may be allowed declare an erasure when it is not sure about the transmitted message. These erasure events are not counted as errors and are usually followed by a retransmission using a decision feedback protocol like Hybrid-ARQ. This subsection extends the earlier result for $E_{\text{md}}(r)$ when such erasures are allowed.

In decoding with erasures, in addition to the message set $\mathcal{M}$, the decoder can map the received sequence Y^n to a virtual message called "erasure". Let P_{erasure} denote the average erasure probability of a code.

$$P_{\text{erasure}} = \Pr\left[\hat{M} = \text{erasure}\right]$$

Previously when there was no erasures, errors were not detected. For errors and erasures decoding, erasures will be detected errors, and rest of the errors will be undetected errors. P_e will denote the undetected error probability. Thus average and conditional (undetected) error probability are given by

$$P_e = \Pr\left[\hat{M} \neq M, \hat{M} \neq \text{erasure}\right] \quad \text{and} \quad P_e(i) = \Pr\left[\hat{M} \neq M, \hat{M} \neq \text{erasure} \middle| M = i\right]$$

An infinite sequence $\mathcal{Q}$ of block codes with errors and erasures decoding is called *reliable*, if its average error probability and average erasure probability, both vanish with n.

$$\lim_{n\to\infty} {P_e}^{(n)} = 0 \qquad \text{and} \qquad \lim_{n\to\infty} {P_{\text{erasure}}}^{(n)} = 0 \tag{14}$$

If the erasure probability is small, then average number of retransmissions needed is also small. Hence this condition of vanishingly small ${P_{\text{erasure}}}^{(n)}$ ensures that the effective data-rate of a decision feedback protocol remains unchanged in spite of retransmissions. We again restrict to reliable $\mathcal{Q}$ whose rate $R_{\mathcal{Q}}$ equals C.

For such decision-feedback (df) scenarios, we could now redefine all previous exponents for reliable codes with erasure decoding. For example, on similar lines of $E_{\text{md}}(r)$, let us define $E^{df}_{\text{md},\mathcal{Q}}(r)$: the best missed-detection exponent achievable uniformly over the special messages.

Definition 6 *For a given $r <$ C, let $E^{df}_{md,\mathcal{Q}}(r)$ denote the missed detection exponent of a capacity-achieving sequence $\mathcal{Q}$ which is achieved uniformly over messages in $\mathcal{M}_s^{(n)}$, where $\liminf_{n\to\infty} \frac{\log|\mathcal{M}_s^{(n)}|}{n} = r$.*

$$E^{df}_{md,\mathcal{Q}}(r) = \liminf_{n\to\infty} \frac{-\log \max_{i\in\mathcal{M}_s^{(n)}} \Pr{}^{(n)}\left[\hat{M}\neq i, \hat{M}\neq erasure | M=i\right]}{n}. \tag{15}$$

Then define $E^{df}_{md}(r) \; = \sup_{\mathcal{Q}} E^{df}_{md,\mathcal{Q}}(r)$.

Next theorem shows that compared to $E_{\text{md}}(r)$ in the no-erasure case, allowing erasures increases the missed-detection exponent for r below critical rate[8].

Theorem 4

$$E^{df}_{md}(r) \geq E_{sp}(r) \quad \forall\; r \in [0, \mathrm{C}).$$

Coding strategy is similar to the no-erasure case. We first start with an erasure code in [8] for $\lceil e^{nr} \rceil$ messages. Then add randomly generated ordinary codewords to it. Again a two-stage decoding is performed where the first stage decides between ordinary and special codewords using a threshold distance. If this first stage chooses special codewords, the second stage applies the decoding rule in [8] amongst special codewords. Otherwise, the second stage chooses the ML ordinary codeword.

The overall missed detection exponent $E^{df}_{\text{md}}(r)$ is bottle-necked by the first stage errors. It is because the first-stage error exponent $E_{\text{sp}}(r)$ is smaller than the second stage error exponent $E_{\text{sp}}(r) + C - r$. This is in contrast with the case without erasures.

4 UEP at Capacity: Variable Length Block Codes with Feedback

In the last section, we analyzed bit wise and message wise UEP problems for fixed length block codes (without feedback) operating at capacity. In this section, we will revisit the same problems for variable length block codes with perfect feedback, operating at capacity. Before going into the discussion of the problems, let us recall variable length block codes with feedback briefly.

A variable length block code with feedback, is composed of a coding algorithm and a decoding rule. Decoding rule determines the decoding time and message that will be decoded then. Possible observations of the receiver can be seen as leaves of $|\mathcal{Y}|$-ary tree, as in [20]. In this tree, all nodes at length 1 from the root denote all $|\mathcal{Y}|$ possible outputs at time $t = 1$. All non-leaf nodes among these will split into further $|\mathcal{Y}|$ branches in the next time $t = 2$ and the branching of the non-leaf nodes will continue like this ever after. Each node of depth t in this tree corresponds to particular sequence, y^t, i.e. a history of outputs until time t. The parent of node y^t is its prefix y^{t-1}. Leaves of this tree will be a prefix free source code, because decision to stop for decoding has to be a casual event. In other words the event $\{\tau = t\}$ will be measurable in the σ-field generated by Y^t. In addition we have $\Pr[\tau < \infty] = 1$ thus τ will be Markov stopping time with respect to receivers observation. The coding algorithm on the other hand will assigning an input letter, $X_{t+1}(Y^t; i)$, to each message, $i \in \mathcal{M}$, at each non-leaf node, Y^t, of this tree. The encoder stops transmission of a message when a leaf has been reached and the decoding is complete.

[8]In all the previous problems except this, the provision of erasures with vanishing probability does not improve the achievable exponents. This implies that decision feedback protocols such as Hybrid-ARQ cannot improve E_{b} and E_{md} by allowing erasures.

Codes we will consider will be block codes in the sense that transmission of each message (packet) will start only after the transmission of the previous one has ended. The error probability and rate of the code will simply be given by

$$P_e = \Pr\left[\hat{M} \neq M\right] \qquad \text{and,} \qquad R = \tfrac{\log \mathcal{M}}{E[n]} \tag{16}$$

A more thorough discussion of variable length block codes with feedback can be found in [19], [20].

Earlier discussion in Section 2.2 about different kinds of errors is still valid as is. Now a reliable sequence of variable decoding time codes with feedback, $\mathcal{Q}$, will be any countably infinite collection of codes indexed by integers, such that

$$\lim_{k\to\infty} P_e{}^{(k)} = 0 \tag{17}$$

In the rate and exponent definitions, we will simply replace block-length n by the average decoding time $E[\tau]$. A capacity achieving sequence with feedback will mean a reliable sequence of variable length block codes with feedback whose rate is C

It is worth noting the importance of our assumption that all the entries of the transition probability matrix, $W_{Y|X}$ are positive. For any channel with a $W_{Y|X}$ which has one or more zero probability transitions, it is possible to have error free codes operating at capacity, [19]. Thus all the exponents discussed below will simply be infinity.

4.1 Special bit

Let us consider a capacity achieving sequence $\mathcal{Q}$ whose message sets are of the form $\mathcal{M}^{(k)} = \mathcal{M}_1 \times \mathcal{M}_2^{(k)}$ where $\mathcal{M}_1 = \{0,1\}$. Then the error exponent of the M_1, *i.e.*, the initial bit b_1, is defined as follows.

Definition 7 *For a capacity achieving sequence with feedback, $\mathcal{Q}$, having the message sets $\mathcal{M}^{(k)}$ of the form $\mathcal{M}^{(k)} = \mathcal{M}_1 \times \mathcal{M}_2^{(k)}$ where $\mathcal{M}_1 = \{0,1\}$, the special bit error exponent is defined as*

$$E^f_{b,\mathcal{Q}} = \liminf_{k\to\infty} \frac{-\log \Pr{}^{(n)}\left[\hat{M}_1 \neq M_1\right]}{E\left[\tau^{(k)}\right]} \tag{18}$$

Then $E^f_b = \sup_{\mathcal{Q}} E^f_{b,\mathcal{Q}}$

Theorem 5 $E^f_b = \tilde{\mathrm{C}}$.

Recall that without feedback, the single bit could not achieve any positive error exponent at capacity, $E_{\mathrm{b}} = 0$. The following strategy shows how feedback connects message-wise UEP with bit-wise UEP: strategy for protecting a special message becomes useful for protecting special bits. This special message indicates incorrect decisions at the receiver.

Optimal strategy: We achieve this exponent using the missed detection exponent of $\tilde{C}$ for a special message (see Fig. 4). This special message notifies the receiver when $\tilde{M}_1$ is incorrect. More specifically we will have length $k + \sqrt{k}$ code with feedback and errors and erasures decoding. Transmitter will first transmit M_1 using a short repetition code of length $\sqrt{k}$. If the temporary decision about M_1, $\tilde{M}_1$, is correct after this repetition code, transmitter will transmit M_2 with a capacity achieving code of length k. If $\tilde{M}_1$ is incorrect after the repetition code, transmitter will transmit the symbol x_r for k time units where x_r is the input letter i maximizing the $D\left(W_{Y|X}(\cdot|i)\middle\| P_Y^*(\cdot)\right)$.

If the output sequence in the second phase, $Y_{\sqrt{k}+1}^{\sqrt{k}+k}$, is not close typical with P_Y^*, an erasure will be declare for the block. And the same message is retransmitted by repeating the same strategy afresh. Else receiver will use an ML decoder to chose $\hat{M}_2$ and $\hat{M} = (\tilde{M}_1, \hat{M}_2)$.

The erasure probability is vanishingly small, as a result, the undetected error probability of M_i in fixed length erasure code is approximately equal to the error probability of M_i in this variable length block code. Furthermore $E[\tau]$ will be approximate $k + \sqrt{k}$ despite the retransmissions. A decoding error for M_1 happens only when $\tilde{M}_1 \neq M_1$ and the empirical distribution of the output sequence in the second phase is close to P_Y^*. Note that latter event happens with probability $\doteq e^{-\tilde{C}E[\tau]}$.

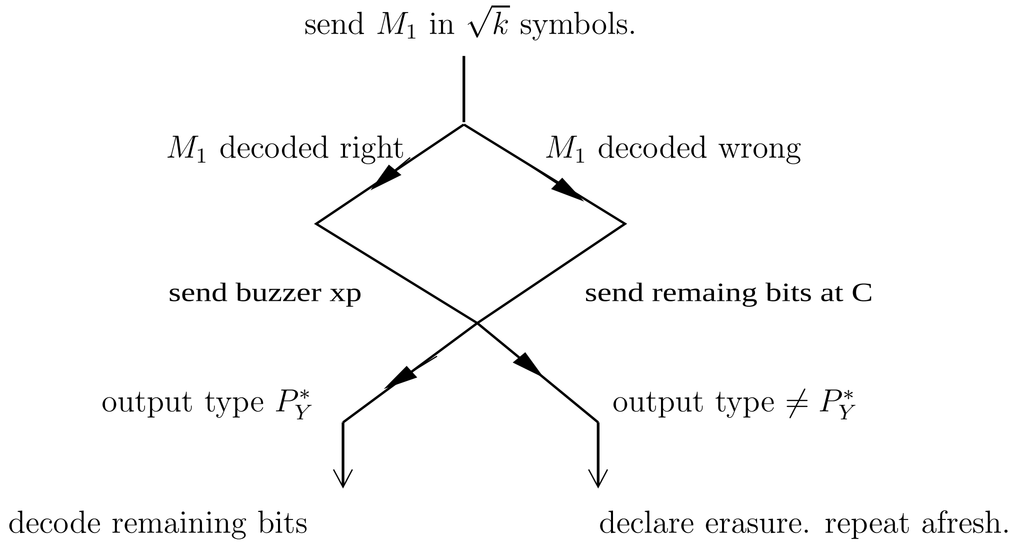


Figure 4: Sending a special bit using a special message

4.2 Many special bits

We now analyze the situation where instead of a single special bit, there are approximately $E[\tau]\, r/\ln 2$ special bits out of the total $E[\tau]\, C/\ln 2$ (approx.) bits. Hence we will again consider the capacity achieving sequences with feedback having message sets of the form $\mathcal{M}^{(k)} = \mathcal{M}_1^{(k)} \times \mathcal{M}_2^{(k)}$. Now unlike the previous subsection where size of $\mathcal{M}_1^{(k)}$ was fixed, we will allow its size to vary with the index of the code. We will restrict ourselves to the cases where $\liminf_{k\to\infty} \frac{\log|\mathcal{M}_1^{(k)}|}{E[\tau^{(k)}]} = r$. This limit will simply give us the rate of the special bits. It is worth noting at this point that even when the rate r of special bits is zero, the number special bits might not be bounded i.e., $\liminf_{k\to\infty} |\mathcal{M}_1^{(k)}|$ might be infinite. The error exponent $E^f_{\text{bits},\mathcal{Q}}$ at a given rate r of special bits will be defined as follows,

Definition 8 *For any capacity achieving sequence with feedback $\mathcal{Q}$ with the message sets $\mathcal{M}^{(k)}$ of the form $\mathcal{M}^{(k)} = \mathcal{M}_1^{(k)} \times \mathcal{M}_2^{(k)}$, where $r_{\mathcal{Q}} = \liminf_{k\to\infty} \frac{\log|\mathcal{M}_1^{(k)}|}{E[\tau^{(k)}]}$. Then*

$$E^f_{bits,\mathcal{Q}} = \liminf_{k\to\infty} \frac{-\log \Pr^{(k)}\left[\hat{M}_1 \neq M_1\right]}{E\left[\tau^{(k)}\right]} \tag{19}$$

Then define $E^{f}_{bits}(r) = \sup_{\mathcal{Q}: r_{\mathcal{Q}} \geq r} E^{f}_{bits,\mathcal{Q}}$

Next theorem shows how this exponent decays linearly with rate r of the special bits.

Theorem 6

$$E^{f}_{bits}(r) = \left(1 - \tfrac{r}{\mathrm{C}}\right) \tilde{\mathrm{C}}$$

Notice that for $r = 0$, the same exponent, $\tilde{C}$, as the single bit case in previous subsection could be achieved, although here the number of bits can be growing to infinity with $E[\tau]$. This linear trade off between rate and reliability reminds us of Burnashev's result [19].

Optimal strategy: Like the single bit case, we will use a fixed length erasure code. We first transmit M_1 using a capacity achieving code of length $\frac{r}{C}k$. If the temporary decision $\tilde{M}_1$ is correct, transmitter will send M_2 with a capacity achieving code of length $(1 - \frac{r}{C})k$. Otherwise transmitter will send the symbol x_r for $(1 - \frac{r}{C})k$ time units. If the output sequence in the second phase is not typical with P^{*}_{Y} an erasure is declared and same strategy will be repeated afresh. Else receiver will use a ML decoder to decide $\hat{M}_2$, and it will decode $\hat{M} = (\tilde{M}_1, \hat{M}_2)$. A decoding error for M_1 happens only when an error happens in the first phase and the output sequence in the second phase is typical with P^{*}_{Y} when the reject codeword is sent. But the probability of the later event is $\doteq e^{k(1-\frac{r}{C})\tilde{C}}$. The factor of $(1 - \frac{r}{C})$ arises because the relative duration of the second phase to the over all communication block. Similar to the single bit case, erasure probability remains vanishingly small.

4.3 Multiple layers of priority

We can generalize this result to the case of multiple levels of priority, where the most important layer contains $E[\tau] r_1/\ln 2$ bits, the second-most important layer contains $E[\tau] r_2/\ln 2$ bits and so on. Hence For an L-layer situation, we will consider the capacity achieving sequences with feedback having message sets of the form $\mathcal{M}^{(k)} = \mathcal{M}^{(k)}_1 \times \mathcal{M}^{(k)}_2 \times \cdots \times \mathcal{M}^{(k)}_L$. We assume that the order of importance of the M_i's will be $M_1 \succ M_2 \succ \cdots \succ M_L$. Hence we require that ${P_e}^{M_1} \leq {P_e}^{M_2} \leq \cdots \leq {P_e}^{M_L}$.

Then for any L-layer capacity achieving sequence with feedback, we define the error exponent of the s^{th} layer as

$$E^{f}_{\mathrm{bits},s,\mathcal{Q}} = \liminf_{k\to\infty} \frac{-\log \Pr{}^{(k)}\left[\hat{M}_s \neq M_s\right]}{E\left[\tau^{(k)}\right]} \tag{20}$$

The achievable error exponent region of L-layered capacity achieving sequences with feedback is the set of all achievable exponent vectors $(E^{f}_{\mathrm{bits},1,\mathcal{Q}}, E^{f}_{\mathrm{bits},2,\mathcal{Q}}, \ldots, E^{f}_{\mathrm{bits},L-1,\mathcal{Q}})$. The following theorem determines that region.

Theorem 7 *Achievable error exponent region of an L-layered capacity achieving sequences with feedback, for rate vector $(r_1, r_2, \ldots, r_{L-1})$ is composed of the vector $(E_1, E_2, \ldots, E_{L-1})$ satisfying following condition,*

$$E_i \leq \left(1 - \frac{\sum_{j=1}^{i} r_j}{\mathrm{C}}\right) \tilde{\mathrm{C}} \qquad \forall i \in \{1, 2, \ldots, (L-1)\} \tag{21}$$

The least important layer cannot achieve any error exponent because we communicating at capacity.

Optimal strategy: Transmitter first sends the most important layer, M_1, using a capacity achieving code of length $\frac{r_1}{C}k$. If it is decoded correctly, then it sends the next layer with a capacity achieving code of length $\frac{r_2}{C}k$. Else it starts sending the input letter x_r for not only $\frac{r_2}{C}k$ time units but also for all remaining $L-2$ phases. Same strategy of will be used for $M_3, M_4, \ldots, M_L$.

Once the block of Y^k symbols is observed. Receivers checks the empirical distribution of the output in all of the phases except the first one. If they are all typical with P_Y^* receiver uses the temporary decisions to decode $\hat{M} = (\tilde{M}_1, \tilde{M}_2, \ldots \tilde{M}_L)$. If one or more of the output sequences are not typical with P_Y^* an erasure is declared for the whole block and transmission starts from scratch.

Thus for each layer i, we can achieve the same exponent as if there were only two kinds of bits (as in Theorem 6):

- Bits in layer i and more important layers $k < i$ are special and
- bits in less important layers than layer i are ordinary.

Hence this could be considered as a successively refinable version of Theorem 6. Figure 5 shows how these simultaneously achievable exponents across layers, which is a successively refinable version of the linear trade off $E_{\text{md}}(r) = \tilde{C}(1 - r/C)$ in Theorem 6.

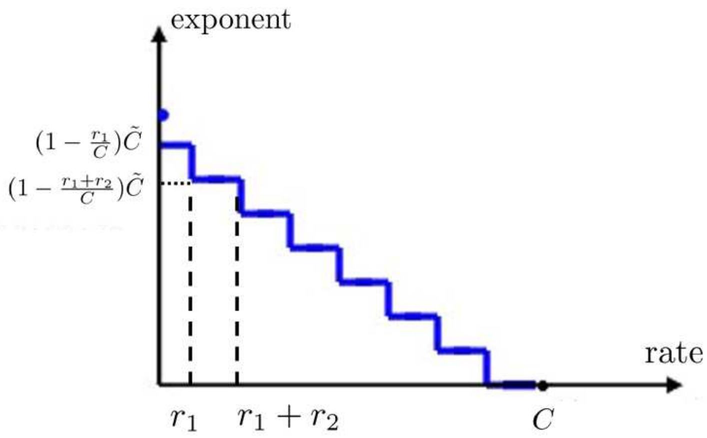


Figure 5: Successive refinability for multiple layers of priority

Note that the most important layer can achieve an exponent close to $\tilde{C}$ if its rate is zero. As we move to across layers with decreasing importance, the achievable error exponent decays gradually.

4.4 A special message

Now consider one particular message, say the first one, which requires small missed-detection probability. Similar to the no-feedback case, define E^f_{md} as its missed-detection exponent at capacity.

Definition 9 *For any capacity achieving sequence with feedback, $\mathcal{Q}$, missed detection exponent is defined as*

$$E^{f}_{md,\mathcal{Q}} \triangleq \liminf_{k\to\infty} \frac{-\log \Pr{}^{(k)}\left[\hat{M}\neq 1 \middle| M=1\right]}{E\left[\tau^{(k)}\right]} \tag{22}$$

Then define $E^{f}_{md} \triangleq \sup_{\mathcal{Q}} E^{f}_{md,\mathcal{Q}}$

It is worth noting at this point that the definition obtained by replacing $\Pr{}^{(k)}\left[\hat{M}\neq 1 \middle| M=1\right]$ by $\min_j \Pr{}^{(k)}\left[\hat{M}\neq j \middle| M=j\right]$ is equivalent to the one given above, since we are taking the supremum over $\mathcal{Q}$ anyway. In short, the message j with smallest conditional error probability could always be relabeled as message 1.

Theorem 8 *Feedback does not improve the missed detection exponent of a single special message:* $E^{f}_{md} = E_{md} = \tilde{\mathrm{C}}$.

If red-alert exponent was defined as the best protection of a special message achievable at capacity, then this result could be thought of as an analog the "feedback does not increase capacity" for the red-alert exponent. Also note that with feedback, E^{f}_{md} for the special message and E^{f}_{b} for the special bit are equal.

4.5 Many special messages

Now let us consider the problem where the first $\lceil e^{E[\tau]r}\rceil$ messages are special. Unlike previous problems, now we will also impose a uniform expected delay constraint as follows.

Definition 10 *For any reliable variable length block code with feedback,*

$$\Gamma \triangleq \frac{\max_{i\in\mathcal{M}} E[\tau|M=i]}{E[\tau]} \tag{23}$$

We will call a reliable sequence with feedback, $\mathcal{Q}$ a uniform delay reliable sequence with feedback if $\lim_{k\to\infty}\Gamma^{(k)}=1$.

This means that the average $E\left[\tau\middle| M=i\right]$ for every message i is essentially equal to $E\left[\tau\right]$ (if not smaller). This uniformity constraint reflects a system requirement for ensuring a robust delay performance, which is invariant of the transmitted message.[9] Let us define the missed-detection exponent $E^{f}_{\mathrm{md}}(r)$ under this uniform delay constraint.

Definition 11 *For any uniform delay capacity achieving sequence with feedback, $\mathcal{Q}$, the missed detection exponent achieved on sequence of subsets $\mathcal{M}_s$ is defined as*

$$E^{f}_{md,\mathcal{Q},\mathcal{M}_s} \triangleq \liminf_{n\to\infty} \frac{-\log \max\limits_{i\in\mathcal{M}_s^{(k)}} \Pr{}^{(k)}\left[\hat{M}\neq i|M=i\right]}{E\left[\tau^{(k)}\right]}.$$

Then for a given $r<C$, we define $E^{f}_{md}(r) \triangleq \sup_{\mathcal{Q},\mathcal{M}_s} E^{f}_{md,\mathcal{Q},\mathcal{M}_s}$ *where maximization is over $\mathcal{M}_s$ such that* $\liminf\limits_{k\to\infty}\dfrac{\log|\mathcal{M}_s^{(k)}|}{E\left[\tau^{(k)}\right]}=r$.

The following theorem shows that the special messages could achieve the minimum of the red-alert exponent and the Burnashev's exponent at rate r.

[9] Optimal exponents in all previous problems remain unchanged irrespective of this uniformity constraint.

Theorem 9

$$E^{f}_{md}(r) \;\triangleq\; \min\left\{\tilde{\mathrm{C}}, (1-\tfrac{r}{\mathrm{C}})\mathrm{D}_{\max}\right\}, \quad \forall\; r < \mathrm{C}.$$

where $\mathrm{D}_{\max} \;\triangleq\; \max_{i,j} D\left(W_{Y|X}(\cdot|i)\middle\| W_{Y|X}(\cdot|j)\right)$.

For r at which $\tilde{C} \leq (1-\frac{r}{C})D_{\max}$, all $\lceil e^{rE[\tau]}\rceil$ special messages achieve the best missed detection exponent $\tilde{C}$ for a single special message. For larger r where $\tilde{C} > (1-\frac{r}{C})D_{\max}$, the special messages achieve the Burnashev's exponent as if the ordinary messages were absent.

The optimal strategy is based on transmitting a special bit first. It again shows how feedback connects bit-wise UEP with message-wise UEP. In the optimal strategy for bit-wise UEP with many bits a special message was used, whereas now in message wise UEP with many messages a special bit is used. The roles of bits and messages, in two optimal strategies are simply swapped between two cases.

Optimal strategy: We combine the strategy for achieving $\tilde{C}$ for a special bit and the Yamamoto-Itoh strategy for achieving Burnashev's exponent [21]. In the first phase, a special bit, b is sent with a repetition code of $\sqrt{k}$ symbols. This is an indicator bit for special messages: it is 1 when a special message is to be sent and 0 otherwise.

If b is decoded incorrectly as $\hat{b}=0$, input letter x_r is sent for the remaining k time unit. If it is decoded correctly as $\hat{b}=0$, then the ordinary message is sent using a codeword from a capacity achieving code. If the output sequence in the second phase is typical with P^{*}_{Y} receiver will use an ML decoder to chose one of the ordinary messages, else an erasure will be declared for length $k+\sqrt{k}$ block.

If $\hat{b}=1$, then a length k two phase code with errors and erasure decoding, like the one used by Yamamoto and Itoh, [21], is used to send the message. In the communication phase a length $\frac{r}{C}$ capacity achieving code is used to send the message, M, if $M \in \mathcal{M}_s$. If $M \notin \mathcal{M}_s$ an arbitrary codeword from the capacity achieving code mentioned, is sent. In the control phase, if $M \in \mathcal{M}_s$ and if it is decode correctly at the end of communication phase, accept letter x_a is sent for $(1-\frac{r}{C})k$ time units, else a reject letter, x_d, will be sent for $(1-\frac{r}{C})k$ time units. If the empirical distribution in the control phase is typical with $W_{Y|X}(\cdot|x_a)$ then special message decoded at the end of the communication phase will be $\hat{M}$, else an erasure will be declared for length $k+\sqrt{k}$ block.

When ever an erasure is declared for the whole block, transmitter and receiver applies above strategy once again from scratch.

5 Avoiding False Alarms

5.1 Block Codes without Feedback

We first consider the no-feedback case where false-alarm of a special message is a critical event, e.g., the "reboot" instruction. Now the false alarm probability $\Pr\left[\hat{M}=1|M\neq 1\right]$ for this message should be minimized. Using Baye's rule and assuming uniformly chosen messages we get,

$$\Pr\left[\hat{M}=1|M\neq 1\right] = \frac{\Pr\left[\hat{M}=1, M\neq 1\right]}{\Pr\left[M\neq 1\right]}$$
$$= \frac{\sum_{j\neq 1}\Pr\left[\hat{M}=1|M=j\right]}{(|\mathcal{M}|-1)} \tag{24}$$

In classical error exponent analysis [2], the error probability for a given message usually means its missed detection probability. However, examples such as the "reboot" message necessitate this notion of false alarm probability.

Definition 12 *For a capacity-achieving sequence $\mathcal{Q}$, false alarm exponent is defined as*

$$E_{fa,\mathcal{Q}} \triangleq \liminf_{n\to\infty} \frac{-\log \Pr{}^{(n)}\left[\hat{M}=1\middle|M\neq 1\right]}{n}.$$

Then E_{fa} is defined as $E_{fa} \triangleq \sup_{\mathcal{Q}} E_{fa,\mathcal{Q}}$. Thus E_{fa} is the best exponential decay rate of false alarm probability with n.

Theorem 10

$$E_{fa}^{l} \leq E_{fa} \leq E_{fa}^{u} \tag{25}$$

The upper and lower bound to the false alarm exponent are given by

$$E_{fa}^{l} \triangleq \max_{i\in\mathcal{X}} \min_{\substack{V_{Y|X}:\\ \sum_j V_{Y|X}(\cdot|j)P_X^*(j)=W_{Y|X}(\cdot|i)}} D\left(V_{Y|X}(\cdot|X)\middle\| W_{Y|X}(\cdot|X)\middle| P_X^*\right) \tag{26}$$
$$E_{fa}^{u} \triangleq \max_{i\in\mathcal{X}} D\left(W_{Y|X}(\cdot|i)\middle\| W_{Y|X}(\cdot|X)\middle| P_X^*\right). \tag{27}$$

The input letters giving the maximum value is denoted as x_{f_l} and x_{f_u} are the maximizers of the corresponding optimization problems.

$$E_{\text{fa}}^{\text{l}} = \min_{V_{Y|X}\in\mathcal{V}_{x_{f_l}}} D\left(V_{Y|X}(\cdot|X)\middle\| W_{Y|X}(\cdot|X)\middle| P_X^*\right) \tag{28}$$
$$E_{\text{fa}}^{\text{u}} = D\left(V_{Y|X}(\cdot|x_{f_u})\middle\| W_{Y|X}(\cdot|X)\middle| P_X^*\right) \tag{29}$$

Strategy to Reach Lower Bound: Codeword for the special message $M=1$ is a repetition sequence of input letter x_{f_l}. Its decoding region $\mathcal{G}(1)$ is the typical 'noise ball' around it, the output sequences whose empirical distribution is approximately equal to $W_{Y|X}(\cdot|x_{f_l})$. For the ordinary messages, we use a capacity achieving code-book where all codewords have the same empirical distribution (approx.) P_X^*. Then for $y^n \notin \mathcal{G}(1)$ receiver uses maximum mutual information decoding amongst ordinary codewords. Alternatively, it could also use ML decoding amongst ordinary codewords when $y^n \notin \mathcal{G}(1)$.

Note the contrast between this strategy for achieving E_{fa}^{l} and the optimal strategy for achieving E_{md}. For achieving E_{md}, output sequences of any type other than the ones close to P_Y^* were assigned to $\mathcal{G}(1)$, whereas for achieving E_{fa} only the output sequences that are close to $W_{Y|X}(\cdot|x_{f_l})$ are in $\mathcal{G}(1)$.

Intuitive Interpretation: A false alarm exponent for the special message corresponds to having the smallest and farthest possible decoding region $\mathcal{G}(1)$ for the special message.

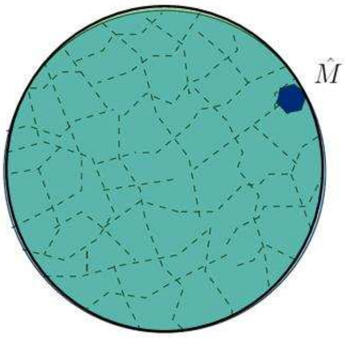


Figure 6: Avoiding false-alarm

This ensures that when some ordinary message is transmitted, probability of landing in $\mathcal{G}(1)$ is exponentially small. We cannot make it too small though, because when the special message is transmitted, the probability of landing outside it should be small too. Hence it should at least contain the typical noise ball around the special codeword. The blue region in Fig. 6 denotes such a region.

Note that E_{fa}^{l} is larger than channel capacity C due to the convexity of KL divergence.

$$
\begin{aligned}
E_{\text{fa}} &= \max_{i \in \mathcal{X}} \min_{V_{Y|X} \in \mathcal{V}_i} D\left(V_{Y|X}(\cdot|X) \middle\| W_{Y|X}(\cdot|X) \middle| P_X^* \right) \\
&> \max_{i \in \mathcal{X}} \min_{V_{Y|X} \in \mathcal{V}_i} D\left(\sum_k P_X^*(k) V_{Y|X}(\cdot|k) \middle\| \sum_{k'} P_X^*(k') W_{Y|X}(\cdot|k') \right) \\
&= \max_{i \in \mathcal{X}} D\left(W_{Y|X}(\cdot|i) \middle\| P_Y^*(\cdot) \right) \\
&= C
\end{aligned}
$$

where P_Y^* denotes the output distribution corresponding to the capacity achieving input distribution P_X^*. As discussed previously, the last equality follows from KKT condition for achieving capacity [4].

Now we can compare our result for a special message with the similar result for classical situation where all messages are treated equally. It turns out that if every message in a capacity-achieving code demands equally good false-alarm exponent, then this uniform exponent cannot be larger than C. This result seems to be directly connected with the problem of identification via channels [18]. We can prove the achievability part of their capacity theorem using an extension of the achievability part of E_{fa}. Perhaps a new converse of their result is also possible using such results. Furthermore we see that reducing the demand of false-alarm exponent to only one message, instead of all, enhances it to at least E_{fa}^{l}.

5.2 Variable Length Block Codes with Feedback

Recall that feedback did not improve the missed-detection exponent for a special message. On the contrary, we will see that the false-alarm exponent for a special message can be improved when feedback is available and variable decoding time is allowed. We again restrict to uniform delay capacity achieving sequences with feedback, i.e. capacity achieving sequences satisfying $\lim_{k \to \infty} \Gamma^{(k)} = 1$.

Definition 13 *For a uniform delay capacity-achieving sequence with feedback, $\mathcal{Q}$, false alarm exponent is defined as*

$$E^f_{fa,\mathcal{Q}} \triangleq \liminf_{k\to\infty} \frac{-\log \Pr{}^{(k)}\left[\hat{M}=1 \middle| M\neq 1\right]}{E\left[\tau^k\right]}.$$

Then E^f_{fa} is defined as $E^f_{fa} \triangleq \sup_{\mathcal{Q}} E^f_{fa,\mathcal{Q}}$.

Theorem 11 $E^f_{fa} = \mathrm{D}_{\max}$.

Note that $D_{\max} > E^{\mathrm{u}}_{\mathrm{fa}} \geq E_{\mathrm{fa}}$, thus feedback strictly improves the false alarm exponent, $E^f_{\mathrm{fa}} > E_{\mathrm{fa}}$.

Optimal strategy: We use the strategy employed in proving Theorem 9 in subsection 4.5. In the first phase, a length $\sqrt{k}$ code is used to convey whether $M = 1$ or not, using a special bit $b = \mathbb{I}_{\{M=1\}}$.

- If $\hat{b} = 0$, we will use a length k block code with errors and erasures decoding similar to the one Yamamoto and Itoh used in [21].
- If $\hat{b} = 1$, we will use a length k code with codewords $(x_a, x_a, \ldots, x_a)$ and $(x_d, x_d, \ldots, x_d)$. Receiver will decode x_a only if output sequence in the second phase is typical with $W_{Y|X}(\cdot|x_a)$. Else receiver will declare an erasure.

Receiver and transmitter starts from scratch if an erasure is declared at the end of second phase.

Note that, this strategy simultaneously achieves the optimal missed-detection exponent $\tilde{C}$ and the optimal false-alarm exponent $D_{\max}$ for this special message.

6 Future directions

This framework for UEP provides a large variety of fundamental problems to be studied. For example, many fundamental limits of bit-wise and message-wise UEP need to be understood for data-rates below capacity. In addition to theoretical understanding, constructing efficient coding mechanisms for achieving these trade-offs is also crucial. One aspect of this task is designing LDPC-like and algebraic codes, which provide extra protection to the high priority information with small computational complexity. Another aspect is addressing the effects of some practical alternatives to classical decoding, *e.g.*, list decoding.

Information networks, such as, two-way channels, broadcast and relay channels provide another rich dimension for future research. Information theoretic understanding of such networks also provides a large set of optimization problems to be studied. Essentially, the interface to the physical layer is no longer bits, but a combination of different levels of protection. The achievable channel resources of reliability and rate need to be efficiently divided amongst these levels, which gives rise to many new resource allocation problems.

7 Block Codes without Feedback: Proofs

Let us first prove why $E_{\mathrm{b}} = 0$ for even a single bit.

7.1 Proof of Theorem 1

This implies that even if we are aiming to protect a single bit with exponential reliability, the data-rate should back-off from capacity.

Proof:

In order to prove that $E_{\text{b}} = 0$, we will first show that any capacity achieving sequence $\mathcal{Q}$ with $E_{\text{b},\mathcal{Q}}$ can be used to construct a capacity achieving sequence, $\mathcal{Q}'$ whose elements are all fixed composition codes with $E_{\text{b},\mathcal{Q}'} = \frac{E_{\text{b},\mathcal{Q}}}{2}$. Than we will show that $E_{\text{b},\mathcal{Q}'} = 0$ for all capacity achieving sequences which only includes fixed composition codes.

Consider a capacity achieving sequence, $\mathcal{Q}$ with message sets $\mathcal{M}^{(n)} = \{0,1\} \times \mathcal{M}_2^{(n)}$. If we group the codewords for the messages of the form $M = (0, M_2)$ according to their empirical distribution at least one of the groups will have more than $\frac{|\mathcal{M}_2^{(n)}|}{(n+1)^{|\mathcal{X}|}}$ messages because the number of different empirical distributions for elements of $\mathcal{X}^n$ is less than $(n+1)^{|\mathcal{X}|}$. Let us spare the first $\frac{|\mathcal{M}_2^{(n)}|}{(n+1)^{|\mathcal{X}|}}$ of the codewords of the this most crowded type of the first half and denote them by $\bar{\mathbf{x}}'_A(\cdot)$ and throw away all the other codewords. We can do the same for the messages of the form $M = (1, M_2)$, let us denote corresponding codewords by $\bar{\mathbf{x}}'_B(\cdot)$.

Now let us consider following length $2n$ code with message set $\mathcal{M}' = \{0,1\} \times \mathcal{M}'_2 \times \mathcal{M}'_3$ where $\mathcal{M}'_2 = \mathcal{M}'_3 = \{1, 2, \dots, \frac{|\mathcal{M}_2|}{(n+1)^{|\mathcal{X}|}}\}$. If $M = (0, M_2, M_3)$ then $\bar{\mathbf{x}}(M) = \bar{\mathbf{x}}'_A(M_2)\bar{\mathbf{x}}'_B(M_3)$. If $M = (1, M_2, M_3)$ then $\bar{\mathbf{x}}(M) = \bar{\mathbf{x}}'_B(M_2)\bar{\mathbf{x}}'_A(M_3)$. Decoder of this new length $2n$ code will decode, y^n and y_{n+1}^{2n} using the decoder of length n code. If the concatenation of the decoded halves corresponds to the codeword of an $i \in \mathcal{M}$ then $\hat{M} = i$. Else it will decode an arbitrary message. One can easily see that the error probability of length $2n$ code is less than the twice of the error probability of the original code. Furthermore bit error probability of the new code is also at most twice of the bit error probability of the original code. Thus using these codes one can obtain a capacity achieving sequence $\mathcal{Q}'$ such that $E_{\text{b},\mathcal{Q}'} = \frac{E_{\text{b},\mathcal{Q}}}{2}$ whose all elements are fixed composition codes.

In the following discussion we will focus on $\mathcal{Q}$'s whose all members are fixed composition codes. Show that $E_{\text{b},\mathcal{Q}} = 0$ for all capacity achieving $\mathcal{Q}$'s whose all elements are fixed composition codes. As a result of the discussion in the above paragraph we will have $E_{\text{b}} = 0$.

We will call empirical distribution of the certain output sequence y^n given the code word $\bar{\mathbf{x}}(i)$ as, the conditional type of y^n given the message i and denote it by $\mathsf{V}(y^n, i)$. Furthermore we will call the set of y^n's whose conditional type with message i is V, as the V-shell of i and denote it by $\mathsf{T}_V(i)$. Similarly we will denote the set of output sequence y^n with the empirical distribution Q_Y, by T_{Q_Y}.

We will denote the empirical distribution of the codewords by $P_X^{(n)}$ and the corresponding output distribution by $P_Y^{(n)}$, i.e.

$$P_Y^{(n)}(\cdot) = \sum_{i \in \mathcal{X}} W_{Y|X}(\cdot|i) P_X^{(n)}(i).$$

In what follows we will simply use P_X and P_Y when ever the value of n is unambiguous from the context. Furthermore $\mathbb{P}_Y^n(\cdot)$ will denote the probability measure on $\mathcal{Y}^n$ such that

$$\mathbb{P}_Y^n(y^n) = \prod_{k=1}^{n} P_Y^{(n)}(y_k).$$

We will denote the set of y^n such that $\hat{M}_1 = 0$ and $\mathsf{V}(y^n, \hat{M}(y^n)) = V$ by $\mathcal{S}^{(n)}_{0,V}$.

$$\mathcal{S}^{(n)}_{0,V} \triangleq \{y^n : \mathsf{V}(y^n, \hat{M}(y^n)) = V \text{ and } \hat{M}(y^n) = (0,k) \text{ for some } k \in \mathcal{M}_2\} \tag{30}$$

In other words, $\mathcal{S}^{(n)}_{0,V}$ is the set of y^n's whose decoded messages V-shell includes y^n itself. It is the set of y^n's such that $y^n \in \mathsf{T}_V\left(\hat{M}(y^n)\right)$. Note that since for each y^n there is a unique $\hat{M}(y^n)$ and for each y^n and message $i \in \mathcal{M}$ there is unique $\mathsf{V}(y^n, i)$; each y^n belongs to a unique $\mathcal{S}^{(n)}_{0,V}$ or a $\mathcal{S}^{(n)}_{1,V}$, i.e. $\mathcal{S}^{(n)}_{0,V}$'s and $\mathcal{S}^{(n)}_{1,V}$'s are disjoint sets.

Let us define the typical neighborhood of $W_{Y|X}$ as $[W]$

$$[W] \triangleq \{V_{Y|X} : |V_{Y|X}(j|i)P^{(n)}_X(i) - W_{Y|X}(j|i)P^{(n)}_X(i)| \leq \sqrt[4]{1/n} \quad \forall i,j\} \tag{31}$$

Let us denote the union of all $\mathcal{S}^{(n)}_{0,V}$'s for typical V's by $\mathcal{S}^{(n)}_0 = \bigcup_{V\in[W]} \mathcal{S}^{(n)}_{0,V}$. We will establish following inequality, inequality (32), later. Let us assume assume for the moment that it holds.

$$\mathbb{P}^n_Y\left(\mathcal{S}^{(n)}_0\right) \geq e^{n(R^{(n)}-C)}\left(\tfrac{1}{2} - \tfrac{|\mathcal{X}||\mathcal{Y}|}{8\sqrt{n}} - P_e\right) \tag{32}$$

As a result of bound given in (32) and the blowing up lemma [4, Ch. 1, Lemma 5.4, page 92], we can conclude that for any capacity achieving sequence $\mathcal{Q}$, there exists a sequence of (ℓ_n, η_n) pairs satisfying $\lim_{n\to\infty} \eta_n = 1$ and $\lim_{n\to\infty} \frac{\ell_n}{n} = 0$ such that

$$\mathbb{P}^n_Y\left(\Gamma^{\ell_n}(\mathcal{S}^{(n)}_0)\right) \geq \eta_n \tag{33}$$

where $\Gamma^{\ell_n}(A)$ is the set of all y^n's for which there exist at least one element $\tilde{y}^n \in A$ which is different at most in ℓ_n entries. Clearly one can repeat the same argument for $\Gamma^{\ell_n}(\mathcal{S}^{(n)}_1)$, thus

$$\mathbb{P}^n_Y\left(\Gamma^{\ell_n}(\mathcal{S}^{(n)}_1)\right) \geq \eta_n \tag{34}$$

Consequently

$$\mathbb{P}^n_Y\left(\Gamma^{\ell_n}(\mathcal{S}^{(n)}_0)\bigcap\Gamma^{\ell_n}(\mathcal{S}^{(n)}_1)\right) \geq 2\eta_n - 1 \tag{35}$$

Note that if $y^n \in \Gamma^{\ell_n}(\mathcal{S}^{(n)}_1)$, then there exist a $\tilde{y}^n \in \mathsf{T}_{P_Y}$ differs from y^n in at most $(\sqrt[4]{n^3} + \ell_n)$ places. Thus we can lower bound its probability by,

$$y^n \in \Gamma^{\ell_n}(\mathcal{S}^{(n)}_1) \Rightarrow \mathbb{P}^n_Y(y^n) \leq e^{-nH(P_Y)-(\sqrt[4]{n^3}+\ell_n)\log\lambda} \tag{36}$$

where $\lambda = \min_{i,j} W_{Y|X}(j|i)$. Thus we have

$$|\Gamma^{\ell_n}(\mathcal{S}^{(n)}_0)\bigcap\Gamma^{\ell_n}(\mathcal{S}^{(n)}_1)| \geq (2\eta_n - 1)e^{nH(P_Y)+(\sqrt[4]{n^3}+\ell_n)\log\lambda} \tag{37}$$

Note that we also know that for all $y^n \in \Gamma^{\ell_n}(\mathcal{S}^{(n)}_0)\bigcap\Gamma^{\ell_n}(\mathcal{S}^{(n)}_1)$ then there exist a $\tilde{y}^n \in \mathsf{T}_W(i)$ for a i of the form $i = (0, M_2)$ which differs from y^n in at most $(\sqrt[4]{n^3} + \ell_n)$ places. Thus we can lower bound the probability of y^n under the hypothesis $M_1 = 0$ as follows

$$\Pr[y^n | M_1 = 0] \geq e^{-n(H(W_{Y|X}|P_X)+R^{(n)})+(\sqrt[4]{n^3}+\ell_n)\log\lambda}$$

Clearly we same holds for $M_1 = 1$ too, thus

$$\Pr[y^n | M_1 = 1] \geq e^{-n(H(W_{Y|X}|P_X)+R^{(n)})+(\sqrt[4]{n^3}+\ell_n)\log\lambda}$$

Consequently

$$
\begin{aligned}
\Pr\left[\hat{M}_1 \neq M_1\right] &\geq \sum_{y^n} \tfrac{1}{2}\min(\Pr\left[y^n \middle| M_1=0\right], \Pr\left[y^n \middle| M_1=0\right]) \\
&\geq \tfrac{1}{2} \sum_{y^n \in \Gamma^{\ell_n}(\mathcal{S}_1^{(n)}) \bigcap \Gamma^{\ell_n}(\mathcal{S}_1^{(n)})} e^{-n(H(W_{Y|X}|P_X)+R^{(n)})+(\sqrt[4]{n^3}+\ell_n)\log\lambda} \\
&\geq \tfrac{1}{2}(2\eta_n-1)e^{nH(P_Y)+(\sqrt[4]{n^3}+\ell_n)\log\lambda}e^{-n(H(W_{Y|X}|P_X)+R^{(n)})+(\sqrt[4]{n^3}+\ell_n)\log\lambda} \\
&= (\eta_n-\tfrac{1}{2})e^{n(I(P_X,W)-R^{(n)})+2(\sqrt[4]{n^3}+\ell_n)\log\lambda} \\
&\geq (\eta_n-\tfrac{1}{2})\tfrac{1}{2}e^{-nR^{(n)}P_e{}^{(n)}+2(\sqrt[4]{n^3}+\ell_n)\log\lambda}
\end{aligned}
$$

Thus

$$
\lim_{n\to\infty} \frac{-\log \Pr{}^{(n)}\left[\hat{M}_1\neq M_1\right]}{n} = 0
$$

Now only thing we are left with is establishing inequality (32). One can write the error probability of the n^{th} code of $\mathcal{Q}$ as

$$
\begin{aligned}
P_e{}^{(n)} &= \sum_{i\in\mathcal{M}^{(n)}} \tfrac{1}{\mathcal{M}} \sum_{y^n\in\mathcal{Y}^n} (1-\mathbb{I}_{\left\{\hat{M}(y^n)=i\right\}}) \Pr\left[y^n \middle| M=1\right] \\
&= \sum_{i\in\mathcal{M}}\sum_{V}\sum_{y^n\in\mathsf{T}_V(i)} (1-\mathbb{I}_{\left\{\hat{M}(y^n)=i\right\}}) e^{-n(D\left(V_{Y|X}(\cdot|X)\middle\|W_{Y|X}(\cdot|X)\middle|P_X\right)+H(V_{Y|X}|P_X)+R^{(n)})} \\
&= \sum_{V} e^{-n(D\left(V_{Y|X}(\cdot|X)\middle\|W_{Y|X}(\cdot|X)\middle|P_X\right)+H(V_{Y|X}|P_X)+R^{(n)})} \sum_{i\in\mathcal{M}}\sum_{y^n\in\mathsf{T}_V(i)} (1-\mathbb{I}_{\left\{\hat{M}(y^n)=i\right\}}) \\
&= \sum_{V} e^{-n(D\left(V_{Y|X}(\cdot|X)\middle\|W_{Y|X}(\cdot|X)\middle|P_X\right)+H(V_{Y|X}|P_X)+R)}(Q_{0,V}+Q_{1,V})
\end{aligned} \tag{38}
$$

where $Q_{k,V} = \sum_{\substack{i=(k,j)\\ j\in\mathcal{M}_2}} \sum_{y^n\in\mathsf{T}_V(i)} (1-\mathbb{I}_{\left\{\hat{M}(y^n)=i\right\}})$ for $k=0,1$.

Note that $Q_{k,V}$ is the sum of the number of the elements in $\mathsf{T}_V(i)$ that are not in the decoding corresponding to message i over the messages which whose $M_1=k$. In a sense it the a measure of the overlap in the V-shells of different codewords. We will use equation (38) to find lower bounds on $\mathbb{P}_Y^n\left(\mathcal{S}_{0,V}^{(n)}\right)$'s.

Let us denote $\sum_i P_X(i)V_{Y|X}(\cdot|i)$ by $(PV)_Y(\cdot)$, then all elements of $\mathcal{S}_{0,V}^{(n)}$ have the same probability under $\mathbb{P}_Y^n(\cdot)$. Thus

$$
\mathbb{P}_Y^n\left(\mathcal{S}_{0,V}^{(n)}\right) = |\mathcal{S}_{0,V}^{(n)}|e^{-n(D((PV)_Y(\cdot)\|P_Y(\cdot))+H((PV)_Y))} \tag{39}
$$

As result of convexity of the KL divergence we get

$$
\begin{aligned}
\mathbb{P}_Y^n\left(\mathcal{S}_{0,V}^{(n)}\right) &\geq |\mathcal{S}_{0,V}^{(n)}|e^{-n(D\left(V_{Y|X}(\cdot|X)\middle\|W_{Y|X}(\cdot|X)\middle|P_X\right)+H((PV)_Y))} \\
&= |\mathcal{S}_{0,V}^{(n)}|e^{-nI(P_X,V_{Y|X})}e^{-n(D\left(V_{Y|X}(\cdot|X)\middle\|W_{Y|X}(\cdot|X)\middle|P_X\right)+H(V_{Y|X}|P_X))} \\
&\geq |\mathcal{S}_{0,V}^{(n)}|e^{-nC}e^{-n(D\left(V_{Y|X}(\cdot|X)\middle\|W_{Y|X}(\cdot|X)\middle|P_X\right)+H(V_{Y|X}|P_X))}
\end{aligned}
$$

Note that

$$
|\mathcal{S}_{0,V}^{(n)}| = |\mathcal{M}_2^{(n)}|\cdot|\mathsf{T}_V(i)| - Q_{V,0} = \tfrac{1}{2}|\mathsf{T}_V(i)|e^{nR} - Q_{V,0} \tag{40}
$$

Since $|\mathcal{M}_2^{(n)}| = \frac{e^{nR^{(n)}}}{2}$ we have,

$$\mathbb{P}_Y^n\left(\mathcal{S}_{0,V}^{(n)}\right) \geq e^{-nC}\left(\tfrac{1}{2}|\mathsf{T}_V(i)|e^{nR^{(n)}} - Q_{V,0}\right)e^{-n\left(D\left(V_{Y|X}(\cdot|X)\|W_{Y|X}(\cdot|X)\middle|P_X\right)+H(V_{Y|X}|P_X)\right)} \tag{41}$$

Since $\mathcal{S}_{0,V}^{(n)}$'s are disjoint sets using the inequality (41) we get,

$$\begin{aligned}\mathbb{P}_Y^n\left(\mathcal{S}_0^{(n)}\right) &\geq \sum_{V\in[W]} e^{-nC}\left(\tfrac{1}{2}|\mathsf{T}_V(i)|e^{nR^{(n)}} - Q_{V,0}\right)e^{-n\left(D\left(V_{Y|X}(\cdot|X)\|W_{Y|X}(\cdot|X)\middle|P_X\right)+H(V_{Y|X}|P_X)\right)}\\ &\geq e^{n(R^{(n)}-C)}\left(\sum_{V\in[W]_{\delta_n}} \tfrac{1}{2}|\mathsf{T}_V(i)|e^{-n\left(D\left(V_{Y|X}(\cdot|X)\|W_{Y|X}(\cdot|X)\middle|P_X\right)+H(V_{Y|X}|P_X)\right)} - P_e\right)\\ &\geq e^{n(R^{(n)}-C)}\left(\tfrac{1}{2} - \tfrac{|\mathcal{X}||\mathcal{Y}|}{8\sqrt{n}} - P_e\right)\end{aligned}$$

where the last inequality simply follows the Chebyshev's inequality. •

7.2 Proof of Theorem 2

7.2.1 Achievability: $E_{\mathbf{md}} \geq \tilde{C}$

Proof:
For each block-length n, the special message is sent with the length-n repetition sequence $\bar{x}^n(1) = (x_r, x_r \cdots, x_r)$ where x_r is the input letter satisfying

$$D\left(P_Y^*(\cdot)\|\, W_{Y|X}(\cdot|x_r)\right) = \max_i D\left(P_Y^*(\cdot)\|\, W_{Y|X}(\cdot|i)\right).$$

The remaining $|\mathcal{M}|-1$ ordinary codewords are generated randomly and independently of each other using capacity achieving input distribution P_X^* i.i.d. over time.

Let us denote the empirical distribution of a particular output sequence y^n by $\mathsf{Q}_y(y^n)$. The receiver will decide that the special message was sent only when the output distribution is not close to P_Y^*, being precise,

$$\mathcal{G}(1) = \{y^n : \|\mathsf{Q}_y(y^n)(i) - P_Y^*(i)\| \geq \sqrt[4]{1/n} \text{ for some } \forall i \in \mathcal{Y}\}$$

Let us denote the set of output sequences close to P_Y^* by $[P_Y^*]$. Since there are at most $(n+1)^{|\mathcal{Y}|}$ different empirical output distribution for elements of $\mathcal{Y}^n$ we get,

$$\Pr{}^{(n)}[y^n \notin \mathcal{G}(1)|\, M=1] \leq (n+1)^{|\mathcal{Y}|}e^{-n\min_{Q_Y\in[P_Y^*]} D\left(Q_Y(\cdot)\|W_{Y|X}(\cdot|x_r)\right)}$$

Thus $\lim_{n\to\infty} \frac{-\log \Pr{}^{(n)}[y^n\notin\mathcal{G}(1)|M=1]}{n} = D\left(P_Y^*(\cdot)\|\, W_{Y|X}(\cdot|x_r)\right) = \tilde{C}$.

Now the only thing we are left to prove is that we can have low enough probability for the remaining messages. For doing that we will first calculate the average error probability of the following random code ensemble. Each entry of the code-book will be generated by using a capacity achieving input distribution P_X^*, independent of all other entries of the codebook. Thus error probability will be same for all $i \neq 1$ in $\mathcal{M}$. Let us calculate the error probability of the message $M=2$.

Assuming that the second message was transmitted, $\Pr[y^n \in \mathcal{G}(1)|\, M=2]$ is vanishingly small. It is because, the output distribution for the random ensemble for ordinary codewords is i.i.d. P_Y^*. Chebyshev's inequality guarantees that probability of the output type being outside a $\sqrt[4]{1/n}$ ball around P_Y^*, i.e. $[P_Y^*]$, is of the order $\sqrt{1/n}$.

Assuming that the second message was transmitted, $\Pr\left[y^n \in \cup_{i>2}\mathcal{G}(i) \middle| M=2\right]$ is vanishingly small due to the standard random coding argument for achieving capacity [1].

Thus for any $P_e > 0$ for all large enough n average error probability of the code ensemble will be smaller than P_e thus we will have at least one code with that P_e. For that code at least half of the codewords will have an error probability less then $2P_e$. •

7.2.2 Converse: $E_{\mathbf{md}} \leq \tilde{C}$

In the section 8.4.2, we will prove that even with feedback and variable decoding time, the missed-detection exponent of a single special message is at most $\tilde{C}$. Thus $E_{\text{md}} \leq \tilde{C}$.

7.3 Proof of Theorem 3

7.3.1 Achievability: $E_{\mathbf{md}} \geq E(r)$

Now we prove the achievability part of missed detection exponents for e^{nr} special messages. The case of a general DMC is considered here, although, some readers may prefer to first read the proof for a BSC in Appendix B. On similar lines of [5], that analysis is based on Hamming distances, which could make it easier to visualize. The general DMC considered ahead essentially replaces those Hamming distances with KL divergences.

Proof:

Special codewords: At any given block length n, we start with a optimum code-book (say $\mathcal{C}_{special}$) for $\lceil e^{nr} \rceil$ messages. Such optimum code-book achieves error exponent $E(r)$ for every message in it.

$$\Pr\left[\hat{M} \neq i | M = i\right] \doteq e^{-nE(r)} \qquad \forall i \in \mathcal{M}_s \equiv \{1, 2, \cdots, \lceil e^{nr} \rceil\}$$

Since there are at most $(n+1)^{|\mathcal{X}|}$ different types, there is at least one type T_{P_X} which has $\frac{\lceil e^{nr} \rceil}{(1+n)^{|\mathcal{Y}|}}$ or more codewords. Throw away all other codewords from $\mathcal{C}_{special}$ and lets call the remaining fixed composition code-book as $\mathcal{C}'_{special}$. Code-book $\mathcal{C}'_{special}$ is used for transmitting the special messages.

As shown in Fig. 3(a), let the ball for special message i be denoted by $\mathcal{B}_i$. These balls need not be disjoint. Now let $\mathcal{B}$ denote the union of these balls of all special messages.

$$\mathcal{B} = \bigcup_{i \in \mathcal{M}_s} \mathcal{B}_i$$

If the output sequence Y^n lies in $\mathcal{B}$, the first stage of the decoder decides a special message was transmitted. The second stage then chooses the ML candidate amongst the messages in $\mathcal{M}_s$.

Let us define $\mathcal{B}_i$ precisely now.

$$\mathcal{B}_i = \{y^n : \mathsf{V}(y^n, i) \in \mathcal{W}(r+\epsilon, P_X)\}$$

where $\mathcal{W}(r+\epsilon, P_X) = \{V_{Y|X} : D\left(V_{Y|X}(\cdot|X) \middle\| W_{Y|X}(\cdot|X) \middle| P_X\right) \leq E_{\text{sp}}(r+\epsilon; P_X)\}$. Recall that the sphere-packing exponent for input type P_X at rate r, $E_{\text{sp}}(r; P_X)$ is given by,

$$E_{\text{sp}}(r; P_X) = \min_{V_{Y|X} : D\left(V_{Y|X}(\cdot|X) \middle\| (PV)_Y(\cdot) | P_X\right) \leq r} D\left(V_{Y|X}(\cdot|X) \middle\| W_{Y|X}(\cdot|X) \middle| P_X\right)$$

The constraint for optimization above is that the mutual information of channel $V_{Y|X}$ under input distribution P_X being less than r, since $(PV)_Y$ denotes the output distribution of channel $V_{Y|X}(\cdot|\cdot)$ for input distribution P_X.

Ordinary codewords: The ordinary codewords will be chosen by random coding, i.i.d. P_X^* over time, where P_X^*. This is the same as Shannon's construction for achieving capacity. The random coding construction provides a simple way to show that in the cavity $\mathcal{B}^c$ (complement of $\mathcal{B}$), we can essentially fit enough typical noise-balls to achieve capacity. This will avoid the complicated task of carefully choosing the ordinary codewords and their decoding regions in the cavity space $\mathcal{B}^c$.

If the output sequence Y^n lies in the cavity $\mathcal{B}^c$, the first stage of the decoder decides an ordinary message was transmitted. The second stage then chooses the ML candidate from ordinary codewords.

Error analysis: First, consider the case that a special codeword $\bar{x}^n(i)$ is transmitted. By Stein's lemma and definition of $\mathcal{B}_i$, the probability of $Y^n \notin \mathcal{B}_i$ has exponent $E_{\text{sp}}(r+\epsilon; P_X)$. Hence the first stage error exponent is at least $E_{\text{sp}}(r+\epsilon; P_X)$.

Assuming correct first stage decoding, the second stage error exponent for special messages equals $E(r)$. Hence the effective error exponent for special messages is

$$\min\{E(r), E_{\text{sp}}(r+\epsilon; P_X)\}$$

Since $E(r)$ is at most the sphere-packing exponent $E_{\text{sp}}(r; P_X)$, [6], choosing arbitrarily small ϵ ensures that missed-detection exponent of each special message equals $E(r)$.

Now consider the situation of a uniformly chosen ordinary codeword being transmitted. We have to make sure the error probability is vanishingly small now. In this case, the output sequence distribution is i.i.d. P_Y^* for the random coding ensemble. The first stage decoding error happens y^n lies in $\bigcup \mathcal{B}_i$. Again by Stein's lemma, this exponent for any particular $\mathcal{B}_i$ equals E_{o}:

$$\begin{aligned}
E_{\text{o}} &= \min_{V_{Y|X} \in \mathcal{W}} D\left(V_{Y|X}(\cdot|X) \middle\| P_Y^*(\cdot) \middle| P_X\right) \\
&= \min_{V_{Y|X} \in \mathcal{W}} D\left(V_{Y|X}(\cdot|X) \middle\| (PV)_Y(\cdot) \middle| P_X\right) + D\left((PV)_Y(\cdot) \| P_Y^*(\cdot)\right) \\
&\geq (r+\epsilon) + D\left((PV)_Y(\cdot) \| P_Y^*(\cdot)\right) \\
&\geq r+\epsilon
\end{aligned}$$

The first step follows since the actual output distribution in this case is i.i.d. P_Y^*. The second step follows by multiplying and dividing by $(PV)_y(\cdot)$ in the log terms in the summed to get $D\left(V_{Y|X}(\cdot|X) \middle\| P_Y^*(\cdot) \middle| P_X\right)$ The third step follows from the definition of $E_{\text{sp}}(r+\epsilon; q_X)$. The forth step is simply the non-negativity of the KL divergence.

Applying union bound over the special messages, the probability of first stage decoding error after sending an ordinary message is at most $\doteq \exp(nr - nE_{\text{o}})$. We have already shown that $E_{\text{o}} \geq r+\epsilon$, which ensures that probability of first stage decoding error for ordinary messages is at most $\doteq e^{-n\epsilon}$ for the random coding ensemble. Recall that for the random coding ensemble, average error probability of the second-stage decoding also vanishes below capacity. To summarize, we have shown these two properties of the random coding ensemble:

1. Error probability of first stage decoding vanishes as $a^{(n)} \doteq \exp(-n\epsilon)$ with n when a uniformly chosen ordinary message is transmitted.

2. Error probability of second stage decoding (say $b^{(n)}$) vanishes with n when a uniformly chosen ordinary message is transmitted.

Since the first error probability is at most $4a^{(n)}$ for some 75% fraction of codes in the random ensemble, and the second error probability is also at most $4b^{(n)}$ for some 75% fraction, there exists a particular code which satisfies both these properties. The overall error probability for ordinary messages is at most $4(a^{(n)} + b^{(n)})$, which vanishes with n. We will use this particular code for the ordinary codewords. This de-randomization completes our construction of a reliable code for ordinary messages to be combined with the code $\mathcal{C}_{special}$ for special messages. •

7.3.2 Converse: $E_{\mathbf{md}} \leq E(r)$

The converse argument for this result is obvious. Removing the ordinary messages from the code can only improve the error probability of the special messages. Even then, (by definition) the best missed detection exponent for the special messages equals $E(r)$.

7.4 Proof of Theorem 4

Let us now address the case with erasures. In this achievability result, the first stage of decoding remains unchanged from the no-erasure case.

Proof:

We use essentially the same strategy as before. Let us start with a good code for $\lceil e^{nr} \rceil$ messages allowing erasure decoding. Forney had shown in [8] that an error exponent equal to $E_{\text{sp}}(r) + C - r$ is achievable while ensuring that erasure probability vanishes with n. We can use that code for these $\lceil e^{nr} \rceil$ codewords. As before, for $Y^n \in \bigcup_i \mathcal{B}_i$, the first stage decides a special codeword was sent. Then the second stage applies the erasure decoding method in [8] amongst the special codewords.

With this decoding rule, when a special message is transmitted, error probability of the two-stage decoding is bottle-necked by the first stage: its error exponent $E_{\text{sp}}(r + \epsilon)$ is smaller than that of the second stage ($E_{\text{sp}}(r) + C - r$). By choosing arbitrarily small ϵ, the special messages can achieve $E_{\text{sp}}(r)$ as their missed-detection exponent.

The ordinary codewords are again generated i.i.d. P_X^*. If the first stage decides in favor of the ordinary messages, choose the ML ordinary codeword. If an ordinary message was transmitted, we can ensure a vanishing error probability as before by repeating earlier arguments for no-erasure case. •

8 Variable Length Block Codes with Feedback: Proofs

In this section we will present a more detailed discussion of bit-wise and message wise UEP for variable length block codes with feedback by proving the Theorems 5, 6, 7, 8 and 9. In the proofs of converse results we will need to discuss issue relating to the conditional entropy of the messages given the observation receiver had. In those discussion we will use the following notation for conditional entropy and conditional mutual information,

$$\mathcal{H}(M|Y^n) = -\sum_{i\in\mathcal{M}} \Pr\left[M=i|\,Y^n\right] \ln \Pr\left[M=i|\,Y^n\right]$$
$$\mathcal{I}\left(M; Y_{n+1}\,|Y^n\right) = \mathcal{H}(M|Y^n) - E\left[\mathcal{H}(M|Y^{n+1})\middle|\,Y^n\right].$$

It is wort noting that this notation is different than widely used one, which includes a further expectation over the the conditioned variable. In the conventional notation "$H(M|Y^n)$" stands for the $E\left[\mathcal{H}(M|Y^n)\right]$ and "$H(M|Y^n=y^n)$" stands for $E\left[\mathcal{H}(M|Y^n)\right]$.

8.1 Proof of Theorem 5

8.1.1 Achievability: $E_{\mathrm{b}}^f \geq \tilde{C}$

As mentioned earlier, this single bit exponent is achieved using the missed detection exponent of a single special message, indicating a decoding error for the special bit. The decoding error for the bit goes unnoticed when this special message is not detected. This shows how feedback connects bit-wise UEP to message-wise UEP in a fundamental manner.

Proof:

We will prove that $E_{\mathrm{b}}^f \geq \tilde{C}$ by constructing a capacity achieving sequence with feedback, $\mathcal{Q}$, such that $E_{\mathrm{b},\mathcal{Q}}^f = \tilde{C}$. For that let $\mathcal{Q}'$ be a capacity achieving sequence such that $E_{\mathrm{md},\mathcal{Q}'} = \tilde{C}$. Note that existence of such a $\mathcal{Q}'$ is guaranteed as a result of Theorem 2. We will first construct a two phase fixed length block code with feedback and erasures. Then using this will obtain the k^{th} element of $\mathcal{Q}$.

In the first phase transmitter will use a length $\lceil\sqrt{k}\rceil$ two message code, for sending M_1. At the end of this phase receiver will have a temporary decision, $\tilde{M}_1$. Note that as a result of [2, Theorem 5.7.1, page 153]

$$\Pr\left[\tilde{M}_1 \neq M_1\right] \leq e^{-\sqrt{k}E_{ex}\left(\frac{\ln 8}{\sqrt{k}}\right)} \tag{42}$$

where $E_{ex}(\cdot)$ stands for the expurgated exponent.

In the second phase transmitter will use the k^{th} member of $\mathcal{Q}'$. The message in the second phase, ϑ, will be determined by M_2 depending on whether M_1 is decoded correctly or not at the end of the first phase.

$$\tilde{M}_1 \neq M_1 \Rightarrow \vartheta = 1$$
$$\tilde{M}_1 = M_1 \text{ and } M_2 = i \Rightarrow \vartheta = i+1 \quad \forall i$$

At the end of the second phase decoder will decode, ϑ using the decoder of $\mathcal{Q}'$, if the decoded message is one, i.e. $\hat{\vartheta} = 1$, it will declare an erasure. Else, $\hat{M}_1 = \tilde{M}_1$ and $\hat{M}_2 = \hat{\vartheta} - 1$.

Note that erasure probability of the two phase fixed length block code is upper bounded as

$$\begin{aligned}\Pr\left[\hat{\vartheta}=1\right] &\leq \Pr\left[\tilde{M}_1 \neq M_1\right] + \Pr\left[\vartheta = 1|\,\vartheta \neq 1\right] \\ &\leq e^{-\sqrt{k}E_{ex}\left(\frac{\ln 4}{\sqrt{k}}\right)} + \tfrac{\mathcal{M}'^{(k)}}{\mathcal{M}'^{(k)}-1} P_e'^{(k)}\end{aligned} \tag{43}$$

where ${P_e}'^{(k)}$ is the error probability of the k^{th} element of $\mathcal{Q}'$.

Similarly we can upper bound the error probabilities of the two phase fixed length block code as follows

$$\Pr\left[\hat{M}_1 \neq M_1 \quad \& \quad \hat{\vartheta} \neq 1\right] \leq {P_e}'^{(k)}(1) \tag{44}$$

$$\Pr\left[\hat{M} \neq M \quad \& \quad \hat{\vartheta} \neq 1\right] \leq \tfrac{\mathcal{M}'^{(k)}}{\mathcal{M}'^{(k)}-1}{P_e}'^{(k)} + {P_e}'^{(k)}(1) \tag{45}$$

where ${P_e}'^{(k)}(1)$ is the conditional error probability of the 1^{st} message in the k^{th} element of $\mathcal{Q}'$.

If there is an erasure the transmitter and the receiver will repeat what they have done again, until they get $\hat{\vartheta} \neq 1$. If we sum all the error probabilities in each step of repetition we get;

$$\Pr\left[\hat{M}_1 \neq M_1\right] \leq \tfrac{\Pr\left[\hat{M}_1 \neq M_1 \quad \& \quad \hat{\vartheta} \neq 1\right]}{1-\Pr\left[\hat{\vartheta}=1\right]} \tag{46}$$

$$\Pr\left[\hat{M} \neq M\right] \leq \tfrac{\Pr\left[\hat{M} \neq M \quad \& \quad \hat{\vartheta} \neq 1\right]}{1-\Pr\left[\hat{\vartheta}=1\right]} \tag{47}$$

Note that expected decoding time of the code will be

$$E\left[\tau\right] \leq \tfrac{k+\lceil\sqrt{k}\rceil}{1-\Pr\left[\hat{\vartheta}=1\right]} \tag{48}$$

Using equations (43), (44), (45), (46), (47) and (48) one can conclude that the resulting sequence of variable length block codes with feedback, $\mathcal{Q}$, is reliable. Furthermore $R_{\mathcal{Q}} = C$ and $E^f_{\mathrm{b},\mathcal{Q}} = \tilde{C}$. •

8.1.2 Converse: $E^f_{\mathrm{b}} \leq \tilde{\boldsymbol{C}}$

We will use a converse result we have not proved yet, namely converse part of Theorem 8, in order to prove that $E^f_{\mathrm{b}} \leq \tilde{C}$.

Proof:

Consider a capacity achieving sequence, $\mathcal{Q}$, with message set sequence $\mathcal{M}^{(k)} = \{0,1\} \times \mathcal{M}_2^{(k)}$. Using $\mathcal{Q}$ we will construct another capacity achieving sequence $\mathcal{Q}'$ with a special message 0, with message set sequence $\mathcal{M}'^{(k)} = \{0\} \cup \mathcal{M}_2^{(k)}$ such that $E^f_{\mathrm{md},\mathcal{Q}'} = E^f_{\mathrm{b},\mathcal{Q}}$. Consequently $E^f_{\mathrm{md}} \geq E^f_{\mathrm{b}}$. As a result of Theorem 8, $E^f_{\mathrm{md}} \leq \tilde{C}$. Thus $E^f_{\mathrm{b}} \leq \tilde{C}$.

Let us denote the message of $\mathcal{Q}$ by M and that of $\mathcal{Q}'$ by ϑ. The k^{th} code of $\mathcal{Q}'$ will be as follow. At time 0 receiver will randomly chose a M_1 for k^{th} element of $\mathcal{Q}$ and send its choice through feedback channel to transmitter. If the message of $\mathcal{Q}'$ is not 0, i.e. $\vartheta \neq 0$ then the transmitter uses the codeword for $M = (M_1, \vartheta)$ to convey ϑ. If $\vartheta = 0$ receiver pick a M_2 with uniform distribution on $\mathcal{M}_2$ and uses the code word for $M = (1-M_1, M_2)$ to convey that $\vartheta = 0$.

Receiver makes decoding using the decoder of $\mathcal{Q}$. If $\hat{M} = (M_1, l)$ it will declare $\vartheta = l$. Else, i.e. if $\hat{M} = (1 - M_1, l)$, it will declare $\vartheta = 0$. One can easily show that expected decoding time and error probability of both of the codes are same. Furthermore error probability of M_1 in $\mathcal{Q}$ is equal to conditional error probability of message 0 in $\mathcal{Q}$ thus, $E^f_{\mathrm{md},\mathcal{Q}'} = E^f_{\mathrm{b},\mathcal{Q}}$. •

8.2 Proof of Theorem 6

8.2.1 Achievability: $E^{f}_{\mathbf{bits}}(r) \geq \left(1 - \frac{r}{\mathbf{C}}\right) \tilde{\boldsymbol{C}}$

Proof:
We will construct the capacity achieving sequence with feedback $\mathcal{Q}$ using a capacity achieving sequence $\mathcal{Q}'$ satisfying $E_{\mathrm{md},\mathcal{Q}'} = \tilde{C}$, as we did in the proof of theorem 5. We know that such a sequence exists, because of Theorem 2.

We will construct a two phase errors and erasures code first. Consider the k^{th} element of the sequence. In the first phase transmitter will use the $\lfloor rk \rfloor^{th}$ element of $\mathcal{Q}'$ to convey M_1. Receiver will have a temporary decision $\tilde{M}_1$. In the second phase transmitter will use the $\lfloor (C-r)k \rfloor^{th}$ element of $\mathcal{Q}'$ to convey M_2 and whether $\tilde{M}_1 = M_1$ or not with a mapping similar to the one we had in the proof of theorem 5.

$$\tilde{M}_1 \neq M_1 \Rightarrow \vartheta = 1$$
$$\tilde{M}_1 = M_1 \text{ and } M_2 = i \Rightarrow \vartheta = i + 1 \quad \forall i$$

Thus $\mathcal{M}_1^{(k)} = \mathcal{M}'^{(\lfloor rk \rfloor)}$ and $\mathcal{M}_2^{(k)} \cup \{|\mathcal{M}_2^{(k)}| + 1\} = \mathcal{M}'^{(\lfloor (C-r)k \rfloor)}$. If we apply a decoding algorithm, like the one one we had in the proof of theorem 5; going through essentially the same analysis with proof of Theorem 5, we can conclude that $\mathcal{Q}$ is a capacity achieving sequence and $E^{f}_{\mathrm{bits},\mathcal{Q}} = \left(1 - \frac{r}{C}\right) \tilde{C}$ and $r_{\mathcal{Q}} = r$. •

8.2.2 Converse: $E^{f}_{\mathbf{bits}}(r) \leq \left(1 - \frac{r}{\mathbf{C}}\right) \tilde{\boldsymbol{C}}$

In establishing the converse we will use a technique that was used previously in [20], together with a lemma we will prove in the converse part Theorem 2.

Proof:
Consider any variable length block code with feedback whose message set $\mathcal{M}$ is of the form $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2$. Let t_δ be the first time instance that an $i \in \mathcal{M}_1$ becomes more likely than $(1-\delta)$ and let $\tau_\delta = t_\delta \wedge \tau$.

For each realization of Y^{τ_δ} we will divide the message set, $\mathcal{M}$ into $|\mathcal{M}_2| + 1$ subsets and index these subsets. For $\ell = 1$ to $|\mathcal{M}_2|$, ℓ^{th} subset will be composed of the message $(\tilde{M}_1(Y^{\tau_\delta}), \ell)$, where $\tilde{M}_1(Y^{\tau_\delta})$ is the most likely message given Y^{τ_δ}. The last subset which we will index by 0, will be composed of the rest of the messages, i.e. all the messages of the form (i, j), where $i \neq \tilde{M}_1(Y^{\tau_\delta})$.

The index of the subset that message, M, is in will be called the axillary-message, ϑ. The decoder for the auxiliary message will decode the index of the decoded message at the decoding time τ, i.e

$$\hat{\vartheta}(Y^\tau) = \ell \Leftrightarrow \hat{M}(Y^\tau) \in \ell(Y^{\tau_\delta}) \tag{49}$$

With these definition we will have;

$$\Pr\left[\hat{M}(Y^\tau) \neq M \middle| Y^{\tau_\delta}\right] \geq \Pr\left[\hat{\vartheta}(Y^\tau) \neq \vartheta(Y^{\tau_\delta}) \middle| Y^{\tau_\delta}\right] \tag{50}$$

$$\Pr\left[\hat{M}_1(Y^\tau) \neq M_1 \middle| Y^{\tau_\delta}\right] \geq \Pr\left[\hat{\vartheta}(Y^\tau) \neq 0 \middle| \vartheta(Y^{\tau_\delta}) = 0, Y^{\tau_\delta}\right] \Pr\left[\vartheta(Y^{\tau_\delta}) = 0 \middle| Y^{\tau_\delta}\right] \tag{51}$$

Now, we apply Lemma 1, which will be proved in the proof of the converse part of

Theorem 8. For the ease of notation we will use following shorthand;

$$P_e^{\vartheta}\{Y^{\tau_\delta}\} = \Pr\left[\hat{\vartheta}(Y^{\tau}) \neq \vartheta(Y^{\tau_\delta})\middle| Y^{\tau_\delta}\right]$$
$$P_e^{\vartheta}\{0, Y^{\tau_\delta}\} = \Pr\left[\hat{\vartheta}(Y^{\tau}) \neq 0\middle| \vartheta(Y^{\tau_\delta}) = 0, Y^{\tau_\delta}\right]$$
$$\xi(Y^{\tau_\delta}) = \Pr\left[\vartheta(Y^{\tau_\delta}) = 0\middle| Y^{\tau_\delta}\right]$$

As a result of Lemma 1, for each realization of Y^{τ_δ} such that $\tau_\delta < \tau$, we have

$$(1-\xi(Y^{\tau_\delta}) - P_e^{\vartheta}\{Y^{\tau_\delta}\}) \ln \tfrac{1}{P_e^{\vartheta}\{0,Y^{\tau_\delta}\}} \leq \ln 2 + E\left[\tau - \tau_\delta\middle| Y^{\tau_\delta}\right] \mathcal{J}\left(\tfrac{\mathcal{H}(\vartheta|Y^{\tau_\delta}) - \ln 2 - P_e^{\vartheta}\{Y^{\tau_\delta}\}\ln|\mathcal{M}_2|}{E[\tau-\tau_\delta|Y^{\tau_\delta}]}\right)$$

If we multiply both sides of the inequality with $\mathbb{I}_{\{\tau_\delta<\tau\}}$, the expression we get holds for all Y^{τ_δ}. Thus

$$\mathbb{I}_{\{\tau_\delta<\tau\}}(1-\xi(Y^{\tau_\delta}) - P_e^{\vartheta}\{Y^{\tau_\delta}\}) \ln \tfrac{1}{P_e^{\vartheta}\{0,Y^{\tau_\delta}\}} \leq \mathbb{I}_{\{\tau_\delta<\tau\}}\left[\ln 2 + E\left[\tau - \tau_\delta\middle| Y^{\tau_\delta}\right] \mathcal{J}\left(\tfrac{\mathcal{H}(\vartheta|Y^{\tau_\delta}) - \ln 2 - P_e^{\vartheta}\{Y^{\tau_\delta}\}\ln|\mathcal{M}_2|}{E[\tau-\tau_\delta|Y^{\tau_\delta}]}\right)\right]$$

Now we will take the expectation of both sides over Y^{τ_δ},

$$\begin{aligned}
L.H.S. &= E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}(1-\xi(Y^{\tau_\delta}) - P_e^{\vartheta}\{Y^{\tau_\delta}\}) \ln \tfrac{1}{P_e^{\vartheta}\{0,Y^{\tau_\delta}\}}\right] \\
&\overset{(a)}{\geq} E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}(1-\xi(Y^{\tau_\delta}) - P_e^{\vartheta}\{Y^{\tau_\delta}\})\right] \ln \tfrac{E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}(1-\xi(Y^{\tau_\delta}) - P_e^{\vartheta}\{Y^{\tau_\delta}\})\right]}{E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}(1-\xi(Y^{\tau_\delta}) - P_e^{\vartheta}\{Y^{\tau_\delta}\})P_e^{\vartheta}\{0,Y^{\tau_\delta}\}\right]} \\
&\overset{(b)}{\geq} (1-\tfrac{P_e}{\lambda\delta}-\delta) \ln \tfrac{E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}(1-\xi(Y^{\tau_\delta}) - P_e^{\vartheta}\{Y^{\tau_\delta}\})\right]}{E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}(1-\xi(Y^{\tau_\delta}) - P_e^{\vartheta}\{Y^{\tau_\delta}\})P_e^{\vartheta}\{0,Y^{\tau_\delta}\}\right]} \\
&\overset{(c)}{\geq} (1-\tfrac{P_e}{\lambda\delta}-\delta) \ln \tfrac{(1-\frac{P_e}{\lambda\delta}-\delta)}{E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}(1-\xi(Y^{\tau_\delta}) - P_e^{\vartheta}\{Y^{\tau_\delta}\})P_e^{\vartheta}\{0,Y^{\tau_\delta}\}\right]}
\end{aligned}$$

where (a) follows log sum inequality. For getting (b) and (c) we use the fact that the log term in (b) is positive and $E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}(1-\xi(Y^{\tau_\delta}) - P_e^{\vartheta}\{Y^{\tau_\delta}\})\right] \geq (1-\frac{P_e}{\lambda\delta}-\delta)$.

Note that

$$E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}(1-\xi(Y^{\tau_\delta}) - P_e^{\vartheta}\{()Y^{\tau_\delta})) P_e^{\vartheta}\{0,Y^{\tau_\delta}\}\right] \leq \tfrac{P_e{}^{M_1}}{\delta\lambda}$$

Thus

$$L.H.S. \geq -\ln 2 - (1-\tfrac{P_e}{\lambda\delta}-\delta)\ln \tfrac{P_e{}^{M_1}}{\lambda\delta} \tag{52}$$

For the right hand side, we use the fact that $\mathcal{J}(R)$ is a decreasing concave function of R to get,

$$\begin{aligned}
R.H.S. &= E\left[\left(\ln 2 + E\left[\tau - \tau_\delta\middle| Y^{\tau_\delta}\right] \mathcal{J}\left(\tfrac{\mathcal{H}(\vartheta|Y^{\tau_\delta}) - \ln 2 - P_e^{\vartheta}\{Y^{\tau_\delta}\}\ln|\mathcal{M}_2|}{E[\tau-\tau_\delta|Y^{\tau_\delta}]}\right)\right)\mathbb{I}_{\{\tau_\delta<\tau\}}\right] \\
&\leq \ln 2 + E\left[E\left[\tau - \tau_\delta\middle| Y^{\tau_\delta}\right] \mathcal{J}\left(\tfrac{\mathcal{H}(\vartheta|Y^{\tau_\delta}) - \ln 2 - P_e^{\vartheta}\{Y^{\tau_\delta}\}\ln|\mathcal{M}_2|}{E[\tau-\tau_\delta|Y^{\tau_\delta}]}\right)\mathbb{I}_{\{\tau_\delta<\tau\}}\right] \\
&\leq \ln 2 + E\left[\tau - \tau_\delta\right] \mathcal{J}\left(E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}\tfrac{\mathcal{H}(\vartheta|Y^{\tau_\delta}) - \ln 2 - P_e^{\vartheta}\{Y^{\tau_\delta}\}\ln|\mathcal{M}_2|}{E[\tau-\tau_\delta]}\right]\right) \\
&\leq \ln 2 + E\left[\tau - \tau_\delta\right] \mathcal{J}\left(\tfrac{E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}\mathcal{H}(\vartheta|Y^{\tau_\delta})\right] - \ln 2 - P_e\ln|\mathcal{M}_2|}{E[\tau-\tau_\delta]}\right)
\end{aligned} \tag{53}$$

Now we will lower bound $E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}\mathcal{H}(\vartheta|Y^{\tau_\delta})\right]$ in terms of $E\left[\mathcal{H}(M|Y^{\tau_\delta})\right]$. Note that for any realization of $\mathcal{Y}^{\tau_\delta}$ we have

$$\begin{aligned}\mathcal{H}(M|Y^{\tau_\delta}) &= \mathcal{H}(\vartheta|Y^{\tau_\delta}) + \Pr\left[M_1 \neq \tilde{M}_1(Y^{\tau_\delta})\middle| Y^{\tau_\delta}\right]\mathcal{H}(M|M_1 \neq \tilde{M}_1(Y^{\tau_\delta}), Y^{\tau_\delta}) \\ &\leq \mathcal{H}(\vartheta|Y^{\tau_\delta}) + \Pr\left[M_1 \neq \tilde{M}_1(Y^{\tau_\delta})\middle| Y^{\tau_\delta}\right]\ln|\mathcal{M}_1||\mathcal{M}_2| \end{aligned} \tag{54}$$

Furthermore for all Y^{τ_δ} such that $\tau > \tau_\delta$, $\Pr\left[\tilde{M}_1(Y^{\tau_\delta} = M_1)\middle| Y^{\tau_\delta}\right] \geq (1-\delta)$. Thus

$$\begin{aligned}E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}\mathcal{H}(\vartheta|Y^{\tau_\delta})\right] &\geq E\left[\mathbb{I}_{\{\tau_\delta<\tau\}}(\mathcal{H}(M|Y^{\tau_\delta}) - \delta\ln|\mathcal{M}_1||\mathcal{M}_2|)\right] \\ &= E\left[(1-\mathbb{I}_{\{\tau_\delta=\tau\}})\mathcal{H}(M|Y^{\tau_\delta})\right] - \delta\ln|\mathcal{M}_1||\mathcal{M}_2| \\ &\geq E\left[\mathcal{H}(M|Y^{\tau_\delta})\right] - \Pr\left[\tau_\delta=\tau\right]\ln|\mathcal{M}_1||\mathcal{M}_2| - \delta\ln|\mathcal{M}_1||\mathcal{M}_2| \end{aligned} \tag{55}$$

Note that $\Pr\left[\tau_\delta=\tau\right] \leq \frac{P_e}{\lambda\delta}$. Inserting this together with equation (55) in the inequality given in (53).

$$\begin{aligned} R.H.S. &\leq \ln 2 + E\left[\tau-\tau_\delta\right]\mathcal{J}\left(\frac{E[\mathcal{H}(M|Y^{\tau_\delta})] - \left(\frac{P_e}{\lambda\delta}+\delta\right)\ln|\mathcal{M}_1||\mathcal{M}_2| - \ln 2 - P_e\ln|\mathcal{M}_2|}{E[\tau-\tau_\delta]}\right) \\ &\leq \ln 2 + E\left[\tau-\tau_\delta\right]\mathcal{J}\left(\frac{\ln|\mathcal{M}_1||\mathcal{M}_2|\left(1-\frac{P_e}{\lambda\delta}-\delta-P_e\right) - \frac{E\left[\mathcal{H}(M|Y^0)-\mathcal{H}(M|Y^{\tau_\delta})\right]}{E[\tau_\delta]}E[\tau_\delta] - \ln 2}{E[\tau-\tau_\delta]}\right)\end{aligned}$$

Now we will use a result that has been previously established, [20],

$$\frac{E\left[\mathcal{H}(M|Y^0)-\mathcal{H}(M|Y^{\tau_\delta})\right]}{E[\tau_\delta]} \leq C \tag{56}$$

Since $\mathcal{J}(\cdot)$ it decreasing in its argument we get,

$$R.H.S. \leq \ln 2 + E\left[\tau-\tau_\delta\right]\mathcal{J}\left(\frac{\ln|\mathcal{M}_1||\mathcal{M}_2|\left(1-\frac{P_e}{\lambda\delta}-\delta-P_e\right) - E[\tau_\delta]C - \ln 2}{E[\tau-\tau_\delta]}\right) \tag{57}$$

Note that $\forall a>0, b>0, C>0$,

$$\frac{d}{dx}(b-x)\mathcal{J}\left(\tfrac{a-Cx}{b-x}\right)|_{x=x_0} = -\mathcal{J}\left(\tfrac{a-Cx_0}{b-x_0}\right) - \left(C - \tfrac{a-Cx_0}{b-x_0}\right)\tfrac{d}{dx}\mathcal{J}(x)\Big|_{x=\frac{a-Cx_0}{b-x_0}} \leq 0$$

Thus using replacing $E[\tau_\delta]$ with a term lower than $E[\tau_\delta]$ itself will in crease the value of the expression given in equation (57), thus using (56) and the fact that $\mathcal{H}(M_1|Y^{\tau_\delta}) \leq \ln 2 + (\delta + \frac{P_e}{\lambda\delta})\ln|\mathcal{M}_1|$ in equation (57) and we get,

$$R.H.S. \leq \ln 2 + \left(E[\tau] - (1-\delta\tfrac{P_e}{\lambda\delta})\tfrac{\ln|\mathcal{M}_1|}{C}\right)\mathcal{J}\left(\frac{\left(1-\frac{P_e}{\lambda\delta}-\delta-P_e\right)\ln|\mathcal{M}_2| - P_e\ln|\mathcal{M}_1| - 2\ln 2}{E[\tau]-(1-\delta-\frac{P_e}{\lambda\delta})\frac{\ln|\mathcal{M}_1|}{C}}\right) \tag{58}$$

Using this together with (52) and choosing $\delta = \sqrt{P_e}$ we can prove that, $E^f_{\text{bits},\mathcal{Q}} \leq \left(1-\frac{r_\mathcal{Q}}{C}\right)\mathcal{J}(C)$. Since $\mathcal{J}(C) = \tilde{C}$ this will also imply $E^f_{\text{bits}}(r) \leq \left(1-\frac{r}{C}\right)\tilde{C}$. •

8.3 Proof of of Theorem 7

8.3.1 Achievability

Proof:
Proof will be almost identical to achievability proof for Theorem 6. Choose a capacity achieving sequence $\mathcal{Q}'$ such that $E^f_{\mathrm{b},\mathcal{Q}'} = \tilde{C}$. The capacity achieving sequence with feedback, $\mathcal{Q}$ will use L elements of $\mathcal{Q}'$ as follows.

For the k^{th} element of code $\mathcal{Q}$, transmitter will use the $\lfloor k \cdot r_1 \rfloor^{th}$ element of $\mathcal{Q}'$ to send the first part of the message M_1. In the remaining phases, $l \geq 2$ transmitter will use $\lfloor k \cdot r_l \rfloor^{th}$ element of $\mathcal{Q}'$. The only difference will be that the special message of the code will be allocated to the error event in previous phases.

$$(\tilde{M}_1, \ldots, \tilde{M}_{(l-1)}) \neq (M_1, \ldots, M_{(l-1)}) \Rightarrow \vartheta_l = 1 \qquad \forall l$$
$$(\tilde{M}_1, \ldots, \tilde{M}_{(l-1)}) = (M_1, \ldots, M_{(l-1)}) \Rightarrow \vartheta_l = M_l \qquad \forall l$$

Thus $\mathcal{M}_1^{(k)} = \mathcal{M}'^{(\lfloor rk \rfloor)}$ and for all $l \geq 1$ $\mathcal{M}_l^{(k)} \cup \{|\mathcal{M}_l^{(k)}| + 1\} = \mathcal{M}'^{(\lfloor (r_l)k \rfloor)}$. If for all $l \in \{2, \ldots, l\}$, $\hat{\vartheta}_l \neq 1$, receiver decode all parts of the information. Else it will declare an erasure. We skip the error analysis because it is essentially identical to that of Theorem 6. •

8.3.2 Converse

Proof:
We will prove the converse of Theorem 7 by contradiction. Evidently

$$\max\{P_e{}^{M_1}, P_e{}^{M_2}, \ldots, P_e{}^{M_j}\} \leq P_e{}^{M_1, M_2, \ldots, M_j} \leq P_e{}^{M_1} + P_e{}^{M_2} + \cdots + P_e{}^{M_j} \tag{59}$$

Thus if there exists a scheme that can reach an error exponent vector outside the region given in Theorem 7, there will be at least one $E_i \geq (1 - \frac{\sum_{j=1}^{i} r_j}{C})\tilde{C}$. Then we can have two super messages as follows,

$$\vartheta_1 = (M_1, M_2, \ldots, M_i) \quad \text{ and } \vartheta_2 = (M_{i+1}, M_{i+2}, \ldots, M_l) \tag{60}$$

Recall that $P_e{}^{M_1} \leq P_e{}^{M_2} \leq \cdots \leq P_e{}^{M_l}$. Thus this new code $\mathcal{Q}'$ will be a capacity achieving code, which has rate $r_{\mathcal{Q}'}$ and $E^f_{\mathrm{bits},\mathcal{Q}'} > E^f_{\mathrm{bits}}(r_{\mathcal{Q}'})$. This is contradicting with the Theorem 6 we have already proved. Thus all the achievable error exponent regions should lie in the region given in Theorem 7. •

8.4 Proof of of Theorem 8

8.4.1 Achievability: $E^f_{\mathbf{md}} \geq \tilde{C}$

Note that any fixed length block code without feedback, is also variable-length block code with feedback, thus $E^f_{\mathrm{md}} \geq E_{\mathrm{md}}$. Using the capacity achieving sequence we have used in the achievability proof of Theorem 2, we get $E^f_{\mathrm{md}} \geq \tilde{C}$.

8.4.2 Converse: $E^f_{\mathbf{md}} \leq \tilde{C}$

Now we will prove that even with feedback and variable decoding time, the best missed detection exponent of single special message is less then or equal to $\tilde{C}$, i.e. $E^f_{\text{md}} \leq \tilde{C}$. Since set of capacity achieving sequences is a subset of capacity achieving sequences with feedback this will also imply that $E_{\text{md}} \leq \tilde{C}$.

Instead of directly proving the converse part of Theorem 8 we will first prove the following lemma.

Lemma 1 *For any variable length block code with feedback with $|\mathcal{M}|$ messages with initial entropy $\mathcal{H}(M|Y^0)$ and with average error probability P_e, the conditional error probability of each message is lower bounded as follows,*

$$\Pr\left[\hat{M} \neq i \middle| M = i\right] \geq e^{-\frac{1}{1-\Pr[M=i]-P_e}\left(\mathcal{J}\left(\frac{\mathcal{H}(M|Y^0)-h(P_e)-P_e\ln(|\mathcal{M}|-1)}{E[\tau]}\right)E[\tau]+\ln 2\right)} \qquad \forall i \quad (61)$$

where $\mathcal{J}(R)$ is given by

$$\mathcal{J}(R) = \max_{\substack{\alpha_{X'}, P_X^1, P_X^2, \ldots, P_X^{|\mathcal{X}|}: \\ \sum_{i\in\mathcal{X}} \alpha_{X'}(i) I(P_X^i, W_{Y|X}) \geq R}} \sum_{i\in\mathcal{X}} D\left((P^iW)_Y(\cdot) \middle\| W(\cdot|i)\right)$$

It is worthwhile remembering the notation we established previously that

$$(P^iW)_Y(\cdot) = \sum_{j\in\mathcal{X}} P_X^i(j) W_{Y|X}(\cdot|j) \quad \text{and} \quad I(P_X^i, W_{Y|X}) = \sum_{j\in\mathcal{X}, k\in\mathcal{Y}} P_X^i(j) W_{Y|X}(k|i) \log \frac{W_{Y|X}(k|i)}{(P^iW)_Y(k)}$$

Note that $\mathcal{J}(R)$ is concave and strictly decreasing function of R defined for $R \leq C$. Considering the achievability proof of Theorem 2 and the following converse proof one can easily see that indeed $\mathcal{J}(R)$ is the best exponent a message can get for a reliable sequence whose rate is R. However we will only make that claim for $R = C$, and use the fact that $\mathcal{J}(C) = \tilde{C}$.

Proof (of Lemma 1):

Now for upper bounding error probability of the special message let us consider the following stochastic sequence,

$$S_n = \ln \frac{\Pr[Y^n]}{\Pr[Y^n|M=i]} - \sum_{t=1}^{n} E\left[\ln \frac{\Pr\left[Y_t|Y^{t-1}\right]}{\Pr[Y_t|M=i, Y^{t-1}]} \middle| Y^{t-1}\right] \qquad (62)$$

Note that clearly $E\left[S_{n+1} \middle| Y^n\right] = S_n$ and since $\min W_{i,j} = \lambda$ we have $E\left[|S_{n+1} - S_n| \middle| Y^n\right] \leq 2\ln\frac{1}{\lambda}$. Thus S_n is a martingale, furthermore since $E[\tau] < \infty$ we can use [22, Theorem 2 p 487], to get

$$E[S_\tau] = S_0 = 0. \qquad (63)$$

Thus

$$E\left[\ln \frac{\Pr[Y^\tau]}{\Pr[Y^\tau|M=1]}\right] = E\left[\sum_{t=1}^{\tau} E\left[\ln \frac{\Pr\left[Y_t|Y^{t-1}\right]}{\Pr[Y_t|M=1, Y^{t-1}]} \middle| Y^{t-1}\right]\right]$$

$$\leq E\left[\sum_{t=1}^{\tau} \mathcal{J}\left(\mathcal{I}\left(M; Y_t \middle| Y^{t-1}\right)\right)\right] \qquad (64)$$

Let $\mathcal{G}(i)$ be decoding region for $\hat{M} = i$ i.e. $\mathcal{G}(i) = \{y^\tau : \hat{M}(y^\tau) = i\}$. Then as a result of data processing inequality for KL divergence we have

$$\begin{aligned} E\left[\ln \tfrac{\Pr[Y^\tau]}{\Pr[Y^\tau|M=i]}\right] &\geq \Pr\left[\mathcal{G}(i)\right] \ln \tfrac{\Pr[\mathcal{G}(i)]}{\Pr[\mathcal{G}(i)|M=i]} + \Pr\left[\overline{\mathcal{G}(i)}\right] \ln \tfrac{\Pr\left[\overline{\mathcal{G}(i)}\right]}{\Pr\left[\overline{\mathcal{G}(i)}\middle|M=i\right]} \\ &\geq -h(\Pr\left[\mathcal{G}(i)\right]) + \Pr\left[\overline{\mathcal{G}(i)}\right] \ln \tfrac{1}{\Pr\left[\overline{\mathcal{G}(i)}\middle|M=i\right]} \end{aligned}$$

Using $h(\Pr\left[\mathcal{G}(i)\right]) \leq \ln 2$ and equation (64) we get

$$\begin{aligned} \Pr\left[\overline{\mathcal{G}(i)}\right] \ln \tfrac{1}{\Pr\left[\overline{\mathcal{G}(i)}\middle|M=i\right]} &\leq \ln 2 + E\left[\ln \tfrac{\Pr[Y^\tau]}{\Pr[Y^\tau|M=i]}\right] \\ &\leq \ln 2 + E\left[\sum_{t=i}^{\tau} \mathcal{J}\left(\mathcal{I}\left(M; Y_t \left|Y^{t-1}\right.\right)\right)\right] \end{aligned} \tag{65}$$

Note that

$$\begin{aligned} \Pr\left[\overline{\mathcal{G}(i)}\right] &= \Pr\left[\overline{\mathcal{G}(i)}\middle| M = i\right] \Pr\left[M = i\right] + \Pr\left[\overline{\mathcal{G}(i)}\middle| M \neq i\right] \Pr\left[M \neq i\right] \\ &\geq (1 - P_e - \Pr\left[M = i\right]) \end{aligned} \tag{66}$$

Thus using the concavity of $\mathcal{J}(\cdot)$ together with equations (65) and (66) we get

$$\Pr\left[\hat{M} \neq i\middle| M = i\right] \geq e^{-\frac{1}{1-P_e-\Pr[M=i]}\left(\mathcal{J}\left(\frac{E\left[\sum_{t=i}^{\tau}\mathcal{I}\left(M;Y_t\left|Y^{t-1}\right.\right)\right]}{E[\tau]}\right)E[\tau]+\ln 2\right)} \tag{67}$$

Since $\mathcal{J}(R)$ is decreasing in R only thing we are left to show is that

$$E\left[\sum_{t=i}^{\tau} \mathcal{I}\left(M; Y_t \left|Y^{t-1}\right.\right)\right] \geq \mathcal{H}(M|Y^0) - h(P_e) - P_e \ln(|\mathcal{M}| - 1 \tag{68}$$

For that consider the stochastic sequence,

$$V_n = \mathcal{H}(M|Y^n) + \sum_{t=1}^{n} \mathcal{I}\left(M; Y_t \left|Y^{t-1}\right.\right).$$

Clearly $E\left[V_{n+1}\middle| Y^n\right] = V_n$ and $E\left[|V_n|\right] < \infty$, thus $\{V_n\}$ is a martingale. Furthermore $E\left[|V_{n+1} - V_n|\middle| Y^n\right] \leq K$ and $E\left[\tau\right] < \infty$ thus using a version of Doob's optional stopping theorem, [22, Theorem 2 p 487], we get.

$$\begin{aligned} V_0 &= E\left[V_\tau\right] \\ &= E\left[\mathcal{H}(M|Y^\tau)\right] + E\left[\sum_{t=1}^{\tau} \mathcal{I}\left(M; Y_t \left|Y^{t-1}\right.\right)\right] \end{aligned} \tag{69}$$

One can write Fano's inequality for every Y^τ as follows,

$$\mathcal{H}(M|Y^\tau) \leq h\left(\Pr\left[\hat{M}(Y^\tau) \neq M\middle| Y^\tau\right]\right) + \Pr\left[\hat{M}(Y^\tau) \neq M\middle| Y^\tau\right] \ln(|\mathcal{M}| - 1)$$

Consequently

$$E\left[\mathcal{H}(M|Y^\tau)\right] \leq E\left[h\left(\Pr\left[\hat{M}(Y^\tau) \neq M\middle| Y^\tau\right]\right)\right] + E\left[\Pr\left[\hat{M}(Y^\tau) \neq M\middle| Y^\tau\right]\right] \ln(|\mathcal{M}| - 1)$$

Using the convexity of binary entropy,

$$E\left[\mathcal{H}(M|Y^{\tau})\right] \leq h(P_e) + P_e \ln(|\mathcal{M}| - 1) \tag{70}$$

Using equation (69) together with equation (70) we get the desired condition given in the equation (68). •

Now we are ready to prove the converse part of the Theorem 8.

Proof (of Converse part of Theorem 8):

In order to prove $E_{\rm md}^{f} \geq \tilde{C}$, first note that for capacity achieving sequences we will consider $\Pr\left[M = i\right] = \frac{1}{|\mathcal{M}^{(k)}|}$. Thus

$$-\frac{\ln({P_e}^M(i))^{(k)}}{E\left[\tau^{(k)}\right]} \leq \frac{1}{1-{P_e}^{(k)}-\frac{1}{|\mathcal{M}^{(k)}|}}\left(\mathcal{J}\left(\frac{\ln|\mathcal{M}^{(k)}|-h(P_e(k))-{P_e}^{(k)}\ln(|\mathcal{M}(k)|-1)}{E\left[\tau^{(k)}\right]}\right)+\frac{\ln 2}{E\left[\tau^{(k)}\right]}\right) \tag{71}$$

Thus for any capacity achieving sequence with feedback

$$\lim_{k\to\infty} -\frac{\ln({P_e}^M(i))^{(k)}}{E\left[\tau^{(k)}\right]} \leq \mathcal{J}\left(C\right) = \tilde{C} \tag{72}$$

•

8.5 Proof of of Theorem 9

In this subsection we will show how the strategy for sending a special bit can be combined with the Yamamoto-Itoh strategy when many special messages demand a missed-detection exponent. However unlike previous results about capacity achieving sequences, Theorems 5, 6, 7, 8, we will have and additional uniform delay assumption.

We will restrict our self to uniform delay capacity achieving sequences.[10] Clearly capacity achieving sequences in general need not to be uniform delay. Indeed many messages, $i \in \mathcal{M}$, can get an expected delay, $E\left[\tau|\, M = i\right]$ much larger than the average delay, $E\left[\tau\right]$. This in return can decrease the error probability of these messages. The potential drawback of such codes, is that their average delay is sensitive to assumption of messages being chosen according to a uniform probability distribution. Expected decoding time, $E\left[\tau\right]$, can increase a lot if the code is used in a system in which the messages are not chosen uniformly.

It is worth emphasizing that all previously discussed exponents (single message exponent $E_{\rm md}^{f}$, single bit exponent $E_{\rm b}^{f}$, many bits exponent $E_{\rm b}^{f}(r)$ and achievable multi-layer exponent regions) remain unchanged whether or not this uniform delay constraint is imposed. Thus the flexibility to provide different expected delays to different messages does not improve these exponents.

However, this is not true for message-wise UEP. Removing the uniform delay constraint can considerably enhance the protection of special messages at rate higher than $(1 - \frac{\tilde{C}}{D_{\max}})C$. Indeed one can make the exponent of all special codes, $\tilde{C}$. The flexibility of providing more resources (decoding delay) to special messages achieves this enhancement. However, we will not discuss those cases in this article and stick to uniform delay codes.

[10]Recall that for any reliable variable length block code with feedback $\Gamma = \frac{\max_{i\in\mathcal{M}} E[\tau|M=i]}{E[\tau]}$ and uniform reliable codes are the ones that satisfy $\lim_{k\to\infty} \Gamma_{\mathcal{Q}}^{(k)} = 1$.

8.5.1 Achievability: $E^f_{\mathbf{md}}(r) \geq \min\{\tilde{\boldsymbol{C}}, (1-\frac{r}{\boldsymbol{C}})\boldsymbol{D}_{\max}\}$

The optimal scheme here reverses the trick for achieving E^f_{b}: now first a special bit tells to the receiver whether the message being transmitted is special one or not. After the decoding of this bit the message itself is transmitted. This further emphasizes how feedback connects bit-wise and message-wise UEP.

Proof:

Like all the previous achievability results, we will construct a capacity achieving sequence, $\mathcal{Q}$, with the desired asymptotic behavior. The k^{th} member $\mathcal{Q}$. A multi phase fixed length errors and erasures code will be used. In the first phase transmitter will use a length $\lfloor\sqrt{k}\rfloor$, zero-rate non-feedback code with two codewords, to tell whether $M \in \mathcal{M}_s^{(k)}$ or not. For the ease of notation let $b = \mathbb{I}_{\{M\in\mathcal{M}_s^{(k)}\}}$. As a result of [2, Theorem 5.7.1, page 153]

$$\Pr\left[\hat{b} \neq 1 \middle| b = 1\right] = \Pr\left[\hat{b} \neq 0 \middle| b = 0\right] \leq e^{-\sqrt{k}E_{ex}\left(\frac{\ln 8}{\sqrt{k}}\right)} \tag{73}$$

where $E_{ex}(\cdot)$ stands for the expurgated exponent.

In the second phase one of two codes will be used depending on $\hat{b}$.

- If $\hat{b} = 0$, in the second phase, transmitter will use the k^{th} member of a capacity achieving sequence, $\mathcal{Q}'$ such that $E_{\mathrm{b},\mathcal{Q}'} = \tilde{C}$. Note that such a sequence exists as a result of Theorem 2. The message, ϑ of the $\mathcal{Q}'$ will be decided according to the following mapping

$$M \in \mathcal{M}_s \Rightarrow \vartheta = 1$$
$$M \notin \mathcal{M}_s \Rightarrow \vartheta = M - |\mathcal{M}_s| + 1$$

  At the end of the second phase, receiver will decode ϑ. If $\hat{\vartheta} = 1$, then it will declare an erasure. If $\vartheta \neq 1$, then it will decode, the message, M, as $\hat{M} = \hat{\vartheta} + |\mathcal{M}_s| - 1$.

- If $\hat{b} = 1$, transmitter will use a two phase code with errors and erasures in the second phase, like the one described by Yamamoto and Itoh in [21]. The two phases of this code are called communication and control phases, respectively. First $|\mathcal{M}_s|$ codewords of the code will be used to convey which special message is transmitted if $M \in \mathcal{M}_s$. The last codeword will be used to tell that $M \notin \mathcal{M}_s$, i.e. the message in the second phase, ϑ, will be

$$M \in \mathcal{M}_s \Rightarrow \vartheta = M$$
$$M \notin \mathcal{M}_s \Rightarrow \vartheta = |\mathcal{M}_s| + 1$$

  In the communication phase a length $\lceil\frac{r}{C}k\rceil$ block code with $\lfloor e^{r_k k}\rfloor = |\mathcal{M}_s| + 1$ messages is used to convey ϑ. At the end of the communication phase receiver will have a temporary decision $\tilde{\vartheta}$. As a result of [2, Corollary , page 140]

$$\Pr\left[\tilde{\vartheta} \neq \vartheta \middle| \vartheta = i\right] \leq 4e^{-\frac{r}{C}kE_r\left(\frac{r_k}{r}C\right)} \quad \forall i \tag{74}$$

  In the control phase, temporary decision $\tilde{\vartheta} = \vartheta$, will be confirmed by sending accept symbol x_a for $k - \lceil\frac{r}{C}k\rceil$ time units, and will be rejected by sending reject symbol

x_d instead. Let $b_c = \mathbb{I}_{\{\hat{\vartheta}=\vartheta\}}$, then following error performance is achievable with Neyman-Pierson kind of decoding,

$$\Pr\left[\hat{b}_c \neq 1 \middle| b_c = 1\right] \leq e^{-(k-\lceil \frac{r}{C}k\rceil)(\mu(s_k)-s_k)\mu'(s_k))} = \xi(1) \tag{75}$$

$$\Pr\left[\hat{b}_c \neq 0 \middle| b_c = 0\right] \leq e^{-(k-\lceil \frac{r}{C}k\rceil)(\mu(s_k)+(1-s_k)\mu'(s_k))} = \xi(0) \tag{76}$$

where

$$\mu(s) = \ln \sum_i W_{Y|X}(i|x_a)^{1-s} W_{Y|X}(i|x_d)^s \qquad \mu'(s) = \tfrac{d}{dx}\mu(s)|_{x=s} \tag{77}$$

and $D_{\max} = \max_{i,j} D\left(W_{Y|x}(\cdot|i)\middle\| W_{Y|X}(\cdot|j)\right) = D\left(W_{Y|x}(\cdot|x_a)\middle\| W_{Y|X}(\cdot|x_d)\right)$.

If $\hat{b}_c = 1$ then receiver will decode, ϑ, as $\hat{\vartheta} = \tilde{\vartheta}$, else it will declare an erasure. If $\hat{\vartheta} = |\mathcal{M}_s| + 1$ or erasure was declared for ϑ, decoder will declare an erasure for the whole block. Else $\hat{M} = \hat{\vartheta}$.

Now we can calculate the error and erasure probabilities of the two phase fixed length block code. Let us denote the erasures by $\hat{M}$ = erasure for each k.

For $i \in \mathcal{M}_s$ using the equations (73), (75), (76) and Bayes rule we get

$$\Pr\left[\tilde{M} = \text{erasure}\middle| M = i\right] \leq e^{-\sqrt{k}E_{ex}\left(\frac{\ln 4}{\sqrt{k}}\right)} + \left(\xi(1) + 4e^{-\frac{r}{C}kE_r\left(\frac{r_k}{r}C\right)}\right) \tag{78}$$

$$\Pr\left[\tilde{M} \neq i, \hat{M} \neq \text{erasure}\middle| M = i\right] \leq P_{e\mathcal{Q}'}(1) + \xi(0) \tag{79}$$

For $i \notin \mathcal{M}_s$ using the equation (73) and Bayes rule we get

$$\Pr\left[\tilde{M} = \text{erasure}\middle| M = i\right] \leq P_{e,\mathcal{Q}'} + e^{-\sqrt{k}E_{ex}\left(\frac{\ln 4}{\sqrt{k}}\right)} \tag{80}$$

$$\Pr\left[\tilde{M} \neq i, \hat{M} \neq \text{erasure}\middle| M = i\right] \leq P_{e,\mathcal{Q}'} + e^{-\sqrt{k}E_{ex}\left(\frac{\ln 4}{\sqrt{k}}\right)} \tag{81}$$

When ever $\hat{M}$ = erasure than transmitter and receiver will try to sent the message once again from scratch using same strategy. Then for any $i \in \mathcal{M}$

$$\Pr\left[\hat{M} \neq i\middle| M = i\right] = \frac{\Pr\left[\tilde{M}\neq i, \hat{M}\neq \text{erasure}\middle| M=i\right]}{1-\Pr\left[\tilde{M}=\text{erasure}\middle| M=i\right]} \tag{82}$$

$$E\left[\tau\middle| M = i\right] = \frac{k+\sqrt{k}}{1-\Pr\left[\tilde{M}=\text{erasure}\middle| M=i\right]} \tag{83}$$

If we choose $r_k = r\left(1 - \frac{1}{\log k}\right)$ and $s_k = \frac{1}{\log k}$, using equations (78), (79), (80), (81), (82) and (83) we can prove that $\mathcal{Q}$ is capacity achieving sequence such that

$$\lim_{k\to\infty} -\frac{\ln \max_{i\in\mathcal{M}_s} \Pr\left[\tilde{M}\neq i, \hat{M}\neq \text{erasure}\middle| M=i\right]}{E[\tau]} = \min\{\tilde{C}, (1-\tfrac{r}{C})D_{\max}\}$$

$$\lim_{k\to\infty} \frac{\ln |\mathcal{M}_s^{(k)}|}{E[\tau]} = r$$

•

8.5.2 Converse: $E_{\mathbf{md}}^{f}(r) \leq \min\{\tilde{\boldsymbol{C}}, (1-\frac{r}{\boldsymbol{C}})\boldsymbol{D}_{\max}\}$

Proof:
Consider any uniform delay capacity achieving sequence, $\mathcal{Q}$. Note that by excluding all $i \notin \mathcal{M}_s^{(k)}$ we will simply get a reliable sequence, $\mathcal{Q}'$ such that

$$P_e{'}^{(k)} \leq \Pr{}^{(k)} \left[\hat{M} \neq M \middle| M \in \mathcal{M}_s \right]$$
$$E\left[\tau'^{(k)}\right] \leq \Gamma^{(k)} E\left[\tau^{(k)}\right]$$

Consequently

$$\frac{-\ln \Pr\left[\hat{M}\neq M \middle| M\in\mathcal{M}_s\right]^{(k)}}{E\left[\tau^{(k)}\right]} \leq -\frac{\ln P_e{'}^{(k)}}{E\left[\tau'^{(k)}\right]}\Gamma^{(k)} \tag{84}$$

Thus $E_{\mathrm{md}}^{f}(r) \leq (1-\frac{r}{C})D_{\max}$. Similarly by excluding all but one of the elements of $\mathcal{M}_s$ we can prove that $E_{\mathrm{md}}^{f}(r) \leq \tilde{C}$. •

9 Avoiding False Alarms: Proofs

9.1 Block Codes without Feedback: Proof of Theorem 10

9.1.1 Lower Bound: $E_{\mathbf{fa}} \geq E_{\mathbf{fa}}^{\mathbf{l}}$

Proof:
As a result of the coding theorem [4, Ch. 2 Corollary 1.3, page 102] we know that there exits a reliable sequence, $\mathcal{Q}'$ fixed composition codes whose rate is C whose n^{th} elements composition, $P_X^{(n)}$, satisfies,

$$\sum_{i\in\mathcal{X}} |P_X^{(n)}(i) - P_X^{*}(i)| \leq \sqrt[4]{\tfrac{1}{n}}.$$

We will use the codewords the n^{th} element of $\mathcal{Q}'$ as the codewords of the ordinary messages in the n^{th} code in $\mathcal{Q}$. For the special message we will use a length-n repetition sequence $\bar{x}^n(1) = (x_{f_l}, x_{f_l}, \cdots, x_{f_l})$.

The decoding region for the special message will be essentially the bare minimum. We will include the typical channel outputs within the decoding region of the special message to ensure small missed detection probability for the special message, but we will not include any other output sequence y^n.

$$\mathcal{G}(1) = \{y^n : \sum_{i\in\mathcal{Y}} |\mathsf{Q}_y\left(y^n\right)(i) - W_{Y|X}(i|x_{f_l})| \leq \sqrt[4]{1/n}\}$$

Note that this definition of $\mathcal{G}(1)$ itself ensures that special message is transmitted reliably when ever it is sent, $\lim_{n\to\infty} P_e{}^{(n)}(1) = 0$.

The decoding regions of the ordinary messages, $j = \{2, 3, \ldots \mathcal{M}^{(n)}\}$, will simply be the intersection of the corresponding decoding region in $\mathcal{Q}'$ with the complement of $\mathcal{G}(1)$.Thus the fact that $\mathcal{Q}'$ is a reliable sequence will imply that,

$$\lim_{n\to\infty} \Pr{}^{(n)} \left[y^n \in \bigcup_{j\notin\{1,i\}} \mathcal{G}(j) \middle| M = i \right] = 0.$$

Consequently the only thing we are left to prove is that decay rate of the $\Pr^{(n)}\left[\hat{M}=1\middle| M\neq 1\right]$ is fast enough. Note the probability of a V-shell of a message i is equal to given by,

$$\Pr\left[\mathsf{T}_V(i)\middle| M=i\right] = e^{-nD\left(V_{Y|X}(\cdot|X)\middle\| W_{Y|X}(\cdot|X)\middle| P_X^{(n)}\right)}$$

Note that also that $\mathcal{G}(1)$ can be written as the union of V-shells of a message i as follow.

$$\mathcal{G}(1) = \bigcup_{V_{Y|X}\in\mathcal{V}^{(n)}} \mathsf{T}_V(i) \qquad \forall i\neq 1$$

where $\mathcal{V}^{(n)} = \{V_{Y|X} : \sum_j |\sum_k V_{Y|X}(j|k)P_X^n(k) - W_{Y|X}(j|x_{fl})| \leq \sqrt[4]{1/n}\}$. Note that since there are at most $(1+n)^{|\mathcal{X}||\mathcal{Y}|}$ different conditional types.

$$\Pr\left[\mathcal{G}(1)\middle| M=i\right] \leq (1+n)^{|\mathcal{X}||\mathcal{Y}|} \max_{V_{Y|X}\in\mathcal{V}^{(n)}} \Pr\left[\mathsf{T}_V(i)\middle| M=i\right]$$

Thus for all $i>1$

$$\lim_{n\to\infty} \frac{-\log \Pr^{(n)}[\mathcal{G}(1)|M=i]}{n} = \min_{V_{Y|X}:\ \sum_j P_X^*(j)V_{Y|X}(\cdot|j)=W_{Y|X}(\cdot|x_{fl})} D\left(V_{Y|X}(\cdot|X)\middle\| W_{Y|X}(\cdot|X)\middle| P_X^*\right)$$

•

9.1.2 Upper Bound: $E_{\mathbf{fa}} \leq E_{\mathbf{fa}}^{\mathbf{u}}$

Proof:
As a result of data processing inequality for KL divergence we have

$$\begin{aligned}\sum_{y^n\in\mathcal{Y}^n} \Pr\left[y^n\middle| M=1\right] \log \tfrac{\Pr[y^n|M=1]}{\Pr[y^n|M\neq 1]} &\geq \Pr\left[\mathcal{G}(1)\middle| M=1\right] \log \tfrac{\Pr[\mathcal{G}(1)|M=1]}{\Pr[\mathcal{G}(1)|M\neq 1]} \Pr\left[\overline{\mathcal{G}(1)}\middle| M=1\right] \log \tfrac{\Pr\left[\overline{\mathcal{G}(1)}\middle|M=1\right]}{\Pr\left[\overline{\mathcal{G}(1)}\middle|M\neq 1\right]} \\ &\geq -\ln 2 - \Pr\left[\mathcal{G}(1)\middle| M=1\right] \log \Pr\left[\mathcal{G}(1)\middle| M\neq 1\right] \end{aligned} \quad (85)$$

Using the convexity of the KL divergence we we get

$$\begin{aligned}\sum_{y^n\in\mathcal{Y}^n} \Pr\left[y^n\middle| M=1\right] \log \tfrac{\Pr[y^n|M=1]}{\Pr[y^n|M\neq 1]} &\leq \sum_{i=2}^{|\mathcal{M}|} \tfrac{1}{|\mathcal{M}|-1} \sum_{y^n\in\mathcal{Y}^n} \Pr\left[y^n\middle| M=1\right] \log \tfrac{\Pr[y^n|M=1]}{\Pr[y^n|M=i]} \\ &= \sum_{i=2}^{|\mathcal{M}|} \tfrac{1}{|\mathcal{M}|-1} \sum_{y^n\in\mathcal{Y}^n} \Pr\left[y^n\middle| M=1\right] \sum_{k=1}^{n} \log \tfrac{\Pr\left[y_k|M=1,y^{k-1}\right]}{\Pr\left[y_k|M=i,y^{k-1}\right]} \\ &= \sum_{k=1}^{n} \sum_{i=2}^{|\mathcal{M}|} \tfrac{1}{|\mathcal{M}|-1} D\left(W_{Y|X}(\cdot|\bar{x}_k(1))\middle\| W_{Y|X}(\cdot|\bar{x}_k(i))\right) \end{aligned} \quad (86)$$

where $x_k(i)$ denotes the input letter for codeword of message i, at time k.

Let us denote the empirical distribution of the $\bar{x}_k(i)$ for a fixed time k, by P_{X_k}.

$$P_{X_k}(i) \triangleq = \frac{\sum_{j\in\mathcal{M}} \mathbb{I}_{\{\bar{x}_k(j)=i\}}}{|\mathcal{M}|} \quad \forall i\in\mathcal{X}$$

Using equation (85) and (86) we get

$$\Pr\left[\mathcal{G}(1)\middle|\, M \neq 1\right] \geq e^{-\frac{1}{\Pr[\mathcal{G}(1)|M=1]}\left(\frac{|\mathcal{M}|}{|\mathcal{M}|-1}\sum_k D\left(W_{Y|X}(\cdot|\bar{x}_k(1))\middle\|W_{Y|X}(\cdot|X_k)\middle|P_{X_k}\right)-\ln 2\right)} \tag{87}$$

We will show below that for all capacity achieving codes, all most all of the k's has a P_{X_k} which is essentially equal to P_X^*. For doing that let us first define the set $\mathcal{P}(\epsilon)$ and $\delta(\epsilon)$

$$\mathcal{P}(\epsilon) \triangleq \{P_X : I(P_X, W_{Y|X}) \geq C - \epsilon\} \quad \text{and} \quad \delta(\epsilon) \triangleq \max_{P_X\in\mathcal{P}(R)} \sum_i |P_X(i) - P_X^*(i)|$$

Note that $\lim_{\epsilon\to 0}\delta(\epsilon) = 0$.

We will now show for almost all of the k's P_{X_k} is essentially P_X^*. Note that as a result of Fano's inequality we have,

$$I(M, Y^n) \geq nR^{(n)}(1 - P_e) + \ln 2 \tag{88}$$

On the other hand using standard manipulations on mutual information we get

$$\begin{aligned} I(M; Y^n) &\leq \sum_{k=1}^{n} I(X_k, Y_k) \\ &= \sum_{k=1}^{n} I(P_{X_k}, W_{Y|X}) \\ &\leq Cn - \epsilon \sum_{k=1}^{n} \mathbb{I}_{\left\{P_{X_k}\in\mathcal{P}(\epsilon)\right\}} \end{aligned}$$

Inserting this in to equation (88) we get,

$$\sum_{k=1}^{n} \mathbb{I}_{\left\{P_{X_k}\notin\mathcal{P}(\epsilon)\right\}} \leq n\frac{(C-R^{(n)}(1-P_e)-\ln 2/n)}{\epsilon}$$

Let us chose $\epsilon^{(n)} = \sqrt{(C - R^{(n)}(1 - P_e) - \ln 2/n)}$. Then for any capacity achieving sequence $\lim_{n\to\infty}\epsilon^{(n)} = 0$. Furthermore we will have,

$$\sum_{k=1}^{n} \mathbb{I}_{\left\{P_{X_k}\notin\mathcal{P}\left(\epsilon^{(n)}\right)\right\}} \leq n\epsilon^{(n)}$$

Note for any $P_X \in \mathcal{P}\left(\epsilon^{(n)}\right)$ we have

$$\begin{aligned} D\left(W_{Y|X}(\cdot|x_k(1))\middle\|W_{Y|X}(\cdot|X_k)\middle|\,P_X\right) &\leq D\left(W_{Y|X}(\cdot|x_k(1))\middle\|W_{Y|X}(\cdot|X)\middle|\,P_X^*\right) + \delta(\epsilon^{(n)})D_{\max} \\ &\leq E_{\text{fa}}^{\text{u}} + \delta(\epsilon^{(n)})D_{\max} \end{aligned} \tag{89}$$

where $E_{\text{fa}}^{\text{u}} = \max_{i\in\mathcal{X}} D\left(W_{Y|X}(\cdot|i)\middle\|W_{Y|X}(\cdot|X)\middle|\,P_X^*\right)$

Using equations (9.1.2) and (89)

$$\sum_k D\left(W_{Y|X}(\cdot|x_k(1))\middle\|W_{Y|X}(\cdot|X_k)\middle|\,P_{X_k}\right) \leq n(E_{\text{fa}}^{\text{u}} + \delta(\epsilon^{(n)})D_{\max} + \epsilon^{(n)}D_{\max}) \tag{90}$$

Inserting this in equation (87) we get

$$\lim_{n\to\infty}\left(\frac{-\log\Pr{}^{(n)}[\mathcal{G}(1)|M\neq 1]}{n} \leq E_{\text{fa}}^{\text{u}} \tag{91}$$

•

9.2 Variable Length Block Codes with Full feedback: Proof of Theorem 11

9.2.1 Achievability: $E^{f}_{\mathbf{fa}} \geq \boldsymbol{D}_{\max}$

Proof:
We will construct a capacity achieving sequence with feedback, $\mathcal{Q}$, by using a construction like the one we have for $E^{f}_{\text{md}}(r)$. In fact, this scheme achieves the false alarm exponent simultaneously with the best missed detection exponent, $\tilde{C}$, for the special message. The k^{th} element of the code will be as follows.

We will first us a fixed length multi phase errors and erasure code. In the first phase of the code, a length $\lceil k \rceil$ two message code will be used to tell whether $M = 1$ or not. Let $b = \mathbb{I}_{\{M=1\}}$. As a result of [2, Theorem 5.7.1, page 153]

$$\Pr\left[\hat{b} \neq 1 \middle| b = 1\right] = \Pr\left[\hat{b} \neq 0 \middle| b = 0\right] \leq e^{-\sqrt{k} E_{ex}\left(\frac{\ln 8}{\sqrt{k}}\right)} \tag{92}$$

where $E_{ex}(\cdot)$ stands for the expurgated exponent.

In the second phase one of two codes will be used depending on $\hat{b}$.

- If $\hat{b} = 0$, transmitter will use the k^{th} member of a capacity achieving sequence, $\mathcal{Q}'$ such that $E_{\text{b},\mathcal{Q}'} = \tilde{C}$ to convey which non-special message is to transmitted. Note that such a sequence exists as a result of Theorem 2.

$$\vartheta = M \tag{93}$$

  If at the end of the second phase $\hat{\vartheta} = 1$, it will declare an erasure, else it will decode M as $\hat{M} = \hat{\vartheta}$.

- If $\hat{b} = 1$, transmitter will use a length k repetition code to convey whether $M = 1$ or not. If $M = 1$, $\vartheta = 1$ and transmitter will send the codeword $(x_a, x_a, \ldots, x_a)$. If $M \neq 1$, $\vartheta = 0$ and transmitter will send the codeword $(x_d, x_d, \ldots, x_d)$. A using a Neyman-Pierson like decoding we can achieve following error probabilities simultaneously,

$$\Pr\left[\hat{\vartheta} \neq 1 \middle| \vartheta = 1\right] \leq e^{-k(\mu(s_k) - s_k \mu'(s_k))} \tag{94}$$

$$\Pr\left[\hat{\vartheta} \neq 0 \middle| \vartheta = 0\right] \leq e^{-k(\mu(s_k) + (1-s_k)\mu'(s_k))} \tag{95}$$

  where

$$\mu(s) = \ln \sum_i W_{Y|X}(i|x_a)^{1-s} W_{Y|X}(i|x_d)^s \qquad \mu'(s) = \tfrac{d}{dx}\mu(s)|_{x=s} \tag{96}$$

  and $D_{\max} = \max_{i,j} D\left(W_{Y|x}(\cdot|i) \middle\| W_{Y|X}(\cdot|j)\right) = D\left(W_{Y|x}(\cdot|x_a) \middle\| W_{Y|X}(\cdot|x_d)\right)$.

  If $\hat{\vartheta} = 1$ then receiver will decode, M, as $\hat{M} = 1$, else it will declare an erasure for the whole block.

Now we can calculate the error and erasure probabilities of the two phase fixed length block code. Let us denote the erasures by $\tilde{M} =$ erasure for each k.

Using the equations (92), (94), (95) and Bayes rule we get

$$\Pr\left[\tilde{M} = \text{erasure}\middle| M = 1\right] \leq e^{-\sqrt{k}E_{ex}\left(\frac{\ln 8}{\sqrt{k}}\right)} + e^{-k(\mu(s_k)-s_k\mu'(s_k))} \tag{97}$$

$$\Pr\left[\tilde{M} = \text{erasure}\middle| M = i\right] \leq P_{e\mathcal{Q}'}^{(k)} + e^{-\sqrt{k}E_{ex}\left(\frac{\ln 8}{\sqrt{k}}\right)} \qquad i \neq 1 \tag{98}$$

$$\Pr\left[\tilde{M} \neq 1, \tilde{M} \neq \text{erasure}\middle| M = 1\right] \leq P_{e\mathcal{Q}'}^{(k)}(1) \tag{99}$$

$$\Pr\left[\tilde{M} \neq i, \tilde{M} \neq \text{erasure}\middle| M = i\right] \leq P_{e\mathcal{Q}'}^{(k)} \qquad i \neq 1 \tag{100}$$

$$\Pr\left[\tilde{M} = 1\middle| M \neq i\right] \leq e^{-k(\mu(s_k)+(1-s_k)\mu'(s_k))} \tag{101}$$

When ever $\tilde{M} =$ erasure than transmitter and receiver will try to sent the message once again from scratch using same strategy. Then all of the above error probabilities will be scaled by $\frac{1}{1-\Pr\left[\tilde{M}=\text{erasure}\middle|M=i\right]}$, furthermore

$$E\left[\tau\middle| M = i\right] = \frac{k+\sqrt{k}}{1-\Pr\left[\tilde{M}=\text{erasure}\middle|M=i\right]} \tag{102}$$

If we chose $s_k = \frac{1}{\log k}$, using equations (97), (98), (99), (100), (101) and (102) we can prove that for $\mathcal{Q}$ is a capacity achieving code with $E^f_{\text{md},\mathcal{Q}} = \tilde{C}$ and $E^f_{\text{fa},\mathcal{Q}} = D_{\max}$. •

9.2.2 Converse: $E^f_{\textbf{fa}} \leq \boldsymbol{D}_{\max}$

Proof:
Let denote the decoding region of each message by $\mathcal{G}(i)$ than

$$\mathcal{G}(i) = \{y^\tau : \hat{M} = i\}$$

Note that as result of convexity of KL divergence we have

$$E\left[\log \frac{\Pr[Y^\tau|M=1]}{\Pr[Y^\tau|M\neq 1]}\middle| M = 1\right] \geq \Pr\left[\mathcal{G}(1)\middle| M = 1\right] \log \frac{\Pr[\mathcal{G}(1)|M=1]}{\Pr[\mathcal{G}(1)|M\neq 1]} + \Pr\left[\overline{\mathcal{G}(1)}\middle| M = 1\right] \log \frac{\Pr\left[\overline{\mathcal{G}(1)}\middle|M=1\right]}{\Pr\left[\overline{\mathcal{G}(1)}\middle|M\neq 1\right]}$$
$$\geq -\ln 2 + \Pr\left[\mathcal{G}(1)\middle| M = 1\right] \log \frac{1}{\Pr[\mathcal{G}(1)|M\neq 1]} \tag{103}$$

It has already been proved in [20] that,

$$E\left[\log \frac{\Pr[Y^\tau|M=1]}{\Pr[Y^\tau|M\neq 1]}\middle| M = 1\right] \leq D_{\max} E\left[\tau\middle| M = 1\right] \tag{104}$$

Note that as a result of definition of Γ we have $E\left[\tau\right] M = 1 \leq E\left[\tau\right]\Gamma$ using this together with equations (103) and (104) the we get,

$$\Pr\left[\mathcal{G}(1)\middle| M \neq 1\right] \geq e^{-\frac{\ln 2+\Gamma D_{\max} E[\tau]}{\Pr[\mathcal{G}(1)|M=1]}} \tag{105}$$

Thus for any uniform capacity delay reliable sequence, $\mathcal{Q}$, we have $E^f_{\text{fa},\mathcal{Q}} \leq D_{\max}$.

•

A Equivalent definitions of UEP exponents

We could define all the UEP exponents in this paper without using the notion of capacity achieving sequences. As an example, we will define the single-bit exponent in this alternate manner. This alternative first defines $\bar{E}_{\rm b}(R)$ as the best exponent for the special bit at a given data-rate R, and then minimizes $\bar{E}_{\rm b}(R)$ over all $R < C$ to obtain $\bar{E}_{\rm b}$.

Definition 14 *For a reliable code sequence $\mathcal{Q}$ of rate $R_{\mathcal{Q}}$, with message sets $\mathcal{M}^{(n)} = \mathcal{M}_1 \times \mathcal{M}_2^{(n)}$ where $\mathcal{M}_1 = \{0,1\}$, the exponent for the special bit error probability $\Pr^{(n)}\left[\hat{M}_1 \neq M_1\right]$ equals*

$$E_{b,\mathcal{Q}} = \liminf_{n\to\infty} \frac{-\log \Pr^{(n)}\left[\hat{M}_1 \neq M_1\right]}{n}. \tag{106}$$

Then define $\bar{E}_b(R) = \sup_{\mathcal{Q}: R_{\mathcal{Q}} \geq R} E_{b,\mathcal{Q}}$. Now the single bit exponent $\bar{E}_b$ is defined as

$$\bar{E}_b = \inf_{R<C} \bar{E}_b(R)$$

This definition of says that no matter how close the rate is to capacity, the special bit can achieve the exponent $\bar{E}_{\rm b}$. We now show briefly why this definition is equivalent to the earlier definition in terms of capacity achieving sequences.

Lemma 2 $\bar{E}_b = E_b$

Proof:
of $E_{\mathbf{b}} \leq \bar{E}_{\mathbf{b}}$: By definition of $E_{\rm b}$, for any given $\delta > 0$, there exists a capacity-achieving sequence $\mathcal{Q}$ whose single bit exponent $E_{{\rm b},\mathcal{Q}}$ satisfies,

$$E_{{\rm b},\mathcal{Q}} \geq E_{\rm b} - \delta.$$

We will use this capacity-achieving sequence $\mathcal{Q}$ to prove $\bar{E}_{\rm b} \geq E_{{\rm b},\mathcal{Q}} \geq E_{\rm b} - \delta$. This is because rate of $\mathcal{Q}$ equals C by definition. Hence definition of $\bar{E}_{\rm b}(R)$ implies

$$\begin{aligned} E_{{\rm b},\mathcal{Q}} &\leq \bar{E}_{\rm b}(R) \quad \text{for any } R < C \\ (E_{\rm b} - \delta \leq)\quad E_{{\rm b},\mathcal{Q}} &\leq \bar{E}_{\rm b}. \end{aligned}$$

The proof follows by choosing arbitrarily small δ.

Proof of $\bar{E}_{\mathbf{b}} \geq E_{\mathbf{b}}$: Let us first fix an arbitrarily small $\delta > 0$. In the table in Figure 7, row k represents a reliable code-sequence $\bar{\mathcal{Q}}_k$ at rate $C - 1/k$, whose single-bit exponent

$$E_{{\rm b},\bar{\mathcal{Q}}_k} \geq \bar{E}_{\rm b}(R) - \delta$$

Let $\bar{\mathcal{Q}}_k(l)$ represent length-l code in this sequence. We construct a capacity achieving sequence $\mathcal{Q}$ from this table as follows. This construction sequentially chooses elements from rows $1, 2, \cdots$.

> **Initialize:** For sequence $\bar{\mathcal{Q}}_1$, let n_1 denote the smallest block length n at which the single bit error probability satisfies
>
> $$\frac{-\log \Pr^{(n)}\left[\hat{M}_1 \neq M_1\right]}{n} \geq \bar{E}_{\rm b}(R) - 2\delta \quad \Leftrightarrow \quad \Pr^{(n)}\left[\hat{M}_1 \neq M_1\right] \leq \exp(-n(\bar{E}_{\rm b}(R) - 2\delta))$$
>
> **Iterate:** For sequence $\bar{\mathcal{Q}}_{i+1}$, choose the smallest $n_{i+1} \geq n_i$ which satisfies above equation.

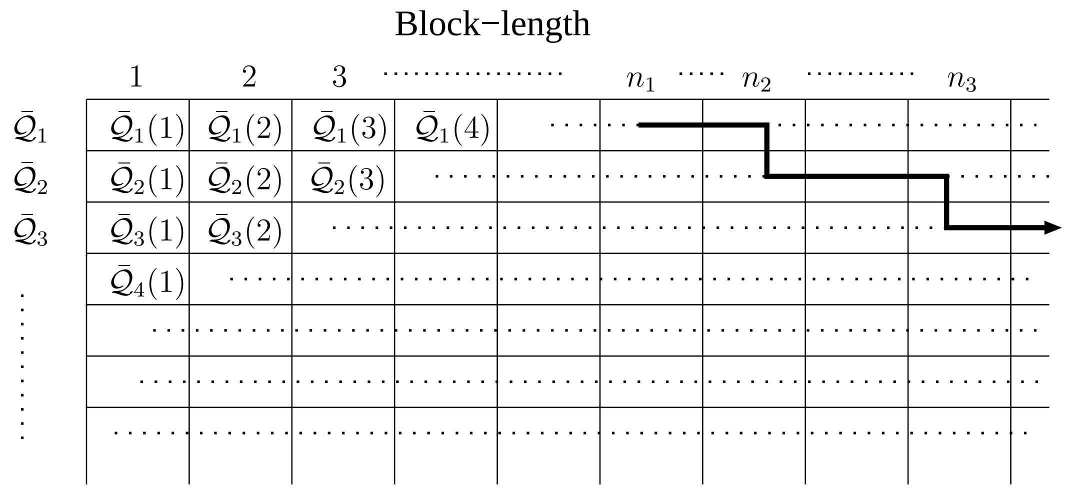


Figure 7: Row k denotes a reliable code sequence at rate $C - 1/k$. Bold path shows capacity achieving sequence $\mathcal{Q}$.

Given the sequence, $n_1, n_2, \cdots$, from each row i, we will choose codes of length n_i to $n_{i+1} - 1$, i.e.,

$$\left(\bar{\mathcal{Q}}_i(n_i), \bar{\mathcal{Q}}_i(n_i+1) \cdots, \bar{\mathcal{Q}}_1(n_{i+1}-1)\right)$$

as members of the capacity-achieving sequence $\mathcal{Q}$. Thus $\mathcal{Q}$ is a sampling of the code-table as shown by the bold path in Figure 7. Note that this choice of $\mathcal{Q}$ is a capacity achieving sequence, moreover it will also achieve a single bit exponent

$$E_{\mathrm{b},\mathcal{Q}} = \inf_{R<C} \{\bar{E}_{\mathrm{b}}(R) - 2\delta\} = \bar{E}_{\mathrm{b}} - 2\delta$$

Choosing arbitrarily small δ proves $E_{\mathrm{b}} \geq \bar{E}_{\mathrm{b}}$. •

B Proof of Theorem 3 for BSC

We will focus on a BSC with crossover probability p in this appendix. Before going further, we state the following lemma for binary hypothesis testing (see [5] for example). Consider binary random sequence Y^n of length n. Under hypothesis $H = 0$, it is i.i.d. over time with distribution Bern(p). Under hypothesis $H = 1$, it is i.i.d. over time with distribution Bern($\frac{1}{2}$). Here Bern(p) denotes the Bernoulli distribution with parameter p.

Lemma 3 *Let E_1 denote the exponent for missed detection probability* $\Pr(\hat{H} = 0|H = 1)$ *and E_0 denote the exponent for false alarm probability* $\Pr(\hat{H} = 1|H = 0)$. *The following implicit equation provides the optimal trade-off between these two exponents, where $D_b(h\|g)$ denotes the KL divergence between two Bernoulli distributions with parameters h and g.*

$$\text{For some } \delta \leq 1/2, \text{ let } E_1 = D_b(\delta\|\frac{1}{2}) \quad \Rightarrow E_0 \leq D_b(\delta\|p) \tag{107}$$

Moreover, this is exponent pair is achievable by a threshold test on the Hamming weight of Y^n, which chooses $\hat{H} = 1$ if the Hamming weight of Y^n exceeds $n\delta$ and vice verse.

Note that if $E_1 = r$, then E_0 denotes the sphere-packing exponent at rate r and δ denotes the Gilbert-Varshamov distance for rate r. To emphasize the dependence on r, we will denote this Gilbert-Varshamov distance by $\delta_{\mathrm{G\,V}}(r)$. Now we are ready to prove the theorem.

Special codewords: At any given block length n, we start with a optimum code-book (say $\mathcal{C}_{special}$) for $\lceil e^{nr} \rceil$ messages. Such optimum code-book achieves error exponent $E(r)$ for every message in it.

$$\Pr\left[\hat{M} \neq i | M = i\right] \doteq \exp(-nE(r)) \qquad \forall i \in \mathcal{M}_s \equiv \{1, 2, \cdots, \lceil e^{nr} \rceil\}$$

This code-book is used for transmitting the special messages. At the decoder, let $\mathcal{B}_i$ denote the set of output sequences within Hamming distance $\Phi = n(\delta_{\mathrm{G\,V}}(r+\epsilon)$ from the i'th codeword $\bar{x}^n(i)$.

$$\mathcal{B}_i = \{y^n : \ |y^n - \bar{x}^n(i)|_H \leq \Phi\}$$

Thus $\mathcal{B}_i$ is a ball of radius Φ around codeword i as shown in Figure 3(a). The radius Φ is essentially the sphere-packing radius. Hence these balls will not be disjoint. Now let $\mathcal{B}$ denote the union of these balls around all special codewords.

$$\mathcal{B} = \bigcup_{i \in \mathcal{M}_s} \mathcal{B}_i$$

If the output sequence Y^n lies in $\mathcal{B}$, the first stage of the decoder decides a special message was transmitted. The second stage then chooses the ML candidate in $\mathcal{M}_s$, i.e., the nearest special codeword from Y^n.

Ordinary codewords: The ordinary codewords will be chosen by random coding: flipping a coin i.i.d. over time. This is the same as Shannon's construction for achieving capacity. The random coding construction provides a simple way to show that in the cavity space $\mathcal{B}^c$ (complement of $\mathcal{B}$), we can essentially fit enough typical noise-balls to achieve capacity. This will avoid the complicated task of carefully choosing the ordinary codewords and their decoding regions in the cavity space.

If the output sequence Y^n lies in $\mathcal{B}^c$, the first stage of the decoder decides an ordinary message was transmitted. The second stage then chooses the ML (nearest) candidate from ordinary codewords.

Error analysis: First, consider the case that a special codeword $\bar{x}^n(i)$ is transmitted. Note that $Y^n \in \mathcal{B}_i$ if and only if $Y^n \oplus \bar{x}^n(i)$ weighs less than Φ. Hence by Stein's lemma, the probability of $Y^n \notin \mathcal{B}_i$ has exponent $D_{\mathrm{b}}(\delta_{\mathrm{G\,V}}(r+\epsilon)\|p)$. It is because channel errors are i.i.d. Bern(p). Since first stage error cannot happen for $Y^n \in \mathcal{B}_i$, first stage error exponent is at least $D_{\mathrm{b}}(\delta_{\mathrm{G\,V}}(r+\epsilon)\|p) = E_{\mathrm{sp}}(r+\epsilon)$ when any special message is sent.

Assuming correct decoding in the first stage, the error exponent for the second stage of decoding between $\lceil e^{nr} \rceil$ codewords equals $E(r)$, which is at most the sphere-packing exponent $E_{\mathrm{sp}}(r)$ (see [3]). Since the first stage exponent equals $E_{\mathrm{sp}}(r+\epsilon)$, the effective error exponent for special messages equals

$$\min\{E(r), E_{\mathrm{sp}}(r+\epsilon)\}$$

By choosing arbitrarily small $\epsilon > 0$, the above two-stage decoding ensures missed detection exponent of $E(r)$ for each special message.

Now consider the situation of a uniformly chosen ordinary codeword being transmitted. We have to make sure the error probability is vanishingly small now. In this case, the output sequence distribution is i.i.d. Bern($\frac{1}{2}$) for the random coding ensemble. The first stage decoding error happens if one of the error sequences weighs less (in Hamming weight) than the threshold Φ. Since the outputs are i.i.d. Bern($\frac{1}{2}$), error sequence $Y^n \oplus \bar{x}^n(j)$ corresponding to any special codeword $\bar{x}^n(j)$ is also i.i.d. Bern($\frac{1}{2}$). Since $Y^n \in \mathcal{B}_j$ if and only if $Y^n \oplus \bar{x}^n(j)$ weighs less than Φ, this probability is at most

$$\exp(-nD_{\mathrm{b}}(\delta_{\mathrm{G\ V}}(r+\epsilon)\|1/2)) = \exp(-n(r+\epsilon)).$$

Applying union bound, the probability of $Y^n \in \bigcup \mathcal{B}_i$ is at most $\exp(-n\epsilon)$. This probability of the first stage error hence vanishes for the random coding ensemble. Recall that for the random coding ensemble, average error probability of the second-stage decoding also vanishes below capacity. To summarize, we have shown these two properties of the random coding ensemble:

1. Error probability of first stage decoding vanishes as $a^{(n)} \doteq \exp(-n\epsilon)$ with n when a uniformly chosen ordinary message is transmitted.

2. Error probability of second stage decoding (say $b^{(n)}$) vanishes with n when a uniformly chosen ordinary message is transmitted.

Since the first error probability is at most $4a^{(n)}$ for some 75% fraction of the random ensemble, and the second error probability is at most $4b^{(n)}$ for some 75% fraction of the random ensemble, there exists a particular code which satisfies both these properties. The overall error probability for ordinary messages is at most $4(a^{(n)}+b^{(n)})$, which vanishes with n. We will use this particular code for the ordinary codewords. This de-randomization completes our construction of a reliable code for ordinary messages to be combined with the code $\mathcal{C}_{special}$ for special messages.

For the special codewords, we had already shown that, probability of first stage decoding error decays exponentially with exponent $E_{\mathrm{sp}}(r)$. This completes the achievability proof for the BSC.

Acknowledgment

The authors are indebted to Bob Gallager for his insights and encouragement for this work in general. In particular, Theorem 3 was mainly inspired from his remarks. Helpful discussions with David Forney are also gratefully acknowledged.

References

[1] C. E. Shannon, "A mathematical theory of communication," Bell System Technical Journal, vol. 27, pp. 379-423 and 623-656, July and October, 1948.

[2] R. Gallager, *Info. Theory and Reliable Communication*, Wiley, 1968.

[3] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower Bounds to Error Probability for Coding on Discrete Memoryless Channels", *Information and Control*, pp. 65-103, December 1966.

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

[5] D. Forney, "On exponential error bounds for random codes on the BSC," unpublished manuscript.

[6] R. Gallager, "Fixed Composition Arguments and Lower Bounds to Error Probability," unpublished manuscript, on-line at `http://web.mit.edu/gallager/www/pages/pubs.html`.

[7] A. Montanari and D. Forney, "On exponential error bounds for random codes on the DMC," unpublished manuscript.

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

[9] T. Cover, *Elements of Information Theory*, Wiley-Interscience, 1991.

[10] S. Borade, B. Nakiboğlu, L. Zheng, "Fundamental limits of UEP: data-rates below capacity," in preparation.

[11] S. Borade and L. Zheng,"Euclidean information theory," Allerton Conference, Monticello, Sept. 2007.

[12] A. Albanese, J. Blomer, J. Edmonds, M. Luby, and M. Sudan, Priority encoding transmission, *IEEE Trans. Inform. Theory*, vol. 42, pp. 1737-1744, Nov. 1996.

[13] B. Masnick, J. Wolf, On linear unequal error protection codes, *IEEE Trans. Inform. Theory*, vol. 13, no. 4, pp. 600-607, Oct. 1967.

[14] A. R. Calderbank, N. Seshadri Multilevel codes for unequal error protection, *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1234-1248, July 1993.

[15] M. Trott, Unequal error protection codes: theory and practice, *Proc. IEEE Information Theory Workshop*, Haifa, June 1996.

[16] I. Csiszár, Joint source-channel error exponent , *Problems of Control and Information Theory*, vol. 9 (5), pp. 315-328, 1980.

[17] S. Diggavi, D. Tse, "On successive refinement of diversity," Allerton Conference, Illinois, September 2004.

[18] R. Ahlswede, G. Dueck Identification via channels, *IEEE Trans. Inform. Theory*, vol. 35, no. 1, pp. 15-29, Jan. 1989.

[19] M. Burnashev, "Data transmission over a discrete channel with feedback, random trans. time", Problems Info. Trans., vol. 12, pp. 10-30, 1976.

[20] P. Berlin, B. Nakiboğlu, B. Rimoldi and E. Telatar, " A Simple Derivation of Burnashev's Reliability Function," preprint, `arXiv:cs/0610145v2`.

[21] H. Yamamoto and K. Itoh, Asymptotic performance of a modified Schalkwijk-Barron scheme for channels with noiseless feedback, *IEEE Trans. Info. Theory*, vol. 25, pp. 729733, November 1979.

[22] Albert N. Shiriaev. *P*robability. Springer-Verlag Inc., New York, NY, USA, 1996.

[23] A. Sahai and S. Mitter, “ The necessity and sufficiency of anytime capacity for control over a noisy communication link: Part II: vector systems,” online preprint: `arXiv: cs.IT/0610146`

[24] A. Sahai and S. Mitter, “Source coding and channel requirements for unstable processes,” online preprint: `arXiv: cs.IT/0610143`