Bit-Wise Unequal Error Protection for Variable-Length Block Codes With Feedback

Nakiboglu, Baris, Siva K. Gorantla, Lihzhong Zheng, and Todd P. Coleman. “Bit-Wise Unequal Error Protection for Variable-Length Block Codes With Feedback.” IEEE Trans. Inform. Theory 59, no. 3 (March 2013): 1475–1504.

As Published
http://dx.doi.org/10.1109/tit.2012.2227671

Publisher
Institute of Electrical and Electronics Engineers (IEEE)

Version
Original manuscript

Accessed
Mon Jan 25 13:50:35 EST 2016

Citable Link
http://hdl.handle.net/1721.1/91015

Terms of Use
Creative Commons Attribution-Noncommercial-Share Alike

Detailed Terms
http://creativecommons.org/licenses/by-nc-sa/4.0/
Bit-wise Unequal Error Protection for Variable Length Block Codes with Feedback
Barış Nakiboğlu  Siva K. Gorantla  Lizhong Zheng  Todd P. Coleman

Abstract
The bit-wise unequal error protection problem, for the case when the number of groups of bits $\ell$ is fixed, is
considered for variable length block codes with feedback. An encoding scheme based on fixed length block codes
with erasures is used to establish inner bounds to the achievable performance for finite expected decoding time. A
new technique for bounding the performance of variable length block codes is used to establish outer bounds to the
performance for a given expected decoding time. The inner and the outer bounds match one another asymptotically
and characterize the achievable region of rates-exponents vectors, completely. The single message message-wise
unequal error protection problem for variable length block codes with feedback is also solved as a necessary step
on the way.

Index Terms
Unequal Error Protection(UEP), Feedback, Variable-Length Communication, Block Codes, Error Exponents,
Burnashev’s Exponent, Yamamoto-Itoh scheme, Kudryashov’s signaling, Errors-and-Erasures Decoding Variable-Length Block Coding, Discrete Memoryless Channels (DMCs)

I. INTRODUCTION
In the conventional formulation of digital communication problem, the primary concern is the correct transmission
of the message; hence there is no distinction between different error events. In other words, there is a tacit assumption
that all error events are equally undesirable; incorrectly decoding to a message $\tilde{m}$ when a message $\tilde{m}$ is transmitted,
is as undesirable as incorrectly decoding to a message $\tilde{m}$ when a message $\tilde{m}$ is transmitted, for any $\tilde{m}$ other than $\tilde{m}$
and $\tilde{m}$ other than $\tilde{m}$. Therefore the performance criteria used in the conventional formulation (minimum distance
between codewords, maximum conditional error probability among messages, average error probability, etc.) are
oblivious to any precedence order that might exist among the error events.

In many applications, however, there is a clear order of precedence among the error events. For example in
Internet communication, packet headers are more important than the actual payload data. Hence, a code used for
Internet communication, can enhance the protection against the erroneous transmission of the packet headers at the
expanse of the protection against the erroneous transmission of payload data. In order to appreciate such a coding
scheme, one may analyze error probability of the packet headers and error probability of payload data separately,
instead of analyzing the error probability of the overall message composed of packet header and payload data. Such
a formulation for Internet communication is an unequal error protection (UEP) problem, because of the separate
calculation of the error probabilities of the parts of the messages.

Problems capturing the disparity of undesirability among various classes of error events, by assigning and
analyzing distinct performance criteria for different classes of error events, are called unequal error protection (UEP)
problems. UEP problems have already been studied widely by researchers in communication theory, coding theory,
and computer networks from the perspectives of their respective fields. In this paper we enhance the information
theoretic perspective on UEP problems [5], [2] for variable length block codes by generalizing the results of [2]
to the rates below capacity.

In information theoretic UEP, error events are grouped into different classes and the probabilities associated with
these different classes of error events are analyzed separately. In order to prioritize protection against one or the
other class of error events, corresponding error exponent is increased at the expense of the other error exponents.

This paper was presented in part at ISIT 2010. [6]
Authors acknowledge the support of the NSF through the NSF Cyberphysical Systems Program (NSF0932410) and the NSF Science & Technology Center (CCF-0939370) and support of DARPA through the ITMANET program (W911NF-07-1-0029)
There are various ways to choose the error event classes but two specific choices of error event classes stand out because of their intuitive familiarity and practical relevance; they correspond to the \textit{message-wise UEP} and the \textit{bit-wise UEP}. Below, we first describe these two types of \textit{UEP} then specify the \textit{UEP} problems we are interested in this manuscript.

In the \textit{message-wise UEP}, the message set $\mathcal{M}$ is assumed to be the union of $\ell$ disjoint sets for some fixed $\ell$, i.e., $\mathcal{M} = \bigcup_{j=1}^{\ell} \mathcal{M}_j$ where $\mathcal{M}_i \cap \mathcal{M}_j = \emptyset$ for all $i \neq j$. For each set $\mathcal{M}_j$, the maximum error probability $P_e(j)$, the rate $R(j)$ and the error exponent $E(j)$ are defined as the corresponding quantities defined in the conventional problem, i.e., $P_e(j) = \max_{\mathcal{M}_j} P(\tilde{\mathcal{M}} \neq \mathcal{M} | \mathcal{M} = \mathcal{M}_j)$, $R(j) = \frac{\ln |\mathcal{M}_j|}{n}$, $E(j) = -\frac{\ln P_e(j)}{n}$, for all $j$ in $\{1, 2, \ldots, \ell\}$ where $n$ is the length of the code. The ultimate aim is calculating the achievable region of rate vector error exponent vector pairs, $(R(\ell), E(\ell))'s$ where $R(\ell) = (R(1), R(2), \ldots, R(\ell))$ and $E(\ell) = (E(1), E(2), \ldots, E(\ell))$. The \textit{message-wise UEP} problem was the first information theoretic \textit{UEP} problem to be considered; it was considered by Csiszár in his work on joint source channel coding [5]. Csiszár showed that for any integer $\ell$, block length $n$ and $\ell$-dimensional rate vector $R(\ell)$ such that $0 \leq R(j) \leq C$ for $j = 1, 2, \ldots, \ell$, there exists a length $n$ block code with message set $\mathcal{M} = \bigcup_{j=1}^{\ell} \mathcal{M}_j$ where $|\mathcal{M}_j| = \epsilon n^{R(\ell) - E(\ell)}$ such that the conditional error probability of each message in each $\mathcal{M}_j$ is less then $e^{-n[\mathcal{E}(j) - \epsilon_n]}$ where $\mathcal{E}(\cdot)$ is the random coding exponent and $\epsilon_n$ converges to zero as $n$ diverges\footnote{This formulation is called the missed detection formulation of the \textit{message-wise UEP} problem in [2]. If $P(\tilde{\mathcal{M}} \neq \mathcal{M} | \mathcal{M} = \mathcal{M}_j)$ is replaced with $P(\tilde{\mathcal{M}} = \mathcal{M} | \mathcal{M} = \mathcal{M}_j)$ we get the false alarm formulation of the \textit{message-wise UEP} problem. In this paper we restrict our discussion to the missed detection problem and use \textit{message-wise UEP} without any qualifications to refer to the missed detection formulation of the \textit{message-wise UEP} problem.}

The \textit{bit-wise UEP} problem is the other canonical form of the information theoretic \textit{UEP} problems. In the \textit{bit-wise UEP} problem the message set $\mathcal{M}$ is assumed to be the Cartesian product of $\mathcal{M}_1$, $\mathcal{M}_2$, $\ldots$, $\mathcal{M}_\ell$ for some fixed $\ell$, i.e., $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2 \times \ldots \times \mathcal{M}_\ell$. Thus the transmitted message $\tilde{\mathcal{M}}$ and the decoded message $\hat{\mathcal{M}}$ are given by $\mathcal{M} = (\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_\ell)$ and $\hat{\mathcal{M}} = (\hat{\mathcal{M}}_1, \hat{\mathcal{M}}_2, \ldots, \hat{\mathcal{M}}_\ell)$, receptively. Furthermore, $\mathcal{M}_j$’s and $\hat{\mathcal{M}}_j$’s are called the transmitted and decoded sub-messages, respectively. The error events of interest in the \textit{bit-wise UEP} problem are the ones corresponding to the missed detection of the sub-messages. The error probability $P_e(j)$, rate $R_j$ and the error exponent $E_j$ of sub-messages are given by $P_e(j) = P(\hat{\mathcal{M}}_j \neq \mathcal{M}_j | \mathcal{M}_j)$, $R_j = \frac{\ln |\mathcal{M}_j|}{n}$, $E_j = -\frac{\ln P_e(j)}{n}$ for all $j$ in $\{1, 2, \ldots, \ell\}$ where $n$ is the block length. As was the case in the \textit{message-wise UEP} problem, the ultimate aim in the \textit{bit-wise UEP} problem is determining the achievable region of the rate vector error exponent vector pairs $(\tilde{R}, \tilde{E})$. The formulation of Internet communication problem we have considered above, with packet header and payload data, is a \textit{bit-wise UEP} problem with two sub-messages, i.e., with $\ell = 2$.

There is some resemblance in the definitions of \textit{message-wise} and \textit{bit-wise UEP} problems, but they have very different behavior in many problems. For example, consider the \textit{message-wise UEP} problem and the \textit{bit-wise UEP} problem with $\ell = 2$, $\mathcal{M}_1 = \{1, 2\}$ and $\mathcal{M}_2 = \{3, 4, \ldots, n^{C-\epsilon_n}\}$ for some $\epsilon_n$ that goes to zero as $n$ diverges. It is shown in [2] Theorem 1 that if $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2$ and $P(\mathcal{M}_2 \neq \hat{\mathcal{M}}_2) \leq \epsilon_n$ for some $\epsilon_n$ that goes to zero as $n$ diverges then $E_1 = 0$. Thus in the \textit{bit-wise UEP} problem every bit can not have a positive error exponent. As result of [5] Theorem 5, on the other hand, if $\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2$ we know that $\mathcal{M}_1$ can have an error exponent $E_{(1)}$ as high as $E_{(0)} > 0$ while having a small error probability for $\mathcal{M}_2$, i.e., $\max_{\mathcal{M}_2} P(\hat{\mathcal{M}} = \mathcal{M}_2 | \mathcal{M} = \mathcal{M}_2) \leq \epsilon_n$ for some $\epsilon_n$ that goes to zero as $n$ diverges. Thus in the \textit{message-wise UEP} problem it is possible to give an error exponent as high as $E_{(0)}$ to $\mathcal{M}_1$.

The \textit{message-wise} and the \textit{bit-wise UEP} problems cover a wide range of problems of practical interest. Yet, as noted in [2], there are many \textit{UEP} problems of practical importance that are neither \textit{message-wise} nor \textit{bit-wise UEP} problems. One of our aims in studying the \textit{message-wise} and the \textit{bit-wise UEP} problems is gaining insights and devising tools for the analysis of those more complicated problems.

\footnote{Here $\ell$ is assumed to be a fixed integer. All rates-exponents vectors, achievable or not, are in the region of $\mathcal{R}^{2\ell}$ in which $R_{(j)} \geq 0$ and $E_{(j)} \geq 0$ for all $1 \leq j \leq \ell$. $\mathcal{R}^{2\ell}$ is the $2\ell$ dimensional real vector space with the norm $\|\tilde{X}\| = \sup_j |x_j|$}

\footnote{Csiszár proved the above result not only for the case when $\ell$ is constant for all $n$ but also for the case when $\ell_n$ is a sequence such that $\lim_{n \to \infty} \frac{\ln \ell_n}{n} = 0$. See [5] Theorem 5.}

\footnote{Similar to the \textit{message-wise UEP} problem discussed above, in the current formulation of \textit{bit-wise UEP} problem we assume $\ell$ to be fixed. Thus all rates-exponents vectors, achievable or not, are in region of $\mathcal{R}^{2\ell}$ in which $R_{j} \geq 0$ and $E_{j} \geq 0$ for all $1 \leq j \leq \ell$, by definition.}

\footnote{The channel is assumed to have no zero probability transition.}
In the above discussion the UEP problems are described for fixed length block codes for the sake of simplicity. One can, however, easily define the corresponding problems for various families of codes: with or without feedback, fixed or variable length, by modifying the definitions of the error probability, the rate and the error exponent appropriately. Furthermore parameter $\ell$ representing the number of groups of bits or messages is assumed to be fixed in the above discussion for simplicity. However, both the message-wise and the bit-wise UEP problems can be defined for $\ell$’s that are increasing with block length $n$ in fixed length block codes and for $\ell$’s that are increasing with expected block length $E[T]$ in variable length block codes. In fact Csiszar’s result discussed above, [5 Theorem 5], is proved not only for constant $\ell$ but also for any $\ell_n$ sequence satisfying $\lim_{n\to\infty} \frac{m\ell_n}{n} = 0$.

In this manuscript we consider two closely related UEP problems for variable length block codes over a discrete memoryless channels with noiseless feedback: the bit-wise UEP problem and the single message message-wise UEP problem.

- In the bit-wise UEP problem there are $\ell$ sub-messages each with different priority and rate. For all fixed values of $\ell$ we characterize the trade-off between the rates and the error exponents of these sub-messages by revealing the region of achievable rate vector, exponent vector pairs. For fixed $\ell$ this problem is simply the variable length code version of the above described bit-wise UEP problem.

- In the single message message-wise UEP problem, we characterize the trade-off between the exponents of the minimum and the average conditional error probability. Thus this problem is similar to the above described message-wise UEP problem for the case $\ell = 2$ and $M_1 = \{1\}$. But unlike that problem we work with variable length codes and average conditional error probability rather than fixed length codes and the maximum error probability.

The bit-wise UEP problem for fixed number of groups of bits, i.e., fixed $\ell$, and the single message message-wise UEP problem were first considered in [2], for the case when the rate is (very close to) the channel capacity; we solve both of these problems for all achievable rates.

In fact, in [2] single message message-wise UEP problem is solved not only at capacity, but also for all the rates below capacity both for fixed length block codes without feedback and for variable length block codes with feedback, but only for case when overall error exponent is zero (see [2 Appendix D]). Recently Wang, Chandar, Chung and Wornell [11] put forward a new proof based on method of types for the same problem. Nazer, Shkel and Draper [9], on the other hand, investigated the problem for fixed length block codes on additive white Gaussian noise channels at zero error exponent and derived the exact analytical expression in terms of rate and power constraints.

Before starting our presentation, let us give a brief outline of the paper. In Section II we specify the channel model and make a brief overview of stopping times and variable length block codes. In Section III we first present the single message message-wise UEP problem and fixed $\ell$ version of the bit-wise UEP problem for variable length block codes; then we state the solutions of these two UEP problems. In Section IV we present inner bounds for both the single message message-wise UEP problem and the bit-wise UEP problem. In Section V we introduce a new technique, Lemma 5 for deriving outer bounds for variable length block codes and apply it to the two UEP problems we are interested in. Finally in Section VI we discuss the qualitative ramifications of our results in terms of the design of communication systems with UEP and the limitations of our analysis. The proofs of the propositions in Sections III IV V are deferred to the Appendices.

II. Preliminaries

As it is customary we use upper case letters, e.g., $M$, $X$, $Y$, $T$ for random variables and lower case letters, e.g., $m$, $x$, $y$, $t$ for their sample values.

We denote discrete sets by capital letters with calligraphic fonts, e.g., $\mathcal{M}$, $\mathcal{X}$, $\mathcal{Y}$ and power sets of discrete sets by $\wp(\cdot)$, e.g., $\wp(M)$, $\wp(X)$, $\wp(Y)$. In order to denote the set of all probability distributions on a discrete set we use $\mathcal{P}(\cdot)$, e.g., $\mathcal{P}(M)$, $\mathcal{P}(X)$, $\mathcal{P}(Y)$.

**Definition 1 (Total Variation):** For any discrete set $Z$ and for any $\mu_1, \mu_2 \in \mathcal{P}(Z)$ the total variation $\Delta(\mu_1, \mu_2)$ is defined as,

$$\Delta(\mu_1, \mu_2) = \frac{1}{2} \sum_{z \in Z} |\mu_1(z) - \mu_2(z)|.$$  

$\dagger$In addition to their new proof in missed-detection problem [11 Theorem 1] Wang, Chandar, Chung and Wornell present a completely new result on the false-alarm formulation of the problem [11 Theorem 5].
We denote the indicator function by $1_{\cdot}$, i.e., $1_{\{\Gamma\}} = 1$ when event $\Gamma$ happens $1_{\{\Gamma\}} = 0$ otherwise. We denote the binary entropy function by $h(\cdot)$, i.e.,
\[ h(s) \triangleq -s \ln s - (1-s) \ln(1-s) \quad \forall s \in [0,1]. \] (2)

A. Channel Model

We consider a discrete memoryless channel (DMC) with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$ and $|\mathcal{X}| - |\mathcal{Y}|$ transition probability matrix $W$. Each row of $W$ corresponds to a probability distribution on $\mathcal{Y}$, i.e., $W_x \in \mathcal{P}(\mathcal{Y})$ for all $x \in \mathcal{X}$. For the reasons that will become clear shortly, in Section II-D, we assume that $W_x(y) > 0$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ and denote the smallest transitions probability by $\lambda$:
\[ \lambda \triangleq \min_{x,y} W_x(y) > 0. \] (3)

The input and output letters at time $\tau$, up to time $\tau$ and between time $\tau_1$ and $\tau_2$ are denoted by $X_\tau$, $Y_\tau$, $X^\tau$, $Y^\tau$, $X^\tau_{\tau_2}$ and $Y^\tau_{\tau_2}$ respectively. DMCs are both memoryless and stationary, hence the conditional probability of $Y_\tau = y$ given $(X^\tau, Y^{\tau-1})$ is given by
\[ P[Y_\tau = y|X^\tau, Y^{\tau-1}] = W_{X_\tau}(y). \]

Definition 2 (Empirical Distribution): For any $\tau_2 \geq \tau_1$ and any sequence $Z^\tau_{\tau_1}$ such that $z_j \in Z$ for all $j \in [\tau_1, \tau_2]$, the empirical distribution $Q_{\{z^\tau_{\tau_1}\}}$ is given by
\[ Q_{\{z^\tau_{\tau_1}\}}(z) = \frac{1}{\tau_2 - \tau_1 + 1} \sum_{\tau=\tau_1}^{\tau_2} 1_{\{z_\tau = z\}} \quad \forall z \in Z. \] (4)

Note that if we replace $Z^\tau_{\tau_1}$ by $Z^\tau$ when the empirical distribution $Q_{\{z^\tau_{\tau_1}\}}(z)$ becomes a random variable for each $z \in Z$.

B. Stopping Times

Stopping times are central in the formal treatment of variable length codes; it is not possible to define or comprehend variable length codes without a solid understanding of stopping times. For those readers who are not already familiar with the concept of the stopping times, we present a brief overview in this section.

In order to make our presentation more accessible, we use the concept of power sets, rather than sigma-fields in the definitions. We can do that only because the random variables we use to define stopping times are discrete random variables. In the general case, when the underlying variables are not necessarily discrete, one needs to use the concept of sigma fields instead of power set.

Let us start with introducing the concept of Markov times. For an infinite sequence of random variables $Z_1, Z_2, \ldots$, a positive, integer valued\(^7\) function $T$ defined on $Z^\infty$ is a Markov time, if for all positive integers $\tau$ it is possible determine whether $T = \tau$ or not by considering $Z^\tau$ only, i.e., if $1_{\{T = \tau\}}$ is not only a function of $Z^\tau$ but also a function of $Z^\tau$ for all positive integers $\tau$. The formal definition is given below.

Definition 3 (Markov Time): Let $Z^\infty_T$ be an infinite sequence of $\mathcal{Z}$ valued random variables $Z_\tau$ for $\tau \in \{1,2,\ldots\}$ and $T$ be a function of $Z^\infty$ which takes values from the set $\{1,2,\ldots,\infty\}$. Then the random variable $T$ is a Markov time with respect to $Z^\tau$ if
\[ \{z^\infty: T = \tau \text{ if } Z^\infty = z^\infty \} \in \mathcal{P}(Z^\tau) \times \{Z^\infty_{\tau+1}\} \quad \forall \tau \in \{1,2,\ldots\}. \]

\(7\)Integer is the set of all integers together with two infinities, i.e., $\{-\infty, -1, 0, 1, \ldots, \infty\}$.\]
and the union of all $Z_{\{T=\tau\}}^\tau$'s by $Z^T$, i.e.,

$$Z_{\{T=\tau\}}^\tau \triangleq \{z^\tau : T = \tau \text{ if } Z^\tau = z^\tau\} \quad \tau \in \{1, 2, \ldots, \infty\} \quad (6a)$$

$$Z^T \triangleq \bigcup_{1 \leq \tau < \infty} Z_{\{T=\tau\}}^\tau \quad (6b)$$

$$Z^T \triangleq Z^{T*} \bigcup Z_{\{T=\infty\}}^\infty \quad (6c)$$

For an arbitrary, positive, integer value function $T$ of $Z^\infty$, however, one can not talk about $Z^T$, because the value of $T$ can in principle depend on $Z_{T+1}^\infty$. For a Markov time $T$, however, the value of $T$ does not depend on $Z_{T+1}^\infty$. That is why we can define $Z^T$, $Z_{\{T=\tau\}}^\tau$, $Z^{T*}$ and $Z^T$ for any Markov time $T$.

Given an infinite sequence of $z^\tau$'s, i.e., $z^\infty$, either $z^\infty \in Z_{\{T=\infty\}}^\infty$ or $z^\infty$ has a unique subsequence $z^\tau$ that is in $Z^{T*}$.

In most practical situations, one is interested in Markov times that are guaranteed to have a finite value; those Markov times are called Stopping times.

**Definition 4 (Stopping Time):** A Markov time $T$ with respect to $Z^\tau$ is a Stopping Time iff $P[T < \infty]$. Note that if $T$ is a stopping time then $P[Z^T \in Z^{T*}] = 1$. Furthermore unlike $Z^T$, $Z^{T*}$ is a countable set for all stopping times $T$ because $|Z|$ is finite.

### C. Variable Length Block Codes

A variable length block code on a DMC is given by a random decoding time $T$, an encoding scheme $\Phi$ and a decoding rule $\Psi$ satisfying $P[T < \infty] = 1$.

- **Decoding time** $T$ is a Markov time with respect to the receiver's observation $Y^\tau$, i.e., given $Y^\tau$ receiver knows whether $T = \tau$ or not. Hence $T$ is a random quantity rather than a constant, thus neither the decoder nor the receiver knows the value of $T$ a priori. But as time passes, both the decoder and the encoder (because of feedback link) will be able to decide whether $T$ has been reached or not, just by considering the current and past channel outputs.

- **Encoding scheme** $\Phi$ is a collection of mappings which determines the input letter at time $(\tau + 1)$ for each message in the finite message set $M$, for each $y^\tau \in Y^\tau$ such that $T > \tau$,

$$\Phi(\cdot, y^\tau) : M \rightarrow \mathcal{X} \quad \forall y^\tau : T > \tau.$$  

- **Decoding Rule** is a mapping from the set of output sequences $y^\tau$ such that $T = \tau$ to the finite message set $M$ which determines the decoded message, $\hat{M}$. With a slight abuse of notation we denote the set of all, possibly infinite, output sequences $y^\tau$ such that $\{T = \tau \text{ if } Y^\tau = y^\tau\}$ by $Y^T$ and write the decoding rule $\Psi$ as,

$$\Psi(\cdot) : Y^T \rightarrow M.$$  

- Note that because of the condition $P[T < \infty] = 1$, decoding time is not only a Markov time, but also a Stopping time.

At time zero the message $M$ chosen uniformly at random from $M$ is given to the transmitter; the transmitter uses the codeword associated $M$, i.e., $\Phi(M, \cdot)$, to convey the message $M$ until the decoding time $T$. Then the receiver chooses the decoded message $\hat{M}$ using its observation $Y^T$ and the decoding rule $\Psi$, i.e., $\hat{M} = \Psi(Y^T)$. The error probability, the rate and the error exponent of a variable length block code are given by

$$P_e = P[\hat{M} \neq M] \quad R = \frac{\ln |M|}{E[T]} \quad E = \frac{-\ln P_e}{E[T]}.$$  

Indeed one can interpret the variable length block codes on DMCs as trees, for a more detailed discussion of this interpretation readers may go over [1, Section II].

---

8 $Z^{T*}$ is a countable set even when $|Z|$ is countably infinite.

9 See equation (6).

10 Having a finite decoding time with probability one, i.e., $P[T < \infty] = 1$, does not imply having a finite expected value for the decoding time, i.e., $E[T] < \infty$. Thus a variable length code can, in principle, have an infinite expected decoding time.
D. Reliable Sequences for Variable Length Block Codes

In order to suppress the secondary terms while discussing the main results, we use the concept of reliable sequences. In a sequence of codes we denote the error probability and the message set of the $\kappa^{th}$ code of the sequence by $P_e^{(\kappa)}$ and $M^{(\kappa)}$, respectively.

Definition 5 (Reliable Sequence): A sequence of variable length block codes $Q$ is reliable if the error probabilities of the codes vanish and the size of the message sets of the codes diverge:

$$\lim_{\kappa \to \infty} \left( P_e^{(\kappa)} + \frac{1}{|M^{(\kappa)}|} \right) = 0,$$

where $P_e^{(\kappa)}$ and $M^{(\kappa)}$ are the error probability and the message set for the $\kappa^{th}$ code of the reliable sequence, respectively.

Note that in a sequence of codes, each code has an associated probability space. We denote the random variables in these probability spaces together with a superscript corresponding to the code. For example the decoding time of the $\kappa^{th}$ code in the sequence is denoted $T^{(\kappa)}$. The expected value of random variables in the probability space associated with the $\kappa^{th}$ code in the sequence is denoted by $E^{(\kappa)}$.

Definition 6 (Rate of a Reliable Sequence): The rate of a reliable sequence $Q$ is the limit infimum of the rates of the individual codes,

$$R_Q \triangleq \liminf_{\kappa \to \infty} \frac{\ln |M^{(\kappa)}|}{E^{(\kappa)} \left[ T^{(\kappa)} \right]}.$$

Definition 7 (Capacity): The capacity of a channel for variable length block codes is the supremum of the rates of the all reliable sequences.

$$C \triangleq \sup_Q R_Q.$$

The capacity of a DMC for variable length block codes is identical to the usual channel capacity, [3]. Hence,

$$C = \max_{\mu \in \mathcal{P}(X)} \sum_{x,y} \mu(x) W_x(y) \ln \frac{W_x(y)}{\bar{\mu}(y)}$$

(8)

where $\bar{\mu}(y) = \sum_x \mu(x) W_x(y)$.

Definition 8 (Error Exponent of a Reliable Sequence): The error exponent of a reliable sequence $Q$ is the limit infimum of the error exponents of the individual codes,

$$E_Q \triangleq \liminf_{\kappa \to \infty} \frac{-\ln P_e^{(\kappa)}}{E^{(\kappa)} \left[ T^{(\kappa)} \right]}.$$

Definition 9 (Reliability Function): The reliability function of a channel for variable length block codes at rate $R \in [0, C]$ is the supremum of the exponents of all reliable sequences whose rate is $R$ or higher.

$$E(R) \triangleq \sup_{Q: R_Q \geq R} E_Q.$$

Burnashev [3] analyzed the performance of variable length block codes with feedback and established inner and outer bounds to their performance. Results of [3] determine the reliability function of variable length block codes on DMCs for all rates. According to [3]:

11Recall that the decoding time of a variable length block code is finite with probability one. Thus $P^{(\kappa)}\left[T^{(\kappa)} < \infty \right] = 1$ for all $\kappa$ for a reliable sequence.

12Evidently it is possible to come up with a probability space that includes all of the codes in a reliable sequence and invoke independence between random quantities associated with different codes. We choose the current convention to emphasize independence explicitly in the notation we use.
• If all entries of $W$ are positive then\footnote{Problem is formulated somewhat differently in [4]. as a result [4] did not deal with the case $E[T] = \infty$. The bounds in [4] does not guarantee that the error probability of a variable length code with infinite expected decoding time is greater than zero, however this is the case if all the transition probabilities are positive. To see that consider a channel with positive minimum transition probability $\lambda$, i.e., $\lambda = \min_{x,y} W_{xy} > 0$. In such a channel any variable length code satisfies $P_u \geq \frac{M-1}{M} E \left[ \left( \frac{1}{1-\lambda} \right)^T \right]$, then $P_u > 0$ as $\lambda > 0$ and $P[T < \infty] = 1$. Consequently both the rate and the error exponent are zero for variable length block codes with infinite expected decoding time. A more detailed discussion of this fact can be found in Appendix [11].}
\[
E(R) = \left( 1 - \frac{R}{C} \right) D \quad \forall R \in [0, C]
\]
where $D$ is maximum Kullback Leibler divergence between the output distributions of any two input letters:
\[
D \triangleq \max_{x,\bar{x} \in \mathcal{X}} \quad D(W_x || W_{\bar{x}}). 
\tag{9}
\]

• If there are one or more zero entries\footnote{Note that in this situation $D = \infty$.} in $W$, i.e., if there are two input letters $x, \bar{x}$ and an output letter $y$ such that, $W_x(y) = 0$ and $W_{\bar{x}}(y) > 0$, then for all $R < C$, for large enough $E[T]$ there are rate $R$ variable length block codes which are error free, i.e., $P_e = 0$.

When $P_e = 0$ all error events can have zero probability at the same time. Consequently all the $UEP$ problems are answered trivially when there is a zero probability transition. This is why we have assumed that $W_x(y) > 0$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

We denote the input letters that get this maximum value of Kullback Leibler divergence by\footnote{This particular naming of letters is reminiscent of the use of these letters in Yamamoto Itoh scheme [12]. Although they are named differently in [12], $a$ is used for accepting and $r$ is used for rejecting the tentative decision in Yamamoto Itoh scheme.} $a$ and $r$:
\[
D = D(W_a || W_r). 
\tag{10}
\]

III. PROBLEM STATEMENT AND MAIN RESULTS

A. Problem Statement

For each $m \in \mathcal{M}$, the conditional error probability is defined as\footnote{Later in the paper we consider block codes with erasures. The conditional error probabilities, $P_{e|m}$ for $m \in \mathcal{M}$, are defined slightly differently for them, see equation [24].}
\[
P_{e|m} \triangleq P \left[ M \neq M \mid M = m \right]. 
\tag{11}
\]

In the conventional setting we are interested in either the average or the maximum of the conditional error probability of the messages. The behavior of the minimum conditional error probability is scarcely investigated. Single message message-wise $UEP$ problem attempts to answer that question by determining the trade-off between exponential decay rates of $P_e$ and $\min_{m \in \mathcal{M}} P_{e|m}$. The operational definition of the problem in terms of reliable sequences is as follows.

**Definition 10 (Single Message Message-wise $UEP$ Problem):** For any reliable sequence $Q$ the missed detection exponent of the reliable sequence $Q$ is defined as
\[
E_{md,Q} = \liminf_{\kappa \to \infty} \frac{-\ln \min_{m \in \mathcal{M}(\kappa)} P_{e|m}^{(\kappa)} E(\kappa) || T(\kappa)}}{E(\kappa)||T(\kappa)} 
\tag{12}
\]
where $P_{e|m}^{(\kappa)}$ is the conditional error probability of the message $m$ for the $\kappa^{th}$ code of the reliable sequence $Q$.

For any rate $R \in [0, C]$ and error exponent $E \in [0, (1 - \frac{R}{C}) D]$, the missed detection exponent $E_{md}(R, E)$ is defined as,
\[
E_{md}(R, E) \triangleq \sup_{Q: E_Q \geq E} E_{md,Q}. 
\tag{13}
\]
In variable length block codes with feedback, the single message message wise UEP problem not only answers a curious question about the decay rate of the minimum conditional error probability of a code, but also plays a key role in the bit-wise UEP problem, which is our main focus in this manuscript.

Though they are central in the message-wise UEP problems, the conditional error probabilities of the messages are not relevant in the bit-wise UEP problems. In the bit-wise UEP problems we analyze the error probabilities of groups of sub-messages. In order to do that, consider a code with a message set \( \mathcal{M} \) of the form
\[
\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2 \times \ldots \times \mathcal{M}_\ell
\]
Then the transmitted message \( M \) and decoded message \( \hat{M} \) of the code are of the form
\[
M = (M_1, M_2, \ldots, M_\ell) \\
\hat{M} = (\hat{M}_1, \hat{M}_2, \ldots, \hat{M}_\ell)
\]
where \( M_j, \hat{M}_j \in \mathcal{M}_j \) for all \( j = 1, 2, \ldots, \ell \). Furthermore \( M_j \) and \( \hat{M}_j \) are called \( j^{th} \) transmitted sub-message and \( j^{th} \) decoded sub-message, respectively.

The error probabilities we are interested in correspond to erroneous transmission of certain parts of the message. In order to define them succinctly let us define \( \mathcal{M}^j \), \( M^j \) and \( \hat{M}^j \) for all \( j \) between one and \( \ell \) as follows:
\[
\mathcal{M}^j \triangleq \mathcal{M}_1 \times \mathcal{M}_2 \times \ldots \times \mathcal{M}_j \\
M^j \triangleq (M_1, M_2, \ldots, M_j) \\
\hat{M}^j \triangleq (\hat{M}_1, \hat{M}_2, \ldots, \hat{M}_j).
\]
Then \( P_e(j) \) is defined\(^{13}\) as the probability of the event that \( \hat{M}^j \neq M^j \)
\[
P_e(j) \triangleq \Pr[\hat{M}^j \neq M^j] \quad \text{for } j = 1, 2, \ldots, \ell. \tag{14}
\]
Note that if \( \hat{M}^j \neq M^j \) then \( \hat{M}^i \neq M^i \) for all \( i \) greater than \( j \). Thus
\[
P_e(1) \leq P_e(2) \leq P_e(3) \leq \ldots \leq P_e(\ell). \tag{15}
\]

**Definition 11** (Bit-wise UEP Problem For Fixed \( \ell \)): For any positive integer \( \ell \) let \( Q \) be a reliable sequence whose message sets \( \mathcal{M}^{(\kappa)} \) are of the form \( \mathcal{M}^{(\kappa)} = \mathcal{M}_1^{(\kappa)} \times \mathcal{M}_2^{(\kappa)} \times \ldots \times \mathcal{M}_\ell^{(\kappa)} \). Then the entries of the rate vector \( \vec{R}_Q \) and the error exponent vector \( \vec{E}_Q \) are defined as
\[
R_{Q,j} \triangleq \liminf_{\kappa \to \infty} \frac{\ln |\mathcal{M}_j^{(\kappa)}|}{E^{(\kappa)} |T^{(\kappa)}} \quad \forall j \in \{1, 2, \ldots, \ell\} \\
E_{Q,j} \triangleq \liminf_{\kappa \to \infty} \frac{-\ln P_e(j^{(\kappa)})}{E^{(\kappa)} |T^{(\kappa)}} \quad \forall j \in \{1, 2, \ldots, \ell\}.
\]
A rates-exponents vector \( (\vec{R}, \vec{E}) \) is achievable if and only if there exists a reliable sequence \( Q \) such that \( (\vec{R}, \vec{E}) = (\vec{R}_Q, \vec{E}_Q) \).

This definition of the bit-wise UEP problem is slightly different than the one described in the introduction, because \( P_e(j) \) is defined as \( \Pr[\hat{M}^j \neq M^j] \) rather than \( \Pr[\hat{M}_j \neq M_j] \). Note that if \( \hat{M}_j \neq M_j \) then \( \hat{M}^j \neq M^j \); consequently \( \Pr[\hat{M}^j \neq M^j] \geq \Pr[\hat{M}_j \neq M_j] \) for all \( j \)'s. In addition, if we assume without loss of generality that \( \Pr[\hat{M}_j \neq M_j] \geq \Pr[\hat{M}_i \neq M_i] \) for all \( j \geq i \), the union bound implies that \( \Pr[\hat{M}^j \neq M^j] \leq j \Pr[\hat{M}_j \neq M_j] \). Thus for the case when \( \ell \) is fixed, both formulations of the problem result in exactly the same achievable region of rates-exponents vectors.

The achievable region of rates-exponents vectors could have been defined as the closure of the points of the form \( (\vec{R}_Q, \vec{E}_Q) \) for some reliable sequence \( Q \). Using the definition of \( (\vec{R}_Q, \vec{E}_Q) \)'s one can easily show that, in this case too both definitions result in exactly the same achievable region of rates-exponents vectors.

\(^{13}\) Similar to the conditional error probabilities, \( P_{e/m} \)’s for \( m \in \mathcal{M}_j \) error probabilities of sub-messages, \( P_e(j) \)'s for \( j = 1, 2, \ldots, \ell \), are defined slightly differently for codes with erasures, see equation \( (10) \).
B. Main Results

For variable length block codes with feedback, the results of both the single message message-wise UEP problem and the bit-wise UEP problem are given in terms of the \( J(R) \) function defined below. The \( J(R) \) function is first introduced by Kudryashov \(^{19} \) while describing the performance of non-block variable length codes with feedback and delay constraints. Later the \( J(R) \) function is used in \(^{2} \) for describing the performance of block codes in single message message-wise UEP problem. It is shown in \(^{2} \) Appendix D) that for both fixed length block codes without feedback and variable length block codes with feedback on DMCs satisfy,

\[
E_{\text{md}}(R, 0) = J(R). \quad (16)
\]

Recently Nazer, Shkel and Draper obtained the closed form expression for \( E_{\text{md}}(R, 0) \) for fixed length block codes on the Additive White Gaussian Noise channel, under certain average and peak power constraints \(^{9} \) Theorem 1. Curiously equality given in (16) holds in that case too.

**Definition 12:** For any \( R \in [-\infty, C] \), \( J(R) \) is defined as

\[
J(R) \triangleq \max_{\alpha, x_1, x_2, \mu_1, \mu_2: 0 \leq \alpha \leq 1, x_1, x_2 \in \mathcal{X}, \mu_1, \mu_2 \in \mathcal{P}(\mathcal{X})} \alpha D(\bar{\mu}_1 \mid W_{x_1}) + (1 - \alpha) D(\bar{\mu}_2 \mid W_{x_2}) \quad (17)
\]

where \( \bar{\mu}_i(y) = \sum_x W_x(y) \mu_i(x) \) for \( i = 1, 2 \).

We have plotted the \( J(R) \) function for Binary Symmetric Channels\(^{21} \) (BSCs) with various cross over probabilities in Figure 1. Note that as the channel becomes noisier, i.e., as the crossover probability becomes closer to 1/2, the value of \( J(R) \) function decreases at all values of rate where it is positive. Furthermore the highest value of rate where it is positive, i.e., the channel capacity, decreases.

\(^{19} \) In \(^{7} \) equation (2.6) there is no optimization over the parameter \( \alpha \). Thus strictly speaking, what is introduced in \(^{7} \) equation (2.6) is \( j(R) \) given in equation (64) rather than \( J(R) \) given in (17).

\(^{20} \) Unlike DMC for these channels it is possible to obtain a closed form expression in terms of the rate and the power constraints.

\(^{21} \) Recall that in a binary symmetric channel with crossover probability probability \( p \), \( \mathcal{X} = \{0, 1\} \), \( \mathcal{Y} = \{0, 1\} \) and \( W_x(y) = (1 - p) \mathbb{I}_{x\neq y} + p \mathbb{I}_{x=y} \).
Lemma 1: The function $J(R)$ defined in equation (17) is a concave, decreasing function such that $J(R) = D$ for $R \leq 0$.

Proof of Lemma 1 is given in Appendix A.

Now let us consider the single message message-wise UEP problem given in Definition 10.

Theorem 1: For any rate $0 \leq R \leq C$ and error exponent $E \leq (1 - \frac{R}{C})D$ the missed detection exponent $E_{md}(R, E)$ defined in equation (13) is equal to

$$E_{md}(R, E) = E + (1 - E)J\left(\frac{R}{1 - \frac{R}{C}}\right)$$  \hspace{1cm} (18)

where $C$, $D$ and $J(\cdot)$ are given in equations (8), (9) and (17), respectively. Furthermore $E_{md}(R, E)$ is jointly concave in $(R, E)$ pairs.

We have plotted $E_{md}(R, E)$ as a function of rate, for various values of $E$ in Figure 2. When rate is zero, the exponent of the average error probability can be made as high as $D$. Thus all the curves meet at $(0, D)$ point. But for all positive rates the exponent of the average error probability makes a difference; as $E$ increases $E_{md}(R, E)$ decreases. Furthermore for any given rate $R$ the exponent of the average error probability can only be as high as $(1 - \frac{R}{C})D$. This is why the curves corresponding to higher values of $E$ have smaller support on rate axis.

Proof of Theorem 1 is presented in Appendix I.

Similar to the single message message-wise UEP problem, the solution of the bit-wise UEP problem is given in terms of the $J(R)$ function.

\footnote{For the case when $R = 0$ and $E = D$ the $\left(1 - \frac{E}{D}\right)J\left(\frac{R}{1 - \frac{R}{C}}\right)$ term should be interpreted as 0, i.e., $\left(1 - \frac{E}{D}\right)J\left(\frac{R}{1 - \frac{R}{C}}\right)\bigg|_{R=0, E=D} = 0$.}
and error exponent\( \eta \). Equation (19c) of a rate vector error exponent vector pair can be turned into an analytical expression for the optimal rates-exponents vectors is convex. Proof of Theorem 2 is presented in Appendix J.

For the special case when there are only two sub-messages the condition given in Theorem 2 for the achievability of a rate vector error exponent vector pair can be turned into an analytical expression for the optimal \( E_1 \) in terms of \( R_1, R_2 \) and \( E_2 \). In order to see why, note that revealing the region of achievable rates-exponents vectors is convex. Furthermore the set of all achievable rates-exponents vectors is convex.

\[ E_i \leq (1 - \sum_{j=1}^{\ell} \eta_j)D + \sum_{j=i+1}^{\ell} \eta_j J \left( \frac{R_j}{\eta_j} \right) \quad \forall i \in \{1, 2, \ldots, \ell\} \] (19a)

\[ R_i \leq C\eta_i \quad \forall i \in \{1, 2, \ldots, \ell\} \] (19b)

\[ \eta_i \geq 0 \quad \forall i \in \{1, 2, \ldots, \ell\} \] (19c)

\[ \sum_{j=1}^{\ell} \eta_j \leq 1 \] (19d)

where \( C, D \) and \( J(\cdot) \) are given in equations (8), (9) and (17), respectively. Furthermore the set of all achievable rates-exponents vectors is convex.

Proof of Theorem 2 is presented in Appendix J.

For the special case when there are only two sub-messages the condition given in Theorem 2 for the achievable rate vector error exponent vector pair can be turned into an analytical expression for the optimal \( E_1 \) in terms of \( R_1, R_2 \) and \( E_2 \). In order to see why, note that revealing the region of achievable \( (R_1, R_2, E_1, E_2) \) vectors is equivalent to revealing the region of achievable \( (R_1, R_2, E_2) \)'s and the value of the maximum achievable \( E_1 \) for all the \( (R_1, R_2, E_2) \)'s in the achievable region.

\[ \text{Corollary 1: } \text{For any rate pair } (R_1, R_2) \text{ such that } R_1 + R_2 \leq C \text{ and error exponent } E_2 \text{ such that } E_2 \leq (1 - \frac{R_1 + R_2}{C})D, \text{ the optimal value of } E_1 \text{ is given by} \]

\[ E_1(R_1, R_2, E_2) = E_2 + \left( 1 - \frac{R_1}{C} - \frac{E_2}{D} \right) J \left( \frac{R_2}{1 - \frac{R_1}{C} - \frac{E_2}{D}} \right) \] (20)

where \( C, D \) and \( J(\cdot) \) are given in equations (8), (9) and (17), respectively. Furthermore \( E_1(R_1, R_2, E_2) \) is concave in \( (R_1, R_2, E_2) \).

\[ \text{For the case when } R_j = 0 \text{ and } \eta_j = 0 \text{ the } \eta_j J \left( \frac{R_j}{\eta_j} \right) \text{ term should be interpreted as } 0, \text{ i.e., } \eta_j J \left( \frac{R_j}{\eta_j} \right) \bigg|_{R_j = 0} = 0. \]

\[ \text{For the case when } R_2 = 0 \text{ and } E_2 = \left( 1 - \frac{R_1}{C} \right)D, \text{ the second term on the right hand side of equation (20) should be interpreted as zero, i.e., } (1 - \frac{R_1}{C} - \frac{E_2}{D}) J \left( \frac{R_2}{1 - \frac{R_1}{C} - \frac{E_2}{D}} \right) \bigg|_{R_2 = 0, E_2 = (1 - \frac{R_1}{C})D} = 0. \]

Fig. 3. \( E_1(R_1, R_2, E_2) \) is drawn for various values rate pairs \( (R_1, R_2) \) as a function error exponent \( E_2 \) for a BSC with crossover probability \( p = 0.01 \). Recall that when \( p = 0.01 \), \( C = 0.6371 \) Nats per channel use and \( D = 4.503 \).
Note that for the \( E_1(R_1, R_2, E_2) \) given in equation (20), \( E_1(R_1, R_2, E_2) \geq E_2 \) for all \((R_1, R_2, E_2)\) triples such that \( R_1 + R_2 \leq C \) and \( E_2 \leq (1 - \frac{R_2 + R_2}{C})D \). Furthermore inequality is strict as long as \( R_2 > 0 \). We have drawn \( E_1(R_1, R_2, E_2) \) for various \((R_1, R_2)\) pairs as a function of \( E_2 \) in Figure 3.

IV. Achievability

In both the single message message-wise UEP problem and the bit-wise UEP problem, the codes that achieve the optimal performance employ a number of different ideas at the same time. In order to avoid introducing all of those ideas at once, we first describe two families of codes and analyze the probabilities of various error events in those two families of codes. Later we use those two families of codes as the building blocks for the codes that achieve the optimal performance in the UEP problems we are interested in. Before going into a more detailed description and analysis of those codes let us first give a birds eye plan for this section.

(a) A Single Message Message-wise UEP Scheme without Feedback: First in Section [V-A] we consider a family of fixed length codes without feedback. We prove that these codes can achieve any rate \( R \) less than channel capacity, with vanishing\(^{25} \) error probability \( P_e \) while having a minimum conditional error probability, \( \min_{m} P_{e|m} \), as low as \( e^{-n(R)} \). The main drawback of this family of codes is that the decay rate of the average error probability \( P_e \) has to be subexponential in this family of codes.

(b) Control Phase and Error-Erasure Decoding: In Section [V-B] in order to obtain non-zero exponential decay for the average error probability, we use an idea introduced by Yamamoto and Itoh in [12]. We append the fixed length codes described in Section [V-A] with a control phase and use an error-erasure decoder. This new family of codes with control phase and error-erasure decoding are shown, in Section [V-B], to achieve any rate \( R \) less than the channel capacity \( C \) with exponentially decaying average error probability \( P_e \), exponentially decaying minimum conditional error probability \( \min_{m} P_{e|m} \) and vanishing erasure probability, \( P_x \).

(c) Single Message Message-wise UEP for Variable Length Codes: In Section [V-C] we obtain variable length codes for single message message-wise UEP problem using the codes described in Section [V-B]. In order to do that we use the fixed length codes with feedback and erasures described in Section [V-B] repetitively until a non-erasure decoding happens. This idea too, was employed by Yamamoto and Itoh in [12].

(d) Bit-wise UEP for Variable Length Codes: In Section [V-D] we first use the codes described in Section [V-A] and the control phase discussed in Section [V-B] to obtain a family of fixed length codes with feedback and erasures which has bit-wise UEP, i.e., which has different bounds on error probabilities for different sub-messages. While using the codes described in Section [V-A] we employ an implicit acceptance explicit rejection scheme first introduced in [7] by Kudrayshov. Once we obtain a fixed length code with erasures and bit-wise UEP, we use a repeat at erasures scheme like the one described in Section [V-C] to obtain a variable length code with bit-wise UEP.

The achievability results we derive in this section are revealed to be the optimal ones, in terms of the decay rates of error probabilities with expected decoding time \( \mathbb{E}[T] \), as a result of the outer bounds we derive in Section [V].

A. A Single Message Message-wise UEP Scheme without Feedback

In this subsection we describe a family of fixed length block codes without feedback that achieves any rate \( R \) less than capacity with small error probability while having an exponentially small \( \min_{m} P_{e|m} \), for sufficiently large block length \( n \). We describe these codes in terms of a time sharing constant \( \alpha \in [0, 1] \), two input letters \( x_1, x_2 \in X \) and two probability distributions on the input alphabet, \( \mu_1, \mu_2 \in \mathcal{P}(X) \).

In order to point out that certain sequence of input letters is a codeword or part of a codeword for message \( m \), we put \( (m) \) after it. Hence we denote the codeword for \( m \) by \( x^n(m) \) in a given code and by \( X^n(m) \) in a code ensemble, as a random quantity.

Let us start with describing the encoding scheme. The codeword of the first message, i.e., \( x^n(1) \), is \( x_1 \) in first \( \lfloor \alpha n \rfloor \) time instances and \( x_2 \) in the rest, i.e., \( x_{\tau}(1) = x_1 \) for \( \tau = 1, \ldots, \lfloor \alpha n \rfloor \) and \( x_{\tau}(1) = x_2 \) for \( \tau = (\lfloor \alpha n \rfloor + 1), \ldots, n \). The codewords of the other messages are described via a random coding argument. In the ensemble of codes we are considering all entries of all codewords other than the first codeword, i.e., \( X_{\tau}(m) \forall \tau \in [1, n], \forall m \neq 1 \), are generated independently of other codewords and other entries of the same codeword. In the first \( \lfloor \alpha n \rfloor \) time instances \( X_{\tau}(m) \) is

\(^{25}\)Vanishing with increasing block length.
generated using $\mu_1$, in the rest using $\mu_2$, i.e., $P[X_r(m) = x] = \mu_1(x)$ for $\tau = 1, \ldots, n_\alpha$ and $P[X_r(m) = x] = \mu_2(x)$ for $\tau = (n_\alpha + 1), \ldots, n$.

Let us begin the description of the decoding scheme, by specifying the decoding region of the first message $G[1]$: it is the set of all output sequences $y^n$ whose the empirical distribution is not typical with $(\alpha, \mu_1, \mu_2)$. More precisely, the decoding region of the first message, $G[1]$, is given by,

$$G[1] = \left\{ y^n : n_\alpha \Delta(Q_{\{y_n^n\}}, \mu_1) + (n - n_\alpha) \Delta(Q_{\{y^n_{n_\alpha+1}\}}, \mu_2) \geq |X||Y|\sqrt{n \ln(1+n)} \right\} \tag{21}$$

where $\Delta$ is the total variation distance defined in equation (1), $Q_{\{y_n^n\}}$ and $Q_{\{y^n_{n_\alpha+1}\}}$ are the empirical distributions of $y_1^n$ and $y_{n_\alpha+1}^n$ defined in equation (2) and $\mu_1$ and $\mu_2$ are probability distributions on $\mathcal{Y}$, i.e., $\mu_1, \mu_2 \in \mathcal{P}(\mathcal{Y})$, such that $\mu_i(y) = \sum_x \mu_i(x) W_x(y)$.

For other messages, $m \neq 1$, decoding regions $G[m]$ are the set of all output sequences for which $Q_{\{x^e(m), y^n\}}$ is typical with $(\alpha, \mu_1 W, \mu_2 W)$ and $Q_{\{x^n(m), y^n\}}$ is not typical with $(\alpha, \mu_1 W, \mu_2 W)$ for any $m \neq m$. To be precise the decoding region of the messages other than the first message are

$$G[m] = B[x^n(m)] \cap \left( \cap_{\hat{m}\neq m} B[x^n(\hat{m})] \right) \quad \forall m \in \{2, 3, \ldots, |\mathcal{M}|\} \tag{22}$$

where for all $x^n \in \mathcal{X}^n$, $B[x^n]$ is the set of all $y^n$'s for which $(x^n, y^n)$ is typical with $(\alpha, \mu_1 W, \mu_2 W)$:

$$B[x^n] = \left\{ y^n : n_\alpha \Delta(Q_{\{x_1^n, y_1^n\}}, \mu_1 W) + (n - n_\alpha) \Delta(Q_{\{x_1^n, y_1^n\}}, \mu_2 W) \leq |X||Y|\sqrt{n \ln(1+n)} \right\} \tag{23}$$

where $\Delta$ is the total variation distance defined in equation (1), $Q_{\{x_1^n, y_1^n\}}$ and $Q_{\{x_1^n, y_1^n\}}$ are the empirical distributions of $(x_1^n, y_1^n)$ and $(x_{n_\alpha+1}^n, y_{n_\alpha+1})$ defined in equation (2) and $\mu_1 W$ and $\mu_2 W$ are probability distributions on $\mathcal{X} \times \mathcal{Y}$, i.e., $\mu_1 W, \mu_2 W \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$.

In Appendix B we have analyzed the conditional error probabilities, $P_{e|m}$ for the above described code and proved Lemma 2 given below.

Lemma 2: For any block length $n$, time sharing constant $\alpha \in [0, 1]$, input letters $x_1, x_2 \in \mathcal{X}$ and input distributions $\mu_1, \mu_2 \in \mathcal{P}(\mathcal{X})$ there exists a length $n$ block code such that

$$|\mathcal{M}| \geq e^{n(\alpha I(\mu_1, W) + (1-\alpha) I(\mu_2, W) - \varepsilon_n)}$$

$$P_{e|1} \leq e^{-n(\alpha D(\mu_1 || W_{x_1}) + (1-\alpha) D(\mu_2 || W_{x_2}) - \varepsilon_n)}$$

$$P_{e|m} \leq \varepsilon_n \quad m = 2, 3, \ldots, |\mathcal{M}|$$

where $\bar{\mu}_1(y) = \sum_x W_x(y) \mu_1(x)$, $\bar{\mu}_2(y) = \sum_x W_x(y) \mu_2(x)$ and $\varepsilon_n = \frac{10|\mathcal{X}||\mathcal{Y}|(\ln \frac{1}{\varepsilon_n}) \ln(1+n)}{\sqrt{n}}$.

Given the channel $W$, if we discard the error terms $\varepsilon_n$, for a given value of rate, $0 \leq R \leq C$, we can can optimize exponent of $P_{e|1}$ over the time sharing constant $\alpha$, the input letters $x_1, x_2$ and input distributions $\mu_1, \mu_2$. Evidently the optimization problem we get is the one given for the definition of $J(R)$, in equation (17). Thus Lemma 2 implies that for any $R \in [0, C]$ and block length $n$ there exists a length $n$ code such that $|\mathcal{M}| \geq e^{n(R - \varepsilon_n)}$, $P_{e|m} \leq \varepsilon_n$ for $m = 2, 3, \ldots, |\mathcal{M}|$ and $P_{e|1} \leq e^{-n(R(1-R) - \varepsilon_n)}$.

One curious question is whether or not the exponent of $P_{e|1}$ can be increased by including more than two phases. Carathéodory’s Theorem answers that question negatively, i.e., to obtain the largest value of $J(R)$ one doesn’t need to do time sharing between more than two input-letter-input-distribution pairs.

B. Control Phase and Error-Erasure Decoding:

The family of codes described in Lemma 2 has a large exponent for the conditional error probability of the first message, i.e., $P_{e|1}$. But the conditional error probabilities of other messages, $P_{e|m}$ for $m \neq 1$, decay subexponentially. In order to facilitate an exponential decay of $P_{e|m}$ for $m \neq 1$ with block length, we append the codes described in Lemma 2 with a control phase and allow erasures. The idea of using a control phase and an error-erasure decoding, in establishing achievability results for variable length code, was first employed by Yamamoto and Itoh in [12].

In order explain what we mean by the control phase, let us describe our encoding scheme and decoding rule briefly. First a code from the family of codes described in Section C is used to transmit $M$ and the receiver makes a tentative decision $\hat{M}$ using the decoder of the very same code. The transmitter knows $\hat{M}$ because of the feedback link. In the remaining time instances, i.e., in the control phase, the transmitter sends the input letter $a$ if
\( \hat{M} \) = \( M \), the input letter \( r \) if \( \hat{M} \neq M \). The input letters \( a \) and \( r \) are described in equation (11). At the end of the control phase, the receiver checks whether or not the output sequence in the control phase is typical with \( W_a \), if it is then \( \hat{M} = \hat{M} \) otherwise an erasure is declared.

Lemma 3 given below states the results of the performance analysis of the above described code. In order to understand what is stated in Lemma 3, accurately, let us make a brief digression and elaborate on the codes with erasure. We have assumed in our models until now that \( \hat{M} \in \mathcal{M} \). However, there are many interesting problems in which this might not hold. In codes with erasures for example, we replace \( \hat{M} \in \mathcal{M} \) with \( \hat{M} \in \mathcal{M} \cup \{ x \} \) and \( x \) is the erasure symbol. Furthermore in codes with erasures for each \( m \in \mathcal{M} \) the conditional error probability \( P_{e|m} \) and conditional erasure probability, \( P_{x|m} \) are defined as follows.

\[
P_{e|m} = P\left[ \hat{M} \notin \{ m, x \} \mid M = m \right] \quad m = 1, 2, \ldots, |\mathcal{M}| \quad (24a)
\]

\[
P_{x|m} = P\left[ \hat{M} = x \mid M = m \right] \quad m = 1, 2, \ldots, |\mathcal{M}| \quad (24b)
\]

Note that definitions of \( P_{e|m} \) and \( P_{x|m} \) given above can be seen as the generalizations of the corresponding definitions in block codes without erasures. In erasure free codes above definitions are equivalent to corresponding definitions there.

**Lemma 3:** For any block length \( n \), rate \( 0 \leq R \leq C \) and error exponent \( 0 \leq E \leq (1 - \frac{R}{C})D \), there exists a length \( n \) block code with erasures such that,

\[
|\mathcal{M}| \geq e^{n(R - \varepsilon_n)}
\]

\[
P_{e | 1} \leq e^{-n\left(E + (1 - \frac{E}{C}) \varepsilon_n \right) \frac{1}{2} \left(n \frac{R}{C} \right)}
\]

\[
P_{e | m} \leq \varepsilon_n \min\{ 1, e^{-n(E - \varepsilon_n)} \} \quad m = 2, 3, \ldots, |\mathcal{M}|
\]

\[
P_{x | m} \leq \varepsilon_n + e^{-n\left((1 - \frac{R}{C}) \varepsilon_n \right) \frac{1}{2} \left(n \frac{R}{C} \right)} \quad m = 1, 2, \ldots, |\mathcal{M}|
\]

where \( \varepsilon_n = \frac{100 |X| |Y| (\ln \frac{1}{\varepsilon}) \sqrt{n}}{\sqrt{1 + n}} \).

Proof of Lemma 3 is given in Appendix C.

Note that in Lemma 3 unlike \( P_{e | m} \)'s which decrease exponentially with \( n \), \( P_{x | m} \)'s decays as \( \frac{\ln n}{\sqrt{n}} \). It is possible to tweak the proof so as to have a non-zero exponent for \( P_{x | m} \)'s, see [8]. But this can only be done at the expense of \( P_{e | m} \)'s. Our aim, however, is achieving the optimal performance in variable length block codes. As we will see in the following subsection, for that what matters is exponents of error probabilities and having vanishing erasure probabilities. The rate at which erasure probability decays does not effect the performance of variable length block codes in terms of error exponents.

**C. Single Message Message-Wise UEP Achievability:**

In this section we construct variable length block codes for the single message message-wise UEP problem using Lemma 3. In first \( n \) time units the variable length encoding scheme uses a fixed length block code with erasures which has the performance described in Lemma 3. If the decoded message of the fixed length code is in the message set, i.e., if \( M \in \mathcal{M} \) then decoded message of the fixed length code becomes the decoded message of the variable length code. If the decoded message of the fixed length code is the erasure symbol, i.e., if \( \hat{M} = x \), then the encoder uses the fixed length code again in the second \( n \) time units. By repeating this scheme until the decoded message of the fixed length code is in \( \mathcal{M} \), i.e., \( \hat{M} \in \mathcal{M} \), we obtain a variable length code.

Let \( L \) be the number of times the fixed length code is used until a \( \hat{M} \in \mathcal{M} \) is observed. Then given the message \( M \), \( L \) is a geometrically distributed random variable with success probability \( 1 - P_{x | M} \) where \( P_{x | M} \) is the conditional erasure probability of the fixed length code given the message \( M \). Then the conditional probability distribution and the conditional expected value of \( L \) given \( M \) are

\[
P[L = l | M] = (1 - P_{x | M}) (P_{x | M})^{l-1} \quad l = 1, 2, \ldots \quad (25a)
\]

\[
E[L | M] = (1 - P_{x | M})^{-1} \quad (25b)
\]
Furthermore the conditional expected value of decoding time and the conditional error probability given the message $M$ are

$$E[T|M] = nE[L|M]$$  

(26a)

$$P\left[\widehat{M} \neq M \bigg| M\right] = P_e|M E[L|M]$$  

(26b)

where $n$ is the block length of the fixed length code and $P_e|M$ is the conditional error probability given the message $M$ for the fixed length code.

Thus as result of equations (25b), (26) and Lemma 3 we know that for any rate $R \in [0, C]$, error exponent $E \in [0, (1 - \frac{R}{C})D]$ there exists a reliable sequence $Q$ such that $R_Q = R$, $E_Q = E$ and

$$E_{mdQ} = E + \left(1 - \frac{E}{R}\right)J\left(\frac{R}{1-E/D}\right).$$

(27)

We show in Section (V-C) that for any reliable sequence $Q$ with rate $R_Q = R$ and error exponent $E_Q = E$, $E_{md,Q}$ is upper bounded by the expression on the right hand side of equation (27).

### D. Bit-Wise UEP Achievability:

In this section we first use the family of codes described in Section [IV-A] and the control phase idea described in Section [IV-B] to construct fixed length block codes with erasures which have *bit-wise UEP*. Then we use them with a repeat until non-erasure decoding scheme, similar to the one described in Section [IV-C] to obtain variable length block codes with *bit-wise UEP*.

Let us start with describing the encoding scheme for the fixed length block code with *bit-wise UEP*. If there are $\ell$ sub-messages, i.e., if $M = (M_1 \times M_2 \times \cdots \times M_\ell)$, then the encoding scheme has $\ell + 1$ phases with lengths $n_1$, $n_2$, ..., $n_{\ell+1}$ such that $n_1 + n_2 + \ldots + n_{\ell+1} = n$.

- In the first phase a length $n_1$ code from the family of codes described in Section [IV-A] is used. The message set of the code $\mathcal{M}_1$ is $\mathcal{M}_1 \cup \{|M_1| + 1\}$ and the message $\mathcal{M}_1$ of the code is determined by the first sub-message: $\mathcal{M}_1 = M_1 + 1$. At the end of first phase receiver uses the decoder of the length $n_1$ code to get a tentative decision $\mathcal{M}_1$ which is known by the transmitter at the beginning of the second phase because of the feedback link.

- In the second phase a length $n_2$ code from the family of codes described in Section [IV-A] with the message set $\mathcal{M}_2 = M_2 \cup \{|M_2| + 1\}$, is used. If $\mathcal{M}$ is decoded correctly at the end of the first phase then the message $\mathcal{M}_2$ of the code used in the second phase is determined by the second sub-message as $\mathcal{M}_2 = M_2 + 1$, else $\mathcal{M}_2 = 1$. At the end of the second phase the receiver uses the decoder of the second phase code to get the tentative decision $\mathcal{M}_2$ which is known by the transmitter at the beginning of the third phase because of the feedback link.

- In phases 3 to $\ell$ above described scheme is used. In phase $i$, a length $n_i$ code, with the message set $\mathcal{M}_i = M_i \cup \{|M_i| + 1\}$, from the family of codes described in Section [IV-A] is used. The message of the length $n_i$ code $\mathcal{M}_i$ is $M_i + 1$ if $\mathcal{M}_{i-1} = M_{i-1}$, 1 otherwise for $i = 3, 4, \ldots, \ell$.

- The last phase is a $n_{\ell+1}$ long control phase, i.e., a $n_{\ell+1}$ long code with the message set $\mathcal{M}_{\ell+1} = \{1, 2\}$ is used in the last phase. The codewords for the first and second messages are $n_{\ell+1}$ long sequences of input letters $r$ and $a$ respectively, where $r$ and $a$ are described in equation (10). The tentative decision in the last phase $\mathcal{M}_{\ell+1}$ is equal to the first message if the output sequence in the last phase is not typical with $W_a$, the second message otherwise. The message of the $n_{\ell+1}$ long code $\mathcal{M}_{\ell+1}$ is equal to 2 if $\mathcal{M}_\ell = \mathcal{M}_\ell$, 1 otherwise.

Note that if we define $\mathcal{M}_0$, $\mathcal{M}_0$ and $M_{\ell+1}$ all to be 1, i.e., $\mathcal{M}_0 = M_0 = M_{\ell+1} = 1$ we can write the following rule for determining the $\mathcal{M}_i$’s for $i = 1$ to $\ell + 1$.

$$\mathcal{M}_i = 1 + I\{\mathcal{M}_{i-1} = M_{i-1}\} M_i \quad \quad \quad i = 1, 2, \ldots, (\ell + 1)$$

(28)

It is important however to keep in mind that the last phase is a control phase and the codes in the first $\ell$ phases are from the family of codes described in Section [IV-A].

Note that during the phases $i = 2$ to $\ell$ erroneous transmission of $\mathcal{M}_{i-1}$ is conveyed using $\mathcal{M}_i = 1$, hence the transmission of $\mathcal{M}_i$ through $\mathcal{M}_i$, i.e., $\mathcal{M}_i = 1 + \mathcal{M}_i$, is a tacit approval of the tentative decision $\mathcal{M}_{i-1}$. Because of this, the above encoding scheme is said to have an implicit acceptance explicit rejection property. The idea of
implicit acceptance explicit rejection was first introduced by Kudryashov in [7] in the context of non-block variable length codes with feedback and delay constraints.

After finishing the description of the encoding scheme, we are ready to describe the decoding scheme. The receiver determines the decoded message using the tentative decisions, $\hat{M}_i$ for $i = 1$ to $\ell + 1$. If one or more of the tentative decisions are equal to 1, then an erasure is declared. If all $\ell + 1$ tentative decision are different from 1 then $\hat{M}_i = \hat{M}_i - 1$ for all $i = 1, 2, \ldots, \ell$. Hence the decoding rule is

$$\left(\hat{M}_1, \hat{M}_2, \ldots, \hat{M}_\ell\right) = \left(\hat{M}_1 - 1, \hat{M}_2 - 1, \ldots, \hat{M}_\ell - 1\right) \times \begin{cases} \prod_{i=1}^{\ell+1} (\hat{M}_i - 1) > 0 & \text{if } \prod_{i=1}^{\ell+1} (\hat{M}_i - 1) = 0. \end{cases} \quad \text{(29)}$$

For bit-wise UEP codes with erasure, the definition of $P_{e}(i)$ is slightly different from the original one given in equation (14)

$$P_{e}(i) = P\left[\hat{M}_i \neq m_i, \hat{M} \neq x\right]. \quad \text{(30)}$$

With this alternative definition in mind let us define $P_{e|m}(i)$ as the conditional probability of the erroneous transmission any one of the first $i$ sub-message when $M = m$:

$$P_{e|m}(i) = P\left[\hat{M}_i \neq m_i, \hat{M} \neq x \mid M = m\right]. \quad \text{(31)}$$

The error analysis of the above described fixed length codes, presented in Appendix D leads to Lemma 4 given below.

**Lemma 4:** For block length $n$, any integer $\ell \leq \frac{n}{\ln(1+n)}$, rate vector $\vec{R}$, and time sharing vector $\vec{\eta}$ such that

$$R_i \leq C\eta_i \quad \forall i \in \{1, 2, \ldots, \ell\} \quad \text{(32a)}$$

$$\eta_i \geq 0 \quad \forall i \in \{1, 2, \ldots, \ell\} \quad \text{(32b)}$$

$$\sum_{i=1}^{\ell} \eta_i \leq 1 \quad \text{(32c)}$$

there exists a length $n$ block code such that:

$$|\mathcal{M}| \geq e^{n(R_i - \varepsilon_{n,i})} \quad \forall i \in \{1, 2, \ldots, \ell\} \quad \text{(33a)}$$

$$|\mathcal{M}| \geq e^{n(-\varepsilon_{n,i} + \sum_{j=1}^{\ell} R_j)} \quad \forall i \in \{1, 2, \ldots, \ell\} \quad \text{(33b)}$$

$$P_{e|m}(i) \leq \varepsilon_{n,\ell} n^{-(\varepsilon_{n,\ell} + \sum_{j=1}^{i-1} \eta_j - \sum_{j=i+1}^{\ell} \frac{\eta_j}{\eta_j})} \quad \forall m \in \mathcal{M}, i \in \{1, 2, \ldots, \ell\} \quad \text{(33c)}$$

where $\eta_{\ell+1} = 1 - \sum_{i=1}^{\ell} \eta_i$, $R_{\ell+1} = 0$, $\varepsilon_{n,\ell} = \frac{\ln(1+n)}{\sqrt{n}} \sqrt{1 + \ell}$.

Recall the repeat at erasures scheme described in Section V-C. If we use that scheme to obtain a variable length code from the fixed length bit-wise UEP code described in Lemma 4, we obtain a variable length code with UEP such that

$$E[T \mid M] = \frac{n}{1-P_{e|m}} \quad \text{(33a)}$$

$$P\left[\hat{M}_i \neq m_i \mid M\right] \leq \frac{P_{e|m}(i)}{1-P_{e|m}} \quad i = 1, 2, \ldots, \ell. \quad \text{(33b)}$$

As result of equation (33) and Lemma 4 we know that for any rate vector $\vec{R}$, error exponent vector $\vec{E}$ and time sharing vector $\vec{\eta}$ such that

$$E_i \leq (1 - \sum_{j=1}^{\ell} \eta_j)D + \sum_{j=i+1}^{\ell} \eta_j J\left(\frac{R_i}{\eta_i}\right) \quad \forall i \in \{1, 2, \ldots, \ell\} \quad \text{(34a)}$$

$$R_i \leq C\eta_i \quad \forall i \in \{1, 2, \ldots, \ell\} \quad \text{(34b)}$$

$$\eta_i \geq 0 \quad \forall i \in \{1, 2, \ldots, \ell\} \quad \text{(34c)}$$

$$\sum_{j=1}^{\ell} \eta_j \leq 1 \quad \text{(34d)}$$

there exists a reliable sequence $Q$ such that $(\vec{R}_Q, \vec{E}_Q) = (\vec{R}, \vec{E})$. Thus the existence of the time sharing vector $\vec{\eta}$ satisfying the constraints given in (34) is a sufficient condition for the achievability of a rates-exponents vector $(\vec{R}, \vec{E})$. We show in Section V-D that the existence of a time sharing vector $\vec{\eta}$ satisfying the constraints given in (34) is also a necessary condition for the achievability of a rates-exponents vector $(\vec{R}, \vec{E})$. 
V. Converse

Berlin et al. [1] used the error probability of a random binary query posed at a stopping time for bounding the error probability of a variable length block code. Later similar techniques have been applied in [2] for establishing outer bounds in UEP problems. Our approach is similar to that of [1] and [2]; we, too, use error probabilities of random queries posed at stopping times for establishing outer bounds. Our approach, nevertheless, is novel because of the error events we choose to analyze and the bounding techniques we use. Furthermore, the relation we establish in Lemma 5 between the error probabilities and the decay rate of the conditional entropy of the messages with time is a brand new tool for UEP problems.

For rigorously and unambiguously generalizing the technique used in [1] and [2] we introduce the concept of anticipative list decoders in Section V-A. Then in Section V-B we bound the probabilities of certain error events associated with anticipative list decoders from below. This bound, i.e., Lemma 5, is used in Sections V-C and V-D to derive tight outer bounds for the performance of variable length block codes in the single message message-wise UEP problem and in the bit-wise UEP problem, respectively.

A. Anticipative List Decoders

In this section we first introduce the concepts of anticipative list decoders and non-trivial anticipative list decoders. After that we show that for a given variable length code, any non-trivial anticipative list decoder \((\tilde{T}, \mathcal{A})\) can be used to define a probability distribution, \(P_{\{\mathcal{A}\}}\) on \(\mathcal{M} \times \mathcal{Y}^{\tilde{T}}\). Finally we use \(P_{\{\mathcal{A}\}}\) to define the probability measure \(P_{\{\mathcal{A}\}}[\cdot]\) for the events in \(\varphi(\mathcal{M} \times \mathcal{Y}^{\tilde{T}})\). Both the non-trivial anticipative list decoders \((\tilde{T}, \mathcal{A})\) and the probability measures \(P_{\{\mathcal{A}\}}[\cdot]\) associated with them play key roles in Lemma 5 of Section V-B.

An anticipative list decoder for a variable length code is a list decoder \(\mathcal{A}\) that decodes at a stopping time \(\tilde{T}\) that is always less than or equal to the decoding time of the code \(T\). The anticipative list decoders are used to formulate questions about the transmitted message or the decoded message, in terms of a subset of the message set \(\mathcal{M}\) that is chosen at a stopping time \(\tilde{T}\). For example let \(\mathcal{A}\) be the set of all \(m \in \mathcal{M}\) whose posterior probability at time one is larger than \(1/|\mathcal{M}|\). Evidently for all values of \(Y_1\), \(\mathcal{A}\) is a subset of \(\mathcal{M}\), but it is not necessarily the same subset for all values of \(Y_1\). Indeed \(\mathcal{A}\) is a function from \(\mathcal{Y}_1\) to the power set of \(\mathcal{M}\) and \((\tilde{T}, \mathcal{A})\) is an anticipative list decoder, for which \(\tilde{T} = 1\). Formal definition, for anticipative list decoders, is given below. In order to avoid separate treatment in certain special cases we include the case when \(\tilde{T} = 0\) and \(\mathcal{A}\) is fixed subset of \(\mathcal{M}\), in the definition.

**Definition 13 (Anticipative List Decoder):** For a variable length code with decoding time \(T\), a pair \((\tilde{T}, \mathcal{A})\) is called an anticipative list decoder (ALD) if

- either \(\tilde{T}\) is the constant random variable \(0\) and \(\mathcal{A}\) is a fixed subset of \(\mathcal{M}\), i.e.,
  \[
  \tilde{T} = 0 \\
  \mathcal{A} \in \varphi(\mathcal{M})
  \]

- or \(\tilde{T}\) is a stopping time, which is smaller than \(T\) with probability one, and \(\mathcal{A}\) is a \(\varphi(\mathcal{M})\) valued function defined on \(\mathcal{Y}^{\tilde{T}}\), i.e.,
  \[
  P[\tilde{T} \leq T] = 1 \\
  \mathcal{A} : \mathcal{Y}^{\tilde{T}} \rightarrow \varphi(\mathcal{M}).
  \]

Definition of ALD does not require \(\mathcal{A}\) to be of some fixed size, nor it requires \(\mathcal{A}\) to include more likely or less likely messages. Thus for certain values of \(Y^{\tilde{T}}\), \(\mathcal{A}\) might not include any \(m \in \mathcal{M}\) with positive posterior probability. In other words for some values of \(Y^{\tilde{T}}\) we might have

\[
P[M \in \mathcal{A}(Y^{\tilde{T}}) \mid Y^{\tilde{T}} = y^{\tilde{T}}] = 0.
\]

The ALD’s in which such \(y^{\tilde{T}}\)’s have zero probability are called nontrivial ALD’s.
Definition 14 (Nontrivial ALD): An anticipative list decoder $\overline{T}, A)$ is called a nontrivial anticipative list decoder (NALD) if $P[M \in A(Y^{\overline{T}}) | Y^{\overline{T}}] > 0$ with probability one, i.e.,

$$P[P[M \in A(Y^{\overline{T}}) | Y^{\overline{T}}] > 0] = 1. \quad (35)$$

Below, for any variable length code and an associated nontrivial anticipative list decoder $\overline{T}, A)$ we define a probability distribution $P_{\{\overline{A}\}}$ on $M \times Y^{T_{\overline{T}}}$ and a probability measure $P_{\{\overline{A}\}}[\cdot]$ for the events in $\varphi(M \times Y^{\overline{T}})$. For doing that first note that the probability measure generated by the code, i.e., $P[\cdot]$, can be used to define a probability distribution $P$ on $M \times Y^{T_{\overline{T}}}$ as follows:

$$P(m, y) \triangleq P[M = m, Y^{\overline{T}} = y] \quad \forall m \in M, y \in Y^{T_{\overline{T}}} \quad \text{(36)}$$

where $Y^{T_{\overline{T}}}$ is a countable set for any stopping time, given in equation (6b).

As $T$ is a stopping time, the probability of any event $\Gamma$ in $\varphi(M \times Y^{T_{\overline{T}}})$ under $P[\cdot]$, i.e., $P[\Gamma]$ is equal to

$$P[\Gamma] = \sum_{(m,y) \in \Gamma} P(m, y) \quad \text{Evidently we can extend the definition of } P \text{ and assume that } P \text{ is zero whenever } y \in Y^{T_{\overline{T}}} = Y_{T_{\overline{T}}} = \{0, \infty\}, \text{ i.e.,}$$

$$P(m, y) \triangleq 0 \quad \forall m \in M, y \in Y^{T_{\overline{T}}} = Y_{T_{\overline{T}}} = \{0, \infty\}. \quad \text{(38)}$$

This extension is neither necessary nor relevant for calculating the probabilities of the events in $\varphi(M \times Y^{T_{\overline{T}}})$, because $T$ is a stopping time, i.e., $P[T < \infty] = 1$.

Definition 15: Given a variable length code with decoding time $T$, for any NALD $\overline{T}, A)$ let $P_{\{\overline{A}\}}$ be\footnote{There is a slight abuse of notation in equation (39), if $\overline{T}$ is not a stopping time but rather a constant random variable $T = 0$, $P_{\{\overline{A}\}}(m, y)$ should be interpreted as}

$$P_{\{\overline{A}\}}(m, y) \triangleq \frac{P(m | y) \mathbb{I}_{\{m \in A(y)\}}}{\sum_{m \in M} P(m | y) \mathbb{I}_{\{m \in A(y)\}}} \quad \text{Evidently we can extend the definition of } P \text{ and assume that } P \text{ is zero whenever } y \in Y^{T_{\overline{T}}} = Y_{T_{\overline{T}}} = \{0, \infty\}, \text{ i.e.,}$$

$$P_{\{\overline{A}\}}(m, y) \triangleq 0 \quad \forall m \in M, y \in Y^{T_{\overline{T}}} = Y_{T_{\overline{T}}} = \{0, \infty\}. \quad \text{(38)}$$

Note that Definition 15 is a parametric definition in the sense that it assigns a $P_{\{\overline{A}\}}$ for all nontrivial anticipative list decoders $\overline{T}, A)$. While proving outer bounds we will employ not one but multiple NALD’s and use them in conjunction with our new result, i.e., Lemma 5. But before introducing Lemma 5 let us elaborate on the relations between marginal and conditional distributions of $P_{\{\overline{A}\}}$ and $P$.

For $P_{\{\overline{A}\}}$ defined in equation (39) we have

$$\sum_{m \in M, y \in Y^{T_{\overline{T}}}} P_{\{\overline{A}\}}(m, y) = 1. \quad \text{Hence } P_{\{\overline{A}\}} \text{ is a probability distribution on } M \times Y^{T_{\overline{T}}}, \text{ i.e., } P_{\{\overline{A}\}} \in \mathcal{P}(M \times Y^{T_{\overline{T}}}).$$

Note that the marginal distributions of $P_{\{\overline{A}\}}$ and $P$ are the same on $Y^{T_{\overline{T}}}$, Furthermore for all $y \in Y^{T_{\overline{T}}}$ and $m \in M$ the conditional distributions of $P_{\{\overline{A}\}}$ and $P$ are the same on $Y_{T_{\overline{T}}}$, The probability distributions $P_{\{\overline{A}\}}$ and $P$ differ only in their conditional distributions on $M$ given $y$. More specifically,

$$P_{\{\overline{A}\}}(y) = P(y) \quad \forall y \in Y^{T_{\overline{T}}} \quad \text{(40a)}$$

$$P_{\{\overline{A}\}}(m | y) = \frac{P(m | y) \mathbb{I}_{\{m \in A(y)\}}}{\sum_{m \in M} P(m | y) \mathbb{I}_{\{m \in A(y)\}}} \quad \forall y \in Y^{T_{\overline{T}}}, m \in M \quad \text{(40b)}$$

$$P_{\{\overline{A}\}}(y_{T_{\overline{T}}} | y, m) = P(y_{T_{\overline{T}}} | y, m) \quad \forall y \in Y^{T_{\overline{T}}}, m \in M. \quad (40c)$$
Using the parametric definition of probability distribution \( \mathcal{P}_{\{A\}} \) on \( \mathcal{M} \times \mathcal{Y}^T \) we define a probability measure \( \mathcal{P}_{\{A\}} (\cdot) \) for the events in \( \varphi(\mathcal{M} \times \mathcal{Y}^T) \) as follows:

\[
\mathcal{P}_{\{A\}} (\Gamma) \triangleq \sum_{(m,y^T) \in \Gamma \cap (\mathcal{M} \times \mathcal{Y}^T)} \mathcal{P}_{\{A\}} (m,y^T) \quad \forall \Gamma \in \varphi(\mathcal{M} \times \mathcal{Y}^T). \tag{41}
\]

Evidently we can extend the definition of \( \mathcal{P}_{\{A\}} \) to \( \mathcal{M} \times \mathcal{Y}^T \) by defining it to be zero on \( \mathcal{M} \times \mathcal{Y}^\infty \)_(\text{T=}\infty)_1, i.e.,

\[
\mathcal{P}_{\{A\}} (m,y^T) \triangleq 0 \quad \forall m \in \mathcal{M}, y^T \in \mathcal{Y}^\infty \_\text{\textit{T=}\infty}_1. \tag{42}
\]

As in the case of \( \mathcal{P} \), this extension is neither necessary nor relevant for calculating the probabilities \( \mathcal{P}_{\{A\}} (\Gamma) \) given in equation \( \text{(41)} \).

### B. Error Probability and Decay Rate of Entropy:

In this section we lower bound the probability of the event that the decoded message \( \tilde{M} \) is in \( \mathcal{A} \) under the probability measure \( \mathcal{P}_{\{A\}} (\cdot) \), i.e., \( \mathcal{P}_{\{A\}} (\tilde{M} \notin \mathcal{A}(Y^T)) \). The bounds we derive depend on the decay rate of the conditional entropy of the messages in the interval between \( \tilde{T} \) and \( T \).

Before even stating our bound, we need to specify what we mean by the conditional entropy of the messages. While defining the conditional entropy, many authors do take an average over the sample values of the conditioned random variable and obtain a constant. We, however, do not take an average over the conditioned random variable and define conditional entropy as a random variable itself, which is a function of the random variable that is conditioned on \( \mathcal{F} \):

\[
H(M|Y^T) \triangleq \sum_{m \in \mathcal{M}} \mathcal{P}[M=m|Y^T] \ln \frac{1}{\mathcal{P}[M=m|Y^T]}.
\]

Using the probability distribution \( \mathcal{P} \) defined in equation \( \text{(36)} \) we see that the conditional entropy defined in \( \text{(43)} \) is equal to,

\[
H(M|Y^T) = \mathbb{E} \left[ \ln \frac{1}{\mathcal{P}[M|Y^T]} \right] Y^T. \tag{44}
\]

**Lemma 5:** For any variable length block code with finite expected decoding time, \( \mathbb{E}[T] < \infty \), let \( (T_1, A_1), (T_2, A_2), \ldots, (T_k, A_k) \) be \( k \) NALD\(^{28}\) such that

\[
\mathcal{P}[\{0 \leq T_1 \leq T_2 \leq \cdots \leq T_k \leq T\}] = 1. \tag{45}
\]

Then for all \( i \) in \( \{1, 2, \ldots, k\} \) such that \( (\mathcal{P}[M \in A_i(Y^{T_1})] + P_e) \leq 1/2 \) we have

\[
\mathcal{P}_{\{A_i\}} (\tilde{M} \notin A_i(Y^{T_1})) \geq \exp \left( -h(P_e + \mathcal{P}[M \in A_i(Y^{T_1})]) - \sum_{j=1}^{k+1} \mathbb{E}[T_j - T_{j-1}] J(r_j) \right) \frac{1 - P_e - \mathcal{P}[M \in A_i(Y^{T_1})]}{1 - P_e + \mathcal{P}[M \in A_i(Y^{T_1})]} \tag{46}
\]

where \( T_0 = 0, T_{k+1} = T \) and for all \( j \) in \( \{1, 2, \ldots, (k+1)\} \), \( r_j \)'s are given by

\[
r_j = \begin{cases} 0 & \text{if } \mathbb{P}[T_j = T_{j-1} = 1] \\ \frac{\mathbb{E}[H(M|Y^{T_{j-1}})-H(M|Y^{T_j})]}{\mathbb{E}[T_j - T_{j-1}]} & \text{if } \mathbb{P}[T_j = T_{j-1} < 1] \end{cases}. \tag{47}
\]

**Proof of Lemma 5** is presented in Appendix \[I\].

Before presenting the application of Lemma 5 in UEP problems, let us elaborate on its hypothesis and ramifications. We assumed that \( (T_i, A_i) \) are all NALD. Thus for each \( (T_i, A_i) \) the set of all \( y^T \in \mathcal{Y}^T \) such that the

\(^{27}\)Recall the standard notation in probability theory about the conditional expectations and conditional probabilities: Let \( H \) be a real valued random variable and \( G \) be a random quantity that takes values from a finite set \( \mathcal{G} \), such that \( \mathcal{P}[G=g] > 0 \) for all \( g \in \mathcal{G} \). Then unlike \( \mathbb{E}[H] \), which is constant, \( \mathbb{E}[H|G] \) is a random variable. Thus an equation of the form \( Z = \mathbb{E}[H|G] \), implies not the equality of two constants but the equality of two random variables, i.e. it means that \( Z = \mathbb{E}[H|G = g] \) for all \( g \in \mathcal{G} \). Similarly let \( H_1 \) be a set of sample values of the random variable \( H \) then, unlike \( \mathbb{P}[H \in H_1] \), which is a constant, \( \mathbb{P}[H \in H_1|G] \) is a random variable. Equations \( \text{(43)} \) and \( \text{(44)} \) are such equations. Explaining conditional expectations and conditional probabilities are beyond the scope of this paper, readers who are not sufficiently fluent with these concepts are encouraged to read \[10\] Chapter I, Section 8 which deal the case where random variables can take finitely many values. Appropriate generalized formal treatment of the subject in terms of sigma fields is presented in \[10\] Chapter II, Section 7.

\(^{28}\)Recall ALD's and NALD's are defined in Definitions \[13\] and \[14\] respectively.
transmitted message is guaranteed to be outside $A_i(y^t)$, has zero probability and there is an associated probability measure $P_{\{A_1\}[\cdot]}$ given in equation (41). Furthermore $P_{\{A_1\}[\cdot]}[\hat{M} \not\in A_i(y^T)]$ is the probability of the event that decoded message $\hat{M}$ is not in $A_i$ under the probability measure $P_{\{A_1\}[\cdot]}$.

Condition given in equation (45) ensures that the decoding times of the $k$ NALD’s we are considering, $T_1$, $T_2$, $\ldots$, $T_k$, are reached in their indexing order and before the decoding time of the variable length code $T$. Any $T_1, T_2, \ldots, T_k$ satisfying equation (45) divides the time interval between 0 and $T$ into $k + 1$ disjoint intervals. The duration of these intervals as well as the decrease of the conditional entropy during them are random. For the $j^{th}$ interval the expected values of the duration and the decrease in the conditional entropy are given by $E[T_j - T_{j-1}]$ and $E[H(M | Y^{T_{j-1}}) - H(M | Y^T)]$, respectively. Hence $r_j$’s defined in equation (47) are rate of decrease of the conditional entropy of the messages per unit time in different intervals.

Lemma 5 bounds the probability of $\hat{M}$ being outside $A_i$ under $P_{\{A_1\}[\cdot]}$ from below in terms of $r_j$’s and $E[T_j - T_{j-1}]$’s for $j > i$. The bound on $P_{\{A_1\}[\cdot]}[\hat{M} \not\in A_i(y^T)]$ also depends on $P[M \in A_i(y^T)]$ and $P_e$. But the particular choice of $A_j$’s for $j \neq i$ has no effect on the bound. This feature of the bound is its main merit over bounds resulting from the previously suggested techniques.

C. Single Message Message-Wise UEP Converse:

In this section we bound the conditional error probabilities of the messages, i.e., $P_e[m]$’s, from below uniformly over the message set $\mathcal{M}$ in a variable length block code with average error probability $P_e$, using Lemma 5. Resulting outer bound reveals that the inner bound we obtained in Section IV-C for the single message message-wise UEP problem is tight.

Consider a variable length block code with finite expected decoding time, i.e., $E[T] < \infty$. In order to bound $P_e[m]$, defined in equation (11), from below we apply Lemma 5 for $k = 2$ with $(T_1, A_1)$, $(T_2, A_2)$ given below.

- Let $T_1$ be zero and $A_1$ be $\{m\}$, i.e.,
  \[ T_1 = 0 \]
  \[ A_1 = \{m\}. \]  
  (48)

- Let $T_2$ be the first time instance before $T$ such that one message, not necessarily the one chosen for $A_1$, i.e., $m$, has a posteriori probability $1 - \delta$ or higher,
  \[ T_2 \triangleq \min\{\tau : \max_m P[M = \tilde{m} | Y^\tau] \geq (1 - \delta) \text{ or } \tau = T\}. \]  
  (50)

Let $A_2$ be the set of all messages whose posterior probability at time $T_2$ is less then $(1 - \delta)$,

\[ A_2(y^{T_2}) \triangleq \{\tilde{m} \in \mathcal{M} : P[M = \tilde{m} | Y^{T_2}] < (1 - \delta)\}. \]  
(51)

We apply Lemma 5 for $(T_1, A_1)$ and $(T_2, A_2)$ given in equations (48), (49), (50) and (51). Then using the fact that $J(\cdot) \leq D$ we get,

\[ \ln P_{e|m} \geq \frac{-h(P_e + |\mathcal{M}|^{-1}) - E[T_2]J\left(\frac{E[H(M | Y^{T_2})]}{R_{T_2}}\right)}{1 - P_e - |\mathcal{M}|^{-1}E[T_2]D} \]  
(52a)

\[ \ln P_{\{A_1\}[\cdot]}[\hat{M} \not\in A_2(y^{T_2})] \geq \frac{-h(P_e + P[M \in A_2(Y^{T_2})] - E[T_2])}{1 - P_e - |\mathcal{M}|^{-1}E[T_2]D}. \]  
(52b)

If $\delta < 1/2$ one can show $P_{\{A_2\}[\cdot]}[\hat{M} \not\in A_2(y^{T_2})]$ is roughly equal to $P_e/\delta$. Thus inequality in (52b) becomes a lower bound on $E[T - T_2]$ in terms of $P_e$. It can be shown that the lower bound (52a) takes its smallest value for the smallest value of $E[T - T_2]$. Then using Fano’s inequality for $E[H(M | Y^{T_2})]$ we obtain Lemma 6 given below.

A complete proof of Lemma 6 for variable length block codes with finite expected decoding time is presented in Appendix [F]. For variable length block codes with infinite expected decoding time, Lemma 6 follows from the lower bounds on $P_e$ and $P_{e|m}$ derived in Appendix [H] and Appendix [J].

**Lemma 6:** For any variable length block code and positive $\delta$ such that $P_e + \frac{\delta}{\delta^2} + |\mathcal{M}|^{-1} \leq 1/2$

\[ -\ln P_{e|m} \leq E + (1 - \frac{E - \varepsilon}{D})J\left(\frac{R - \varepsilon}{1 - \varepsilon^2}\right) \]  
(53)
\[ R = \frac{|M|}{E[T]} \]
\[ E = -\ln P_e, \quad \hat{\epsilon} = \frac{\hat{t}_1 + \hat{t}_2}{1 - \hat{t}_2}, \quad \hat{t}_1 = P_e + \frac{P_e}{\delta} + |M|^{-1} \quad \text{and} \quad \hat{t}_2 = \frac{K(\hat{\epsilon}) - \ln \lambda}{E[T]} \]

Lemma 5 is a generalization of [2, Theorem 8] and [2, Lemma 1]. While deriving bounds given in [2, Theorem 8] and [2, Lemma 1], no attention is paid to the fact that the rate of decrease of the conditional entropy of the messages can be different in different time intervals. As result both [2, Theorem 8] and [2, Lemma 1] are tight only when the error exponent is very close to zero. While deriving the bound given in Lemma 6, on the other hand, the variation in the rate the conditional entropy decreases in different intervals is taken into account. Hence the outer bound given in Lemma 6 matches the inner bound given in Section IV-C for all achievable values of error exponent, \( 0 \leq E \leq (1 - \frac{1}{\lambda}) D \).

Consider a reliable sequence of codes \( Q \) with rate \( R_Q \) and error exponent \( E_Q \). Then if we apply Lemma 6 with \( \delta = \frac{1}{\ln(1/P_e)} \) we get
\[ E_{\text{end}, Q} \leq E_Q + (1 - \frac{E_Q}{D}) J \left( \frac{R_Q}{1 - E_Q/D} \right). \]
(54)

Note that the upper bound on \( E_{\text{end}, Q} \)'s given in equation (54) is achievable by at least one \( Q \) described in Section IV-C.

D. Bit-Wise UEP Converse:

In this section we apply Lemma 5 to a variable length block code with a message set \( M \) of the form \( M = M_1 \times M_2 \times \ldots \times M_{\ell} \), in order to obtain lower bounds on \( P_e(i) \)'s for \( i = 1, 2, \ldots, \ell \) in terms of the sizes of the sub-message sets \( |M_1|, |M_2|, \ldots, |M_{\ell}| \) and the expected decoding time \( E[T] \). When applied to reliable code sequences these bounds on \( P_e(i) \)'s in terms of \( |M_i| \)'s and \( E[T] \) gives a necessary condition for the achievability of a rate vector and error exponent vector pair \( (\hat{R}, \hat{E}) \) that matches the sufficient condition for the achievability derived in Section IV-D.

In order to bound \( P_e(i) \)'s we use Lemma 5 with \( \ell \) NALD's, \((T_1, A_1), \ldots, (T_\ell, A_\ell)\). Let us start with defining \( T_i \)'s and \( A_i(Y^{T_i}) \)'s.

- For any \( i \in \{1, 2, \ldots, \ell\} \), let \( T_i \) be the first time instance that a member of \( M_i \) gains a posterior probability larger than or equal to \((1 - \delta)\) if it happens before \( T \), \( T \) otherwise:
  \[ T_i = \min\{\tau : \max_m P[M^i = m^i | Y^\tau] \geq 1 - \delta \text{ or } \tau = T\}. \]
  (55)

- For any \( i \in \{1, 2, \ldots, \ell\} \), let \( A_i(Y^{T_i}) \) be the set of all messages of the form \( m = (m_i, m_{i+1}, \ldots, m_{\ell}) \) for which posterior probability of \( m^i \) is less than \((1 - \delta)\) at \( T_i \):
  \[ A_i(Y^{T_i}) = \{ (m_i, m_{i+1}, \ldots, m_{\ell}) : P[M^i = m^i | Y^{T_i}] < 1 - \delta\}. \]
  (56)

If we apply Lemma 5 for \((T_1, A_1), \ldots, (T_\ell, A_\ell)\) defined in equations (55) and (56), we obtain lower bounds on \( P_{\{A_i\}}[\hat{M} \notin A_i(Y^{T_i})] \)'s in terms of \( P[M \in A_i(Y^{T_i})] \)'s and \( r_j \)'s and \( E[T_j - T_{j+1}] \)'s for \( j > i \). In order to turn these bounds into bounds on \( P_e(i) \)'s we bound \( P_{\{A_i\}}[\hat{M} \notin A_i(Y^{T_i})] \)'s and \( P[M \in A_i(Y^{T_i})] \)'s from above.

- The probability of a message at time \( \tau + 1 \) can not be smaller than \( \lambda \) times its value at time \( \tau \) because \( \min_{x \in X, y \in Y} W_x(y) = \lambda \). Thus if \( \delta < 1/2 \) one can bound \( P_{\{A_i\}}[\hat{M} \notin A_i(Y^{T_i})] \)'s from above:
  \[ P_{\{A_i\}}[\hat{M} \notin A_i(Y^{T_i})] < \frac{P_e(i)}{\delta} \quad \forall i \in \{1, 2, \ldots, \ell\}. \]
  (57)

Note that if at \( T_i \) there is a \( m^i \) with posterior probability \((1 - \delta)\) then \( P[M \in A_i(Y^{T_i})] Y^{T_i}] \leq \delta \). If at \( T_i \) there is no \( m^i \) with posterior probability \((1 - \delta)\) then \( P[M \in A_i(Y^{T_i})] \geq \delta \). Using these facts one can bound \( P[M \in A_i(Y^{T_i})] \)'s from above:
\[ P[M \in A_i(Y^{T_i})] \leq \frac{P_e(i)}{\delta} + \delta \quad \forall i \in \{1, 2, \ldots, \ell\}. \]
(58)

More detailed derivations of the inequalities given in (57) and (58) can be found in Appendix G.

Using equations (57) and (58) together with Lemma 5 we can conclude that,
\[ \ln P_e(i) \geq \ln(\lambda \delta) + \frac{K(\epsilon_{i+1}) - \ln \lambda}{E[T_i - T_{i+1}, |A_i]} \quad \forall i \in \{1, 2, \ldots, \ell\}. \]
(59)
provided that $P_e + \delta + P_e / \delta \leq 1/2$, where $r_j$’s are defined in (37).

Note that the lower bound on $P_e(i)$’s given in equation (59) takes different values depending on the rate of decrease of the conditional entropy of the messages in different intervals, i.e., $r_j$’s, and the expected duration of different intervals, i.e., $E[T_j - T_{j-1}]$’s. Making a worst case assumption on the rate of decrease of entropy and the durations of the intervals one can obtain Lemma 7 given below.

A complete proof of Lemma 7 for variable length block codes with finite expected decoding time is presented in Appendix G. For variable length block codes with infinite expected decoding time, Lemma 7 follows from the lower bounds on $P_e$ and $P_e(i)$’s derived in Appendix H1 and Appendix H3.

**Lemma 7:** For any variable length block code with feedback with a message set $\mathcal{M}$ of the form $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2 \times \ldots \times \mathcal{M}_\ell$ and for any positive $\delta$ such that $P_e + \delta + P_e / \delta \leq 1/2$, we have

$$
(1 - \tilde{c}_3)E_i - \tilde{c}_5 \leq (1 - \sum_{j=1}^\ell \eta_j)D + \sum_{j=i+1}^\ell \eta_j J\left(\frac{(1 - \tilde{c}_3)R_j}{\eta_j}\right)
$$

$$
(1 - \tilde{c}_3)R_i - \tilde{c}_4 \mathbb{I}_{\{i=1\}} \leq C\eta_i
$$

for some time sharing vector $\vec{\eta}$ such that

$$
\eta_i \geq 0, \quad i = 1, 2, \ldots, \ell
$$

$$
\sum_{j=1}^\ell \eta_j \leq 1
$$

where $R_i = \frac{\log |\mathcal{M}_i|}{E[T_i]}$, $E_i = \frac{\log P_e(i)}{E[T_i]}$, $\tilde{c}_3 = P_e + \delta + \frac{P_e}{\delta}$, $\tilde{c}_4 = \frac{\beta(\tilde{c}_3) - \log \lambda}{E[T_i]}$, $\tilde{c}_5 = \frac{\beta(\tilde{c}_3) - \log \lambda}{E[T_i]}$.

For any reliable sequence $Q$ whose message sets $\mathcal{M}^{(k)}$ are of the form $\mathcal{M}^{(k)} = \mathcal{M}_1^{(k)} \times \mathcal{M}_2^{(k)} \times \ldots \times \mathcal{M}_\ell^{(k)}$ if we set $\delta$ to $\delta = \frac{1}{\log |\mathcal{M}|}$ Lemma 7 implies that there exists a $\vec{\eta}$ such that $\vec{R}, \vec{E}$ is achievable only if there exists a reliable code sequence $Q$ such that $(R_Q, E_Q) = (\vec{R}, \vec{E})$. Thus a rates-exponents vector $(\vec{R}, \vec{E})$ is achievable only if there exists a time sharing vector $\vec{\eta}$ satisfying equation (34). In other words the sufficient condition for the achievability of $(\vec{R}, \vec{E})$ we have derived in Section IV.D is also a necessary condition.

**VI. CONCLUSIONS**

We have considered the single message message-wise and the fixed $\ell$ bit-wise UEP problems and characterized the achievable rate error exponent regions completely for both of the problems.

In bit-wise UEP problem we have observed that encoding schemes decoupling the communication and bulk of the error correction both at the transmitter and at the receiver can achieve optimal performance. This result is extending the similar observations made for conventional variable length block coding schemes without UEP. However, for doing that one needs to go beyond the idea of communication phase and control phase introduced in [12], and harness the implicit confirmation explicit rejection schemes, introduced by Kudryashov in [7].

For the converses results, we have introduced a new technique for establishing outer bounds to the performance of the variable length block codes, that can be used in both message-wise and bit-wise UEP problems [31].

We were only interested in bit-wise UEP problem in this paper. We have analyzed single-message message-wise UEP problem, because it is closely related to bit-wise UEP problem and its analysis allowed us to introduce the

---

29 We tacitly assume, without loss of generality, that $|\mathcal{M}_1| \geq 2$.

30 This fact is far from trivial, yet it is intuitive to all who has worked with sequences of vectors in a bounded subset of $\mathbb{R}^\ell$ where $\mathbb{R}^\ell$ is the $\ell$ dimensional real vector space with the norm $\|X\| = \sup |x_j|$ For details see Appendix I.

31 We have not employed the bound in any hybrid problem but it seems result is abstract enough to be employed even in those problems with judicious choice of NALD’s.
ideas we use for bit-wise UEP, gradually. However it seems using the technique employed in \cite[Theorem 9]{2} on the achievability side and Lemma \ref{lemma:achievability} on the converse side, one might be able to determine the achievable region of rates-exponents vectors for variable length block codes in message-wise UEP problem. Such a work would allow us to determine the gains of feedback and variable length decoding, because Csiszár \cite{5} had already solved the problem for fixed length block codes.

Arguably, the most important shortcoming of our bit-wise UEP result is that it only addresses the case when the number of groups of bits \( \ell \) is a fixed integer. However this has more to do with the formal definition of the problem we have chosen in Section \ref{section:problem} than our analysis and non-asymptotic results given in Sections \ref{section:main_results} and \ref{section:converse} i.e., Lemma \ref{lemma:achievability} and Lemma \ref{lemma:converse}.

Using the rates-exponents vectors for representing the performance of a reliable sequence with bit-wise UEP, is apt only when the number of groups of bits are fixed or bounded. When the number of groups of bits \( \ell \) in a reliable sequence diverge with increasing \( \kappa \), i.e., when \( \lim_{\kappa \to \infty} \ell_{\kappa} = \infty \), the rates-exponents vector formulation becomes fundamentally inapt. Consider, for example, a reliable sequence in which \( |\mathcal{M}_{i}^{(\kappa)}| = \lfloor e^{\frac{E_{T}(\kappa)}{I(\kappa)}} \rfloor \). The rate of this reliable sequence is \( R \), yet the rate of all of the sub-messages are zero. Thus when \( \ell_{\kappa} \) diverges the rates-exponents vector does not have the same operational relevance or meaning it has when \( \ell_{\kappa} \) is fixed or bounded. In order to characterize the change of error performance among sub-messages in the case when \( \ell_{\kappa} \) diverges, one needs to come up with an alternative formulation of the problem, in terms of cumulative rate of sub-messages.

Our non-asymptotic results are useful to some extend even when \( \ell_{\kappa} \) diverges. Although infinite dimensional rates-exponents vectors falls short of representing all achievable performances one can still use Lemma \ref{lemma:achievability} of Section \ref{section:main_results} and Lemma \ref{lemma:converse} of Section \ref{section:converse} to characterize the set of achievable rate vector error exponent vector pairs.

- As a result of Lemma \ref{lemma:converse} the necessary condition given in equation \eqref{equation:necessary_condition} is still a necessary condition for the achievability of rates-exponents vector.
- Using Lemma \ref{lemma:achievability} we see that the sufficient condition given in equation \eqref{equation:sufficient_condition} is still a sufficient condition as long as the number of sub-messages in the reliable sequence satisfy \( \lim_{\kappa \to \infty} \ell_{\kappa}(\kappa) \leq \frac{\ell_{\kappa}(\kappa)}{\ln n} = 0 \).

Thus for the case when \( \ell \sim o\left(\frac{E_{T}(\kappa)}{\ln E_{T}(\kappa)}\right) \), i.e., \( \lim_{\kappa \to \infty} \frac{\ell_{\kappa}(\kappa)}{\ln E_{T}(\kappa)} = 0 \), the condition given in equation \eqref{equation:necessary_condition} is still a necessary and sufficient condition for the achievability of a rates-exponents vector.

**ACKNOWLEDGMENT**

Authors would like to thank Gerhard Kramer and Reviewer II for pointing out the importance of the case of varying the number of groups of bits with expected block length. Reviewer II has also coined the term rates-exponents vector which replaced authors’ previous attempt: rate vector error exponent vector pair. Authors would also like to thank all of the reviewers for their meticulous reviews and penetrating questions.

**APPENDIX**

\textbf{A. Proof of Lemma \ref{lemma:converse}}

\textbf{Proof:} Note that \( J(R) \) defined in equation \eqref{equation:j_of_R} is also equal to
\[
J(R) = \max_{0 \leq \alpha < 1} \max_{\alpha_x, \alpha_2, \mu_1, \mu_2, R_1, R_2 \in \mathcal{R}} D(\hat{\mu}_1 \parallel W_{x_1}) + (1 - \alpha)D(\hat{\mu}_2 \parallel W_{x_2})
\]
\[
\leq \max_{0 \leq \alpha < 1} \max_{\alpha, R_1, R_2} \alpha j(R_1) + (1 - \alpha)j(R_2)
\]
\[
\leq \alpha \max_{0 \leq \alpha < 1} \max_{\alpha, R_1, R_2} j(R_1) + (1 - \alpha)\max_{0 \leq \alpha < 1} \max_{\alpha, R_1, R_2} j(R_2)
\]

where \( j(R) \) is given by
\[
j(R) \triangleq \max_{\alpha, \mu \in \mathcal{R}} D(\hat{\mu} \parallel W_x) \quad \forall R \in \mathcal{C}.
\]
Note that \( j(R) \) is a bounded real valued function of a real variable. Therefore, Carathéodory’s Theorem implies that considering two point convex combinations suffices in order make \( j(R) \) a concave function. In other words for any \( k \) we have,

\[
\max_{\alpha R_1,R_2: \alpha R_1+(1-\alpha)R_2=R} \alpha j(R_1) + (1-\alpha)j(R_2) = \max_{0<\alpha_1,\ldots,\alpha_k \leq 1 \forall i} \sum_{i=1}^k \alpha_i j(R_i). \tag{65}
\]

Then the concavity of \( J(R) \) follows from the equations (63), (64) and (65).

Evidently if the constraint set in a maximization is curtailed than resulting maximum value can not increase. Hence \( J(R) \) function defined in equation (17) is a decreasing function of \( R \).

As a result of the definition of \( D \) given in equation (2) and the convexity of Kullback-Leibler divergence, we have \( D \geq J(0) \). On the other hand \( D(\mu \parallel W_x) = D \) and \( I(\mu, W) \geq 0 \) for \( x = r \) and \( \mu(\cdot) = \mathbb{1}_{\{a\}} \) where \( a \) and \( r \) described in equation (10). Therefore we have \( J(0) \geq D \). Using the fact that \( J(R) \geq j(R) \) we conclude that \( J(0) = j(0) = D \).

**B. Proof of Lemma 2**

Proof: We prove the lemma for a slightly more general setting and establish a result that will be easier to make use of in the proofs of other achievability results. Let \( G_\gamma[1], G_\gamma[m] \) and \( B_\gamma[x^n] \) be

\[
G_\gamma[1] = \left\{ y^n : n_\alpha \Delta(Q_{y^n_{\alpha+1}}, \bar{\mu}_1) + (n-n_\alpha) \Delta(Q_{y^n_{\alpha+1}}, \bar{\mu}_2) \geq \gamma \right\}
\]

\[
G_\gamma[m] = B_\gamma[x^n(m)] \quad \forall m \in \{2, 3, \ldots, |M|\}
\]

\[
B_\gamma[x^n] = \{ y^n : n_\alpha \Delta(Q_{\{y^n_\alpha\}, y^n_{\alpha+1}}, \mu_1 W) + (n-n_\alpha) \Delta(Q_{\{y^n_\alpha\}, y^n_{\alpha+1}}, \mu_2 W) < \gamma \}. \tag{66c}
\]

Note that \( G_\gamma[1], G_\gamma[m] \) and \( B_\gamma[x^n] \) given equations (21), (22) and (23) are simply the \( G_\gamma[1], G_\gamma[m] \) and \( B_\gamma[x^n] \) for \( \gamma = |X|^3 \sqrt{n \ln(1+n)} \).

For all \( y^n \notin G_\gamma[1] \) we have,

\[
n_\alpha D(Q_{\{y^n_\alpha\}} \parallel W_{x_1}) + (n-n_\alpha) D(Q_{y^n_{\alpha+1}} \parallel W_{x_2}) \]

\[
= n_\alpha D(Q_{\{y^n_\alpha\}} \parallel \bar{\mu}_1) + (n-n_\alpha) D(Q_{y^n_{\alpha+1}} \parallel \bar{\mu}_2) \]

\[
+ n_\alpha \sum_y Q_{\{y^n_\alpha\}}(y) \ln \frac{\bar{\mu}_1(y)}{W_{x_1}(y)} + (n-n_\alpha) \sum_y Q_{y^n_{\alpha+1}}(y) \ln \frac{\bar{\mu}_2(y)}{W_{x_2}(y)} \]

\[
\geq n_\alpha \sum_y Q_{\{y^n_\alpha\}}(y) \ln \frac{\bar{\mu}_1(y)}{W_{x_1}(y)} + (n-n_\alpha) \sum_y Q_{y^n_{\alpha+1}}(y) \ln \frac{\bar{\mu}_2(y)}{W_{x_2}(y)} \]

\[
\geq n_\alpha D(\bar{\mu}_1 \parallel W_{x_1}) + (n-n_\alpha) D(\bar{\mu}_2 \parallel W_{x_2}) + 2\gamma \ln \lambda. \tag{66d}
\]

Inequality (a) follows from the non-negativity of the Kullback-Leibler divergence. In order to see why (b) holds, first recall that \( \min_{x,y} W_x(y) = \lambda \). Hence \( \ln \frac{\bar{\mu}_1(y)}{W_{x_1}(y)} \leq \frac{\lambda}{X} \) and \( \ln \frac{\bar{\mu}_2(y)}{W_{x_2}(y)} \leq \frac{\lambda}{X} \). Then the inequality (b) follows from the definitions of total variation \( \Delta \) and \( G_\gamma[1] \), given in equations (11) and (66a) and the fact that \( y^n \notin G_\gamma[1] \).

Note that the conditional probability of the first message is given by

\[
P_{e_1} = P[M \neq 1 | M = 1] = \sum_{y^n \notin G_\gamma[1]} P[Y^n = y^n | M = 1].
\]

Recall that, the codeword of the message \( M = 1 \) is the concatenation of \( n_\alpha \) \( x_1 \)'s and \( (n-n_\alpha) \) \( x_2 \)'s where \( n_\alpha = \lfloor n \alpha \rfloor \). Hence the probability of all \( y^n \)'s whose empirical distribution in first \( n_\alpha \) times instances is \( Q_{\{y^n_\alpha\}} \) and whose empirical distribution in \( ((n_\alpha + 1), n] \) is \( Q_{y^n_{\alpha+1}} \) is upper bounded by \( e^{-n_\alpha D(Q_{\{y^n_\alpha\}} \parallel W_{x_1}) - (n-n_\alpha) D(Q_{y^n_{\alpha+1}} \parallel W_{x_2})} \).
Furthermore, there are less than \((n_\alpha + 1)\mathcal{P}\) distinct empirical distributions in the first phase and there are less than \((n - n_\alpha + 1)\mathcal{P}\) distinct empirical distributions in the second phase. Thus

\[
P_{e[1]} \leq (n_\alpha + 1)\mathcal{P}(n - n_\alpha + 1)\mathcal{P}e^{-nD(\bar{\mu}_1 \parallel W_{x_1}) + (n - n_\alpha)D(\bar{\mu}_2 \parallel W_{x_2}) - 2\gamma \ln \lambda}
\]

where \(\varepsilon_2(\gamma, n) = \frac{-2\gamma \ln \lambda + D + 2\mathcal{P}\ln(n+1)}{n}\).

The codewords and the decoding regions of the remaining messages are specified using a random coding argument together with an empirical typicality decoder. Consider an ensemble of codes in which first \(n_\alpha\) entries of all the codewords are independent and identically distributed (i.i.d.) with input distribution \(\mu_1\) and the rest of the entries are i.i.d. with the input distribution \(\mu_2\).

For any message \(m\) other than the first one, i.e., \(m \neq 1\), the decoding region is \(G_1[m]\) given in (66b). In other words for any message \(m\), other than the first one, the decoding region is the set of output sequences for which \((x^n(m), y^n)\) is typical with \((\alpha, \mu_1 W, \mu_2 W)\), i.e., \(y^n \in B_1[x^n(m)]\), and \((x^n(m), y^n)\) is not typical with \((\alpha, \mu_1 W, \mu_2 W)\), i.e., \(y^n \in B_2[x^n(m)]\), for any \(m \neq 1\).

Since the decoding regions of different messages are disjoint, above described code does not decode to more than one message. Disjointness of decoding regions of messages \(2, 3, \ldots, |M|\) follows from the definitions of \(G_\gamma[2], G_\gamma[3], \ldots, G_\gamma[|M|]\), given in equation (66b). In order to see why \(G_\gamma[1] \cap (\bigcup_{m \neq 1} G_\gamma[m]) = \emptyset\) holds, note that for any pair probability of distributions, the total variation between them is lower bounded by the total variation between their marginals. In particular,

\[
\Delta(Q_{(x^n(m), y^n)}; \mu_1 W) \geq \Delta(Q_{(y^n)}, \bar{\mu}_1)
\]

\[
\Delta(Q_{(x^n(m), y^n_{n+1})}; \mu_2 W) \geq \Delta(Q_{(y^n_{n+1})}, \bar{\mu}_2).
\]

Then as results of definitions of \(G_\gamma[1], B_1[x^n]\) and \(G_\gamma[m]\) for \(m \neq 1\) given in equations (60a), (60c) and (60b) we have

\[
G_\gamma[1] \cap G_\gamma[m] = \emptyset \quad m = 2, 3, \ldots, |M|.
\]

Then for \(m \in \{2, 3, \ldots, |M|\}\) the average of the conditional error probability of \(m^{th}\) message over the ensemble is upper bounded as

\[
E[P_{e[m]}] \leq P[Y^n \notin B_\gamma[x^n(m)] | M = m] + \sum_{m \neq m} P[Y^n \in B_\gamma[x^n(m)] | M = m]. \quad (67)
\]

Let us start with bounding \(P[Y^n \notin B_\gamma[x^n(m)] | M = m]\). Let \(S_1(x, y)\) and \(S_2(x, y)\) be

\[
S_1(x, y) \triangleq n_\alpha |Q_{(x^n(m), y^n)}(x, y) - \mu_1(x) W_y(y)|,
\]

\[
S_2(x, y) \triangleq (n - n_\alpha) |Q_{(x^n(m), y^n_{n+1})}(x, y) - \mu_2(x) W_y(y)|. \quad (68a)
\]

As a result of the definition of total variation distance given in equation (1) and above definitions we have

\[
n_\alpha \Delta(Q_{(x^n(m), y^n)}; \mu_1 W) + (n - n_\alpha) \Delta(Q_{(x^n(m), y^n_{n+1})}; \mu_2 W) = \frac{1}{2} \sum_{x,y} [S_1(x, y) + S_2(x, y)]
\]

Thus the definition of \(B_\gamma[x^n(m)]\) given in equation (66c) implies that

\[
P[Y^n \notin B_\gamma[x^n(m)] | M = m] = P\left[\sum_{x,y} [S_1(x, y) + S_2(x, y)] \geq 2\gamma | M = m\right]. \quad (69)
\]

If for all \(x \in X, y \in Y\) and \(j \in \{1, 2\}\), \(S_j(x, y) \leq \gamma|X|^{-1}|Y|^{-1}\) then \(\sum_{x,y} [S_1(x, y) + S_2(x, y)] \leq 2\gamma\). Thus if \(Y \notin B_\gamma[x^n(m)]\) then for at least one \((x, y, j)\) triple \(S_j(x, y) \geq \gamma|X|^{-1}|Y|^{-1}\). Using the union bound we get

\[
P\left[\sum_{x,y} [S_1(x, y) + S_2(x, y)] \geq 2\gamma | M = m\right] \leq \sum_{x,y} P[S_j(x, y) \geq \frac{\gamma}{|X| |Y|} | M = m]. \quad (70)
\]

For bounding \(P[S_j(x, y) \geq \frac{\gamma}{|X| |Y|} | M = m]\), we can simply use Chebyshev’s inequality, however in order to get better error terms we use a standard concentration result about the sums of bounded random variables, [4, Theorem 5.3].
Lemma 8: Let $Z_1, Z_2, \ldots, Z_k$ be independent random variables satisfying $|Z_i - \mathbb{E}[Z_i]| \leq c_i$ for all $1 \leq i \leq k$. Then

$$P\left[ \sum_{i=1}^{k} (Z_i - \mathbb{E}[Z_i]) > \gamma \right] \leq 2e^{-\frac{\gamma^2}{2\sum_{i=1}^{k} c_i}}.$$ 

For all $\mu_1 \in \mathcal{P}(\mathcal{X})$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$ we have $c_i = 1$ for all $i = 1, 2, \ldots, n_a$; thus

$$P\left[ S_1(x,y) \geq \frac{\gamma}{|\mathcal{X}||\mathcal{Y}|} | M = m \right] \leq 2e^{-\frac{\gamma^2}{2|\mathcal{X}|^2|\mathcal{Y}|^2n_a}} \leq 2e^{-\frac{\gamma^2}{2|\mathcal{X}|^2|\mathcal{Y}|^2n}}. \tag{71}$$

Similarly,

$$P\left[ S_2(x,y) \geq \frac{\gamma}{|\mathcal{X}||\mathcal{Y}|} | M = m \right] \leq 2e^{-\frac{\gamma^2}{2|\mathcal{X}|^2|\mathcal{Y}|^2n}} \tag{72}$$

Using equations (69), (70), (71) and (72) we get

$$P[Y \notin B_\gamma[X^n(m)] | M = m] \leq 4|\mathcal{X}||\mathcal{Y}|e^{-\frac{\gamma^2}{2|\mathcal{X}|^2|\mathcal{Y}|^2n}} \tag{73}$$

Now we focus on $P[Y^n \in B_\gamma[X^n(m)] | M = m]$ terms. Note that all $y^n$ in $B_\gamma[x^n(m)]$ satisfy

$$n_\alpha \Delta(Q_{i_1^n(m),y_1^n}) + (n - n_\alpha) \Delta(Q_{i_1^n(m),y_1^n}) + \mu_1 W \leq \gamma. \tag{74}$$

On the other hand, when $M = m$, $X^n(m)$ and $Y^n$ are independent and their distribution is given by,

$$P[(X^n(m), Y^n) = (x^n(m), y^n) | M = m] = \prod_{i=1}^{n} \mu_1(x_i(m)) \mu_1(y_i) \prod_{j=n_\alpha+1}^{n} \mu_2(x_j(m)) \mu_2(y_j) = e^{-n_\alpha D(Q_{i_1^n(m),y_1^n})} \mu_1 \mu_2(e^{-n \alpha H(Q_{i_1^n(m),y_1^n})})$$

$$e^{-(n-n_\alpha) D(Q_{i_1^n(m),y_1^n})} \mu_2(e^{-(n-n_\alpha) H(Q_{i_1^n(m),y_1^n})}). \tag{75}$$

Furthermore the number of $(x_1^n(m), y_1^n)$ sequences with an empirical distribution $Q_{x_1^n(m),y_1^n}$ is upper bounded as $e^{n \alpha H(Q_{i_1^n(m),y_1^n})}$. In addition there are at most $(n_\alpha + 1)|\mathcal{X}||\mathcal{Y}|$ different empirical distributions. Using these two bounds and their counter parts for $(X^n(m), Y^n)$ together with equations (74) and (75) we get

$$P[Y^n \in B_\gamma[X^n(m)] | M = m] \leq (n_\alpha + 1)|\mathcal{X}||\mathcal{Y}|(n-n_\alpha + 1)|\mathcal{X}||\mathcal{Y}|e^{-n_\alpha D(\mu_1 W \| \mu_1) - (n-n_\alpha) D(\mu_2 W \| \mu_2) - 2\gamma \ln \lambda}$$

$$= (n_\alpha + 1)|\mathcal{X}||\mathcal{Y}|(n-n_\alpha + 1)|\mathcal{X}||\mathcal{Y}|e^{-n \alpha H(\mu_1, W) - (n-n_\alpha) H(\mu_2, W) - 2\gamma \ln \lambda} \leq e^{-n(\alpha H(\mu_1, W) + (1-\alpha)H(\mu_2, W))} e^{C+2|\mathcal{X}||\mathcal{Y}| \ln(n+1) - 2\gamma \ln \lambda}. \tag{76}$$

Hence if $|\mathcal{M}\setminus\{1\}| = 4|\mathcal{X}||\mathcal{Y}|e^{-\frac{\gamma^2}{2|\mathcal{X}||\mathcal{Y}|^2n}} e^{n(\alpha H(\mu_1, W) + (1-\alpha)H(\mu_2, W))} e^{C-2|\mathcal{X}||\mathcal{Y}| \ln(n+1) + 2\gamma \ln \lambda}$ then

$$\sum_{m \notin m} P[Y^n \in B_\gamma[X^n(m)] | M = m] \leq 4|\mathcal{X}||\mathcal{Y}|e^{-\frac{\gamma^2}{2|\mathcal{X}||\mathcal{Y}|^2n}}. \tag{77}$$

Thus the average $P_e$ over the ensemble can be bounded using (57), (70) and (77) as

$$\mathbb{E}[P_e] \leq 8|\mathcal{X}||\mathcal{Y}|e^{-\frac{\gamma^2}{2|\mathcal{X}||\mathcal{Y}|^2n}}. \tag{78}$$

But if the ensemble average of the error probability is upper bounded like this, there is at least one code that has this low error probability. Furthermore half of its messages have conditional error probabilities less then twice this average. Thus for any block length $n$, time sharing constant $\alpha \in [0,1]$, input letters $x_1, x_2 \in \mathcal{X}$, input distributions $\mu_1, \mu_2 \in \mathcal{P}(\mathcal{X})$ there exists a length $n$ code such that

$$|\mathcal{M}\setminus\{1\}| \geq e^{n(\alpha H(\mu_1, W) + (1-\alpha)H(\mu_2, W) - \varepsilon_1(\gamma, n))} \tag{78a}$$

$$P_e\|_1 \leq e^{-n(\alpha D(\mu_1 \| W_{x_1}) + (1-\alpha)D(\mu_2 \| W_{x_2}) - \varepsilon_2(\gamma, n))} \tag{78b}$$

$$P_e\|_m \leq \varepsilon_3(\gamma, n) \tag{78c}$$

$m = 2, 3, \ldots, |\mathcal{M}|$
\[ \epsilon_1(\gamma, n) = \frac{C - \ln(2\kappa X|Y| + 2|X|Y| \ln(n+1) - 2\gamma \ln \lambda)}{\ln n} + \frac{\gamma^2}{2|X|^2|Y|^2 n^2} \]  

(79a)

\[ \epsilon_2(\gamma, n) = \frac{D + 2|Y| \ln(n+1) - 2\gamma \ln \lambda}{\ln n} \]  

(79b)

\[ \epsilon_3(\gamma, n) = 16|X| |Y| e^{-\frac{\gamma^2}{2|X|^2|Y|^2 n^2}} \]  

(79c)

Lemma 2 follows from equation (78) and the fact that

\[ \epsilon_i(\gamma, n) = |X| |Y| \sqrt{n \ln(1+n)} \leq \frac{9|X| |Y| |1-\ln(\frac{1}{n})| \sqrt{\ln(1+n)}}{\sqrt{n}} \]  

for \( \epsilon_1(\gamma, n) \), \( \epsilon_2(\gamma, n) \) and \( \epsilon_3(\gamma, n) \) given in equation (79).

C. Proof of Lemma 3

Proof: Let \( n_1 \) be \( n_1 \geq [(1 - \frac{E}{D})n] \). Recall that we have assumed \( E \leq (1 - \frac{R}{C})D \), then we have \( \frac{R}{C} \leq 1 - \frac{E}{D} \). Consequently \( \frac{R}{C} \leq \frac{R}{n_1} \) and \( \frac{R}{n_1} \leq C \). On the other hand as a result of equation (78) and the definition of \( J(\cdot) \) given in equation (77), for any positive integer \( n_1 \), positive real number \( \gamma_1 \), rate \( R \leq C \) there exists a length \( n_1 \) code such that,

\[ |\mathcal{M}| - 1 \geq e^{n_1 [\frac{R}{C} - \epsilon_1(\gamma_1, n_1)]} \]  

(81a)

\[ P_e[1] \leq e^{-n_1 [J(\frac{R}{C}) - \epsilon_2(\gamma_1, n_1)]} \]  

(81b)

\[ P_e[m] \leq \epsilon_3(\gamma_1, n_1) \quad m = 2, 3, \ldots, |\mathcal{M}| \]  

(81c)

where \( \epsilon_1(\gamma_1, n_1) \), \( \epsilon_2(\gamma_1, n_1) \), \( \epsilon_3(\gamma_1, n_1) \) are given in equation (79).

We use such a code in the first phase with \( R = \frac{n_1}{n} R \) and call its decoded message \( \hat{M} \), the tentative decision. Then as a result of equation (81) and the fact that 32\( n_1 J\left(\frac{n_1 R}{C}\right) \geq n_1 (1 - \frac{E}{D}) J\left(\frac{R}{1 - \frac{E}{D}}\right) \) we get

\[ |\mathcal{M}| - 1 \geq e^{n_1 R - n_1 \epsilon_1(\gamma_1, n_1)} \]  

(82a)

\[ P\left[ \hat{M} \neq M \mid M = 1 \right] \leq e^{-n_1 (1 - \frac{E}{D}) J\left(\frac{R}{1 - \frac{E}{D}}\right) + n_1 \epsilon_2(\gamma_1, n_1)} \]  

(82b)

\[ P\left[ \hat{M} \neq m \mid M = m \right] \leq \epsilon_3(\gamma_1, n_1) \quad m = 2, 3, \ldots, |\mathcal{M}|. \]  

(82c)

The transmitter knows what the tentative decision is and determines the channel inputs in the last \( (n - n_1) \) time instances depending on its correctness. If \( \hat{M} = M \) the channel inputs in the last \( (n - n_1) \) time instances are all \( a \), if \( \hat{M} \neq M \) the channel inputs in the last \( (n - n_1) \) time instances are all \( r \).

After observing \( Y^n \), receiver checks whether the empirical distribution of the channel output in the last \( (n - n_1) \) time units is typical with \( W_a \), if it is then \( \hat{M} = \hat{M} \) otherwise \( \hat{M} = x \). Hence the decoding region for erasures is given by

\[ \mathcal{G}_\gamma [x] = \{ y^n : (n - n_1) \Delta \left( Q_{y^n_{n_1+1}}, W_a \right) \geq \gamma_2 \}. \]

Let us start with bounding \( P\left[ \hat{M} = x \mid \hat{M} = m, M = m \right] \), i.e., the probability of erasure for correct tentative decision. First note that

\[ (n - n_1) \Delta \left( Q_{y^n_{n_1+1}}, W_a \right) = \frac{1}{2} \sum_y S(y) \]

where \( S(y) = (n - n_1) Q_{y^n_{n_1+1}}(y) - W_a(y) \). Then following an analysis similar to that one presented between equations (69) and (73) we get

\[ P\left[ \hat{M} = x \mid \hat{M} = m, M = m \right] \leq 2 |Y| e^{-\frac{\gamma_2^2}{2|Y|^2(n-n_1)}} = \epsilon_3(\gamma_2, n-n_1) \quad \forall m \in \mathcal{M}. \]  

(83)

32Recall that \( n_1 \geq (1 - E/D)n \) and \( J(\cdot) \) is a non-increasing and positive function.
In order to bound the probability of non-erasure decoding when tentative decision is incorrect, note that
\[
P_{\hat{y}_{n+1}^n = y_{n+1}^n \mid \hat{M} \neq m, M = m} = \prod_{j=n+1}^{n} W_r(y_j)
= e^{-(n-n_1)D(Q(r^{n+1}_n) \| \mu_2)} e^{-(n-n_1)H(Q(r^{n+1}_n))}.
\]
Then following an analysis similar to the one between (75) and (76) we get
\[
P\left[\hat{M} \neq x \mid \hat{M} \neq m, M = m\right] \leq \min\{(n-n_1 + 1)|\mathcal{Y}| e^{-(n-n_1)D-2\gamma_2 \ln \lambda}, 1\}
\leq \min\{e^{-nE+|\mathcal{Y}|\ln(n-n_1)+D-2\gamma_2 \ln \lambda}, 1\}
\leq \min\{e^{-nE+(n-n_1)\varepsilon_2(\gamma_2, n-n_1)}, 1\} \quad \forall m \in \mathcal{M}. (84)
\]
Furthermore the conditional error and erasure probabilities can be bounded in terms of \(P\left[\hat{M} \neq m \mid M = m\right], P\left[\hat{M} \neq x \mid \hat{M} \neq m, M = m\right]\) and \(P\left[\hat{M} = x \mid \hat{M} = m, M = m\right]\) as follows.
\[
P_{e|m} = P\left[\hat{M} \neq m \mid M = m\right] P\left[\hat{M} \neq x \mid \hat{M} \neq m, M = m\right] \quad \forall m \in \mathcal{M}. (85a)
\]
\[
P_{x|m} \leq P\left[\hat{M} \neq m \mid M = m\right] + P\left[\hat{M} = x \mid \hat{M} = m, M = m\right] \quad \forall m \in \mathcal{M}. (85b)
\]
Using the equations (82), (83), (84) and (85) we get
\[
|\mathcal{M}| - 1 \geq e^{nR-n_1 \varepsilon_1(\gamma_1, n_1)} (86a)
\]
\[
P_{e|1} \leq e^{-n(1-\frac{\gamma_1}{\gamma_2, n_1})} \min\{e^{-nE+n_2 \varepsilon_2(\gamma_2, n_2)}, 1\} (86b)
\]
\[
P_{x|1} \leq e^{-n(1-\frac{\gamma_1}{\gamma_2, n_1})+n_2 \varepsilon_2(\gamma_2, n_1)} + \varepsilon_3(\gamma_2, n_2) (86c)
\]
\[
P_{e|m} \leq \varepsilon_3(\gamma_1, n_1) \min\{e^{-nE+|\mathcal{Y}|\ln(n+1)+n_2 \varepsilon_2(\gamma_2, n_2)}, 1\} \quad m \neq 1 (86d)
\]
\[
P_{x|m} \leq \varepsilon_3(\gamma_1, n_1) + \varepsilon_3(\gamma_2, n_2) \quad m \neq 1. (86e)
\]
where \(n_2 = n - n_1\).
We set \(\gamma_j = |\mathcal{X}| |\mathcal{Y}| \sqrt{5n_j \ln(1+n)}\) for \(j = 1, 2\) and obtain
\[
n_j \varepsilon_1(\gamma_j, n_j) \leq 2|\mathcal{X}| |\mathcal{Y}|(\ln(n+1) - \sqrt{5n_j \ln(1+n) \ln \lambda}) + (5/2) \ln(1+n) (87a)
\]
\[
n_j \varepsilon_2(\gamma_j, n_j) \leq 2|\mathcal{X}| |\mathcal{Y}|(\ln(n+1) - \sqrt{5n_j \ln(1+n) \ln \lambda}) + D (87b)
\]
\[
\varepsilon_3(\gamma_j, n_j) \leq 16|\mathcal{X}| |\mathcal{Y}|/(1+n)^{5/2} (87c)
\]
Lemma 3 follows from the identities \(|\mathcal{X}| \geq 2, |\mathcal{Y}| \geq 2, D \leq \ln(\frac{1}{\lambda})\), \(n \geq 1\) and the equations (86) and (87). \(\square\)

D. Proof of Lemma 4

Proof: Note that given the encoding scheme summarized in equation (28) and the decoding rule given in equation (29), if \(\hat{M} = x\) then there is a \(i \leq \ell + 1\) such that \(\hat{M}_j = i M_j\) for all \(j < i\) and \(\hat{M}_i \neq i M_i\). Thus the conditional erasure probability \(P_{x|m}\) is upper bounded as
\[
P_{x|m} \leq \sum_{i=1}^{\ell+1} P\left[\hat{M}_i \neq (1 + m_i) \mid M = m, \hat{M}_1 = i M_1, \ldots, \hat{M}_{i-1} = i M_{i-1}\right]
= \sum_{i=1}^{\ell+1} P\left[\hat{M}_i \neq i M_i \mid i M_i = 1 + m_i\right] (88)
\]
Similarly if \(\hat{M} \neq x\) and \(\hat{M}^i \neq M^i\) then for all \(j > i\), \(\hat{M}_j = i M_j\) and \(\hat{M}_i \neq i M_i\); furthermore there is a \(k \leq i\) such that \(\hat{M}_j = i M_j\) for all \(j < k\) and \(\hat{M}_k \neq i M_k\). Hence one can bound \(P_{e|m}(i)\) as
\[
P_{e|m}(i) \leq \left[\sum_{j=1}^{i} P\left[\hat{M}_j \neq i M_j \mid i M_j = 1 + m_j\right]\right] \prod_{j=i+1}^{\ell+1} P\left[\hat{M}_j \neq 1 \mid i M_j = 1\right]. (89)\]
In the first $\ell$ phases, we use $n_i = \lfloor \eta_i n \rfloor$ long codes with rate $R_{\eta_i}$ with the performance given in equation (81). Thus for $1 \leq i \leq \ell$ we have,

$$|\cM_i| \geq 1 + e^{nR_i - n_i \varepsilon_1(\gamma, n_i)}$$  \hfill (90a)

$$P \left[ \hat{\cM}_i \neq \cM_i \mid \cM_i = 1 \right] \leq e^{-n_i J \left( \frac{R_i}{\eta_i} \right) + D - n_i \varepsilon_2(\gamma, n_i)}$$  \hfill (90b)

$$P \left[ \hat{\cM}_i \neq \cM_i \mid \cM_i = 1 + m_i \right] \leq \varepsilon_3(\gamma, n_i) \quad m_i = 1, 2, 3, \ldots, (|\cM_i| - 1)$$  \hfill (90c)

where $\varepsilon_1(\gamma, n_i), \varepsilon_2(\gamma, n_i), \varepsilon_3(\gamma, n_i)$ are given in equation (79).

In order derive bounds corresponding to the ones given in equation (90) for the last phase let us give the decoding regions for 1 and 2 for the length $n_{\ell+1}$ code employed between $(n + 1 - n_{\ell+1})$ and $n$.

$$\cG_\gamma[1] = \{y_n^{n+1-n_{\ell+1}} : n_{\ell+1} \Delta (Q_{y_n^{n+1-n_{\ell+1}}}, W_a) \geq \gamma_{\ell+1} \}$$

$$\cG_\gamma[2] = \{y_n^{n+1-n_{\ell+1}} : n_{\ell+1} \Delta (Q_{y_n^{n+1-n_{\ell+1}}}, W_a) \leq \gamma_{\ell+1} \}$$

Following an analysis similar to the one leading to equations (83) and (84) we get,

$$P \left[ \hat{\cM}_{\ell+1} \neq \cM_{\ell+1} \mid \cM_{\ell+1} = 1 \right] \leq e^{-n_{\ell+1} \Delta + n_{\ell+1} \varepsilon_2(\gamma_{\ell+1}, n_{\ell+1})}$$  \hfill (91a)

$$P \left[ \hat{\cM}_{\ell+1} \neq 2 \mid \cM_{\ell+1} = 2 \right] \leq \varepsilon_3(\gamma_{\ell+1}, n_{\ell+1})$$  \hfill (91b)

Using equations (88), (89), (90) and (91) we get,

$$|\cM_i| \geq e^{nR_i - n_i \varepsilon_1(\gamma, n_i) - C} \quad \forall i = 1, 2, \ldots, \ell$$  \hfill (92a)

$$|\cM_i| \geq e^{n \sum_{j=1}^{i} R_{j} \varepsilon_{j}(\gamma, n_j) + C} \quad \forall i = 1, 2, \ldots, \ell$$  \hfill (92b)

$$P_{e|m}(i) \leq \frac{i}{\varepsilon_3(\gamma, n_j)} \min \left\{ 1, e^{-\sum_{j=1}^{\ell+1} n_j \varepsilon_{j}(\gamma, n_j) + D} \right\} \quad \forall i = 1, 2, \ldots, \ell, \forall m \in \cM$$  \hfill (92c)

$$P_{x|m} \leq \sum_{j=1}^{\ell+1} \varepsilon_3(\gamma, n_j) \quad \forall m \in \cM$$  \hfill (92d)

If we set $\gamma_i = |\cX||\cY| \sqrt{4n_i \ln(1 + n)}$ for $i = 1, 2, \ldots, (\ell + 1)$ for $\varepsilon_1(\gamma, n_i), \varepsilon_2(\gamma, n_i)$ and $\varepsilon_3(\gamma, n_i)$ given in equation (79) we have

$$n_i \varepsilon_1(\gamma, n_i) + C \leq 2|\cX||\cY|(\ln(n_i + 1) - \sqrt{4n_i \ln(1 + n) \ln \lambda}) + 2 \ln(1 + n) + C$$

$$n_i \varepsilon_2(\gamma, n_i) + D \leq 2|\cX||\cY|(\ln(n_i + 1) - \sqrt{4n_i \ln(1 + n) \ln \lambda}) + 2D$$

$$\varepsilon_3(\gamma, n_i) \leq 16|\cX||\cY|/(1 + n)^2$$

Using the concavity of $\sqrt{x}$ function we can conclude that,

$$\sum_{i=1}^{\ell+1} \frac{n_i \varepsilon_3(\gamma, n_i) + D}{n} \leq 2|\cX||\cY|(\frac{(\ell+1) \ln(1+n)}{n} - \frac{\ell+1}{n} 2 \sqrt{n \ln(1 + n) \ln \lambda}) + \frac{2(\ell+1) \ln(1+n)}{n} + \frac{(\ell+1) C}{n}$$  \hfill (93a)

$$\sum_{i=1}^{\ell+1} \frac{n_i \varepsilon_3(\gamma, n_i) + D}{n} \leq 2|\cX||\cY|(\frac{(\ell+1) \ln(1+n)}{n} - \frac{\ell+1}{n} 2 \sqrt{n \ln(1 + n) \ln \lambda}) + \frac{\ell+1}{n} 2D$$  \hfill (93b)

$$\sum_{i=1}^{\ell+1} \varepsilon_3(\gamma, n_i) \leq 8|\cX||\cY|\frac{\ell+1}{1+n}$$  \hfill (93c)

Then Lemma 4 follows from equations (92) and (93) for any $\ell \leq \frac{n}{\ln(n+1)}$.  \hfill ■
E. Proof of Lemma 5

Proof: For \( P \) defined in equation (36) as a result of equation (37) we have

\[
P \left[ \hat{M} \in A_i( Y_{T_i}^t) \right] = \sum_{y^t \in \{ \hat{y} : \hat{M} \in A_i( Y_{T_i}^t) \} \cap Y_{T_i}^t} P( y^t )
\]  
(94a)

\[
P \left[ \hat{M} \notin A_i( Y_{T_i}^t) \right] = \sum_{y^t \in \{ y^t : \hat{M} \notin A_i( Y_{T_i}^t) \} \cap Y_{T_i}^t} P( y^t )
\]  
(94b)

For \( P_{\{A\}} \) defined in equation (39) as a result of equation (41) we have

\[
P_{\{A\}} \left[ \hat{M} \in A_i( Y_{T_i}^t) \right] = \sum_{y^t \in \{ y^t : \hat{M} \in A_i( Y_{T_i}^t) \} \cap Y_{T_i}^t} P_{\{A\}}( y^t )
\]  
(95a)

\[
P_{\{A\}} \left[ \hat{M} \notin A_i( Y_{T_i}^t) \right] = \sum_{y^t \in \{ y^t : \hat{M} \notin A_i( Y_{T_i}^t) \} \cap Y_{T_i}^t} P_{\{A\}}( y^t )
\]  
(95b)

Using equations (94) and (95) together with the data processing inequality for Kullback-Leibler divergence, we get

\[
\sum_{y^t \in Y_{T_i}^t} P( y^t ) \ln \frac{P(y^t)}{P_{\{A\}}( y^t )} \geq -h \left( P \left[ \hat{M} \in A_i( Y_{T_i}^t) \right] \right) + \left( 1 - P \left[ \hat{M} \in A_i( Y_{T_i}^t) \right] \right) \ln \frac{1}{P_{\{A\}}( M \in A_i( Y_{T_i}^t) )}.
\]  
(96)

Since \( 0 \leq P_{\{A\}} \left[ \hat{M} \in A_i( Y_{T_i}^t) \right] \leq 1 \) we have

\[
\sum_{y^t \in Y_{T_i}^t} P( y^t ) \ln \frac{P(y^t)}{P_{\{A\}}( y^t )} \geq -h \left( P \left[ \hat{M} \in A_i( Y_{T_i}^t) \right] \right) + \left( 1 - P \left[ \hat{M} \in A_i( Y_{T_i}^t) \right] \right) \ln \frac{1}{P_{\{A\}}( M \in A_i( Y_{T_i}^t) )}.
\]  
(97)

Note that if \( \hat{M} \in A_i( Y_{T_i}^t) \) and \( M \notin A_i( Y_{T_i}^t) \) then \( \hat{M} \neq M \). Consequently

\[
P \left[ \hat{M} \in A_i( Y_{T_i}^t) \right] = P \left[ \{ \hat{M} \in A_i( Y_{T_i}^t), M \notin A_i( Y_{T_i}^t) \} \right] + P \left[ \{ \hat{M} \in A_i( Y_{T_i}^t), M \in A_i( Y_{T_i}^t) \} \right]
\]  
\[ \leq P_e + P \left[ M \in A_i( Y_{T_i}^t) \right]. \]  
(98)

Let \( B, B^* \) and \( B_\tau \) be

\[
B \triangleq \ln \frac{P( y^t )}{P_{\{A\}}( y^t )},
\]  
(99a)

\[
B^* \triangleq \ln \frac{P( y^t )}{P_{\{A\}}( y^t )} 1_{\{ T < \infty \}}
\]  
(99b)

\[
B_\tau \triangleq \ln \frac{P( y^t )}{P_{\{A\}}( y^t )} 1_{\{ T \wedge \tau \leq \tau \}}, \quad \forall \tau \in \{ 1, 2, \ldots \}
\]  
(99c)

where \( T \wedge \tau \) is the minimum of \( T \) and \( \tau \).

Note that as \( \tau \) goes to infinity, \( B_\tau \to B \) and \( B_\tau \to B^* \) with probability one. Since \( |B_\tau| \leq T \ln \frac{1}{\tau} \) and \( E[T] < \infty \), we can apply the dominated convergence theorem [10] Theorem 3 p 187] to obtain

\[
E[B] = E[B^*] = \lim_{\tau \to \infty} E[B_\tau].
\]  
(100)

Finally for \( B \) and \( B^* \) defined in equation (99) we have

\[
E[B] = E \left[ \ln \frac{P( Y_t^i )}{P_{\{A\}}( Y_t^i )} \right]
\]  
(101a)

\[
E[B^*] = \sum_{y^t \in Y_{T_i}^t} P( y^t ) \ln \frac{P(y^t)}{P_{\{A\}}( y^t )}.
\]  
(101b)

Thus as a result of equations (100) and (101) we have

\[
E \left[ \ln \frac{P( Y_t^i )}{P_{\{A\}}( Y_t^i )} \right] = \sum_{y^t \in Y_{T_i}^t} P( y^t ) \ln \frac{P(y^t)}{P_{\{A\}}( y^t )}.
\]  
(102)
Furthermore using the definition of $P\{\mathcal{A}\}$ given in equation (39) we get
\[
E\left[\ln \frac{p(Y^\tau)}{P(\mathcal{A}_j) \cdot p(Y^\tau)}\right] = E\left[\ln \frac{p(Y^\tau_{j+1}|Y^\tau_j)}{P(\mathcal{A}_j) \cdot p(Y^\tau_{j+1}|Y^\tau_j)}\right] = \sum_{j=1}^k \xi_{i,j} \tag{103}
\]
where for all $i \geq 1$ and $j > i$
\[
\xi_{i,j} \overset{\Delta}{=} \begin{cases} 0 & \text{if } P[T_{j+1} = T_j] = 1 \\ \mathbb{E}\left[\ln \frac{p(Y^\tau_{j+1}|Y^\tau_j)}{P(\mathcal{A}_j) \cdot p(Y^\tau_{j+1}|Y^\tau_j)}\right] & \text{if } P[T_{j+1} = T_j] < 1 \end{cases}. \tag{104}
\]

Assume for the moment that,
\[
\xi_{i,j} \leq E[T_{j+1} - T_j] J(r_{j+1}) \tag{105}
\]
where $T_{k+1} = T$ and $r_j$ is defined in equation (47).

Then Lemma 5 follows from equations (98), (102), (103) and (105).

Above, we have proved Lemma 5 by assuming that the inequality given in (105) holds for all $i \in \{1, 2, \ldots, k\}$ and $j \in \{(i + 1), \ldots, (k + 1)\}$; below we prove that fact.

First note that if $P[T_{j+1} = T_j] = 1$ then as result of equations (47) and (103) equation (105) is equivalent to $0 \leq 0J(0)$ which holds trivially. Thus we assume hence forth that $P[T_{j+1} = T_j] < 1$, which implies $E[T_{j+1} - T_j] > 0$.

Let us consider the stochastic sequence
\[
U_\tau = \left[\left(- \ln \frac{p(Y^\tau_{j+1}|Y^\tau_j)}{P(\mathcal{A}_j) \cdot p(Y^\tau_{j+1}|Y^\tau_j)} + \sum_{k=T+1}^{\tau} J\left(I(M; Y_k|Y^{k-1})\right)\right) \right] \mathbb{1}_{\{\tau \geq T_j\}} \tag{106}
\]
where $I(M; Y_k|Y^{k-1})$ is the conditional mutual information between $M$ and $Y_k$ given $Y^{k-1}$, defined as
\[
I(M; Y_k|Y^{k-1}) \overset{\Delta}{=} \mathbb{E}\left[\ln \frac{p(Y_k|M, Y^{k-1})}{p(Y_k|Y^{k-1})} \right] Y^{k-1} \tag{107}
\]
Note that as it was the case for conditional entropy, while defining the conditional mutual information we do not take the average over the conditioned random variable. Thus $I(M; Y_k|Y^{k-1})$ is itself a random variable.

For $U_\tau$ defined in equation (106) we have
\[
U_{\tau+1} - U_\tau = \left(- \ln \frac{p(Y^\tau_{j+1}|Y^\tau_j)}{P(\mathcal{A}_j) \cdot p(Y^\tau_{j+1}|Y^\tau_j)} + J(I(M; Y_{\tau+1}|Y^\tau)\right) \mathbb{1}_{\{\tau \geq T_j\}}. \tag{108}
\]
Conditioned on $Y^\tau$ random variables $M - X_{\tau+1} - Y_{\tau+1}$ form a Markov chain: thus as a result of the data processing inequality for the mutual information we have $I(X_{\tau+1}; Y_{\tau+1}|Y^\tau) > I(M; Y_{\tau+1}|Y^\tau)$. Since $J(\cdot)$ is a decreasing function this implies that
\[
J(I(X_{\tau+1}; Y_{\tau+1}|Y^\tau)) \geq J(I(M; Y_{\tau+1}|Y^\tau)) \geq J(I(M; Y_{\tau+1}|Y^\tau)) \tag{109}
\]
Furthermore, because of the definitions of $J(\cdot)$, $P$ and $P(\mathcal{A})$ given in equations (17), (36) and (39), the convexity of Kullback Leibler divergence and Jensen’s inequality we have
\[
J(I(X_{\tau+1}; Y_{\tau+1}|Y^\tau)) \geq \mathbb{E}\left[\ln \frac{p(Y_{\tau+1}|Y^\tau)}{P(\mathcal{A}_j) \cdot p(Y_{\tau+1}|Y^\tau)} \right] Y^\tau. \tag{110}
\]
Using equations (107), (108) and (109) we get
\[
E[U_{\tau+1} | Y^\tau] \geq U_\tau. \tag{111}
\]
Recall that $\min_{x,y} W_x(y) = \lambda$ and $|J(\cdot)| \leq D$. Thus as a result of equation (107) we have
\[
E[U_{\tau+1} - U_\tau | Y^\tau] \leq \ln \frac{1}{\lambda} + D. \tag{112}
\]
As a result of (110), (111) and the fact that $U_0 = 0$, $U_\tau$ is a submartingale.

Recall that we have assumed that $P[T_{j+1} \leq T] = 1$ and $E[T] < \infty$; consequently
\[
E[T_{j+1}] < \infty. \tag{112}
\]
Because of (111) and (112) we can apply a version of Doob’s optional stopping theorem [10, Theorem 2, p 487] to the submartingale U_\tau and the stopping time T_{j+1} to obtain E[U_{T_{j+1}}] \geq E[U_0] = 0. Consequently,

$$E \left[ \ln \frac{p(Y_{T_{j+1}}^{j+1} | Y_j^{j+1})}{p(A_j) (Y_{T_{j+1}}^{j+1} | Y_j^{j+1})} \right] \leq E \left[ \sum_{\tau=T_{j+1}}^{T_{j+1}} J(I(M; Y_\tau | Y_{\tau-1})) \right].$$  \hspace{1cm} (113)

Note that as a result of the concavity of J(\cdot) and Jensen’s inequality we have

$$E \left[ \sum_{\tau=T_{j+1}}^{T_{j+1}} J(I(M; Y_\tau | Y_{\tau-1})) \right] = E[T_{j+1} - T_j] E \left[ \sum_{\tau \geq 1} I_{\{T_{j+1}+1 > \tau \geq T_j\}} \frac{J(I(M; Y_\tau | Y_{\tau-1}))}{E[T_{j+1} - T_j]} \right] \leq E[T_{j+1} - T_j] J \left( E \left[ \sum_{\tau \geq 1} I_{\{T_{j+1}+1 > \tau \geq T_j\}} \frac{J(I(M; Y_\tau | Y_{\tau-1}))}{E[T_{j+1} - T_j]} \right] \right).$$  \hspace{1cm} (114)

In order to calculate the argument of J(\cdot) in (114) consider the stochastic sequence

$$V_\tau = H(M|Y^\tau) + \sum_{j=1}^{\tau} I(M; Y_j | Y_j^{-1}).$$  \hspace{1cm} (115)

Clearly E[V_{\tau+1} | Y^\tau] = V_\tau and E[|V_\tau|] \leq \ln |\mathcal{M}| + C\tau < \infty. Hence V_\tau is a martingale.

Furthermore,

$$E[|V_{\tau+1} - V_\tau| | Y^\tau] \leq \ln |\mathcal{M}| + C.$$  \hspace{1cm} (116)

Recall that we have assumed that P[T_j \leq T_{j+1} \leq T] = 1 and E[T] < \infty; consequently

$$E[T_j] \leq E[T_{j+1}] < \infty.$$  \hspace{1cm} (117)

As a result of equations (116) and (117) we can apply Doob’s optimal stopping theorem, [10, Theorem 2, p 487] to V_\tau both at stopping time T_j and at stopping time T_{j+1}, i.e., E[V_{T_{j+1}}] = E[V_0] and E[V_{T_j}] = E[V_0]. Consequently,

$$E \left[ \sum_{\tau \geq 1} I_{\{T_{j+1}+1 > \tau \geq T_j\}} I(M; Y_\tau | Y_{\tau-1}) \right] = E \left[ H(M|Y^{T_j}) - H(M|Y^{T_{j+1}}) \right].$$  \hspace{1cm} (118)

Using equations (113), (114) and (118)

$$E \left[ \ln \frac{p(Y_{T_{j+1}}^{j+1} | Y_j^{j+1})}{p(A_j) (Y_{T_{j+1}}^{j+1} | Y_j^{j+1})} \right] \leq E[T_{j+1} - T_j] J \left( \frac{E[H(M|Y^{T_j}) - H(M|Y^{T_{j+1}})]}{E[T_{j+1} - T_j]} \right).$$  \hspace{1cm} (119)

Hence inequality given in (105) not only when P[T_{j+1} = T_j] = 1 but also when P[T_{j+1} = T_j] < 1. \hspace{1cm} \blacksquare

\textbf{F. Proof of Lemma \ref{lem:bound} for The Case E[T] < \infty}

\textbf{Proof:} In order to bound P_{e|m} from below we apply Lemma \ref{lem:bound} for (T_1, A_1) and (T_2, A_2) given in equations (48), (49), (50) and (51) and use the fact that J(\cdot) \leq D we get

$$\ln P_{e|m} \geq \frac{-b(P_e + |\mathcal{M}|^{-1}) - E[T_2] J \left( \frac{E[H(M) - H(M)|Y^{T_2}])}{E[T_2]} \right)}{1 - P_e - |\mathcal{M}|^{-1}}$$  \hspace{1cm} (120a)

$$\ln P_{\{A_3\} [\tilde{M} \notin A_2(Y^{T_2})]} \geq \frac{-b(P_e + P M \in A_2(Y^{T_2})) - E[T_2] D}{1 - P_e - |\mathcal{M}|^{-1}}$$  \hspace{1cm} (120b)

provided that |\mathcal{M}|^{-1} + P_e \leq 1/2 and P[M \in A_2(Y^{T_2})] + P_e \leq 1/2.

We start with bounding P_{\{A_3\} [\tilde{M} \notin A_2(Y^{T_2})]} from above and P[M \notin A_2(Y^{T_2})] from below.

- Since min_{x \in X, y \in \mathcal{Y}} W(x) = \lambda the posterior probability of a message at time \tau + 1 can not be smaller than \lambda times the posterior probability of the same message at time \tau. Hence for the stopping time T_2 defined in equation (50), random set A_2 defined in equation (51) and \delta < \frac{1}{2} we have

$$P[M \in A_2(Y^{T_2}) | Y^{T_2} = y^{T_2}] > \lambda \delta \hspace{1cm} \forall y^{T_2} \in Y^{T_2*}.$$  \hspace{1cm} (121)

\textsuperscript{33}The set A_2 is random in the sense that it depends on previous channel outputs.
As a result of the definition of $P_{\{A\}}(m, y^t)$ given in equation (39) we have,

$$P_{\{A\}}(m, y^t) < P(m, y^t) \frac{\mathbb{I}\{m \in A_2(y^{2})\}}{\lambda_0} \quad \forall m \in M, y^t \in Y^T. \quad (122)$$

If the decoded message $\hat{M}(y^t)$ is not in $A_2(y^{2})$ and message $m$ is in $A_2(y^{2})$ then $\hat{M}(y^t) \neq m$:

$$\mathbb{I}\{\hat{M}(y^t) \notin A_2(y^{2})\} \leq \mathbb{I}\{m \in A_2(y^{2})\} \quad \forall m \in M, y^t \in Y^T. \quad (123)$$

Using equations (122) and (123) we get

$$P_{\{A\}}(m, y^t) \mathbb{I}\{\hat{M}(y^t) \notin A_2(y^{2})\} < P(m, y^t) \frac{\mathbb{I}\{m \in A_2(y^{2})\}}{\lambda_0} \quad \forall m \in M, y^t \in Y^T. \quad (124)$$

If we sum over all $(m, y^t)$’s in $M \times Y^T$ and use equations (37) and (41) we get,

$$P_{\{A\}}[\hat{M} \notin A_2(Y^{2})] < \frac{P[\hat{M} \neq M]}{\lambda_0} = \frac{P_e}{\lambda_0}. \quad (125)$$

- The probability of an event $\Gamma_1$ is lower bounded by the probability of its intersection with any event $\Gamma_2$, i.e., $P[\Gamma_1] \geq P[\{\Gamma_1, \Gamma_2\}]$:

$$P_e = P[\hat{M} \neq M]$$

$$\geq P\left[\{\hat{M} \neq M, A_2(Y^{2}) = M\}\right]$$

$$= P[\hat{M} \neq M] P[A_2(Y^{2}) = M] P[A_2(Y^{2}) = M $$

Note that if $A_2(y^{2}) = M$ then $T$ is reached before any of the messages reach a posterior probability of $1 - \delta$. Thus

$$P[\hat{M} \neq M] P[A_2(Y^{2}) = M] > \delta \quad (126)$$

Thus as a result of equations (126) and (127) we have

$$P[A_2(Y^{2}) = M] < \frac{P_e}{\delta}. \quad (128)$$

On the other hand if $A_2(y^{2}) \neq M$, then the most likely message with a probability at least $(1 - \delta)$ is excluded from $A_2(y^{2})$. Thus

$$P[\hat{M} \neq M] P[A_2(Y^{2}) \neq M] \leq \delta \quad (129)$$

Using equations (128) and (129) together with total probability formula we get

$$P[\hat{M} \notin A_2(Y^{2})] \quad (125)$$

$$P[\hat{M} \notin A_2(Y^{2})] \quad (125)$$

$$P\left[M \in A_2(Y^{2}) \mid A_2(Y^{2}) \neq M\right] \quad (125)$$

$$P\left[M \in A_2(Y^{2}) \mid A_2(Y^{2}) = M\right] \quad (125)$$

$$+ P\left[M \in A_2(Y^{2}) \mid A_2(Y^{2}) \neq M\right] \quad (125)$$

$$P\left[M \in A_2(Y^{2}) \mid A_2(Y^{2}) \neq M\right] \quad (125)$$

$$< \frac{P_e}{\delta} + \delta. \quad (130)$$

We plug the bounds on $P_{\{A\}}[\hat{M} \notin A_2(Y^{2})]$ and $P[\hat{M} \notin A_2(Y^{2})]$ given in equations (125) and (130) in equation (120) to get

$$\ln P_{e|m} \geq \frac{-h(\hat{e}_1) - E[T_2] f\left(\frac{E[H(M|Y^{2})]}{E[T_2]}\right)}{1 - \epsilon_1} \quad (131a)$$

$$\ln \frac{P_e}{\lambda_0} \geq \frac{-h(\hat{e}_1) - E[T_2] D}{1 - \epsilon_1} \quad (131b)$$

provided that $\hat{e}_1 \leq 1/2$ where $\hat{e}_1 = P_e + \delta + P_e + |M|^{-1}$.

Now we bound $E[H(M|Y^{2})]$ from below. Note that $\mathbb{I}\{m \in A_2(y^{2})\}$ is a discrete random variable that is either zero or one; its conditional entropy given $Y^{2}$ is given by

$$H(\mathbb{I}\{m \in A_2(y^{2})\} \mid Y^{2}) = h\left(P\left[M \in A_2(Y^{2}) \mid Y^{2}\right]\right). \quad (132)$$
Furthermore since $\mathbb{I}_{\{M \in \mathcal{A}_2(Y^{T_2})\}}$ is a function of $Y^{T_2}$ and $M$, chain rule entropy implies that

$$H(M|Y^{T_2}) = H(\mathbb{I}_{\{M \in \mathcal{A}_2(Y^{T_2})\}}|Y^{T_2}) + E\left[H(M|Y^{T_2}, \mathbb{I}_{\{M \in \mathcal{A}_2(Y^{T_2})\}}|Y^{T_2}\right].$$ (133)

Since $\mathcal{A}_2(Y^{T_2})$ has at most $|M|$ elements and its complement, $\mathcal{M} \setminus \mathcal{A}_2(Y^{T_2})$, has at most one element, we can bound the conditional entropy of the messages as follows

$$H(M|Y^{T_2}, \mathbb{I}_{\{M \in \mathcal{A}_2(Y^{T_2})\}}) \leq \mathbb{I}_{\{M \in \mathcal{A}_2(Y^{T_2})\}} \ln |M|. \tag{134}$$

Thus using equations (132), (133) and (134) we get

$$H(M|Y^{T_2}) \leq h\left(P\left[M \in \mathcal{A}_2(Y^{T_2}) \mid Y^{T_2}\right]\right) + P\left[M \in \mathcal{A}_2(Y^{T_2}) \mid Y^{T_2}\right] \ln |M|. \tag{135}$$

Then using concavity of the binary entropy function $h(\cdot)$ together with equations (130) and (135) we get

$$E\left[H(M|Y^{T_2})\right] < h(\delta + \frac{P_x}{\delta}) + (\delta + \frac{P_x}{\delta}) \ln |M|. \tag{136}$$

provided that $\delta + \frac{P_x}{\delta} \leq 1/2$.

If we plug in equation (136) and the identity $H(M) = \ln |M|$ in equation (131) we get,

$$(1 - \tilde{\epsilon}_1)\ln \frac{P_x}{E[T]} \geq -\frac{h(\tilde{\epsilon}_1)}{E[T]} - \eta J\left((1-\tilde{\epsilon}_1)R - h(\tilde{\epsilon}_1)/E[T]\right) - (1 - \eta)D \tag{137a}$$

$$-(1 - \tilde{\epsilon}_1)E \geq -\frac{h(\tilde{\epsilon}_1)+\ln \lambda \delta}{E[T]} - (1 - \eta)D \tag{137b}$$

provided that $\tilde{\epsilon}_1 \leq 1/2$ where $\eta = \frac{E[T]}{\tilde{\epsilon}_1}, \tilde{\epsilon}_1 = P_x + \delta + \frac{P_x}{\delta} + |M|^{-1}$, $R = \frac{\ln P_x}{E[T]}$ and $E = -\ln P_x$.

Note that the inequality given in equation (137b) bounds the value of $\eta$ from above,

$$\eta \leq 1 - \frac{(1-\tilde{\epsilon}_1)E}{D} \tag{138}$$

where $\tilde{\epsilon}_2 = \frac{h(\tilde{\epsilon}_1)\ln \eta \delta}{E[T]}$.

Furthermore for any $\eta_1 \leq \eta_2 \leq \frac{\tilde{\epsilon}_1}{\tilde{\epsilon}_1}$ as a result of concavity of $J(\cdot)$ we have

$$\eta_1 J\left(\frac{R}{\eta_1}\right) + (1 - \eta_1)D = \eta_1 J\left(\frac{R}{\eta_2}\right) + (\eta_2 - \eta_1)J(0) + (1 - \eta_2)D \leq \eta_2 J\left(\frac{R}{\eta_2}\right) + (1 - \eta_2)D. \tag{139}$$

Using equations (138), (139) we see that the bound in equation (137a) is lower bounded by its value at $\eta = 1 - \frac{(1-\tilde{\epsilon}_1)E}{D}$ if $E \geq \frac{\tilde{\epsilon}_2}{1-\tilde{\epsilon}_1}$ and by its value at $\eta = 1$ otherwise, i.e.,

$$\ln \frac{P_x}{E[T]} \geq \left\{ -E - \frac{(1 - E - \tilde{\epsilon}_1)}{D} J\left(\frac{R - \tilde{\epsilon}_1}{1 - \tilde{\epsilon}_1}\right) \quad \text{if} \quad E \geq \frac{\tilde{\epsilon}_2}{1-\tilde{\epsilon}_1} \right\}$$

$$-\frac{\tilde{\epsilon}_2}{1-\tilde{\epsilon}_1} = \frac{\tilde{\epsilon}_2}{1-\tilde{\epsilon}_1} = J((1 - \tilde{\epsilon}_1)R - \tilde{\epsilon}_2) \quad \text{if} \quad E < \frac{\tilde{\epsilon}_2}{1-\tilde{\epsilon}_1}$$

where $\tilde{\epsilon} = \frac{\tilde{\epsilon}_1 D + \tilde{\epsilon}_2}{1-\tilde{\epsilon}_1}$.

Then, for the case $E \geq \frac{\tilde{\epsilon}_2}{1-\tilde{\epsilon}_1}$ Lemma 6 follows from the fact that $J(\cdot)$ is a non-negative decreasing function. For the case $E < \frac{\tilde{\epsilon}_2}{1-\tilde{\epsilon}_1}$ in Lemma 6 follows from the fact that $J(\cdot)$ is a concave non-negative decreasing function.

\[\square\]

**G. Proof of Lemma 7 for The Case $E[T] < \infty$**

**Proof:** We start with proving the bounds given in equations (57) and (58).

- Let us start with the bound on $P_{\{A_i\}}\left[M \notin A_i(Y^{T_i})\right]$ given in equation (57). Since $\min_{x \in \mathcal{X}, y \in \mathcal{Y}} W_x(y) = \lambda$, the posterior probability of a $m^i \in \mathcal{M}^\lambda$ at time $\tau + 1$ can not be smaller than $\lambda$ times its value at time $\tau$. Hence for $\delta < 1/2$, as a result definitions of $T_i$ and $A_i(Y^{T_i})$ given in equations (55) and (56), we have

$$P\left[M \in A_i(Y^{T_i}) \mid Y^{T_i} = y^\tau\right] > \lambda \delta \quad \forall y^\tau \in \mathcal{Y}^{T_i}, i \in \{1, 2, \ldots, \ell\}. \tag{58}$$
Then as a result of the definition of $P_{\{A_i\}}(m, y^i)$ given in equation (39), we have,

$$
P_{\{A_i\}}(m, y^i) = \frac{P(m) \prod_{i=1}^{\ell} \mathbb{I}_{\{m \in A_i(y^{i-1})\}}}{\lambda_0} \quad \forall m \in M, y^i \in Y^T, \ i \in \{1, 2, \ldots, \ell\}. \tag{140}
$$

For $A_i(y^i)$ given in equation (56), if the decoded message $\hat{M}(y^i)$ is not in $A_i(y^i)$ but $m$ is in $A_i(y^i)$ then $\hat{M}(y^i) \neq m^i$:

$$
\mathbb{I}_{\{\hat{M}(y^i) \notin A_i(y^i)\}} \prod_{i=1}^{\ell} \mathbb{I}_{\{m \in A_i(y^{i-1})\}} \leq \mathbb{I}_{\{\hat{M} \neq m^i\}} \quad \forall m \in M, y^i \in Y^T, \ i \in \{1, 2, \ldots, \ell\}. \tag{141}
$$

Using equations (140) and (141) we get

$$
P_{\{A_i\}}(m, y^i) \mathbb{I}_{\{\hat{M}(y^i) \notin A_i(y^i)\}} < \frac{P(m) \mathbb{I}_{\{\hat{M} \neq m^i\}}}{\lambda_0} \quad \forall m \in M, y^i \in Y^T, \ i \in \{1, 2, \ldots, \ell\}. \tag{142}
$$

If we sum over all $(m, y^i)$’s in $M \times Y^T$ and use equations (37) and (41) we get,

$$
P_{\{A_i\}}(\hat{M} \notin A_i(Y^T)) \leq \frac{P(M \neq M^i)}{\lambda_0} \quad \forall i \in \{1, 2, \ldots, \ell\}. \tag{144}
$$

Let us now prove the bound on $P\left[M \in A_i(Y^T)\right]$ given in equation (58).

- If $A_i(Y^T) \neq M$, then at $T_i$ there is a $m^i$ with posterior probability $(1 - \delta)$ and all the messages $m$ of the form $m = (m^1, m_{i+1}, \ldots, m_k)$ are excluded from $A_i$. Consequently we have

$$
P\left[M \in A_i(Y^T) \mid A_i(Y^T) \neq M\right] < \delta. \tag{143}
$$

- If $A_i(Y^T) = M$, then at $T_i$ there is no $m^i$ with posterior probability $(1 - \delta)$ and $T_i = T$. Since $\hat{M}^i \neq M^i$ implies that $\hat{M} \neq M$ we have

$$
P\left[\hat{M} \neq M \mid A_i(Y^T) = M\right] \geq \delta. \tag{144}
$$

As a result of total probability formula for $P[\hat{M} \neq M]$ we have

$$
P_e = P[\hat{M} \neq M | A_i(Y^T) = M] P[A_i(Y^T) = M] + P[\hat{M} \neq M | A_i(Y^T) \neq M] P[A_i(Y^T) \neq M] \geq \delta P[A_i(Y^T) = M]. \tag{145}
$$

If use the total probability formula for $P[M \in A_i(Y^T)]$ together with equations (143) and (145) we get

$$
P[M \in A_i(Y^T)] = P\left\{M \in A_i(Y^T), A_i(Y^T) \neq M\right\} + P\left\{M \in A_i(Y^T), A_i(Y^T) = M\right\} \leq P[M \in A_i(Y^T), A_i(Y^T) \neq M] + P[A_i(Y^T) = M] \leq \delta + \frac{P_e}{\delta}. \tag{146}
$$

We apply Lemma 5 for $(T_1, A_1), \ldots, (T_k, A_k)$ defined in equations (55) and (56); use the bounds on $P_{\{A_i\}}(\hat{M} \notin A_i(Y^T))$ and $P[M \in A_i(Y^T)]$ given in (57) and (58). Then we can conclude that if $P_e + \delta + P_e/\delta \leq 1/2$ then

$$
(1 - \tilde{c}_3)E_i \leq \tilde{c}_5 + \sum_{j=i+1}^{\ell+1} \nu_j J(r_j) \quad i = 1, 2, \ldots, \ell \tag{146}
$$

where $R_i$, $E_i$, $\tilde{c}_3$ and $\tilde{c}_5$ are defined in Lemma 7, $r_j$’s are defined in equation (47) of Lemma 5 and $\nu_j$’s are defined as follows.$^{34}$

$$
\nu_j \triangleq \frac{E[T_j] - E[T_{j-1}]}{E[T]} \quad \forall j \in \{1, 2, \ldots, \ell + 1\} \tag{147}
$$

$^{34}$We use the convention $T_0 = 0$ and $T_{\ell+1} = T$. 

Depending on the values of \( \nu_j \) and \( r_j \) the bound in equation (146) takes different values. However \( \nu_j \) and \( r_j \) are not changing freely. As a result of equation (118) and the fact that binary entropy function is an increasing function on the interval \( [0, 1] \), we have

\[
  r_j \leq C \quad j \in \{1, 2, \ldots, (\ell + 1)\}. \tag{148}
\]

In addition \( \nu_j \)'s and \( r_j \)'s are constrained by the definitions of \( T_j \) and \( A_j(Y^{T_j}) \) given in equations (55) and (56). At \( T_j \) with high probability one element of \( \mathcal{M} \) has a posterior probability \( (1 - \delta) \). Below we use this fact to bound \( \mathbb{E}[H(M|Y^{T_j})] \) from above. Then we turn this bound into a constraint on the values of \( \nu_j \)'s and \( r_j \)'s and use that constraint together with equations (146), (148) to bound \( E_i \)'s from above.

For all \( j \in \{1, 2, \ldots, \ell\} \), \( 1 \{M \in A_j(Y^{T_j})\} \) is a discrete random variable that is either zero or one; its conditional entropy given by

\[
  H(1 \{M \in A_j(Y^{T_j})\}|Y^{T_j}) = h \left( P \left[ M \in A_j(Y^{T_j}) \right] Y^{T_j} \right), \tag{149}
\]

Furthermore since \( 1 \{M \in A_j(Y^{T_j})\} \) is a function of \( Y^{T_j} \) and \( M \), the chain rule entropy implies that

\[
  H(M|Y^{T_j}) = H(1 \{M \in A_j(Y^{T_j})\}|Y^{T_j}) + \mathbb{E}[H(M|Y^{T_j}, 1 \{M \in A_j(Y^{T_j})\})|Y^{T_j}] \tag{150}
\]

Note that \( A_i(Y^{T_i}) \) has at most \( |\mathcal{M}| \) elements and its complement, \( \mathcal{M} \setminus A_i(Y^{T_i}) \), has at most \( |\mathcal{M}| - |\mathcal{M}^i| \) elements. We can bound the conditional entropy of the messages \( H(M|Y^{T_j}, 1 \{M \in A_j(Y^{T_j})\}) \) as follows

\[
  H(M|Y^{T_j}, 1 \{M \in A_j(Y^{T_j})\}) \leq 1 \{M \in A_j(Y^{T_j})\} \log |\mathcal{M}| + 1 \{M \in \mathcal{M} \setminus A_j(Y^{T_j})\} \log |\mathcal{M}^i| \nonumber
\]

thus using equations (149), (150) and (151) we get

\[
  H(M|Y^{T_j}) \leq h \left( P \left[ M \in A_j(Y^{T_j}) \right] Y^{T_j} \right) + \log \frac{|\mathcal{M}| + P \left[ M \in \mathcal{M} \setminus A_j(Y^{T_j}) \right] Y^{T_j}}{|\mathcal{M}|} \tag{152}
\]

If we take the expectation of both sides of the inequality (152) and use the concavity of the binary entropy function, we get

\[
  \mathbb{E}[H(M|Y^{T_j})] \leq h \left( P \left[ M \in A_j(Y^{T_j}) \right] Y^{T_j} \right) + \log \frac{|\mathcal{M}| + P \left[ M \in \mathcal{M} \setminus A_j(Y^{T_j}) \right] Y^{T_j}}{|\mathcal{M}|} \log |\mathcal{M}^i| \nonumber
\]

Using the inequality given (58) and the fact that binary entropy function is an increasing function on the interval \([0, 1/2]\) we see that

\[
  \mathbb{E}[H(M|Y^{T_j})] < h \left( P_e + \delta + \frac{P_e}{\delta} \right) + \log \frac{|\mathcal{M}| + P \left[ M \in \mathcal{M} \setminus A_j(Y^{T_j}) \right] Y^{T_j}}{|\mathcal{M}|} \log |\mathcal{M}^i| \tag{153}
\]

provided that \( P_e + \delta + \frac{P_e}{\delta} \leq 1/2 \).

Note that as a result of Fano’s inequality for \( \mathbb{E}[H(M|Y^T)] \) we have

\[
  \mathbb{E}[H(M|Y^T)] < h(P_e) + P_e \log |\mathcal{M}|. \tag{154}
\]

If we divide both sides of the inequalities (153) and (154) to \( \mathbb{E}[T] \), we see that following bounds holds

\[
  \frac{\mathbb{E}[H(M|Y^{T_j})]}{\mathbb{E}[T]} \leq \bar{c} + R - \sum_{j=1}^i R_i + \bar{c}_3 \sum_{j=1}^i R_i. \quad i = 1, 2, \ldots, \ell \tag{155a}
\]

\[
  \mathbb{E} \left[ H(M|Y^T) \right] \leq h(P_e) + P_e R. \tag{155b}
\]

Note that

\[
  \frac{\mathbb{E}[H(M|Y^{T_j})]}{\mathbb{E}[T]} = R - \sum_{j=1}^i \nu_j x_j \quad i = 1, 2, \ldots, (\ell + 1) \tag{156}
\]

Using equations (155) and (156) we get,

\[
  \sum_{j=1}^i \nu_j x_j \geq (1 - \bar{c}_3) \sum_{j=1}^i R_j - \bar{c}_4 \quad i = 1, 2, \ldots, \ell \tag{157a}
\]

\[
  \sum_{j=1}^{\ell + 1} \nu_j x_j \geq (1 - P_e) R - \frac{h(P_e)}{\mathbb{E}[T]} \tag{157b}
\]
where $r_j$'s and $\nu_j$'s given in equation (47) and (47) respectively.

Thus using equations (47), (146), (147), (148) and (157) we reach the following conclusion. For any variable length block code satisfying the hypothesis of the Lemma 7 and for any positive $\delta$ such that $P_e + \delta + \frac{P}{\delta} \leq \frac{1}{2}$

\[
(1 - \tilde{e}_3)E_i - \tilde{e}_5 \leq \sum_{j=i+1}^{\ell+1} \nu_j J(r_j) \quad i = 1, 2, \ldots, \ell \quad (158a)
\]
\[
(1 - \tilde{e}_3) \sum_{j=1}^{i} R_j - \tilde{e}_4 \leq \sum_{j=1}^{i} \nu_j r_j \quad i = 1, 2, \ldots, \ell \quad (158b)
\]
\[
(1 - P_e)R - \frac{h(P_e)}{E[1]} \leq \sum_{j=1}^{\ell+1} \nu_j r_j \quad (158c)
\]

for some $(\nu_1^{\ell+1}, \nu_1^{\ell+1})$ such that

\[
\begin{align*}
    r_i &\in [0, C] \quad i = 1, 2, \ldots, (\ell + 1) \\
    \nu_i &\geq 0 \quad i = 1, 2, \ldots, (\ell + 1) \\
    \sum_{i=1}^{\ell+1} \nu_i &= 1
\end{align*}
\]

(159a) (159b) (159c)

We show below if the constraints given in equation (158) is satisfied for some $(\nu_1^{\ell+1}, \nu_1^{\ell+1})$ satisfying (159), constraints given in (60) is satisfied for some $(\eta_1^{\ell+1})$ satisfying (61).

One can confirm numerically that

\[
(1 - \tilde{e}_3) \ln 2 > h(\tilde{e}_3) \quad \forall \tilde{e}_3 \in [0, \frac{1}{2}]
\]

Recall that we have assumed that $P_e + \delta + \frac{P}{\delta} \leq \frac{1}{2}$, i.e., $\tilde{e}_3 \leq \frac{1}{2}$. Thus,

\[
(1 - \tilde{e}_3)R_1 - \tilde{e}_4 > 0.
\]

(160)

Let $\eta_1, \tilde{r}_1, \tilde{v}_2$ and $\tilde{r}_2$ be

\[
\begin{align*}
    \eta_1 &= \frac{(1 - \tilde{e}_3)R_1 - \tilde{e}_4}{r_1} \\
    \tilde{r}_1 &= r_1 \\
    \tilde{v}_2 &= \nu_2 + \nu_1 - \eta_1 \\
    \tilde{r}_2 &= \frac{r_2 \nu_2 + (\nu_1 - \eta_1) \tilde{r}_1}{\nu_2}
\end{align*}
\]

Note that $(\eta_1, \tilde{v}_2, \nu_1^{\ell+1}, \tilde{r}_1, \tilde{r}_2, \tilde{r}_3^{\ell+1})$ satisfies (158b), (158c) and (159) by construction. Furthermore as a result of concavity of $J(\cdot)$ we have,

\[
\nu_1 J(r_1) + \nu_2 J(r_2) \leq \eta_1 J(\tilde{r}_1) + \tilde{v}_2 J(\tilde{r}_2).
\]

Thus $(\eta_1, \tilde{v}_2, \nu_1^{\ell+1}, \tilde{r}_1, \tilde{r}_2, \tilde{r}_3^{\ell+1})$ also satisfies (158a).

For $j \geq 2$ we use $\tilde{v}_j$ and $\tilde{r}_j$ to define $\eta_j$, $\nu_j^{\ell+1}$ and $\tilde{r}_j+1$ as follows:

\[
\begin{align*}
    \eta_j &= \frac{(1 - \tilde{e}_3)R_i - \tilde{e}_4}{r_i} \quad (161a) \\
    \tilde{v}_j+1 &= \nu_j+1 + \tilde{v}_j - \eta_j \quad (161b) \\
    \tilde{r}_{j+1} &= \frac{r_j \nu_j+1 + (\nu_j - \eta_j) \tilde{r}_j}{\nu_{j+1}} \quad (161c)
\end{align*}
\]

Using the fact that $(\eta_1^{j-1}, \tilde{v}_j, \nu_1^{j+1}, \tilde{r}_1^{j+1}, \tilde{r}_2^{j+1})$ satisfies (158) and (159) and the concavity of $J(\cdot)$ we can show that $(\eta_1^{j}, \tilde{v}_j+1, \nu_1^{j+1}, \tilde{r}_1^{j+1}, \tilde{r}_2^{j+1})$ also satisfies (158) and (159). We repeat the iteration given in equation (161) until we reach $\tilde{v}_{\ell+1}$ and $\tilde{r}_{\ell+1}$ and we let $\eta_{\ell+1} = \tilde{v}_{\ell+1}$.

Then we conclude that for any variable length block code satisfying the hypothesis of the Lemma 7 and for any positive $\delta$ such that $P_e + \delta + \frac{P}{\delta} \leq \frac{1}{2}$

\[
\begin{align*}
    (1 - \tilde{e}_3)E_i - \tilde{e}_5 &\leq \sum_{j=i+1}^{\ell+1} \eta_j J(\tilde{r}_j) \quad i = 1, 2, \ldots, \ell \quad (162a) \\
    (1 - \tilde{e}_3)R_i - \tilde{e}_4 I_{(i=1)} &= \tilde{r}_i \eta_i \quad i = 1, 2, \ldots, \ell \quad (162b) \\
    (\tilde{e}_3 - P_e)R + \frac{h(P_e) - h(P_e^*)}{E[1]} &\leq \tilde{r}_{\ell+1} \eta_{\ell+1} \quad (162c)
\end{align*}
\]

for some \((\eta_1, \ldots, \eta_{\ell+1}, \bar{r}_1, \ldots, \bar{r}_{\ell+1})\) such that\(^{35}\)
\[
\begin{align*}
\bar{r}_i &\in [0, C] \quad i = 1, 2, \ldots, (\ell + 1) \\
\eta_i &\geq 0 \quad i = 1, 2, \ldots, (\ell + 1) \\
\sum_{i=1}^{\ell+1} \eta_i &= 1.
\end{align*}
\]

The Lemma 7 follows from the fact that \(J(\cdot) \leq D\).

**H. Codes with Infinite Decoding Time on Channels with Positive Transition Probabilities**

In this section we consider variable length block codes on discrete memoryless channels with positive transition probabilities, i.e., \(\min_{x \in \mathcal{X}, y \in \mathcal{Y}} W_x(y) > 0\), and derive lower bounds to the probabilities of various error events. These bounds, i.e., equations (166), (172) and (173), enable us to argue that Lemma 6 and Lemma 7 hold for variable length block codes with infinite expected decoding time, i.e., \(E[T] = \infty\).

1) \(P_e > 0\): On discrete memoryless channel such that \(\min_{x \in \mathcal{X}, y \in \mathcal{Y}} W_x(y) = \lambda\) the posterior probability of any message \(m \in \mathcal{M}\) at time \(\tau\) is lower bounded as
\[
P[M = m | Y^\tau] \geq \left( \frac{\lambda}{1 - \lambda} \right)^\tau \frac{1}{|\mathcal{M}|}.
\]

Then conditioned on the event \(\{T = \tau\}\) the probability of erroneous decoding is lower bounded as
\[
P[\hat{M} \neq M | T = \tau] \geq \frac{|\mathcal{M}|-1}{|\mathcal{M}|} \left( \frac{\lambda}{1 - \lambda} \right)^\tau.
\]

Note that since \(P[T < \infty] = 1\), the error probability of any variable length code satisfies
\[
P_e = \sum_{\tau=1}^{\infty} P[M \neq \hat{M} | T = \tau] P[T = \tau].
\]

Using equation (164) and (165) we get
\[
P_e \geq \frac{|\mathcal{M}|-1}{|\mathcal{M}|} E \left[ \left( \frac{\lambda}{1 - \lambda} \right)^T \right].
\]

Note that equation (166) implies that for a variable length code with infinite expected decoding time not only the rate \(R\) but also the error exponent \(E\) is zero.

2) If \(P_e + \frac{1}{|\mathcal{M}|} < 1\) then \(\min_m P_{e|m} > 0\): Note that since \(P[T < \infty] = 1\) and \(|\mathcal{M}| < \infty\),
\[
P[T < \infty | M = m] = 1 \quad \forall m \in \mathcal{M}.
\]

For any variable length block code such that \(P_e + \frac{1}{|\mathcal{M}|} < 1\), let \(\tau^*\) be
\[
\tau^* = \min \left\{ \tau : \max_{m \in \mathcal{M}} P[T > \tau | M = m] \leq \frac{|\mathcal{M}|-1}{|\mathcal{M}|} - P_e \right\}.
\]

Since \(P[T < \infty | M = m] = 1\) for all \(m\) in \(\mathcal{M}\) and \(\mathcal{M}\) is finite, \(\tau^*\) is finite.

Note that for any \(\tau, m\) and \(\hat{m}\) we have,
\[
P[Y^\tau = y^\tau | M = m] \geq \left( \frac{\lambda}{1 - \lambda} \right)^\tau P[Y^\tau = y^\tau | M = \hat{m}]
\]

Then using equation (168) we get,
\[
P_{e|m} \geq \sum_{\hat{m} \neq m} P \left[ \{ \hat{M} = \hat{m}, T \leq \tau^* \} \mid M = m \right]
\]
\[
= (\frac{1}{1 - \lambda})^{\tau^*} \sum_{\hat{m} \neq m} P \left[ \{ \hat{M} = \hat{m}, T \leq \tau^* \} \mid M = \hat{m} \right]
\]
\[
\geq (\frac{1}{1 - \lambda})^{\tau^*} \sum_{\hat{m} \neq m} \left[ P \left[ \hat{M} = \hat{m} \mid M = \hat{m} \right] - P[T > \tau^* | M = \hat{m}] \right] \quad (169)
\]

\(^{35}\)One can replace the inequality in equation (162c) by equality because \(J(\cdot)\) is a decreasing function.
Note that as a result of equation (167) we have,
\[ P[T > \tau^* | M = \tilde{m}] \leq \left( \frac{|M|-1}{|M|} - P_e \right) \quad \forall \tilde{m} \in M \]

Furthermore
\[ \sum_{\tilde{m} \neq m} P[\tilde{M} = \tilde{m} | M = m] \geq |M|(1 - P_e) - 1 \]

Thus using equations (169), (170) and (171) we get
\[ \min_{m \in M} P_e |m| \geq (\frac{\lambda}{1-\lambda})^{\tau^*} \left( 1 - \frac{1}{|M|} - P_e \right) \]

where \( \tau^* \) is a finite integer defined in equation (167).

3) For all \( i \in \{1, 2, \ldots, \ell\} \), \( P_e(i) > 0 \): For a variable length block code with message set \( M \) of the form \( M = M_1 \times M_2 \times \ldots \times M_k \) on a discrete memoryless channel such that \( \min_{x \in X, y \in Y} W(x) = \lambda \) the posterior probability of any element of \( M_i \) at time \( \tau \) is lower bounded as
\[ P[M^i = m^i | Y^\tau] \geq \left( \frac{\lambda}{1-\lambda} \right) \frac{1}{|M_i|} \quad \forall m^i \in M_i, \forall i \{1, 2, \ldots, \ell\}. \]

Then conditioned on the event \( \{T = \tau\} \) the probability of decoding the \( i \)th sub-message erroneously is lower bounded as
\[ P[\tilde{M}^i \neq M^i | T = \tau] \geq \frac{|M^i|-1}{|M^i|} \left( \frac{\lambda}{1-\lambda} \right)^\tau. \]

Since \( P[T < \infty] = 1 \), \( P_e(i) \) satisfies
\[ P_e = \sum_{\tau=1}^\infty P[M^i \neq \tilde{M}^i | T = \tau] P[T = \tau]. \]

Using equation (173) and (174) we get
\[ P_e(i) \geq \frac{|M^i|-1}{|M^i|} E \left( \left( \frac{\lambda}{1-\lambda} \right)^\tau \right) \quad \forall i \{1, 2, \ldots, \ell\}. \]

Equation (175) implies that for any variable length code with infinite expected decoding time on a DMC without any zero probability transition, not only the rates but also the error exponents of the sub-messages are zero.

I. Proof of Theorem 7

Proof: In Section IV-C it is shown that for any rate \( R \in [0, C] \), error exponent \( E \in [0, (1 - R/T)D] \) there exists a reliable sequence \( Q \) such that \( R_Q = R, E_Q = E, E_{md,Q} = E + (1 - \frac{E}{D})J\left( \frac{R}{1-E/D} \right) \). Thus as a result of the definition of \( E_{md}(R, E) \) given in equation (13) we have
\[ E_{md}(R, E) \geq E + (1 - \frac{E}{D})J\left( \frac{R}{1-E/D} \right). \]

In Section IV-C we have shown that any reliable sequence of codes \( Q \) with rate \( R_Q \) and error exponent \( E_Q \) satisfies
\[ E_{md,Q} \leq E_Q + (1 - \frac{E_Q}{D})J\left( \frac{R_Q}{1-E_Q/D} \right). \]

Thus, using the fact that \( J(\cdot) \) is a decreasing concave function we can conclude that
\[ \max_{R_Q \geq R} E_{md,Q} \leq E + (1 - \frac{E}{D})J\left( \frac{R}{1-E/D} \right). \]

Consequently as a result of the definition of \( E_{md}(R, E) \) given in equation (13) we have
\[ E_{md}(R, E) \leq E + (1 - \frac{E}{D})J\left( \frac{R}{1-E/D} \right). \]
Thus using equations (175) and (177) we can conclude that

\[ E_{md}(R, E) = E + \left(1 - \frac{R}{D}\right) J\left(\frac{R}{1-E/D}\right). \]  

(178)

In order to prove the concavity of \( E_{md}(R, E) \) in \((R, E)\) pair, let \((R_a, E_a)\) and \((R_b, E_b)\) be two pairs such that

\[
R_a \in [0, C] \quad E_a \leq (1 - \frac{R_a}{C})D \quad (179a)
\]

\[
R_b \in [0, C] \quad E_b \leq (1 - \frac{R_b}{C})D. \quad (179b)
\]

Then for any \( \alpha \in [0, 1] \) let \( R_\alpha \) and \( E_\alpha \) be

\[
R_\alpha = \alpha R_a + (1 - \alpha)R_b
E_\alpha = \alpha E_a + (1 - \alpha)E_b. \quad (180a)
\]

From equations (179) and (180) we have

\[
R_\alpha \in [0, C] \quad E_\alpha \leq (1 - \frac{R_\alpha}{C}). \quad (181)
\]

Furthermore using the concavity of \( J(\cdot) \) we get,

\[
\alpha E_{md}(R_a, E_a) + (1 - \alpha)E_{md}(R_b, E_b)
= \alpha \left(E_a + (1 - \frac{E_a}{D}) J\left(\frac{R_a}{1-E_a/D}\right)\right) + (1 - \alpha) \left(E_b + (1 - \frac{E_b}{D}) J\left(\frac{R_b}{1-E_b/D}\right)\right)
= E_a + \alpha \left(1 - \frac{E_a}{D}\right) J\left(\frac{R_a}{1-E_a/D}\right) + (1 - \alpha) \left(1 - \frac{E_b}{D}\right) J\left(\frac{R_b}{1-E_b/D}\right)
\leq E_a + (1 - \frac{E_a}{D}) J\left(\frac{\alpha R_a + (1 - \alpha)R_b}{1-E_a/D}\right)
= E_{md}(R_\alpha, E_\alpha). \quad (182)
\]

Thus \( E_{md}(R, E) \) is jointly concave in rate exponent pairs.

\[ \square \]

\( \text{J. Proof of Theorem 2} \)

\textbf{Proof:} In Section (IV-D) it is shown that for any positive integer \( \ell \) a rates-exponents vector \((\vec{R}, \vec{E})\) is achievable if there exists a time sharing vector \( \vec{n} \) such that,

\[
E_i \leq (1 - \sum_{j=1}^{\ell} \eta_{j,i})D + \sum_{j=i+1}^{\ell} \eta_{j,i} J\left(\frac{R_{j,i}}{\eta_{j,i}}\right) \quad \forall i \in \{1, 2, \ldots, \ell\} \quad (183a)
\]

\[
R_i \leq C\eta_{i,i} \quad \forall i \in \{1, 2, \ldots, \ell\} \quad (183b)
\]

\[
\eta_{i,i} \geq 0 \quad \forall i \in \{1, 2, \ldots, \ell\} \quad (183c)
\]

\[
\sum_{j=1}^{\ell} \eta_{j,i} \leq 1 \quad (183d)
\]

Thus the existence of a time sharing vector \( \vec{n} \) satisfying (183) is a sufficient condition for the achievability of a rates-exponents vector \((\vec{R}, \vec{E})\).

For any reliable code sequence \( Q \) whose message sets are of the form \( \mathcal{M}^{(\kappa)} = \mathcal{M}_1^{(\kappa)} \times \mathcal{M}_2^{(\kappa)} \times \ldots \times \mathcal{M}_\ell^{(\kappa)} \), Lemma 7 with \( \delta = \frac{1}{\ln P_a} \) implies that there exists a sequence \( \vec{\eta}_\kappa \) such that

\[
(1 - \tilde{\epsilon}_{3,\kappa})E_{i,\kappa} - \tilde{\epsilon}_{5,\kappa} \leq (1 - \sum_{j=1}^{\ell} \eta_{j,i,\kappa})D + \sum_{j=i+1}^{\ell} \eta_{j,i,\kappa} J\left(\frac{(1-\tilde{\epsilon}_{3,\kappa})R_{i,\kappa}}{\eta_{j,i,\kappa}}\right) \quad i = 1, 2, \ldots, \ell \quad (184a)
\]

\[
(1 - \tilde{\epsilon}_{3,\kappa})R_{i,\kappa} - \tilde{\epsilon}_{4,\kappa}1_{\{i=1\}} \leq C\eta_{i,i,\kappa} \quad \eta_{i,i,\kappa} \geq 0 \quad \forall i = 1, 2, \ldots, \ell \quad (184b)
\]

\[
\sum_{j=1}^{\ell} \eta_{j,i,\kappa} \leq 1 \quad (184c)
\]

where \( R_{i,\kappa} = \frac{\ln \mathcal{M}^{(\kappa)}_i}{E(\kappa)|1(\kappa)} \), \( E_{i,\kappa} = \frac{\ln P_a^{(i)}}{E(\kappa)|1(\kappa)} \), \( \tilde{\epsilon}_{3,\kappa} = \frac{P_a^{(i)\kappa} + 1}{P_a^{(i)}} \ln P_a^{(i)} \), \( \tilde{\epsilon}_{4,\kappa} = \frac{h(\tilde{\epsilon}_{3,\kappa})}{E(\kappa)|1(\kappa)} \), \( \tilde{\epsilon}_{5,\kappa} = \frac{h(\tilde{\epsilon}_{3,\kappa})}{E(\kappa)|1(\kappa)} - E(\kappa)|1(\kappa). \)
Note that as a result of equation (184) all members of the sequence $\tilde{\eta}_\ell$ are from a compact metric space. Thus there exists a convergent subsequence, converging to a $\tilde{\eta}$. Using equation (184), definitions of $R_{Q,i}$ and $E_{Q,i}$ given in Definition 11 we can conclude that $\tilde{\eta}$ satisfies

$$
E_{Q,i} \leq (1 - \sum_{j=1}^{\ell} \eta_j) D + \sum_{j=i+1}^{\ell} \eta_j J\left(\frac{R_{Q,j}}{\eta_j}\right) \quad \forall i \in \{1, 2, \ldots, \ell\} \tag{185a}
$$

$$
R_{Q,i} \leq C\eta_i \quad \forall i \in \{1, 2, \ldots, \ell\} \tag{185b}
$$

$$
\eta_i \geq 0 \quad \forall i \in \{1, 2, \ldots, \ell\} \tag{185c}
$$

$$
\sum_{j=1}^{\ell} \eta_j \leq 1. \tag{185d}
$$

According to Definition 11 describing the bit-wise UEP problem a rates-exponents vector $(R, \bar{E})$ is achievable only if there exists a reliable code sequence $Q$ such that $(R_Q, E_Q) = (R, \bar{E})$. Consequently the existence of a time sharing vector satisfying (183) is also a necessary condition for the achievability of a rates-exponents vector $(R, \bar{E})$.

Thus we can conclude that a rates-exponents vector $(R, \bar{E})$ is achievable if and only if there exists a $\tilde{\eta}$ satisfying (183).

In order to prove the convexity of region of achievable rates-exponents vectors, let $(\bar{R}_a, \bar{E}_a)$ and $(\bar{R}_b, \bar{E}_b)$ be two achievable rates-exponents vectors. Then there exist triples $(\bar{R}_a, \bar{E}_a, \bar{\eta}_a)$ and $(\bar{R}_b, \bar{E}_b, \bar{\eta}_b)$ satisfying (183).

For any $\alpha \in [0, 1]$ let $\bar{R}_\alpha$, $\bar{E}_\alpha$ and $\bar{\eta}_\alpha$ be

$$
\bar{R}_\alpha = \alpha \bar{R}_a + (1 - \alpha) \bar{R}_b
$$

$$
\bar{E}_\alpha = \alpha \bar{E}_a + (1 - \alpha) \bar{E}_b
$$

$$
\bar{\eta}_\alpha = \alpha \bar{\eta}_a + (1 - \alpha) \bar{\eta}_b.
$$

As $J(\cdot)$ is concave and the triples $(\bar{R}_\alpha, \bar{E}_\alpha, \bar{\eta}_\alpha)$ and $(\bar{R}_b, \bar{E}_b, \bar{\eta}_b)$ satisfy the constraints given in (183), the triple $(\bar{R}_\alpha, \bar{E}_\alpha, \bar{\eta}_\alpha)$ also satisfies the constraints given in (183). Consequently the rates-exponents vector $(\bar{R}_\alpha, \bar{E}_\alpha)$ is achievable and the region of achievable rates-exponents vectors is convex.

**REFERENCES**

[1] P. Berlin, B. Nakiboğlu, B. Rimoldi, and E. Telatar. A simple converse of Burnashev’s reliability function. *Information Theory, IEEE Transactions on*, 55(7):3074–3080, July 2009.

[2] S. Borade, B. Nakiboğlu, and L. Zheng. Unequal error protection: An information-theoretic perspective. *Information Theory, IEEE Transactions on*, 55(12):5511–5539, Dec. 2009.

[3] M. V. Burnashev. Data transmission over a discrete channel with feedback, random transmission time. *Problemy Perdachi Informatsii*, 12(4):10–30, 1976.

[4] F. Chung and L. Lu. Concentration inequalities and martingale inequalities: A survey. *Internet Mathematics*, 3(1):79–127, 2006.

[5] I. Csiszár. Joint source-channel error exponent. *Problems of Control and Information Theory*, Vol. 9, Iss.5:315–328, 1980.

[6] S.K. Gorantla, B. Nakiboğlu, T.P. Coleman, and L. Zheng. Bit-wise unequal error protection for variable length blockcodes with feedback. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*: 241–245, June 2010.

[7] B. D. Kudryashov. On message transmission over a discrete channel with noiseless feedback. *Problemy Perdachi Informatsii*, 15(1):3–13, 1979.

[8] B. Nakiboğlu and L. Zheng. Errors-and-erasures decoding for block codes with feedback. *Information Theory, IEEE Transactions on*, 58(1):24–49, Jan. 2012.

[9] B. Nazer, Y. Shkel, and S. C. Draper. The AWGN red alert problem. [arXiv:1102.4411[cs.IT]] DOI:10.1109/TIT.2012.2235120.

[10] A. N. Shiriaev. *Probability*. Springer-Verlag Inc., New York, NY, USA, 1996.

[11] D. Wang, V. Chandar, S. Y. Chung, and G.W. Wornell. On reliability functions for single-message unequal error protection. In *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*: 2934–2938, July 2012.

[12] H. Yamamoto and K. Itah. Asymptotic performance of a modified Schalkwijk-Barron scheme for channels with noiseless feedback. *Information Theory, IEEE Transactions on*, 25(6):729 – 733, Nov 1979.

---

36 Let the metric be $\|\tilde{\eta} - \tilde{\nu}\| = \max_j |\eta_j - \nu_j|$. 