Variable Block Length Coding for Channels with Feedback and Cost Constraints

by

Barış Nakiboğlu

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2005

Author . . . .
Department of Electrical Engineering and Computer Science
Aug 17, 2005

Certified by . . . .
Robert R. Gallager
Professor (Emeritus)
Thesis Supervisor

Certified by . . . .
Moe Z. Win
Associate Professor
Thesis Supervisor

Accepted by . . . .
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Variable Block Length Coding for Channels with Feedback and Cost Constraints

by

Barış Nakiboğlu

Submitted to the Department of Electrical Engineering and Computer Science
on Aug 17, 2005, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

Variable-decoding-time/generalized block-coding schemes are investigated for discrete memoryless channels (DMC) with perfect feedback (error free, delay free, infinite capacity) under cost constraints. For a given number of messages and average error probability, upper and lower bounds are found for expected decoding time. These coincide with each other up to a proportionality constant which approaches one in a certain asymtotic sense. A resulting reliability function is found for variable decoding time DMC's with perfect feedback under a cost constraint. The results in this work generalize Burnashev's results,[2] to the cost constrained case.

Thesis Supervisor: Robert R. Gallager
Title: Professor (Emeritus)

Thesis Supervisor: Moe Z. Win
Title: Associate Professor

Acknowledgments

I am grateful to my adviser Robert G. Gallager for his questions and comments which made me understand and handle the problem at a deeper level, and for his innumerable suggestions and corrections on my writing, which helped me to be more clear and understandable in my writing. I would give thanks to my adviser Moe Z. Win for his support. I also would like to thank Constantine Caramanis, Anant Sahai, Alp Şimşek, Abdol-Reza Mansouri, Weepeng Tay, Shashibhushan Borade, Chung Chan, Emmanuel Auguste Abbe, and Murtaza Zafer for their comments and help about various parts of the thesis.

"Once upon a time they used to represent victory as winged. But her feet were heavy and blistered, covered with blood and dust."

The Fall of Paris

Ilya Ehrenburg

Contents

List of Figures

Chapter 1

Introduction

The effect of feedback in communication is a problem that has been studied from the early days of information theory. We will start with an overview of the results from these studies in the introduction. Since these results depend highly on the model used and the constraints imposed on the model, we will state each result with an explicit statement of the model and constraint. We will use the shorthand 'DMC' for finite input alphabet, finite output alphabet channel with fixed transition probabilities. 'AWGNC' will stand for additive white Gaussian noise channel; definitions of these channels are standard. The feedback is assumed to be perfect: infinite capacity, error free, delay free. It is evident in many cases that the assumption of infinite capacity and instantaneous feedback are not necessary for the corresponding results.

All of the results considered here are 'generalized block coding' results, i.e., the channel is used to send 'information' about only one message at a time, and the time intervals used for sending successive messages are disjoint. Disjoint time interval allocation for successive messages would immediately imply conventional block coding when there is no feedback. In other words, the transmitter would transmit a fixed length codeword for each message. In a block-coding scheme with feedback, the transmitter is allowed to have codewords of fixed length whose elements might depend on previous channel outputs as well as the message. In the case of 'generalized block coding' the disjointness of the time interval allocated to each message is preserved, but the duration of this interval is not necessarily constant. In other words the receiver decides when to make a decision about the transmitted message.

When we look at the maximum achievable rate or minimum expected time to send a 'large amount of information' with diminishing error probability, feedback does not yield an improvement. For block coding with feedback, Shannon [15], showed that channel capacity does not increase with feedback in DMC. Although it is not stated specifically in [15], one can generalize this result to the 'generalized block-coding' case, using the weak law of large numbers.

However, the story is very different for zero-error capacity. Shannon showed in [15] that for a set of channels the zero-error capacity can increase with feedback even if we are using block codes. Also it is shown in [15] that if a DMC does not have a zero transition probability then, even with feedback, its zero-error capacity should be zero if we are restricted to use block codes. We will extend this result to generalized block codes. Furthermore we will show that if any zero transition probabilities exist then zero-error capacity is equal to the channel capacity for generalized block codes.[1]

Another widely accepted quality criterion for block codes is the error exponent. The error exponent is the rate of decrease of the logarithm of error probability with increasing block length. Dobrushin [5], showed that for symmetric channels[2] the sphere packing exponent is still a valid upper bound for the error exponent for block codes with feedback. It has been long conjectured but never proved that this is true for non-symmetric channels also. The best known upper bound for block codes with feedback is in [7], by Haroutunian, which coincides with the sphere packing bound for symmetric channels. However there does not exist an achievability proof for this exponent, except in the symmetric case for rates above the *critical rate*. A similar result for AWGNC is given by Pinsker [10]. He showed that the sphere packing exponent is still an upper bound on error exponent even with feedback. In addition to a constant decoding time assumption, Pinsker also used a constant power assumption, i.e., for each message and channel realization, the total amount of energy spent is at most the average power constraint times the block length.

[1]This result is due to Burnashev, [2], we will just extend this to the cost constraint case.

[2]Channels with a transition probability matrix whose columns are permutations of each other, and whose rows are also permutation of each other.

A first relaxation would be having a block code with a constraint on the expected energy. Schalkwijk and Kailath [14], [12], considered the case where the power constraint is in the form of expected power.[3] They showed that the error probability can be made to decay as a two-fold exponential. Indeed Kramer [9], proved that error probability can be made to decay n-fold exponentially. In fact no lower bound to error probability is known if there is no peak power limit or total energy constraint together with the average power constraint. Under various conditions one can prove various performance results, but without a lower bound on error probability we have no clue about the relative performance of these compared to what is ultimately achievable.

'Generalized block-coding schemes' (i.e. schemes with variable decoding time), allow a corresponding relaxation for DMC. In contrast to the case of channel capacity, where Shannon's result in [15] can be extended to variable decoding schemes, the error exponent of variable decoding time systems with feedback can not be extended from the ones corresponding to fixed decoding time systems with feedback. Indeed they are strictly better in almost all non-trivial cases. Although the error exponent for block coding schemes are not known completely, the error exponent for generalized block codes are known. Burnashev calculates the reliability function for generalized block-coding schemes for all values of rate in [2]. He assumed that the feedback has infinite available capacity, but it is evident that noiseless feedback of $\ln|\mathcal{Y}|$ nats per channel use is enough.[4] Indeed as shown by Sahai and Şimşek in [11] feedback rate equal to the capacity of the forward channel is enough.

The main contribution of this work is finding the expression for the reliability function of generalized block-coding schemes on DMC with cost constraints. The flow of the argument will be as follows. The next chapter is devoted to the depiction of the primary model of interest which was initially described by Burnashev in [2]. In the chapters 3, 4 and 5 we will derive the known results about generalized block codes. In chapter 6 we will review cost constraint capacity and detection exponent and describe the cost constraint for variable decoding time systems. In chapter 7

[3]The expected value of the total energy that will be needed to send a message, over possible noise realizations and over possible messages, divided by block length.

[4]$|\mathcal{Y}|$ is the size of the channel output set.

we will prove lemmas that establish bounds on the change of entropy together with costs. Chapter 8 and 9 contain derivations of lower and upper bounds to the minimum expected decoding time under a cost constraint, in terms of size of message set and probability of error. Using these two bounds we will find the reliability function for generalized block-coding schemes under a cost constraint.[5] Finally in chapter 10 we discuss DMC which have one or more zero transition probabilities and extend the result of Burnashev in [2] to the cost constraint case i.e., prove that the zero error capacity under cost constraints is equal to the cost constraint capacity for generalized block codes.

[5]For DMC which does not have any zero transition probability.

Chapter 2

Model And Notation

A communication problem can be posed by defining the probabilistic nature of the forward and feedback channels and the constraints that the receiver and the transmitter are subject to. Coding and decoding schemes are ways of using the channel under those constraints to reach specific reliability measures. We will start by describing the channel models and feedback to be considered here, and then continue with the depiction of possible coding and decoding algorithms, finally we will explain how we will define equivalent macroscopic performance measures and constraints in variable decoding time systems. We will be considering discrete time systems only.

2.1 Forward Channel

The forward channel is described by the stochastic input/output relation that the transmitter and the receiver are operating under. We will assume that the channel is stationary and memoryless, which means that this relation is independent of time, and of previous uses of the channel. We will denote the channel input and output at time n, by X_n and Y_n.

Our forwards channel will be a finite input finite output discrete memoryless channel, i.e., X_n will take values from an input alphabet $\mathcal{X} = \{1, \ldots, K\}$ and Y_n will take values from an output alphabet $\mathcal{Y} = \{1, \ldots, L\}$. The channel is described by the conditional probabilities of the output letters given the input letter; these are called transition probabilities.

$$P_{i,j} = \mathbf{P}\left[Y_n = j | X_n = i\right] \qquad \forall i = 1 \dots K \ \forall j = 1 \dots L \ \forall n \tag{2.1}$$

It is assumed that there are no rows that are replicas of others and no columns that are all zero. Also till the chapter 10 we will assume that all of the transition probabilities are non zero.

We will also discuss cases where an additive cost constraint exists on the codewords. We will denote the cost associated with the i^{th} element of $\mathcal{X}$ as ρ_i for each i. We will denote the average cost constraint by $\mathcal{P}$.

2.2 Feedback Channel

We will denote the input and output of the feedback channel at time n by Z_n and $Z_n^{'}$, where Z_n is a random variable generated at the receiver, and $Z_n^{'}$ is a random variable observed at the transmitter. We will only deal with the case of error free feedback.

$$\mathbf{P}\left[Z_n' = \phi | Z_n = \phi\right] = 1 \qquad \forall n, \phi$$

Thus we will use Z_n to describe both the feedback symbol send and the feedback symbol received.

We will assume that the feedback channel is perfect, namely instantaneous and infinite[1] in capacity in addition to being error-free.

2.3 Coding Algorithm

All of the schemes that will be considered here are generalized block-coding schemes. This means that the transmitter is given one of M equiprobable[2] messages, and until

[1]We are allowed to send an $\mathfrak{n}$-tuple of letters from a size $\mathfrak{r}$ alphabet, where $\mathfrak{r}$ can be as large as required.

[2]This assumption is by no means vital or necessary. It will be evident how to generalize to the case where messages are not equiprobable when we finish the proof.

transmission of that specific message is done, it is not given a new message; when it is given a new message, it can not send any further information about the previous one.[3] Let θ be the message that is to be sent; it takes values from the set $\mathcal{M} = \{1, \ldots, M\}$.

A coding scheme is an assignment of messages to the input letters, at each time depending on feedback. In other words a coding scheme is a sequence of M functions. The k^{th} function, $1 \leq k \leq M$, is given as[4]

$$X_n(k) = \mathfrak{C}_n(k, Z^{n-1}) \qquad \forall Z^{n-1} \tag{2.2}$$

where $Z^{n-1} = \{Z_1, Z_2, \ldots, Z_{n-1}\}$.

Then the transmitted code sequence, given the message, will be

$$X_n = \mathfrak{C}_n(\theta, Z^{n-1}) \qquad \forall Z^{n-1}, \forall \theta \tag{2.3}$$

The knowledge of the receiver at time n is the σ-field generated by all the random variables observed at the receiver, $\mathcal{F}_n$.

$$\mathcal{F}_n = \text{Minimum } \sigma\text{-algebra generated the random variable } Y^n,\ Z^n \text{ and } \Gamma^n$$

where Γ^n is the vector of random variables that are generated at the receiver but not send back to the transmitter via Z^n. When we have perfect feedback we can include all of these random variables in Z^n. Thus

$$\mathcal{F}_n = \text{Minimum } \sigma\text{-algebra generated the random variable } Z^n$$

The sequence of $\mathcal{F}_n$'s forms a filtration, $\mathcal{F}$, in the sense that they are nested, $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_3 \subset \cdots$

[3]The only work that considers non-block algorithms with feedback is by Horstein, [8]. That case seems to be harder to analyze systematically than generalized block coding.

[4]The dependences on Z^{n-1} should be replaced by a dependence in Z'^{n-1} in the case where feedback is noisy. Also there might be a dependence on some other variables that are generated at the transmitter, like the assignments of the other messages till time n etc., if the feedback capacity is not infinite.

In order to find a bound on best possible performance, it is necessary to find the bound on a comprehensive set which includes randomized coding schemes also. In that respect it is worth mentioning that we are not restricted to deterministic coding schemes as a result of the above description. We will assume perfect feedback in all of the proofs that find a performance bound. Thus any random assignment done using the feedback at the transmitter at time n can also be done at the receiver[5] at time $n-1$ and sent to the transmitter through the feedback channel via[6] Z_{n-1}. Since disregarding that knowledge at the receiver can only degrade the performance, our performance bound will be valid for randomized algorithms also.

At each time n the receiver observes an element of the corresponding σ-field, which will be denoted by $\mathfrak{f}_n$. Note that even the overall σ-field generated by all $\mathcal{F}_n$'s, $\mathcal{F}_\infty = \cup_{n=0}^{\infty}\mathcal{F}_n$, is not equal to the σ-field that governs the overall probabilistic structure of the system, since it does not include the random selection of message, θ. However if we consider the σ-field $\mathcal{G}_n$ generated by $\mathcal{F}_n$ and θ, it will summarize everything[7] up to and including time n.

2.4 Decoding Criteria

A decoding criterion is a decision rule about continuing or stopping the process depending on the observations up to that time. In other words it should be a Markov stopping time with respect to the filtration $\mathcal{F}$. At each time instance n, the random variable corresponding to the decision ζ_n takes one of $M+1$ possible values. The first M of them will correspond to elements of $\mathcal{M}$, and will stop the transmission process. The last one will correspond to continuing the transmission. It is evident that the only form of ζ^n's possible are, $\{M+1, M+1, \ldots, M+1\}$ or

[5]It might be necessary to make the corresponding random assignment M times for all M possible messages.

[6]Since we are talking about discrete time systems we need to be specific about the causality relation of X_n, Y_n and Z_n, which will be assumed to be $X_n \to Y_n \to Z_n$, as expected.

[7]*In the most general case where feedback is not necessarily error free or infinite, $\mathcal{F}_n$ and θ will not be sufficient to describe the over all probabilistic structure. In that case $\mathcal{G}_n$ will be the σ-field generated by all the random variables that can be observed at either at receiver or transmitter*

$\{M+1, M+1, \ldots, M+1, k, k, \ldots, k\}$. The decoding time τ is then given as

$$\tau = \min\{k | \zeta_k \neq (M+1)\} \tag{2.4}$$

One can partition the receiver observation into $M+1$ disjoint sets, according to the first element in ζ^n which is not equal to $M+1$. The first M of them,$\chi_1, \chi_2, \ldots, \chi_M$ will correspond to decoding the corresponding message. The last one, χ_{M+1}, will correspond to the event that decoding never occurs. The probability of error, P_e is then given by

$$P_e = \frac{1}{M} \sum_{i=1}^{M} \mathbf{P}\left[e | \theta = i\right] \tag{2.5}$$

where $\mathbf{P}\left[e | \theta_i\right] = 1 - \mathbf{P}\left[\chi_i | \theta = i\right]$ for $i = 1 \ldots M$.

The expected transmission time[8] is given by

$$\bar{\tau} = \mathbf{E}\left[\tau | \mathcal{F}_0\right] = \frac{1}{M} \sum_{i=1}^{M} \mathbf{E}\left[\tau | \theta = i, \mathcal{F}_0\right] \tag{2.6}$$

Note that at time zero each message will have an a posteriori probability of $1/M$. After the observation of the first channel output, this a posteriori probability will change. This change for each message will be a function of the channel output under the specific coding scheme. One can continue to calculate these a posteriori probabilities as more and more observations are made. Consequently the a posteriori probability of a message is a function of $\mathfrak{f}_n$ and thus, as a random variable, it is a measurable function on $\mathcal{F}_n$. Thus the corresponding entropy of this a posteriori distribution is also a random variable measurable in $\mathcal{F}_n$.

$$H_n = H(p(\mathfrak{f}_n)) = -\sum_{i=1}^{M} p_i(\mathfrak{f}_n) \ln p_i(\mathfrak{f}_n) \tag{2.7}$$

where $p(\mathfrak{f}_n) = (p_1(\mathfrak{f}_n), p_2(\mathfrak{f}_n), \ldots, p_M(\mathfrak{f}_n))$ is the a posteriori distribution of the messages for the given $\mathfrak{f}_n$.

[8]The expectation is over possible channel realizations and over possible messages.

Indeed the value of this random variable is nothing but

$$H_n = H(\theta \mid \mathcal{F}_n = \mathfrak{f}_n) \tag{2.8}$$

2.5 Performance Measure and Constraints

Note that because of the variable nature of block lengths, it is not possible to define a fixed operating rate. Instead one needs to make a new definition of Rate and Error Exponent which is consistent with existing ones for fixed-length codes. The definitions we use are

$$R = \frac{\ln M}{\mathbf{E}[\tau]} \tag{2.9}$$

$$E(R) = \lim_{P_e \to 0} \frac{-\ln P_e}{\mathbf{E}[\tau]} \tag{2.10}$$

These definitions are not only consistent with the definitions for fixed-length codes, but they are also the average quantities to which the system converges after many successive uses.

We will give some definitions which will be used in the proofs. $D(p \mid\mid q)$ will denote the Kullback-Leibler divergence of two probability distributions.

$$D(p \mid\mid q) = \sum_i p_i \ln \frac{p_i}{q_i} \tag{2.11}$$

We will denote the indicator function for the event q by $\mathbb{I}_{\{q\}}$.

We will use ϕ for probability mass functions on the input letter set, and ψ for the probability mass functions on the output letters.

Chapter 3

Basic Lemmas

Lemma 1 (Generalized Fano Inequality(for variable decoding time))**.** *For any coding algorithm and decoding rule such that* $\mathbf{P}[\tau < \infty] = 1$*,*

$$\mathbf{E}[H_\tau] \leq \mathfrak{h}(P_e) + P_e \ln(M-1) \tag{3.1}$$

where $\mathfrak{h}(x) = -x\ln(x) - (1-x)\ln(1-x)$

Note that indeed this lemma has nothing to do with the model of channel. Thus it can be used in the AWGNC case also.

Proof:

Since τ is a stopping time, the event $\tau = n$ is measurable in $\mathcal{F}_n$. H_n is also measurable in $\mathcal{F}_n$, so that $(H_n, \mathcal{F}_n)$ is a stochastic sequence.

$$H_\tau = \sum_{n=0}^{\infty} H_n \mathbb{I}_{\{\tau=n\}}$$

Thus $\mathbf{E}[H_\tau]$ can be written as a limit.

$$\mathbf{E}[H_\tau] = \lim_{N\to\infty} \sum_{n=0}^{N} \mathbf{E}[H_n | \tau = n]\, \mathbf{P}[\tau = n] \tag{3.2}$$

Since H_n is a bounded random variable and $\mathbf{P}[\tau < \infty] = 1$ this limit is well defined.

Note that at each element, $\mathfrak{f}_n$, of the σ-field $\mathcal{F}_n$, there exists a probability mass function associated with the message set. One can use the conventional Fano inequality to upper bound H_n for a given element $\mathfrak{f}_n$ of $\mathcal{F}_n$.

$$H_n \leq \mathfrak{h}(P_e(\mathfrak{f}_n)) + P_e(\mathfrak{f}_n)\ln(M-1) \tag{3.3}$$

where $P_e(\mathfrak{f}_n)$ is the probability of error of a detector for a source with probability mass function $\bar{p}(n)$ on the message set.
Since τ is a stopping time, the observation up to time n will determine whether $\tau = n$ or not. Then we can define a set A_n that corresponds to the elements of $\mathcal{F}_n$ at which the decoding time will be n, $A_n = \{\mathfrak{f}_n \in \mathcal{F}_n | \tau = n\}$. As a result

$$\mathbf{E}\left[H_n \middle| \tau = n\right] = \sum_{\mathfrak{f}_n \in A_n} H_n \mathbf{P}\left[\mathfrak{f}_n \middle| \tau = n\right] \qquad \mathbf{E}\left[H_\tau\right] = \sum_n \mathbf{E}\left[H_n \middle| \tau = n\right]\mathbf{P}\left[\tau = n\right] \tag{3.4}$$

$$\mathbf{E}\left[P_e[n] \middle| \tau = n\right] = \sum_{\mathfrak{f}_n \in A_n} P_e(\mathfrak{f}_n)\mathbf{P}\left[\mathfrak{f}_n \middle| \tau = n\right] \qquad P_e = \sum_n \mathbf{E}\left[P_e[n] \middle| \tau = n\right]\mathbf{P}\left[\tau = n\right] \tag{3.5}$$

Note that $P_e = \mathbf{E}\left[P_e(\mathfrak{f}_n)\right]$. Using equations (3.2), (3.3), (3.4), (3.5), together with the concavity of the entropy of a binary random variable we get equation (3.1).

QED

As a result of Lemma 1, we know that the expected value of entropy at the decoding time is upper bounded in terms of the expected error probability. When proving non-existence (converse) results, the Fano inequality is generally used as a lower bound on probability of error in terms of the conditional entropy. Our approach will be a little bit different; we will use average error probability to find an upper bound on the expected value of entropy over decoding instances. Then we will find lower bounds on the expected time to reach those expected values.

It is important to remember that conditional expectations are indeed functions in terms of the conditioned random variables. In other words $\mathbf{E}\left[X \middle| Y\right] < A$ means that for every value y of the random variable Y, $f(y) = \mathbf{E}\left[X \middle| y\right] < A$. Equivalently $f(Y) = \mathbf{E}\left[X \middle| Y\right] < A$. The following lemmas about the change of entropy can best be understood with this interpretation.

Lemma 2. $\forall n \geq 0$, *we have the inequality,*[1]

$$\mathbf{E}\left[H_n - H_{n+1} \middle| \mathcal{F}_n\right] \leq \mathbf{C}$$

where $\mathbf{C}$ *is the channel capacity, given by*

$$\mathbf{C} = \max_{\phi} \sum_{i=1,j=1}^{K,L} \phi_i P_{i,j} \ln \frac{P_{i,j}}{\sum_{r=1}^{K} \phi_r P_{rj}}$$

The expected value inherently includes an averaging over the possible messages along with the received symbol Y_{n+1}. Being more explicit, as a result of Bayes theorem, we can say that the probability of $\mathfrak{f}_n$ given the message $\theta = i$ i.e. $\mathbf{P}\left[\mathfrak{f}_n \middle| \theta = i\right]$, is given by $\mathbf{P}\left[\mathfrak{f}_n\right] p_i(\mathfrak{f}_n) M$, where $\mathbf{P}\left[\mathfrak{f}_n\right]$ is the probability of being at $\mathfrak{f}_n$ and $p_i(\mathfrak{f}_n)$, is the a posteriori probability of the i^{th} message given that the realization of $\mathcal{F}_n$ is $\mathfrak{f}_n$.

The expected decrease we are bounding here is indeed averaged over different possible messages using $p_i(\mathfrak{f}_n)$. In other words, at a specific a posteriori probability distribution on messages, one can propose a coding method that will decrease the entropy, on average[2] for a specific source message much more than $\mathbf{C}$. However this method will have a poorer performance in the case when one of the other messages is sent. If one weights these cases with the corresponding probabilities of the messages then the weighted sum is less then $\mathbf{C}$.

Proof:

What this lemma says is that the expected entropy difference above for a given $\mathfrak{f}_n \in \mathcal{F}_n$, is the conditional mutual information between the messages and the channel output at time $n+1$. θ has the conditional probability distribution $p(\mathfrak{f}_n)$ on the possible message set. The feedback Z_n, together with the coding for time $n+1$, can be considered as a method to assign elements of the message set to the input

[1] Interpreting conditioned expectation as function of the conditioned quantity, it is evident that this relation is valid for any realization, $\mathfrak{f}_n$ of the σ-field $\mathcal{F}_n$.

[2] Averaged over different possible channel outputs.

alphabet. The conditional probability mass function $p(\mathfrak{f}_{n+1})$ is the a posteriori probability distribution of θ at time $n+1$ given the channel output and the coding method. So the expected value of $p(\mathfrak{f}_{n+1})$ for some specific value of Z_n is the entropy of θ given Y_{n+1} and $\mathfrak{f}_n$. Thus

$$\begin{aligned}\mathbf{E}\left[H_n - H_{n+1} \middle| \mathcal{F}_n = \mathfrak{f}_n\right] &= H(\theta|\mathfrak{f}_n) - H(\theta|\mathfrak{f}_n, Y_{n+1}) \\ &= I(\theta; Y_{n+1}|\mathfrak{f}_n)\end{aligned}$$

As a result of Markov relation between implied by our assignment $\theta \leftrightarrow X_{n+1} \leftrightarrow Y_{n+1}$, and data processing inequality;

$$\mathbf{E}\left[H_n - H_{n+1} \middle| \mathcal{F}_n = \mathfrak{f}_n\right] \leq I(X_{n+1}; Y_{n+1}) \leq \mathbf{C}$$

A more algebraic proof is given in appendix B.

QED

Note that Lemma 2 is rather strong. It states that for all possible realizations of the observations up to and including time n, the expected decrease at time $n+1$ is less then $\mathbf{C}$, i.e., as a random variable measurable in $\mathcal{F}_n$, the expected decrease in one time unit is bounded. It is not a result in terms of an expectation over the realizations of the σ-field $\mathcal{F}_n$. It is also important to note that it is a result in terms of an average over messages, with the corresponding a posteriori probabilities.

Lemma 3. $\forall n \geq 0$ *we have the inequality,*

$$\mathbf{E}\left[\ln H_n - \ln H_{n+1} \middle| \mathcal{F}_n\right] \leq \mathbf{D} \tag{3.6}$$

where

$$\mathbf{D} = \max_{i,k} \sum_{l=1}^{L} P_{i,l} \ln \frac{P_{i,l}}{P_{k,l}} \tag{3.7}$$

Proof:

For arbitrary non-negative a_i,b_{il},c_l, the log sum inequality states that

$$\ln \frac{\sum_i a_i}{\sum_i b_{il}} \leq \sum_i \frac{a_i}{\sum_j a_j} \ln \frac{a_i}{b_{il}}$$

Multiplying both sides by c_l and summing over l,

$$\sum_l c_l \ln \frac{\sum_i a_i}{\sum_i b_{il}} \leq \sum_l c_l \sum_i \frac{a_i}{\sum_j a_j} \ln \frac{a_i}{b_{il}}$$

Then evidently, for arbitrary non-negative a_i,b_{il},c_l,

$$\sum_l c_l \ln \frac{\sum_i a_i}{\sum_i b_{il}} \leq \max_i \sum_l c_l \ln \frac{a_i}{b_{il}} \tag{3.8}$$

Using the short hand

$$f_i = p_i(\mathfrak{f}_n) \qquad f_i(l) = \mathbf{P}\left[\theta = i \middle| Y_{n+1} = l, \mathcal{F}_n = \mathfrak{f}_n\right]$$

$$w(k|i) = \mathbf{P}\left[X_{n+1} = k \middle| \mathcal{F}_n = \mathfrak{f}_n, \theta = i\right] \qquad p(l|i) = \mathbf{P}\left[Y_{n+1} = l \middle| \theta = i, \mathcal{F}_n = \mathfrak{f}_n\right]$$

making the substitution $c_l = p(l)$, $a_i = -f_i \ln f_i$, $b_{il} = -f_i(l) \ln f_i(l)$ in equation (3.8), we get

$$\begin{aligned}
\mathbf{E}\left[\ln(H_n) - \ln(H_{n+1}) \middle| \mathcal{F}_n = \mathfrak{f}_n\right] &= \sum_{l=1}^{L} p(l) \ln \frac{-\sum_{i=1}^{M} f_i \ln f_i}{-\sum_{i=1}^{M} f_i(l) \ln f_i(l)} \\
&\leq \max_i (\sum_l p(l) \ln \frac{-f_i \ln f_i}{-f_i(l) \ln f_i(l)}) \\
&= \max_i \sum_l p(l) \ln \left(\frac{f_i}{f_i(l)} \frac{\ln 1/f_i}{\ln 1/f_i + \ln \frac{f_i}{f_i(l)}} \right)
\end{aligned}$$

Using the relation $f_i(l) = \frac{f_i p(l|\theta=i)}{p(l)}$, in the above expression we get

$$\mathbf{E}\left[\ln(H_n) - \ln(H_{n+1}) \middle| \mathcal{F}_n = \mathfrak{f}_n\right] \leq \max_i \sum_l p(l) \left(\ln \frac{p(l)}{p(l|\theta = i)} + \ln \frac{1}{1 + \frac{\ln \frac{p(l)}{p(l|\theta=i)}}{\ln 1/f_i}} \right)$$

Using the identity $\ln x \le x - 1$

$$\begin{aligned}
\mathbf{E}\left[\ln(H_n) - \ln(H_{n+1}) \middle| \mathcal{F}_n = \mathfrak{f}_n\right] &\le \max_i \sum_l p(l) \left(\ln \frac{p(l)}{p(l|\theta = i)} - \frac{\frac{\ln \frac{p(l)}{p(l|\theta=i)}}{\ln 1/f_i}}{1 + \frac{\ln \frac{p(l)}{p(l|\theta=i)}}{\ln 1/f_i}} \right) \\
&= \max_i \sum_l p(l) \ln \frac{p(l)}{p(l|\theta = i)} \left(1 - \frac{1}{\ln 1/f_i + \ln \frac{p(l)}{p(l|\theta=i)}} \right) \\
&= \max_i \sum_l p(l) \ln \frac{p(l)}{p(l|\theta = i)} \left(1 - \frac{1}{\ln \frac{p(l)}{f_i p(l|\theta=i)}} \right)
\end{aligned}$$

$$\le \max_i \sum_l p(l) \ln \frac{p(l)}{p(l|\theta = i)} \tag{3.9}$$

Using the convexity of the Kullback-Leibler divergence (2.11), it is evident from the above inequality that $\mathbf{E}\left[\ln(H_n) - \ln(H_{n+1}) \middle| \mathcal{F}_n = \mathfrak{f}_n\right] \le \mathbf{D}$

QED

Lemma 4. *For any $n \ge 0$, $Y_{n+1} = l$*

$$\ln H_n - \ln H_{n+1} \le \max_{i,k} \ln \frac{p_{kl}}{p_{il}} \le \max_{i,k,l} \ln \frac{p_{kl}}{p_{il}} = \mathbf{F} \tag{3.10}$$

Proof:

Note that

$$\begin{aligned}
\ln(H_n) - \ln(H_{n+1}) &= \ln \frac{-\sum_{i=1}^M f_i \ln f_i}{-\sum_{i=1}^M f_i(l) \ln f_i(l)} \\
&\le \max_i \ln \frac{-f_i \ln f_i}{-f_i(l) \ln f_i(l)}
\end{aligned}$$

Doing almost the same calculation with the previous lemma one can find

$$\begin{aligned}\ln(H_n) - \ln(H_{n+1}) &\le \max_i \ln \frac{p(l)}{p(l|\theta = i)} \\ &\le \max_i \ln \frac{\sum_{j=1}^{M} p(l|\theta = j)}{p(l|\theta = i)} \\ &\le \max_{i,j} \ln \frac{P_{jl}}{P_{il}}\end{aligned}$$

QED

Finally we have the following lemma relating stochastic sequences of certain properties, stopping times and expected values of stooped stochastic sequences.

Lemma 5. *Assume the sequence* $\Gamma_0, \Gamma_1, \Gamma_2 ...$ *of random variables are measurable in the sigma fields* $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 ...$*, and that, for some* K *and* R

$$|\Gamma_n| < Kn \qquad \forall n \tag{3.11}$$

$$\mathbf{E}\left[\Gamma_n - \Gamma_{n+1} | \mathcal{F}_n\right] \le R \qquad \forall n \tag{3.12}$$

Assume τ_i *and* τ_f *are stopping times with respect to the filtration* $\mathcal{F}$*, such that* $\mathbf{E}[\tau_f] < \infty$ *and* $\tau_i(w) \le \tau_f(w) \quad \forall w \in \mathcal{F}$*. Let* $\nu_n = \Gamma_n + Rn$*. Then the following are true*

1. $(\nu_n, \mathcal{F}_n)$ *is a submartingale and* $|\nu_n| < K'n \quad \forall n$
2. $\xi_n = \nu_n - \nu_{n \wedge \tau_i}$ *is a submartingale and* $|\xi_n| < K''n \quad \forall n$
3. $R\mathbf{E}[\tau_i | \mathcal{F}_0] \ge \mathbf{E}[\Gamma_0 - \Gamma_{\tau_i} | \mathcal{F}_0]$
4. $R\mathbf{E}[\tau_f | \mathcal{F}_0] \ge \mathbf{E}\left[\Gamma_0 - \Gamma_{\tau_f} \middle| \mathcal{F}_0\right]$
5. $R\mathbf{E}[\tau_f - \tau_i | \mathcal{F}_0] \ge \mathbf{E}\left[\Gamma_{\tau_i} - \Gamma_{\tau_f} \middle| \mathcal{F}_0\right]$

Proof of the lemma 5 is given in the appendix A.

Chapter 4

Lower Bound For The Expected Time

Generally the error exponent is interpreted as the rate of increase of $-\ln P_e$ at constant communication rate, R, with increasing block length, l. An alternative approach is to view the error exponent as the rate of change of block-length with increasing $-\ln P_e$ at a fixed rate R.

$$E(R) = \lim_{P_e \to 0} \frac{-\ln P_e}{l(P_e, R)}$$

where $l(P_e, R)$ is the minimum block-length needed in order to operate at rate $R = \frac{\ln M}{l(P_e,R)}$ with probability of error P_e. The converse discussed here is the extension of this approach to generalized block schemes where block length is replaced by expected block length.

4.1 Converse

Theorem 1. *For any transmission method over a DMC with feedback,* $\forall P_e > 0$ *and* $\forall M > e^B$*, the expected number of observations* $\mathbf{E}[\tau]$ *satisfies the inequality*

$$\mathbf{E}[\tau] \geq \frac{\ln M}{\mathbf{C}} - \frac{\ln P_e}{\mathbf{D}} - \frac{\ln(\ln M - \ln P_e + 1)}{\mathbf{D}} - \frac{P_e \ln M}{\mathbf{C}} + \mathbf{\Delta} \tag{4.1}$$

where

$$\mathbf{C} = \max_f \sum_{k=1}^{K} \sum_{l=1}^{L} f_k P_{kl} \ln\left(\frac{P_{kl}}{\sum_{j=1}^{K} f_j P_{jl}}\right)$$

$$\mathbf{D} = \max_{i,j} \sum_{l=1}^{L} P_{i,l} \ln \frac{P_{i,l}}{P_{j,l}} \qquad \mathbf{F} = \max_{i,j,l} \ln \frac{P_{i,l}}{P_{j,l}}$$

and $\mathbf{\Delta}$ *and* B *are constants determined by the channel transition probabilities, satisfying* $B < \mathbf{F} + 1$.

Before going into the proof let us elaborate on what this equation tells us. In order to calculate error exponent at a rate R, we need to consider the limit as P_e goes to zero[1]. For $M > 2$ and $P_e < 1/e$ we can write equation (4.1) as

$$\mathbf{E}[\tau] \geq \frac{\ln M}{\mathbf{C}} \left(1 - P_e - \frac{\mathbf{C}\ln(1+\ln M)}{\mathbf{D}\ln M}\right) + \frac{-\ln P_e}{\mathbf{D}} \left(1 + \frac{\ln(1-\ln P_e)}{\ln P_e}\right) + \mathbf{\Delta}$$

If we divide both sides by $\mathbf{E}[\tau]$ and calculate $\liminf_{P_e \to 0}$ on both sides we get

$$\liminf_{P_e \to 0} \frac{-\ln P_e}{\mathbf{D}\mathbf{E}[\tau]} \left(1 + \frac{\ln(1-\ln P_e)}{\ln P_e}\right) \leq 1 - \limsup_{P_e \to 0} \frac{\ln M}{\mathbf{C}\mathbf{E}[\tau]} \left(1 - P_e + \frac{\mathbf{C}\mathbf{D}\mathbf{\Delta} - \mathbf{C}\ln(1+\ln M)}{\mathbf{D}\ln M}\right) \tag{4.2}$$

For any sequence of coding decoding algorithms to have rate R, $\limsup_{P_e \to 0} \frac{\ln M}{\mathbf{E}[\tau]} \geq R$. Using equation (4.2)

$$\liminf_{P_e \to 0} \frac{-\ln P_e}{\mathbf{E}[\tau]} \leq \mathbf{D}\left(1 - \frac{R}{\mathbf{C}}\right)$$

Repeating same calculations for lim sup, considering the condition $\liminf_{P_e \to 0} \frac{\ln M}{\mathbf{E}[\tau]} \geq R$ we get

$$\limsup_{P_e \to 0} \frac{-\ln P_e}{\mathbf{E}[\tau]} \leq \mathbf{D}\left(1 - \frac{R}{\mathbf{C}}\right)$$

[1] For ϵ-capacity, the theorem extends the known subtlety about the feedback case, to generalized block coding. If we fix a constant error probability, $P_e = \epsilon$, the fourth term does not become negligible as we increase M. Thus we can not prove that ϵ-capacity is not improved by feedback using this result. Indeed it is the same subtlety mentioned in [5], referring to [15] about feedback channels. We know that for DMC, the ϵ-capacity, $\mathbf{C}_\epsilon$ is equal to channel capacity for every epsilon less than one and greater than zero. However, with feedback, the value of $\mathbf{C}_\epsilon$ is not known in general. What is known for block-coding schemes is that if the rate is strictly greater then $\mathbf{C}$ the error probability is lower bounded away from zero. Theorem 1 says that this is true for generalized block coding schemes also, but nothing more.

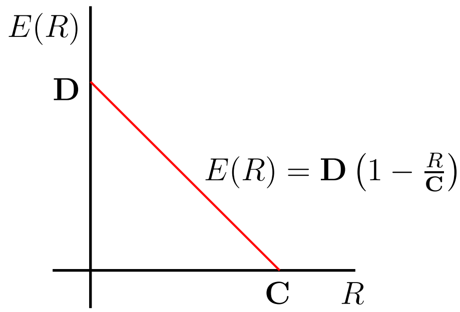

Figure 4-1: Error-Exponent vs Rate for DMC

Thus

$$E(R) \leq \mathbf{D}\left(1 - \frac{R}{\mathbf{C}}\right)$$

For the case of the error exponent at $R = 0$, we need M to go infinity as we decrease P_e, but it should be slower than any exponential function of expected decoding time. Thus the zero-rate exponent will not be higher than $\mathbf{D}$.

Proof:

The generalized Fano inequality implies that the expected value of the entropy[2] at decoding time is upper bounded by a function of the average error probability. Also the expected change in entropy is bounded as a result of lemmas 2, 3 and 4. If we measure the time required for a sufficient decrease of entropy in some way we will be able to bound the minimum expected time in terms of the change in entropy.

Let us consider the stochastic sequence $(\xi_n, \mathcal{F}_n)$ such that

$$\xi_n = \begin{Bmatrix} \mathbf{C}^{-1} H_n + n & \text{if} \quad H_n \geq B \\ \mathbf{D}^{-1} \ln H_n + a + n & \text{if} \quad H_n < B \end{Bmatrix} \tag{4.3}$$

where B and a are constants to be selected later. This can be written as

$$\xi_n = n + \mathbf{C}^{-1} H_n \mathbb{I}_{\{H_n \geq B\}} + (\mathbf{D}^{-1} \ln H_n + a)\mathbb{I}_{\{H_n < B\}} \tag{4.4}$$

where $\mathbb{I}_{\{\cdot\}}$ is the indicator function.

[2]The expectation is over possible messages and possible decoding times.

First we will assume that ξ_n is a submartingale without proof.[3] After finishing the proof of the theorem we will verify that ξ_n is a submartingale.

It is a known result that if $(\xi_n, \mathcal{F}_n)$ is a submartingale, and τ is a stopping time with respect to the filtration $\mathcal{F}$, then $(\xi_{n\wedge\tau}, \mathcal{F}_n)$ also forms a submartingale, this can be found in [6](pp248, theorem 4).

$$
\begin{aligned}
\xi_0 &\leq \mathbf{E}\left[\xi_{n\wedge\tau} \middle| \mathcal{F}_0\right] \leq \lim_{n\to\infty} \mathbf{E}\left[\xi_{n\wedge\tau} \middle| \mathcal{F}_0\right] \\
&= \lim_{n\to\infty} \mathbf{E}\left[(n\wedge\tau) + \mathbf{C}^{-1} H_{n\wedge\tau} \mathbb{I}_{\{H_{n\wedge\tau}\geq B\}} + (\mathbf{D}^{-1}\ln H_{n\wedge\tau} + a)\mathbb{I}_{\{H_{n\wedge\tau}<B\}} \middle| \mathcal{F}_0\right] \\
&= \lim_{n\to\infty} \mathbf{E}\left[(n\wedge\tau) + \mathbf{D}^{-1}\ln H_{n\wedge\tau} + a + (\mathbf{C}^{-1} H_{n\wedge\tau} - \mathbf{D}^{-1}\ln H_{n\wedge\tau} - a)\mathbb{I}_{\{H_{n\wedge\tau}\geq B\}} \middle| \mathcal{F}_0\right]
\end{aligned}
$$

where $n \wedge m$ is the minimum of n and m and we have used equation (4.4).

Using the positivity of entropy,

$$
\xi_0 \leq \lim_{n\to\infty} \mathbf{E}\left[(n\wedge\tau) + \mathbf{C}^{-1} H_{n\wedge\tau} + \mathbf{D}^{-1}\ln H_{n\wedge\tau} \middle| \mathbf{F}_0\right] + |a| + \frac{|\ln B|}{\mathbf{D}}
$$

Using the concavity of $\ln(\cdot)$ together with Jensen's inequality

$$
\xi_0 \leq \lim_{n\to\infty} \mathbf{E}\left[(n\wedge\tau) + \mathbf{C}^{-1} H_{n\wedge\tau} \middle| \mathcal{F}_0\right] + \mathbf{D}^{-1}\ln \mathbf{E}\left[H_{n\wedge\tau} \middle| \mathcal{F}_0\right] + |a| + \frac{|\ln B|}{\mathbf{D}}
$$

Since $\mathbf{P}[\tau < \infty] = 1$, $\lim_{n\to\infty} \mathbf{E}[n\wedge\tau | \mathcal{F}_0] = \mathbf{E}[\tau | \mathcal{F}_0]$. Then we can use the boundedness of H_n to see[4]

$$
\mathbf{E}\left[\tau \middle| \mathcal{F}_0\right] \geq \xi_0 - \mathbf{C}^{-1}\mathbf{E}\left[H_\tau \middle| \mathcal{F}_0\right] - \mathbf{D}^{-1}\ln \mathbf{E}\left[H_\tau \middle| \mathcal{F}_0\right] - |a| - \frac{|\ln B|}{\mathbf{D}}
$$

Since $H_0 = \ln M > B$, we have $\xi_0 = \mathbf{C}^{-1} H_0$ so

$$
\mathbf{E}\left[\tau \middle| \mathcal{F}_0\right] \geq \mathbf{C}^{-1}\ln M - \mathbf{C}^{-1}\mathbf{E}\left[H_\tau \middle| \mathcal{F}_0\right] - \mathbf{D}^{-1}\ln \mathbf{E}\left[H_\tau \middle| \mathcal{F}_0\right] - |a| - \frac{|\ln B|}{\mathbf{D}}
$$

[3]Namely we will assume that $\mathbf{E}[H_n] < \infty$ and that there exists a and B such that $\mathbf{E}[\xi_{n+1} | \mathcal{F}_n] \geq \xi_n$.

[4]If $\mathbf{P}[\tau < \infty] \neq 1$, then $\mathbf{E}[\tau] = \infty$, and the theorem holds.

Inserting the generalized Fano inequality, and bounding the binary entropy by $\mathfrak{h}(P_e) \leq -P_e \ln P_e + P_e$,

$$\mathbf{E}\left[\tau|\,\mathcal{F}_0\right] \geq \mathbf{C}^{-1}\ln M - \mathbf{C}^{-1}(P_e \ln M + \mathfrak{h}(P_e)) - \mathbf{D}^{-1}(\ln P_e + \ln(\ln M - \ln P_e + 1)) - |a| - \frac{|\ln B|}{\mathbf{D}}$$

Bounding the binary entropy, $-\mathfrak{h}(P_e) > -\ln 2$, and defining $\mathbf{\Delta}$ to be $\mathbf{\Delta} = -|a| - \frac{|\ln B|}{\mathbf{D}} - \mathbf{C}^{-1}\ln 2$

$$\mathbf{E}\left[\tau|\,\mathcal{F}_0\right] \geq \mathbf{C}^{-1}\ln M - P_e\mathbf{C}^{-1}\ln M - \mathbf{D}^{-1}\ln P_e - \mathbf{D}^{-1}\ln(\ln M - \ln P_e + 1) + \mathbf{\Delta} \quad (4.5)$$

Now we need to prove that ξ_n is a submartingale. We will start with proving that $\mathbf{E}\left[|\xi_n|\right] < \infty$.

Consider the following two stochastic sequences

$$\xi'_n = \mathbf{C}^{-1}H_n + n \qquad \xi''_n = \mathbf{D}^{-1}\ln H_n + n + a$$

Using the boundedness of entropy, $0 \leq H_n \leq \ln M$, we get $n < \xi'_n \leq \frac{\ln M}{\mathbf{C}} + n$ consequently $|\xi'_n| \leq n + \frac{\ln M}{\mathbf{C}}$. As a result of this boundedness

$$\mathbf{E}\left[|\xi'_n|\right] < \infty \quad (4.6)$$

Using Lemma 4 we can see that, $\ln H_0 - n\mathbf{F} \leq \ln H_n \leq \ln\ln M$ and consequently $|\xi''_n| \leq |\frac{\ln\ln M}{\mathbf{D}} + n| + |\frac{\ln H_0}{\mathbf{D}} - \frac{\mathbf{F}-\mathbf{D}}{\mathbf{D}}n|$. Using this boundedness one can conclude that

$$\mathbf{E}\left[|\xi''_n|\right] < \infty \quad (4.7)$$

Note that an alternative way of writing ξ_n is

$$\xi_n = \xi'_n \mathbb{I}_{\{H_n > B\}} + \xi''_n \mathbb{I}_{\{H_n \leq B\}} \quad (4.8)$$

As a result

$$\mathbf{E}\left[|\xi_n|\right] \leq \mathbf{E}\left[|\xi'_n|\right] + \mathbf{E}\left[|\xi''_n|\right]$$

Using equation 4.6, 4.7, 4.8 we get $\mathbf{E}[|\xi_n|] < \infty$

Now we will look at the change from ξ_n to ξ_{n+1} and try to bound it, using lemma 2,3 and corollary 4 and setting the constants B and a accordingly. First consider the two processes we described before

$$\mathbf{E}\left[\xi'_{n+1} - {\xi'}_n \middle| \mathcal{F}_n\right] = \mathbf{E}\left[1 - \frac{H_{n+1} - H_n}{\mathbf{C}} \middle| \mathcal{F}_n\right] \tag{4.9}$$

Using lemma 2

$$\mathbf{E}\left[{\xi'}_{n+1} - {\xi'}_n \middle| \mathcal{F}_n\right] > 0 \tag{4.10}$$

Together with equation (4.6), this implies that ${\xi'}_n$ is a submartingale.

$$\mathbf{E}\left[{\xi''}_{n+1} - {\xi''}_n \middle| \mathcal{F}_n\right] = \mathbf{E}\left[1 - \frac{\ln H_{n+1} - \ln H_n}{\mathbf{D}} \middle| \mathcal{F}_n\right] \tag{4.11}$$

Using lemma 3

$$\mathbf{E}\left[{\xi'}_{n+1} - {\xi'}_n \middle| \mathcal{F}_n\right] > 0 \tag{4.12}$$

Together with equation (4.7), this implies that ${\xi''}_n$ is a submartingale.

Now we need to find the values for B and a such that, $\mathbf{E}\left[\xi_{n+1} - \xi_n \middle| \mathcal{F}_n\right] > 0$ holds. Consider the functions

$$f_1(H) = \mathbf{C}^{-1}H \qquad f_2(H) = \mathbf{D}^{-1}\ln H + a \qquad f(H) = \begin{Bmatrix} \mathbf{C}^{-1}H & \text{if} \quad H \geq B \\ \mathbf{D}^{-1}\ln H + a & \text{if} \quad H < B \end{Bmatrix} \tag{4.13}$$

It is evident from the graph that

$$f(H) \geq f_1(H) \qquad \forall H \geq A$$
$$f(H) \geq f_2(H) \qquad \forall H$$

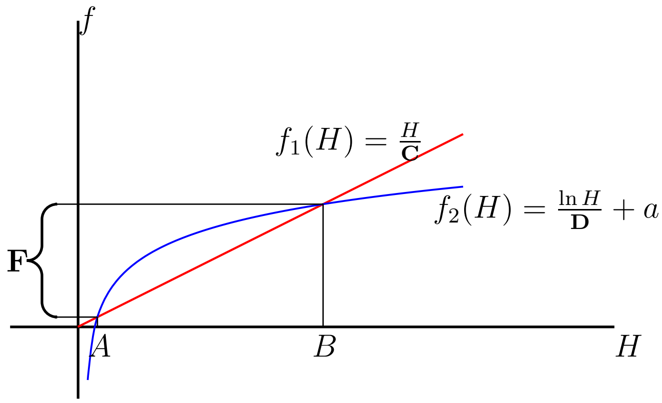

Figure 4-2: f1(H) & f2(H) vs H

Considering the relations $\xi_n = f(H_n) + n$, ${\xi'}_n = f^1(H_n) + n$, and ${\xi''}_n = f(H_n) + n$.

$$\xi_n \geq {\xi'}_n \qquad \text{if} \quad H_n \geq A$$

$$\xi_n \geq {\xi''}_n$$

Let us consider two cases

Case 1: $H_n < B$:

Since $\xi_{n+1} \geq {\xi''}_{n+1}$, $\mathbf{E}\left[\xi_{n+1} \middle| \mathcal{F}_n\right] \geq \mathbf{E}\left[{\xi''}_{n+1} \middle| \mathcal{F}_n\right]$. Since ${\xi''}_n$ is a submartingale, $\mathbf{E}\left[\xi_{n+1} \middle| \mathcal{F}_n\right] \geq {\xi''}_n$. Using the fact that $H_n < B$ we get $\mathbf{E}\left[\xi_{n+1} \middle| \mathcal{F}_n\right] \geq \xi_n$.

Case 2: $H_n > B$:

If we know that $\mathbf{P}\left[H_{n+1} \geq A \middle| \mathcal{F}_n\right] = 1$ than we can argue that $\xi_{n+1} \geq {\xi'}_{n+1}$ and then $\mathbf{E}\left[\xi_{n+1} \middle| \mathcal{F}_n\right] \geq \mathbf{E}\left[{\xi'}_{n+1} \middle| \mathcal{F}_n\right]$. Since ${\xi'}_n$ is a submartingale, $\mathbf{E}\left[\xi_{n+1} \middle| \mathcal{F}_n\right] \geq {\xi'}_n$. Using the condition about the case $H_n > B$ we get $\mathbf{E}\left[\xi_{n+1} \middle| \mathcal{F}_n\right] \geq \xi_n$.

So what we need is to ensure that $\mathbf{P}\left[H_{n+1} \geq A \middle| \mathcal{F}_n\right] = 1$ whenever $H_n > B$. But we know that $\mathbf{P}\left[H_{n+1} \geq H_n e^{-\mathbf{F}} \middle| \mathcal{F}_n\right] = 1$. So solving equations,

$$\mathbf{C}^{-1} A = \mathbf{D}^{-1} \ln A + a$$

$$\mathbf{C}^{-1} B = \mathbf{D}^{-1} \ln B + a$$

$$A = B e^{-\mathbf{F}}$$

we get the values of B and a that will guarantee that ξ_n is a submartingale.

$$A = \frac{\mathbf{CF}}{\mathbf{D}(1-e^{-\mathbf{F}})} e^{-\mathbf{F}}$$
$$B = \frac{\mathbf{CF}}{\mathbf{D}(1-e^{-\mathbf{F}})}$$
$$a = \frac{1}{\mathbf{D}} \left(\frac{\mathbf{F}e^{\mathbf{F}}}{e^{\mathbf{F}}-1} + \ln \frac{\mathbf{D}(1-e^{-\mathbf{F}})}{\mathbf{CF}} \right)$$

QED

We will give an alternative proof of converse in the next section, which reallies on a conjecture we have made. Conjecture depends on an intuitive assumption that we failed to prove. Apart from the material presented in the next sub-section, all of the discussions in the thesis is independent of validity of the conjecture.

4.2 Alternative Converse & A Conjecture

The motivation of the following calculation is two fold. The first is to find a bound that is asymptotically as tight as the one found by Burnashev using simpler probabilistic tools. The second is to understand the connection between the converse proof done by Burnashev and the converse proof claimed in [13].

Although the earlier proof dealt quite a bit with the entropy, it did not impose any structure on the coding algorithm or decoding rule. It only used the bound on the expected value of the entropy, which is valid for any coding algorithm decoding rule pair that has an expected error P_e because of the generalized Fano inequality. The only converse proof in the literature for generalized block-coding schemes before Burnashev's work is in [13] for the infinite bandwidth AWGNC channel with an amplitude limit. However they restricted themselves to the set of decoders that only decodes when the a posteriori probability of one of the messages goes above $1-P_e$. As mentioned by Burnashev it is by no means obvious that the optimal decoder should be of this form. Let us call the restricted set of decoding rules $\mathfrak{D}_R$ and the general set $\mathfrak{D}$. What we conjectured is

$$\min_{\mathfrak{C},\mathfrak{D}} \mathbf{E}\left[\tau \middle| \mathcal{F}_0\right] \geq \lambda(P_e) \min_{\mathfrak{C},\mathfrak{D}_R} \mathbf{E}\left[\tau \middle| \mathcal{F}_0\right] \tag{4.14}$$

where $\lim_{P_e \to 0} \lambda(P_e) = 1$. Thus the restriction from $\mathfrak{D}$, to $\mathfrak{D}_R$ for the set of possible decoders, does not change results about error exponent or channel capacity.

The following is the theorem and partial proof of this conjecture. There is a monotonicity assumption which is taken for granted in the proof of the theorem. Although the assumption seems intuitive we have no proof of it at this time.[5]

Theorem 2. *For any transmission method over a DMC with feedback,* $\forall P_e > 0$, $\forall M > 2$ *and for all* $e^{-\mathbf{F}} > \delta > P_e$ *the expected number of observations is bounded as follows.*

$$\mathbf{E}\left[\tau\right] \geq \left(1 - \frac{P_e(\ln(M-1) - \ln P_e + 1)}{\delta}\right)\left(\frac{\ln M}{\mathbf{C}} - \frac{\ln \delta}{\mathbf{D}} - \frac{1}{\mathbf{C}} - \frac{\mathbf{F}}{\mathbf{D}}\right) \tag{4.15}$$

Proof:

We will first start with finding a bound on the probability of the event that $\{H_\tau > \delta\}$, for any δ, where H_τ is the value of the entropy[6] at the decoding time. Then we will define a method to obtain a 'modified stopping rule' for any coding algorithm decoding rule pair such that the expected value of the 'modified stopping time' is proportional to a lower bound on the expected decoding time of the pair. After that we will bound the expected value of 'modified stopping time'. Finally we will combine these to find a lower bound on expected decoding time to propose a lower bound on reliability function.

Let $\mathcal{F}_\infty$ be the σ-field that includes all of the $\mathcal{F}_n$'s i.e., $\mathcal{F}_\infty = \cup_{k=0}^{\infty} \mathcal{F}_k$. Then $H_\tau \leq \delta$ is a well defined event in $\mathcal{F}_\infty$. Then we can write the expectation of the decoding

[5]Since we do not rely on this monotonicity assumption in any of the calculations other then the alternative proof of converse all of our results are valid independent of validity of this assumption.

[6]Indeed the conditional entropy given the observation $\mathfrak{f}_n$.

time in terms of conditional expectations as follows

$$\begin{aligned}\mathbf{E}\left[H_\tau\right] &= \mathbf{E}\left[H_\tau \mid H_\tau \leq \delta\right]\mathbf{P}\left[H_\tau \leq \delta\right] + \mathbf{E}\left[H_\tau \mid H_\tau > \delta\right]\mathbf{P}\left[H_\tau > \delta\right] \\ &\geq \delta \mathbf{P}\left[H_\tau > \delta\right]\end{aligned}$$

Using the generalized Fano inequality, $\mathbf{E}\left[H_\tau\right] \leq \mathfrak{h}(P_e) + P_e \ln\left(M-1\right)$

$$\mathbf{P}\left[H_\tau > \delta\right] \leq P_e \frac{\ln(M-1) - \ln P_e + 1}{\delta} \tag{4.16}$$

In order to be able describe the modified scheme, and write bounds on the expected value of modified scheme we need to set the notation for threshold crossing times of entropy. Let the initial value of the entropy be A, and the threshold for stopping be B, and the first time instance with an H_n below threshold be $\mathcal{T}_{A\to B} = \min\{n : H_n \leq B\}$. Then we can define the minimum of expected value of $\mathcal{T}_{A\to B}$ over all coding algorithms.[7]

$$\alpha(A, B) = \min_{\mathfrak{C}} \mathbf{E}\left[\mathcal{T}_{A\to B}\right] \tag{4.17}$$

For any coding/decoding pair, and for any $\delta > 0$ we can define the following stopping time. We run the coding decoding algorithm once without any interruption.If $H_\tau \leq \delta$ we stop, else we start the coding algorithm which has the minimum expected time to reach the threshold δ.[8]

$$\begin{aligned}\mathbf{E}\left[\tau'(\delta)\right] &= \mathbf{E}\left[\tau \mid H_\tau \leq \delta\right]\mathbf{P}\left[H_\tau \leq \delta\right] + (\mathbf{E}\left[\tau + \mathcal{T}_{H_\tau\to\delta} \mid H_\tau > \delta\right])\mathbf{P}\left[H_\tau > \delta\right] \\ &= \mathbf{E}\left[\tau \mid H_\tau \leq \delta\right]\mathbf{P}\left[H_\tau \leq \delta\right] + \mathbf{E}\left[\tau \mid H_\tau > \delta\right]\mathbf{P}\left[H_\tau > \delta\right] + \mathbf{E}\left[\mathcal{T}_{H_\tau\to\delta} \mid H_\tau > \delta\right]\mathbf{P}\left[H_\tau > \delta\right] \\ &= \mathbf{E}\left[\tau\right] + \alpha(H_\tau \to \delta)\mathbf{P}\left[H_\tau > \delta\right]\end{aligned}$$

If we assume $\alpha(\cdot \to \delta)$ to be monotonic function, since $H_\tau \leq H_0$, we can conclude

[7] There might be different probability distributions that correspond to the same entropy value, we stick with the one at hand to define α, but we will later lower bound it using function that is blind to the initial distribution other then its entropy value.

[8] Note that we are not imposing a unique algorithm for different decoding points, at every decoding point with $H_\tau > \delta$ transmitter will pick the coding algorithm with least expected time to reach δ.

that $\alpha(H_\tau \to \delta) \leq \alpha(H_0 \to \delta)$. Thus

$$\mathbf{E}\left[\tau'(\delta)\right] \leq \mathbf{E}\left[\tau\right] + \alpha(H_0 \to \delta)\mathbf{P}\left[H_\tau > \delta\right]$$

Using the fact $\mathbf{E}\left[\tau'(\delta)\right] \geq \alpha(H_0 \to \delta)$

$$\alpha(H_0 \to \delta) \leq \mathbf{E}\left[\tau\right] + \alpha(H_0 \to \delta)\mathbf{P}\left[H_\tau > \delta\right]$$

Which will immediately lead to

$$\mathbf{E}\left[\tau\right] \geq \alpha(H_0 \to \delta)\mathbf{P}\left[H_\tau \leq \delta\right] \tag{4.18}$$

We have already bounded $\mathbf{P}\left[H_\tau \leq \delta\right]$ by 4.16, we will bound the $\alpha(H_0 \to \delta)$ using lemmas 2, 3,4 and 5.

Because of Lemma 4 the decrease of $\ln H_n$ in one time unit is upper bounded by $\mathbf{F}$. As a result H_{n+1} can not go below $e^{-\mathbf{F}}$ if H_n is greater then 1. We can write $\mathcal{T}_{H_0 \to \delta}$ in two parts. The first part is from the starting time to the first time that entropy goes below 1. The second is the time from the first value of H_n less then 1 to δ. Since this first value is greater than $e^{-\mathbf{F}}$ we can lower bound the expectation of second phase with the expected time from $e^{-\mathbf{F}}$ to δ

$$\alpha(H_0, \delta) \geq \alpha(H_0, 1) + \alpha(e^{-\mathbf{F}}, \delta)$$

Note that $|H_n| \leq \ln M$, $\mathbf{E}\left[H_n - H_{n+1} | \mathcal{F}_n\right] \leq \mathbf{C}$, thus we can apply lemma 5 part 4, for H_n and first passage time of 1,

$$\alpha(H_0, 1) \geq \frac{H_0 - 1}{\mathbf{C}}$$

consequently

$$\alpha(H_0, 1) \geq \frac{\ln M - 1}{\mathbf{C}}$$

Note that $|\ln H_n| \leq \mathbf{F}n + \ln\ln M$, $\mathbf{E}\left[\ln H_n - \ln H_{n+1} | \mathcal{F}_n\right] \leq \mathbf{D}$, thus we can apply lemma 5 part 4, for $\ln H_n$ and first passage time of δ,

$$\alpha(e^{-\mathbf{F}}, \delta) \geq \frac{-\ln \delta}{\mathbf{D}}$$

consequently

$$\alpha(e^{-\mathbf{F}}, \delta) \geq \frac{-\mathbf{F} - \ln \delta}{\mathbf{D}}$$

Using 4.18, we get the required relation

QED

We can make various substitutions, for δ in theorem 2 to get the bound on reliability function, we already have as result of theorem 1. Note that if $E(R) \neq 0$ then $\lim_{P_e \to 0} \frac{\ln M}{-\ln P_e} < \infty$ as a result for small enough P_e values we can make the substitution $\delta = P_e(\ln M - \ln P_e + 1)(-\ln P_e) \leq e^{-\mathbf{F}}$, to get the relation,

$$\mathbf{E}[\tau] \geq \left(1 + \frac{1}{\ln P_e}\right)\left(\frac{\ln M}{\mathbf{C}} - \frac{\ln P_e}{\mathbf{D}} - \frac{\ln(-\ln P_e) + \ln(\ln M - \ln P_e + 1)}{\mathbf{D}} - \frac{\mathbf{F}}{\mathbf{D}} - \frac{1}{\mathbf{C}}\right) \tag{4.19}$$

Indeed Equation (4.19) tells us the reason why the converse proofs in [13], which restricts the set of possible decoders to $\mathfrak{D}_R$, can be transfered to general case, where decoder set is $\mathfrak{D}_R$.[9] Because (4.19) mean that for any coding/decoding algorithm pair, with a average probability of error, P_e; there exist a coding/decoding algorithm pair such that the decoding occurs only when the entropy of the messages goes below a threshold, $P_e(-\ln P_e)(\ln M - \ln P_e + 1)$, and expected decoding time of the modified scheme is proportional to a lower bound on the expected decoding time of the original scheme. Furthermore the proportionality constant goes to 1 as P_e goes to 0.

[9] In that part of the discussion we did not use anything specific to DMC, channel could perfectly be AWGNC. There are issues about cost constraints but those can be worked out.

Chapter 5

Achievability

In the last section we found lower bounds on the decoding time, depending on the size of message set and the average error probability. It will be shown in this section that there exist coding algorithms with a decoding criterion such that their expected decoding time is upper bounded by a function whose main terms coincide with those of the lower bound.

We will give two separate proofs of achievability. The first will be the proof used by Burnashev in [2] for a special class of channels.[1] The second will be a more precise and detailed version of the proof by Yamamoto and Itoh in [17]. It is striking that although Yamamoto and Itoh were proposing an asymptotically optimal method for generalized block-coding schemes, they were not aware of it. They make a comparison with Horstein's work, [8], and conclude that their scheme is suboptimal. The comparison is unfair, however, because the scheme proposed by Horstein is not a generalized block-coding scheme.

Both of the proofs will require high-rate feedback, even higher than the evident bound $\ln|\mathcal{Y}|$, where $|\mathcal{Y}|$ is the size of output alphabet. But the Yamamoto-Itoh scheme will then be modified to give a coding scheme that will only need a feedback rate equal to the forward channel capacity.

[1] We won't give his proof for general case.

5.1 Burnashev Proof

Considering the definition of $\mathbf{D}$, we can see that at least one (i_0, j_0) pair exits such that

$$\mathbf{D} = \max_{i,j} \sum_l P_{il} \ln \frac{P_{il}}{P_{jl}} = \sum_l P_{i_0 l} \ln \frac{P_{i_0 l}}{P_{j_0 l}}$$

Then we will define $\mathbf{D}^*$ of the channel as

$$\mathbf{D}^* = \sum_l P_{j_0 l} \ln \frac{P_{j_0 l}}{P_{i_0 l}}$$

If more than one pair of (i, j)'s exists with a corresponding Kullback Leibler divergence equal to $\mathbf{D}$, we will define $\mathbf{D}^*$ to be the maximum of all reverse Kullback Leibler divergences, i.e.,

$$\mathbf{D}^* = \max_{(i_0, j_0); \sum_l P_{i_0 l} \ln \frac{P_{i_0 l}}{P_{j_0 l}} = \mathbf{D}} \sum_l P_{j_0 l} \ln \frac{P_{j_0 l}}{P_{i_0 l}}$$

Theorem 3. *For a DMC with infinite, noiseless and instantaneous feedback*

- *if* $\mathbf{D}^* > \mathbf{C}$ *a coding algorithm decoding, criterion pair exists such that*

  $$\mathbf{E}[\tau] < \frac{\ln M}{\mathbf{C}} - \frac{\ln P_e}{\mathbf{D}} + \mathbf{\Delta}$$

  where $\mathbf{\Delta}$ *is a constant determined by the transition probabilities.*

- *for all values of* $\mathbf{D}^*$ *and* $\forall \epsilon > 0$*, coding/decoding pairs exist such that*

  $$\mathbf{E}[\tau] < \frac{\frac{\ln M}{\mathbf{C}} - \frac{\ln P_e}{\mathbf{D}} + \frac{(\mathbf{D} - \mathbf{C}) \ln \epsilon}{\mathbf{CD}}}{1 - \epsilon} + \mathbf{\Delta}$$

  where $\mathbf{\Delta}$ *is a constant determined by the transition probabilities.*

We will prove only the first statement of the theorem, which is for the channels such that $\mathbf{D}^* > \mathbf{C}$. We will prove a theorem which is almost equivalent[2] to the second

[2]Except that it only works efficiently for equiprobable messages; it is unable to make use of

statement of the theorem, in the next section. The coding scheme that will be used will only require a finite delay noiseless feedback channel whose capacity is equal to the forward channel capacity, **C**.

Proof:

Our decoding rule will be log-likelihood decoding; for any given $\delta > 0$, we will decode at the first time instant that the log likelihood ratio of one of the messages goes above $\ln 1/\delta$. It is easy to show that the probability of error is less than or equal to δ with log-likelihood decoding.

In proving achievability we will work with log-likelihood ratios rather than the entropy. The log-likelihood ratio of the m^{th} message at time n is defined as

$$\Lambda_m(\mathfrak{f}_n) = \ln \frac{p_m(\mathfrak{f}_n)}{1 - p_m(\mathfrak{f}_n)} \tag{5.1}$$

Note that $p_m(\mathfrak{f}_n)$ and $\Lambda_m(\mathfrak{f}_n)$ are measurable in the σ-field generated by $\mathcal{F}_{n-1}$ and Y_n, i.e., we do not need to know the assignments of the messages for the next time instant in order to calculate the log-likelihood ratios of the messages. Log-likelihood decoding can be summarized by the stopping time,

$$\tau(\delta) = \min\{n : \max_j \Lambda_j(\mathfrak{f}_n) \geq \ln \frac{1}{\delta}\} \tag{5.2}$$

We will denote the log-likelihood ratio of the true message at time n by $\Lambda(\mathfrak{f}_n)$. Then we can define another stopping time as

$$\tau_{tr}(\delta) = \min\{n : \Lambda(\mathfrak{f}_n) \geq \ln \frac{1}{\delta}\} \tag{5.3}$$

For any coding algorithm, $\tau(\delta)$ will be less then or equal to $\tau_{tr}(\delta)$ for all realizations of the experiment.[3] Consequently for all coding algorithms the expected value of stopping time $\tau_{tr}(\delta)$ is an upper bound for the expected decoding time with log-likelihood decoding; $\mathbf{E}[\tau(\delta)] \leq \mathbf{E}[\tau_{tr}(\delta)]$.

smaller starting entropy values with M messages.

[3]Note that here we are talking about $\mathcal{G}$ rather than $\mathcal{F}$, since $\tau_{tr}(\delta)$ is not measurable in $\mathcal{F}$.

It is evident that a decoding time should be a stopping time with respect to filtration $\mathcal{F}$, i.e., it should only depend on the observation of the receiver. Nevertheless a mathematical stopping time, as opposed to a decoding time for a coding algorithm, is not necessarily a stopping time with respect to the filtration $\mathcal{F}$; rather it can be a stopping time with respect to $\mathcal{G}$. Indeed this is the difference between $\tau(\delta)$ and $\tau_{tr}(\delta)$

We will start with developing a coding algorithm such that the sequence of $\Lambda(\mathfrak{f}_n)$'s forms a submartingale of the form,

$$\mathbf{E}\left[\Lambda(\mathfrak{f}_{n+1}) - \Lambda(\mathfrak{f}_n)\middle|\, \mathcal{F}_{n-1}, Y_n, \theta\right] \geq \left\{ \begin{array}{ll} \mathbf{C} & \text{if } \Lambda(\mathfrak{f}_n) < \ln \frac{p_0}{1-p_0} \\ \mathbf{D} & \text{if } \Lambda(\mathfrak{f}_n) \geq \ln \frac{p_0}{1-p_0} \end{array} \right\} \tag{5.4}$$

where p_0 is a constant to be determined later. Then we will use this submartingale to establish an upper bound on $\mathbf{E}\left[\tau_{tr}(\delta)\right]$.

Our coding algorithm will involve a random part, i.e., the assignments that we send back to the transmitter via Z_n, to be used at time $n+1$, will not necessarily be a deterministic function of Y_n and $\mathcal{F}_{n-1}$. The relation will be just a probabilistic one as follows.

We will set the a posteriori probability of the k^{th} input letter given $\mathcal{F}_{n-1} = \mathfrak{f}_{n-1}$, $Y_n = y_n$, $\theta = m$ to be $\phi_k(\mathfrak{f}_{n-1}, y_n, m)$.

$$\sum_{k=1}^{K} \phi_k(\mathfrak{f}_{n-1}, y_n, m) = 1 \qquad \forall y_n, \forall \mathfrak{f}_{n-1}, \forall m$$

$\phi_k(\mathfrak{f}_{n-1}, y_n, m)$ will be such that a posteriori probability of the k^{th} input letter given $\mathcal{F}_{n-1} = \mathfrak{f}_{n-1}$, $Y_n = y_n$ will be φ^0_k, where $\varphi^0 = (\varphi^0_1, \varphi^0_2, \ldots, \varphi^0_K)$ is the capacity achieving probability distribution.

$$\sum_{m=1}^{M} \phi_k(\mathfrak{f}_{n-1}, y_n, m) p_m(\mathfrak{f}_n) = \varphi^0_k \qquad \forall k, \forall y_n, \forall \mathfrak{f}_{n-1} \tag{5.5}$$

At each time n, for all $m = 1, \ldots, M$, the receiver will make an assignment according to the probability mass function $\phi(\mathfrak{f}_{n-1}, y_n, m)$. The assignment of the

m^{th} message will be denoted by $\kappa(m)$. Evidently each $\kappa(m)$ is a part of Z_n and will be used at time $n+1$, in the calculation of $\Lambda(\mathfrak{f}_{n+1})$. Once these assignments are made we can calculate the a posteriori probability of the k^{th} input letter, given $\mathcal{F}_n$, at time $n+1$ as follows:

$$\varphi_k(\mathfrak{f}_n) = \sum_{\kappa(m)=k} p_m(\mathfrak{f}_n) \tag{5.6}$$

We will need to find the expected value of $\varphi_k(\mathfrak{f}_n)$, conditioned on $\mathcal{F}_{n-1}$,Y_n,θ,$\kappa(\theta)$. Using (5.6),

$$\begin{aligned}
\mathbf{E}\left[\varphi_k(\mathfrak{f}_n)\middle|\, \mathcal{F}_{n-1}, Y_n, \theta, \kappa(\theta)\right] &= \mathbf{E}\left[\sum_{\kappa(m)=k} p_m(\mathfrak{f}_n)\middle|\, \mathcal{F}_{n-1}, Y_n, \theta, \kappa(\theta)\right] \\
&= \mathbf{E}\left[\sum_{m} \mathbb{I}_{\{\kappa(m)=k\}} p_m(\mathfrak{f}_n)\middle|\, \mathcal{F}_{n-1}, Y_n, \theta, \kappa(\theta)\right] \\
&= \sum_{m} \mathbf{P}\left[\kappa(m)=k\middle|\, \mathcal{F}_{n-1}, Y_n, \theta, \kappa(\theta)\right] p_m(\mathfrak{f}_n)
\end{aligned}$$

For each $m \neq \theta$, $\mathbf{P}\left[\kappa(m)=k\middle|\, \mathcal{F}_{n-1}, Y_n, \theta, \kappa(\theta)\right] = \phi_k(\mathfrak{f}_{n-1}, y_n, m)$,

Where as for $m=\theta$, $\mathbf{P}\left[\kappa(m)=k\middle|\, \mathcal{F}_{n-1}, Y_n, \theta, \kappa(\theta)\right] = \delta_{\kappa(\theta),k}$. Thus the conditional expected value of the a posteriori probability of k^{th} input letter at time $n+1$, given $\mathcal{F}_{n-1}$,Y_n,θ,$\kappa(\theta)$ is

$$\begin{aligned}
\mathbf{E}\left[\varphi_k(\mathfrak{f}_n)\middle|\, \mathcal{F}_{n-1}, Y_n, \theta, \kappa(\theta)\right] &= \sum_{m\neq\theta} \phi_k(\mathfrak{f}_{n-1}, y_n, m) p_m(\mathfrak{f}_n) + p_\theta(\mathfrak{f}_n)\delta_{\kappa(\theta),k} && (5.7)\\
&= \varphi_k^0 + (\delta_{k,\kappa(\theta)} - \phi_k(\mathfrak{f}_{n-1}, y_n, \theta))p_\theta(\mathfrak{f}_n) && (5.8)
\end{aligned}$$

where we have used (5.5) in the last step

The a posteriori probability, $p_m(\mathfrak{f}_{n+1})$, of the m^{th} message at time $n+1$ can be written as follows:

$$\begin{aligned}\mathbf{P}\left[\theta=m\right|\mathcal{F}_{n+1}=\mathfrak{f}_{n+1}] &= \mathbf{P}\left[\theta=m\right|\mathcal{F}_n=\mathfrak{f}_n, Y_{n+1}=l] \\ &= \frac{\mathbf{P}\left[\theta=m\right|\mathcal{F}_n=\mathfrak{f}_n]\,\mathbf{P}\left[Y_{n+1}=l\right|\mathcal{F}_n=\mathfrak{f}_n,\theta=m]}{\mathbf{P}\left[Y_{n+1}=l\right|\mathcal{F}_n=\mathfrak{f}_n]} \\ &= \frac{\mathbf{P}\left[\theta=m\right|\mathcal{F}_n=\mathfrak{f}_n]\,P_{\kappa(m)l}}{\sum_{j=1}^{M}\mathbf{P}\left[\theta=j\right|\mathcal{F}_n=\mathfrak{f}_n]\,P_{\kappa(j)l}}\end{aligned}$$

Using equation (5.6),

$$p_m(\mathfrak{f}_{n+1}) = \frac{p_m(\mathfrak{f}_n)P_{\kappa(m)l}}{\sum_{j=1}^{K}\varphi_j(\mathfrak{f}_n)P_{jl}}$$

$$\Lambda(\mathfrak{f}_{n+1})-\Lambda(\mathfrak{f}_n) = \ln\frac{(1-p_\theta(\mathfrak{f}_n))P_{\kappa(\theta)l}}{\sum_{j=1}^{K}\varphi_j(\mathfrak{f}_n)P_{jl}-p_\theta(\mathfrak{f}_n)P_{\kappa(\theta)l}} \tag{5.9}$$

Using the convexity of $\ln(\frac{1}{x-\phi})$ in x with equation (5.9)

$$\mathbf{E}\left[\Lambda(\mathfrak{f}_{n+1})-\Lambda(\mathfrak{f}_n)\right|\mathcal{F}_{n-1},Y_n,\theta,\kappa(\theta)] \geq \sum_{l=1}^{L}P_{\kappa(\theta)l}\ln\frac{(1-p_\theta(\mathfrak{f}_n))P_{\kappa(\theta)l}}{\sum_{k=1}^{K}\mathbf{E}\left[\varphi_k(\mathfrak{f}_n)\right|\mathcal{F}_{n-1},Y_n,\theta,\kappa(\theta)]\,P_{kl}-p_\theta(\mathfrak{f}_n)P_{\kappa(\theta)l}}$$

Using equation (5.8)

$$\begin{aligned}&\geq \sum_l P_{\kappa(\theta)l}\ln\frac{(1-p_\theta(\mathfrak{f}_n))P_{\kappa(\theta)l}}{\sum_k(\varphi_k^0+(\delta_{k,\kappa(\theta)}-\phi_k(\mathfrak{f}_{n-1},y_n,\theta))p_\theta(\mathfrak{f}_n))P_{kl}-p_\theta(\mathfrak{f}_n)P_{\kappa(\theta)l}} \\ &= \sum_l P_{\kappa(\theta)l}\ln\frac{(1-p_\theta(\mathfrak{f}_n))P_{\kappa(\theta)l}}{\sum_k(\varphi_k^0-\phi_k(\mathfrak{f}_{n-1},y_n,\theta)p_\theta(\mathfrak{f}_n))P_{kl}} \\ &= \sum_l P_{\kappa(\theta)l}\left[\ln\frac{(1-p_\theta(\mathfrak{f}_n))P_{\kappa(\theta)l}}{\sum_k\varphi_k^0P_{kl}}-\ln\left(1-\frac{\sum_k\phi_k(\mathfrak{f}_{n-1},y_n,\theta)P_{kl}}{\sum_k\varphi_k^0P_{kl}}p_\theta(\mathfrak{f}_n)\right)\right]\end{aligned}$$

Using the fact that $\mathbf{C}=\ln\frac{P_{k,l}}{\sum_{i=1}^{K}\varphi_i^0P_{i,l}}$ $\forall\varphi_k^0>0$, for the capacity achieving probability distribution, φ^0.

$$= \mathbf{C}+\ln(1-p_\theta(\mathfrak{f}_n))-\sum_l P_{\kappa(\theta)l}\ln\left(1-\frac{\sum_k\phi_k(\mathfrak{f}_{n-1},y_n,\theta)P_{k,l}}{\sum_k\varphi_k^0P_{k,l}}p_\theta(\mathfrak{f}_n)\right)$$

Using iterated expectations on $\Lambda(\mathfrak{f}_{n+1}) - \Lambda(\mathfrak{f}_n)$

$$\begin{aligned}\mathbf{E}\left[\Lambda(\mathfrak{f}_{n+1}) - \Lambda(\mathfrak{f}_n)\right| \mathcal{F}_{n-1}, Y_n, \theta] &= \mathbf{E}\left[\mathbf{E}\left[\Lambda(\mathfrak{f}_{n+1}) - \Lambda(\mathfrak{f}_n)\right| \mathcal{F}_{n-1}, Y_n, \theta, \kappa(\theta)]\right| \mathcal{F}_{n-1}, Y_n, \theta] \\ &\geq \mathbf{C} + \ln(1 - p_\theta(\mathfrak{f}_n)) \\ &\quad - \sum_{k,l} \phi_k(\mathfrak{f}_{n-1}, y_n, \theta) P_{kl} \ln\left(1 - \frac{\sum_{i=1}^K \phi_i(\mathfrak{f}_{n-1}, y_n, \theta) P_{il}}{\sum_{i=1}^K \varphi_i^0 P_{il}} p_\theta(\mathfrak{f}_n)\right)\end{aligned}$$

Note that both $\psi(l) = \sum_k \phi_k(\mathfrak{f}_{n-1}, y_n, \theta) P_{kl}$ and $\hat{\psi}(l) = \sum_k \varphi_k^0 P_{kl}$ give us probability mass functions on the output set $\mathcal{Y}$. we then can write the last term of the sum as follows,

$$\begin{aligned}\sum_{k,l} \phi_k(\mathfrak{f}_{n-1}, y_n, \theta) P_{kl} \ln\left(1 - \frac{\sum_i \phi_i(\mathfrak{f}_{n-1}, y_n, \theta) P_{il}}{\sum_i \varphi_i^0 P_{il}} p_\theta(\mathfrak{f}_n)\right) &= \sum_l \psi_l \ln\left(1 - \frac{\psi_l}{\hat{\psi}_l} p_\theta(\mathfrak{f}_n)\right) \\ &= \sum_l \hat{\psi}_l \frac{\psi_l}{\hat{\psi}_l} \ln\left(1 - \frac{\psi_l}{\hat{\psi}_l} p_\theta(\mathfrak{f}_n)\right)\end{aligned}$$

Using the convexity of the function $x \ln(1 - \delta x)$, $\forall \delta \geq 0$ in x, we get

$$\begin{aligned}\sum_{k,l} \phi_k(\mathfrak{f}_{n-1}, y_n, \theta) P_{kl} \ln\left(1 - \frac{\sum_i \phi_n(\mathfrak{f}_{i-1}, y_i, \theta) P_{il}}{\sum_i \varphi_i^0 P_{il}} p_\theta(\mathfrak{f}_n)\right) &\geq (\sum_l \hat{\psi}_l \frac{\psi_l}{\hat{\psi}_l}) \ln\left(1 - \sum_l \hat{\psi}_l \frac{\psi_l}{\hat{\psi}_l} p_\theta(\mathfrak{f}_n)\right) \\ &= \ln(1 - p_\theta(\mathfrak{f}_n))\end{aligned}$$

As a result the expected increase in log-likelihood ratio of the true message is greater then channel capacity, $\mathbf{C}$, with this randomized coding algorithm.[4]

$$\mathbf{E}\left[\Lambda(\mathfrak{f}_{n+1}) - \Lambda(\mathfrak{f}_n)\right| \mathcal{F}_{n-1}, Y_n, \theta] \geq \mathbf{C} \tag{5.10}$$

Now we know how to code in order to obtain part of the submartingale in equation (5.4) when $p_\theta < p_0$. But we need to find the appropriate threshold p_0 and also the coding for the second part. Note that if we assign the true message to i_0 and all of

[4]We can not argue right away that a deterministic coding algorithm exists which satisfies equation (5.10). Any random coding method is a weighted sum of deterministic ones, but it is possible that none of the deterministic algorithms satisfies equation (5.10) for all possible θ, while their average does.

the others to j_0 as a result of equation (5.9) we know that the expected increase in log likelihood ratio of the true message is $\mathbf{D}$. But a coding algorithm can not assign messages to input letters in a way that will depend on the true message, i.e., symbol assigned to a message can only depend on feedback and the message itself. Thus we can not assume that receiver assigns the true message to i_0 and all others to j_0 it can only assign the messages depending on the realization of $\mathcal{F}_n$ not depending on θ. The way we deal with this problem is as follows.

When the a posteriori probability of one message is over a threshold,p_0 then we will assign the most likely message to i_0 and all others to j_0. Consequently if the message with the highest a posteriori probability is the same as the true message, i.e.,$\hat{\theta} = \theta$

$$\mathbf{E}\left[\Lambda(\mathfrak{f}_{n+1}) - \Lambda(\mathfrak{f}_n) \middle| \mathcal{F}_{n-1}, Y_n, \theta, \hat{\theta} = \theta\right] = \mathbf{D} \tag{5.11}$$

otherwise, some $\theta \neq \hat{\theta}$ is the most likely message and

$$\begin{aligned}\mathbf{E}\left[\Lambda(\mathfrak{f}_{n+1}) - \Lambda(\mathfrak{f}_n) \middle| \mathcal{F}_{n-1}, Y_n, \theta, \hat{\theta} \neq \theta\right] &= \sum_{l=1}^{L} p_{j_0 l} \ln \frac{p_{j_0 l}(1 - p_\theta(\mathfrak{f}_n))}{p_{j_0 l}(1 - p_\theta(\mathfrak{f}_n)) + p_{\hat{\theta}}(\mathfrak{f}_n)(p_{i_0 l} - p_{j_0 l})} \\ &= \mathbf{D}^* + \sum_{l=1}^{L} p_{j_0 l} \ln \frac{p_{i_0 l}(1 - p_\theta(\mathfrak{f}_n))}{p_{j_0 l}(1 - p_\theta(\mathfrak{f}_n) - p_{\hat{\theta}}(\mathfrak{f}_n)) + p_{\hat{\theta}}(\mathfrak{f}_n) p_{i_0 l}}\end{aligned}$$

Note that second term in the expression can be made as close to zero as desired by setting higher and higher threshold values for phase change of the coding. Using the assumption $\mathbf{D}^* > \mathbf{C}$ we can guarantee an expected increase of $\mathbf{C}$ by setting an appropriate threshold value p_0, i.e.,

$$\mathbf{E}\left[\Lambda(\mathfrak{f}_{n+1}) - \Lambda(\mathfrak{f}_n) \middle| \mathcal{F}_{n-1}, Y_n, \theta, \hat{\theta} \neq \theta\right] \geq \mathbf{C}$$

Thus, independent of the phase of the coding the expected increase in the log-likelihood ratio of the true message is greater then or equal to $\mathbf{C}$, if the a posteriori probability of the true message is less then p_0. Also when the a posteriori probability of the true message is greater then p_0 the expected increase in the log-likelihood of the true message is $\mathbf{D}$.

$$\mathbf{E}\left[\Lambda(\mathfrak{f}_{n+1}) - \Lambda(\mathfrak{f}_n) \middle| \mathcal{F}_{n-1}, Y_n, \theta\right] \geq \left\{ \begin{array}{ll} \mathbf{C} & \text{if } p_\theta(\mathfrak{f}_n) < p_0 \\ \mathbf{D} & \text{if } p_\theta(\mathfrak{f}_n) \geq p_0 \end{array} \right\}$$

Using the following lemma from, [3], for the filtration whose n^{th} element will be the σ-field generated by $\mathcal{F}_{n-1}$ and Y_n, and $\Lambda(\mathfrak{f}_n) - \ln \frac{p_0}{1-p_0}$ we get.

$$\mathbf{E}\left[\tau_{tr}(\delta)\right] < \frac{\ln M}{\mathbf{C}} - \frac{\ln(\delta)}{\mathbf{D}} + \mathbf{\Delta} \tag{5.12}$$

where $\mathbf{\Delta}$ is a constant determined by transition probabilities.[5]

Lemma 6. *Assume that the sequence $(\xi_n, \mathcal{F}_n)$,$n = 0, 1, 2\ldots$ forms a submartingale, where*

$$\begin{aligned} &\mathbf{E}\left[\xi_{n+1}\middle| \mathcal{F}_n\right] = \xi_n + K_1, \qquad \text{if } \xi_n < 0, \qquad \text{where } K_1 > 0 \\ &\mathbf{E}\left[\xi_{n+1}\middle| \mathcal{F}_n\right] = \xi_n + K_2, \qquad \text{if } \xi_n \geq 0, \qquad \text{where } K_2 > K_1 \\ &|\xi_{n+1} - \xi_n| \leq K_3, \qquad \xi_0 < 0 \end{aligned}$$

and Markov moment τ given by the condition

$$\tau = \min\{n : \xi_n \geq B\} \qquad \text{where } B > 0$$

Then we have the inequality

$$\mathbf{E}\left[\tau\middle| \mathcal{F}_0\right] \leq \frac{B}{K_2} + \frac{|\xi_0|}{K_1} + \mathbf{\Delta}(K_1, K_2, K_3)$$

where $\mathbf{\Delta}(K_1, K_2, K_3)$ only depends on K_1,K_2 and K_3.

QED

[5] The expectation is over channel realizations, and the random choice of coding symbols. There is no expectation over possible messages here.

5.2 Yamamoto & Itoh Proof

The scheme proposed is a two phase scheme, like the one in Burnashev's paper [2], but instead of using log-likelihood ratios, and bounding the average error probability by bounding the error probability at each and every decoding time, more conventional methods are used, i.e., given θ the probability of $\hat{\theta} \neq \theta$, $\mathbf{P}\left[\hat{\theta} \neq j \middle| \theta = j\right]$ is bounded. The time used for each phase is fixed in each 'trial'. In other words $l_0 = l_1 + l_2$ where l_1, l_2 are the corresponding times used for each phase in one trial, and l_0 is the over all length of one trial.

Phase 1:

A plain (non-feedback) code of rate $\mathbf{C}(1-\epsilon)$ is used to send the message. The coding theorem states that for any $\epsilon > 0$,any error probability $P_{e1} > 0$; any sufficiently large M, a code of block-length $\frac{\ln M}{\mathbf{C}(1-\epsilon)}$ or less exist with probability of error P_{e1} or less. In other words for $\forall P_{e1} > 0$, $\forall \epsilon > 0$, and $\forall M$ a code of length l_1 exists, such that

$$l_1 < \frac{\ln M}{\mathbf{C}(1-\epsilon)} + \Delta_1(P_{e1}, \epsilon) \tag{5.13}$$

where $\Delta_1(P_{e1}, \epsilon)$ is a function that is determined by the transition probabilities.

At the end of this phase the receiver has an estimate about the message, with an a posteriori probability. The likelihood corresponding to this a posteriori probability was used in Burnashev's Proof.

Phase 2:

As a result of perfect feedback, the transmitter knows the receivers estimate of the message instantaneously.[6] If this estimate is true the transmitter will send an acknowledgment message; if it is false it will send a rejection message. An acknowledgment will end the transmission for the current message, and the transmitter will start sending the next message. A rejection will lead to a retransmission of the current message again. There are two kinds of errors in the second phase. The first is that the receiver might interpret an acknowledgment as a rejection, $A \to R$, which will

[6] Momentarily we are assuming, infinite, instantaneous, error-free feedback that will allow us to let the receiver tell the estimate to the transmitter instantly. Better methods which will decrease the required feedback will be presented soon after the basic proof.

increase the expected decoding time. The second is interpreting a rejection as an acknowledgment, $R \to A$, which will result in an immediate error. Since our demand on probability of error is much more stringent,[7] we want the probability of $R \to A$ type errors, P_{RA}, to decay as fast as possible with time, under the constraint that the probability of $A \to R$ type errors P_{AR} is kept below some predefined value.

In order to understand the asymptotic behavior of these two errors as a function of signaling duration, we quote the following corollary in Csiszár and Körner, [4](pp 19, Corollary 1.2) Suppose one needs to choose between two probability distributions $P(x)$ and $Q(x)$. The output set will be divided into two parts,A & B one corresponding to detecting distribution P, A; one corresponding to the distribution as Q, B.

Lemma 7. *For any* $0 < \delta < 1$

$$\lim_{k \to \infty} \frac{1}{k} \log \beta(k, \delta) = -D(P||Q)$$

where

$$\beta(k, \delta) = \min_{A, P^k(A) > 1-\delta} Q^k(A)$$

Indeed lemma 7 is equivalent to saying $\forall \epsilon > 0$, $\delta > 0$ $\exists k_0(\delta, \epsilon)$ such that $\forall k > k_0(\epsilon, \delta)$

$$Q^k(A) \le e^{-kD(P||Q)(1-\epsilon)} \qquad \text{and} \qquad P^k(A) \ge 1 - \delta$$

After a little algebra, one can show that, the minimum time needed for detection between two probability distributions, with the condition that $\mathbf{P}\left[P \to Q\right] \le \delta$ and $\mathbf{P}\left[Q \to P\right] \le \varphi$, is upper bounded as

$$k(\varphi, \delta) \le \frac{-\ln \varphi}{D(P||Q)(1-\epsilon)} + \mathbf{\Delta}(\epsilon, \delta) \qquad \forall \epsilon > 0 \tag{5.14}$$

[7]We are trying to maximize $-\frac{\ln P_e}{\mathbf{E}[\tau]}$. A small increase in P_{RA} will increase $\mathbf{E}\left[\tau\right]$ slightly, where a small increase in P_{RA} will decrease $-\ln P_e$ drastically.

Going back to our problem,

$$\mathbf{D} = \max_{i,j} D(P(y|x_i||P(y|x_j) = D(P(y|x_{i_0}||P(y|x_{j_0}))$$

If we use x_{i_0} for acknowledgment and x_{j_0} for rejection, we know that

$$l_2 < \frac{-\ln P_{RA}}{\mathbf{D}(1-\epsilon)} + \mathbf{\Delta}(P_{AR}, \epsilon) \qquad \forall \epsilon > 0, \forall P_{RA} > 0, \forall P_{AR} > 0 \tag{5.15}$$

where $\mathbf{\Delta}$ is a function determined by the transition probabilities.

In other words for any $P_{AR} > 0$ and $\epsilon > 0$

$$l_0 = l_1 + l_2$$
$$l_0 \leq \left(\frac{\ln M}{\mathbf{C}} - \frac{\ln P_{RA}}{\mathbf{D}} \right) \frac{1}{1-\epsilon} + \mathbf{\Delta}(P_{AR}, P_{e1}, \epsilon)$$

After the first trial, there are four possible outcomes. The possible events with corresponding probabilities are

1. True message was estimated, acknowledgment was successful, $(1-P_{e1})(1-P_{AR})$
2. True message was estimated, acknowledgment was unsuccessful, $(1-P_{e1})P_{AR}$
3. Estimate was wrong, rejection was successful, $P_{e1}(1-P_{RA})$
4. Estimate was wrong, rejection was unsuccessful, $P_{e1}P_{RA}$

It is evident that the first case corresponds to successful decoding at the first trial, and the last case corresponds to erroneous decoding at the first trial. The other two cases correspond to retransmission. Let us call the retransmission probability at a trial P_η. Then the probability of error, P_e, is

$$P_e = P_{e1}P_{RA} \left(\sum_{k=0}^{\infty} P_\eta^k \right) = \frac{P_{e1}P_{RA}}{1-P_\eta}$$

In [17] this expression is given as $P_e = P_{e1}P_{RA}$ which is not true, but is a reasonable approximation. Similarly the expected decoding time can be written as

$$\mathbf{E}[\tau] = l_0 \sum_{k=0}^{\infty} P_\eta^k = \frac{l_0}{1 - P_\eta}$$

Finally the retransmission probability is

$$P_\eta = P_{e1} P_{RR} + (1 - P_{e1}) P_{AR}$$

If we fix $P_{e1} = P_{AR} = \delta$, then $P_\eta < 2\delta$, and the expected decoding time is

$$\mathbf{E}[\tau] \leq \frac{l_0}{1 - 2\delta} \qquad P_e < \frac{\delta}{(1 - 2\delta)} P_{RA} \tag{5.16}$$

This will mean that for any $\delta > 0$ and $\epsilon > 0$ a coding/decoding method exists whose expected decoding time satisfies the following inequality

$$\mathbf{E}[\tau] < \frac{1}{1 - 2\delta} \left[\left(\frac{\ln M}{\mathbf{C}} - \frac{\ln P_{RA}}{\mathbf{D}} \right) \frac{1}{1 - \epsilon} + \Delta'(P_{AR}, P_{e1}, \epsilon) \right] \tag{5.17}$$

Using equations (5.15) and (5.16)

$$\mathbf{E}[\tau] < \frac{1}{1 - 2\delta} \left[\left(\frac{\ln M}{\mathbf{C}} - \frac{\ln P_e}{\mathbf{D}} + \frac{\frac{\delta}{(1-2\delta)}}{\mathbf{D}} \right) \frac{1}{1 - \epsilon} + \Delta'(\delta, \epsilon) \right] \tag{5.18}$$

$$\mathbf{E}[\tau] < \frac{1}{1 - 2\delta - \epsilon} \left(\frac{\ln M}{\mathbf{C}} - \frac{\ln P_e}{\mathbf{D}} + \Delta'(\delta, \epsilon) \right) \tag{5.19}$$

This immediately proves the achievability of Burnashev's exponent.

5.3 What is actually needed

Throughout all of the calculations it was assumed that an instantaneous, infinite and error-free feedback is available. It is evident that the converse result (lower bound on expected decoding time) will still hold for any feedback which is deprived of any of these qualities. A first look at the system would suggest that, knowing the channel output exactly instantaneously should be enough for achieving the best possible reliability function. In other words an error-free, finite delay feedback path with a channel capacity of $\ln |Y|$ should be enough. Instead of proving this result we will prove a stronger result.[8]

Let us assume that constant delay is $\mathbf{T}$. Let the size of message set be $M = N_1^{N_2}$. As a similar approach with the *phase1* of previous section we can have codes of error probability $\frac{P_{e1}}{N_2}$, size N_1, and block-length $\frac{\ln N_1}{\mathbf{C}(1-\epsilon)}$. We can use this code N_2 times successively in order to send a member of super code, of size M, with a probability of error less then P_{e1}. If the transmitter sends each sub-message right after when it is decoded then the duration of of l_1, the time receiver needs to estimate the message sent, and inform the transmitter about its estimate, will be given by the following expression.

$$l_1 < \frac{N_2+1}{N_2}\frac{\ln M}{\mathbf{C}(1-\epsilon)} + \mathbf{T} + \boldsymbol{\Delta}_1(P_{e1}, \epsilon)$$

where $\boldsymbol{\Delta}_1(P_{e1}, \epsilon)$ is a function determined by transition probabilities. We can write this result by including $\frac{N_2+1}{N_2}$ in the coefficient of $\frac{\ln N}{\mathbf{C}}$ and the constant $\mathbf{T}$ in $\boldsymbol{\Delta}_1(P_{e1}, \epsilon)$. By doing this we are fixing N_2 which means any further asymptotic will be using N_1,which turns out to be enough.

$$l_1 < \frac{\ln M}{\mathbf{C}(1-2\epsilon)} + \boldsymbol{\Delta}_1'(P_{e1}, \epsilon, \mathbf{T})$$

In the second phase will be asking for the correctness of the super message rather than parts of it. We would have corresponding duration expression as

$$l_2 < \frac{-\ln P_{RA}}{\mathbf{D}(1-\epsilon)} + \boldsymbol{\Delta}_2(P_{AR}, \epsilon) + \mathbf{T}$$

So the duration of a trial can be bounded as

$$l_0 < \left(\frac{\ln M}{\mathbf{C}} - \frac{\ln P_{RA}}{\mathbf{D}}\right)\frac{1}{1-2\epsilon} + \boldsymbol{\Delta}(P_{AR}, \epsilon, \mathbf{T}) \tag{5.20}$$

[8]The Author was not initially aware of the work [11]; the results are derived independently.

where the function $\mathbf{\Delta}(P_{AR}, \epsilon, \mathbf{T})$ is function determined by transition probabilities. Following almost same steps we got the same result as equation (5.18)

Theorem 4. *For any DMC, with an error free feedback channel of rate* $\mathbf{C}$ *or higher of finite delay* $\mathbf{T}$, $\forall \delta > 0$, *there exist a coding scheme such that for all* $M > 1$ *and* $\forall P_e > 0$ *the expected decoding time* $\mathbf{E}[\tau]$ *is upper bounded as follows*

$$\mathbf{E}[\tau] \leq \frac{1}{1-\delta}\left(\frac{\ln M}{\mathbf{C}} - \frac{\ln P_e}{\mathbf{D}} + \mathbf{\Delta}(\delta, \mathbf{T})\right) \tag{5.21}$$

where $\mathbf{\Delta}$ *is a function determined by transition probabilities.*

This result shows the sufficiency of a noiseless feedback capacity $\mathbf{C}$. We have no result saying that feedback channels with smaller capacity will not be able to reach the same exponent. But it is plausible that $\mathbf{C}$ is also the necessary amount. Before starting the second phase the transmitter needs to know estimate at the receiver, and this requirement will introduce extra delay unless the feedback link has a rate equal to $\mathbf{C}$, at least for the set of generalized coding schemes we have used here to prove the achievability. For rates smaller then $\mathbf{C}$, one can make an optimization with the sphere packing exponent, by demanding a rate that is strictly smaller then $\mathbf{C}$, for *phase 1.*

Chapter 6

Preliminaries Concepts for Cost Constraint Case

6.1 Capacity and Detection Exponent Under Cost Constraint

The mutual information between input and output of a DMC channel with probability transition matrix $\mathfrak{P}$ and input distribution ϕ can be written as the minimum of a convex function over a compact set as follows.

$$\mathcal{I}_{(\phi,\mathfrak{P})} = \sum_{k=1,l=1}^{K,L} \phi_k P_{kl} \ln \frac{P_{kl}}{\sum_{j=1}^{K} \phi_j P_{jl}} = \min_{\psi} \sum_{k=1,l=1}^{K,L} \phi_k P_{kl} \ln \frac{P_{kl}}{\psi_l} \tag{6.1}$$

where ψ is constrained to be a probability mass function.

The capacity is $\mathbf{C} = \max_{\phi} \mathcal{I}_{(\phi,\mathfrak{P})}$. Consequently it can be written as

$$\mathbf{C} = \max_{\phi} \min_{\psi} \mathcal{L}(\phi, \psi) \qquad \mathcal{L}(\phi, \psi) = \sum_{k,l} \phi_k P_{kl} \ln \frac{P_{kl}}{\psi_l}$$

Note that $\mathcal{L}$ is convex in ψ and concave in ϕ. Since both ϕ and ψ are probability mass functions we also have compact set constraints. Thus it is known[1] that $\mathcal{L}$ has a

[1] Bertsekas [16] Saddle Point Theorem, Proposition 2.6.9 (case 1), pp151

saddle point and the min-max is equal to the max-min, i.e.,

$$\mathbf{C} = \min_{\psi} \max_{\phi} \sum_{k=1,l=1}^{K,L} \phi_k P_{kl} \ln \frac{P_{kl}}{\psi_l}$$

Using the fact that ϕ is a probability mass function

$$\mathbf{C} = \min_{\psi} \max_{k} \sum_{l} P_{kl} \ln \frac{P_{kl}}{\psi_l}$$

It is easy to see that $\mathbf{C} = \sum_l P_{kl} \ln \frac{P_{kl}}{\sum_j {\phi^*}_j P_{jl}}$ for all values of k such that ${\phi^*}_k > 0$ for the capacity achieving distribution ϕ^*.

Similar results exists for the cost constraint case. The cost constrained capacity is

$$\mathbf{C}(\mathcal{P}) = \max_{\substack{\phi \\ \sum_{k=1} K \phi_k \rho_k \leq \mathcal{P}}} \sum_{k=1,l=1}^{K,L} \phi_k P_{kl} \ln \frac{P_{kl}}{\sum_{r=1}^{K} \phi_r P_{rl}}$$

We can include cost constraint in the maximized expression it self,

$$\mathbf{C}(\mathcal{P}) = \max_{\phi} \min_{\gamma \geq 0} \left(\sum_{k=1,l=1}^{K,L} \phi_k P_{kl} \ln \frac{P_{kl}}{\sum_{r=1}^{K} \phi_r P_{rl}} + \gamma(\mathcal{P} - \sum_{k=1}^{K} \phi_k \rho_k) \right)$$

Using equation (6.1)

$$\mathbf{C}(\mathcal{P}) = \max_{\phi} \min_{\gamma \geq 0} \min_{\psi} \sum_{k,l} \phi_k \left(P_{kl} \ln \frac{P_{kl}}{\psi_l} + \gamma(\mathcal{P} - \rho_k) \right)$$

$$\begin{aligned} \mathbf{C}(\mathcal{P}) &= \max_{\phi} \min_{\gamma \geq 0, \psi} \mathcal{L}(\phi, \psi, \gamma) \\ \mathcal{L}(\phi, \psi, \gamma) &= \sum_{k} \phi_k \left(\sum_{l} P_{kl} \ln \frac{P_{kl}}{\psi_l} + \gamma(\mathcal{P} - \rho_k) \right) \end{aligned}$$

Since the set of $\gamma \geq 0$ is not compact we can not apply the argument used for capacity. But it is known[2] that $\mathcal{L}$ has a saddle point if ϕ has a compact set constraint,

[2]Bertsekas [16] Saddle Point Theorem, Proposition 2.6.9 (case 2), pp151

and there exists a ϕ^0 with a corresponding R such that $\{(\psi,\gamma)|\mathcal{L}(\phi^0,\psi,\gamma) < R\}$ is a non-empty compact set. Indeed this condition is just the existence of a probability mass function that satisfies the cost constraint. Provided that we do not have an impossible cost constraint $\mathcal{P}$, i.e., provided that $\mathcal{P} \geq \rho_{min}$, this condition is satisfied. Thus we can write the max-min as a min-max

$$\begin{aligned}
\mathbf{C}(\mathcal{P}) &= \min_{\psi} \min_{\gamma \geq 0} \max_{\phi} \sum_k \phi_k \left(\sum_l P_{kl} \ln \frac{P_{kl}}{\psi_l} + \gamma(\mathcal{P} - \rho_k) \right) \\
\mathbf{C}(\mathcal{P}) &= \min_{\psi} \min_{\gamma \geq 0} \max_{k} \sum_l P_{kl} \ln \frac{P_{kl}}{\psi_l} + \gamma(\mathcal{P} - \rho_k)
\end{aligned}$$

Also for every value of cost constraint $\mathcal{P}$, there exists a $\gamma_{\mathcal{P}}$ such that

$$\mathbf{C}(\mathcal{P}) = \min_{\psi} \max_{k} \sum_l P_{kl} \ln \frac{P_{kl}}{\psi_l} + \gamma_{\mathcal{P}}(\mathcal{P} - \rho_k) \tag{6.2}$$

If $\mathcal{P}$ is a non-trivial constraint,[3] i.e. $\mathbf{C}(\mathcal{P}) < \mathbf{C}$ then $\gamma_{\mathcal{P}} > 0$. Otherwise it will be zero.

One can define the following parameters for the problem of binary detection using a DMC with cost constraints.

Let us define $\mathbf{D}_i$ as,

$$\mathbf{D}_i = \max_j \sum_{l=1}^{L} P_{il} \ln \frac{P_{il}}{P_{jl}} \tag{6.3}$$

This is the error exponent of a constrained binary detection problem in which we are obliged to use the i^{th} input letter for the hypothesis with higher probability. Recall that D in the non-constrained case is given by,

$$\mathbf{D} = \max_i \mathbf{D}_i$$

Similar to the cost constrained capacity one can define a cost-constrained detection

[3] This is equivalent to stating that none of the capacity achieving distributions satisfy the cost constraint.

problem.[4]

$$\mathbf{D}(\mathcal{P}) = \max_{\substack{\phi \\ \sum_{k=1}^{K} \phi_k \rho_k \leq \mathcal{P}}} \sum_{k}^{K} \phi_k \mathbf{D}_k \tag{6.4}$$

The convexity concavity discussions are evident because of linearity, thus we can write

$$\begin{aligned}
\mathbf{D}(\mathcal{P}) &= \min_{\gamma \geq 0} \max_{\phi} \sum_k \phi_k \mathbf{D}_k + \gamma(\mathcal{P} - \sum_k \phi_k \rho_k) \\
\mathbf{D}(\mathcal{P}) &= \min_{\gamma \geq 0} \max_{k} (\mathbf{D}_k - \gamma \rho_k) + \gamma \mathcal{P}
\end{aligned}$$

Also for every value of cost constraint $\mathcal{P}$, there exist a $\gamma_{\mathcal{P}}$ such that

$$\mathbf{D}(\mathcal{P}) = \max_{k} (\mathbf{D}_k - \gamma_{\mathcal{P}} \rho_k) + \gamma_{\mathcal{P}} \mathcal{P} \tag{6.5}$$

Similar to the cost constrained capacity case, if $\mathcal{P}$ is a non-trivial constraint, i.e. $\mathbf{D}(\mathcal{P}) < \mathbf{D}$ then $\gamma_{\mathcal{P}} > 0$, else $\gamma_{\mathcal{P}}$ will be zero.

6.2 Cost Constraint For Variable Decoding Time Systems

In order to be able to adopt the variable decoding time nature of the system we will relax our cost constraint from a fixed total power constraint to an expected power constraint. Thus we will put restrictions on the expected energy that is used in terms of the expected time of the transmission. The expected energy needed for transmission of one message will be $\mathbf{E}\left[\mathcal{S}_\tau \middle| \mathcal{F}_0\right]$. So the cost constraint transmission will mean the set of transmissions satisfying $\mathbf{E}\left[\mathcal{S}_\tau \middle| \mathcal{F}_0\right] \leq \mathcal{P}\mathbf{E}\left[\tau \middle| \mathcal{F}_0\right]$. We write the expected energy spend as the expectation of the energy per use of the channel; this allows an uneven distribution of the expected energy and its time average between the different decoding times. Thus some of the possible decoding paths might have

[4] Although we are using the words 'capacity', or 'detection', indeed we are just defining minimization/maximization problems for probability mass functions with set constraints their operational meaning will be clear when we prove the relevant theorems for the corresponding problems.

an average power that might be very large or small when compared to the average constraint we had.

In order to measure the energy spent up to the decoding time, we will define a stochastic sequence, $\mathcal{S}_n$ which is the sum of energies of all the input symbols used until the decoding time. Evidently we need to know what the true message is in order to calculate the value of $\mathcal{S}_n$. Thus $\mathcal{S}_n$, is not a random variable that is measurable in $\mathcal{F}_n$, however it is a random variable that is measurable in $\mathcal{G}_n$. The constraint on $\mathcal{S}_n$ is in terms of the expected value of $\mathcal{S}_n$, $\mathbf{E}\left[\mathcal{S}_\tau\right]$, which is measurable in the filtration $\mathcal{F}$. Being more specific the expectation will average over the messages at each decoding point, with the weights being a posteriori probabilities, consequently resulting quantity will be measurable in $\mathcal{F}$. In short although the actual energy used is not a quantity that is known at the receiver, i.e., it is not measurable in the filtration $\mathcal{F}$, we are not interested in that since we are considering constraints on expected power. A more detailed explanation of this fact can be as follows.

Let the random variable $\mathcal{S}_n(i)$ be the cost for the codeword that corresponds to i^{th} message up to and including time n. Since we know the part of the codewords up to time n for each messages at the receiver, when we know $\mathfrak{f}_n$, we can calculate $\mathcal{S}_n(i)$ at the receiver at time n. Thus $\mathcal{S}_n(i)$'s are measurable in $\mathcal{F}_n$ and the expected cost at some $\mathfrak{f}_n \in \mathcal{F}_n$ can be written as

$$\mathbf{E}\left[\mathcal{S}_n \middle| \mathcal{F}_n = \mathfrak{f}_n\right] = \sum_{i}^{M} p_i(\mathfrak{f}_n)\mathcal{S}_n(i)$$

where $p_i(\mathfrak{f}_n)$ is the a posteriori probability of the i^{th} message defined previously. Since we can be at $\mathfrak{f}_n$ with different θ's, the expected energy given $\mathfrak{f}_n$ is an average over messages. Similarly we can define the total expected energy $\mathbf{E}\left[\mathcal{S}_\tau\right]$ where τ is the decoding time.

Let us calculate the expected value of $\mathcal{S}_{n+1}$ given $\mathfrak{f}_n$ Note that

$$\mathbf{E}\left[\mathcal{S}_{n+1} \middle| \mathfrak{f}_{n+1}\right] = \sum_{m=1}^{M} p_m(\mathfrak{f}_{n+1})\mathcal{S}_{n+1}(m)$$

$$\mathbf{E}\left[\mathcal{S}_{n+1}\middle|\,\mathfrak{f}_{n+1}\right] = \sum_m \mathcal{S}_{n+1}(m)\frac{p_m(\mathfrak{f}_n)p(l|m)}{p(l)}$$

$$\mathbf{E}\left[\mathcal{S}_{n+1}\middle|\,\mathfrak{f}_{n}\right] = \sum_l p(l)\left(\sum_m \mathcal{S}_{n+1}(m)\frac{p_m(\mathfrak{f}_n)p(l|m)}{p(l)}\right)$$

$$\mathbf{E}\left[\mathcal{S}_{n+1}\middle|\,\mathfrak{f}_{n}\right] = \sum_l \left(\sum_m \mathcal{S}_{n+1}(m)p_m(\mathfrak{f}_n)p(l|m)\right)$$

$$\mathbf{E}\left[\mathcal{S}_{n+1}\middle|\,\mathfrak{f}_{n}\right] = \sum_m p_m(\mathfrak{f}_n)\mathcal{S}_{n+1}(m)$$

$$\mathbf{E}\left[\mathcal{S}_{n+1}\middle|\,\mathfrak{f}_{n}\right] = \mathbf{E}\left[\mathcal{S}_{n}\middle|\,\mathfrak{f}_{n}\right] + \sum_m p_m(\mathfrak{f}_n)\sum_k w(k|m)\rho_k$$

Thus

$$\mathbf{E}\left[\mathcal{S}_{n+1}\middle|\,\mathcal{F}_{n}\right] = \mathbf{E}\left[\mathcal{S}_{n}\middle|\,\mathcal{F}_{n}\right] + \sum_m p_m(\mathfrak{f}_n)\sum_k w(k|m)\rho_k \tag{6.6}$$

Chapter 7

Preliminary Discussions For Cost Constrained Case

7.1 Basic Lemmas For Cost Constraint Case

We have established constraints on the 'average' change of entropy, but these constraints are blind to any cost that might be associated with the input letters. Now we will establish bounds on the expected change of entropy together with the expected change in the energy that has been used. Similar to the case without costs, we should interpret the conditional expectations as functions of the conditioned quantity, thus the following lemmas are valid for all realizations, $\mathfrak{f}_n$ of $\mathcal{F}_n$.

Let us define a stochastic sequence as follows

$$V_n^{\mathcal{P}} = H_n + \gamma_{\mathcal{P}} \mathbf{E}\left[\mathcal{S}_n \middle| \mathcal{F}_n\right] \tag{7.1}$$

Note that $V_n^{\mathcal{P}}$ is measurable in $\mathcal{F}_n$ and using equation (6.2), one can bound the expected decrease in $V_n^{\mathcal{P}}$ in one unit of time.

Lemma 8. *$\forall n \geq 0$, and $\forall \mathcal{P} \geq \rho_{min}$ we have the inequality,*

$$\mathbf{E}\left[V_n^{\mathcal{P}} - V_{n+1}^{\mathcal{P}} \middle| \mathcal{F}_n\right] \leq \mathbf{C}(\mathcal{P}) - \gamma_{\mathcal{P}} \mathcal{P} \tag{7.2}$$

where the cost constrained capacity $\mathbf{C}(\mathcal{P})$ *is given by*

$$\mathbf{C}(\mathcal{P}) = \max_{\substack{\phi \\ \sum_{k=1}^{K} \phi_k \rho_k \leq \mathcal{P}}} \sum_{k=1,l=1}^{K,L} \phi_k P_{kl} \ln \frac{P_{kl}}{\sum_{r=1}^{K} \phi_r P_{rl}}$$

Proof:

Using equation (B.1) which is proved in the course of proving Lemma 2,

$$\mathbf{E}\left[H_n - H_{n+1} \middle| \mathcal{F}_n = \mathfrak{f}_n\right] \leq \min_{\psi} \sum_{l=1}^{L} \sum_{i=1}^{M} \sum_{k=1}^{K} p_i(\mathfrak{f}_n) w(k|i) P_{kl} \ln \frac{P_{kl}}{\psi_l}$$

Together with Equation (6.6), this will lead to

$$\mathbf{E}\left[V_n^{\mathcal{P}} - V_{n+1}^{\mathcal{P}} \middle| \mathcal{F}_n = \mathfrak{f}_n\right] \leq \min_{\psi} \sum_{i} p_i(\mathfrak{f}_n) \left(\sum_{l} \sum_{k} w(k|i) P_{kl} \ln \frac{P_{kl}}{\psi_l} - \gamma_s \sum_{k} w(k|i) \rho_k \right)$$

$$\begin{aligned}
&= \min_{\psi} \sum_{i} \sum_{k} p_i(\mathfrak{f}_n) w(k|i) \left(\sum_{l} P_{kl} \ln \frac{P_{kl}}{\psi_l} - \gamma_s \rho_k \right) \\
&\leq \min_{\psi} \max_{k} \left(\sum_{l} P_{kl} \ln \frac{P_{kl}}{\psi_l} - \gamma_s \rho_k \right) \\
&= \mathbf{C}(\mathcal{P}) - \gamma_{\mathcal{P}} \mathcal{P}
\end{aligned}$$

where we have used equation (6.2) in the last equality.

QED

Let us define a stochastic sequence that will be a more accurate tool to handle small values of entropy as follows

$$W_n^{\mathcal{P}} = \ln H_n + \gamma_{\mathcal{P}} \mathbf{E}\left[\mathcal{S}_n \middle| \mathcal{F}_n\right] \tag{7.3}$$

Note that $W_n^{\mathcal{P}}$ is measurable in $\mathcal{F}_n$ and using the equation (6.5) one can bound the expected decrease in $W_n^{\mathcal{P}}$ in one unit of time.

Lemma 9. *$\forall n \geq 0$, and $\forall \mathcal{P} \geq \rho_{min}$ we have the inequality*

$$\mathbf{E}\left[W_n^{\mathcal{P}} - W_{n+1}^{\mathcal{P}}\middle|\, \mathcal{F}_n\right] \leq \mathbf{D}(\mathcal{P}) - \gamma_{\mathcal{P}}\mathcal{P} \tag{7.4}$$

where

$$\mathbf{D}(\mathcal{P}) = \max_{\substack{\phi \\ \sum_{k=1}^{K} \phi_k \rho_k \leq \mathcal{P}}} \sum_{k=1}^{K} \phi_k \mathbf{D}_k \qquad \textit{and} \qquad \mathbf{D}_k = \max_j \sum_{l=1}^{L} P_{kl} \ln \frac{P_{kl}}{P_{jl}}$$

Proof:

Using equation (3.9)

$$\begin{aligned}
\mathbf{E}\left[\ln H_n - \ln H_{n+1}\right| \mathcal{F}_n] &\leq \max_i \sum_{l=1}^{L} p(l) \ln \frac{p(l)}{p(l|i)} \\
&= \max_i \sum_l \left(\sum_{k,j} p_j(\mathfrak{f}_n) w(k|j) \right) P_{kl} \ln \frac{\sum_{k,j} p_j(\mathfrak{f}_n) w(k|j) P_{kl}}{p(l|i) \sum_{k,j} p_j(\mathfrak{f}_n) w(k|j)}
\end{aligned}$$

where $\sum_k w(k|j) = 1$, $\sum_j p_j(\mathfrak{f}_n) = 1$ thus $\sum_{k,j} p_j(\mathfrak{f}_n) w(k|j) = 1$.

Using the log sum inequality

$$\begin{aligned}
&\leq \max_i \sum_l \sum_{k,j} p_j(\mathfrak{f}_n) w(k|j) P_{kl} \ln \frac{p_j(\mathfrak{f}_n) w(k|j) P_{kl}}{p(l|i) p_j(\mathfrak{f}_n) w(k|j)} \\
&= \max_i \sum_l^{L} \sum_{k,j}^{K,M} p_j(\mathfrak{f}_n) w(k|j) P_{kl} \ln \frac{P_{kl}}{p(l|i)} \\
&= \sum_{k,j} p_j(\mathfrak{f}_n) w(k|j) \max_i \sum_l P_{kl} \ln \frac{P_{kl}}{p(l|i)}
\end{aligned}$$

Using the convexity of Kullback-Leibler divergence together with the fact

$p(l|i) = \sum_k w(k|i)P_{kl}$

$$\leq \sum_{k,j} p_j(\mathfrak{f}_n)w(k|j) \max_r \left(\sum_l P_{kl} \ln \frac{P_{kl}}{P_{rl}} \right)$$
$$= \sum_{k=1,j=1}^{K,M} p_j(\mathfrak{f}_n)w(k|j)\mathbf{D}_k$$

Using the definition of $W^{\mathcal{P}}{}_n$,together with equation (6.6) we get

$$\begin{aligned}\mathbf{E}\left[W_n^{\mathcal{P}} - W_{n+1}^{\mathcal{P}})\middle|\, \mathcal{F}_n = \mathfrak{f}_n\right] &\leq \sum_{k=1}^{K}\sum_{j=1}^{M} p_j(\mathfrak{f}_n)w(k|j)(\mathbf{D}_k - \gamma_{\mathcal{P}}\rho_k) \\ &\leq \max_k \mathbf{D}_k - \gamma_{\mathcal{P}}\rho_k \\ &= \mathbf{D}(\mathcal{P}) - \gamma_{\mathcal{P}}\mathcal{P}\end{aligned}$$

where we have used equation (6.5) in the last equality.

QED

7.2 Trivial Extensions of the case without cost constraint

Before starting the analysis of the case with cost constraint we will discuss some trivial cases and exclude them from the analysis for both converse and achievability.

Let us first note that if $\mathcal{P} \geq \rho_{max}$ then the cost constraint does not introduce any restriction on the coding algorithms or decoding rules that can be used, it does not effect the converse either. Thus this case will be equivalent to that there are no-cost constraint.

The other evident case is when $\mathcal{P} = \rho_{min}$, i.e. $\mathbf{E}[\mathcal{S}_\tau] \leq \rho_{min}\mathbf{E}[\tau]$. It is evident that no letter with $\rho_i > \rho_{min}$ can ever be used. Thus we are strictly restricted to the set of input letters whose costs are ρ_{min}. This means that we have an equivalent of the problem without cost constraint for a smaller set of input letters, i.e., using one or other element of the set does not cost us more or less energy. It is worth noting that this 'awkward' discontinuity of the characteristics at $\mathcal{P} = \rho_{min}$, is a result of the way we impose the cost constraint. If we had allowed a 'negligible' extra cost that vanishes 'asymptotically' we would reach a characteristics that is equal to the limits of the characteristics as $\mathcal{P}$ goes to zero.

Our last consideration will be on the costs, ρ_i, and the constraint $\mathcal{P}$. Our constraint is an additive constraint of the form $\mathbf{E}[\mathcal{S}_\tau] \leq \mathcal{P}\mathbf{E}[\tau]$. Thus following two constraints are equivalent

- $\mathbf{E}[\mathcal{S}_\tau] \leq \mathcal{P}\mathbf{E}[\tau]$
- $\mathbf{E}[\mathcal{S}_\tau - \rho_{min}\tau] \leq (\mathcal{P} - \rho_{min})\mathbf{E}[\tau]$,

In other words if we subtract ρ_{min} both from the letter costs, and constraint we get an equivalent problem. Because of this equivalence, we henceforth assume that $\rho_{min} = 0$.

Chapter 8

Converse With Cost Constraint

Theorem 5. *Consider any generalized block code with* $M > 2$*,* $P_e > 0$*, and* $\mathcal{P} > 0$*, for any DMC with feedback, such that* [1]$\mathbf{E}\left[\mathcal{S}_\tau\right] \leq \mathcal{P}\mathbf{E}\left[\tau\right]$ *the expected number of observations* $\mathbf{E}\left[\tau\right]$ *satisfies the inequality*

$$\mathbf{E}\left[\tau \middle| \mathcal{F}_0\right] \geq \min_{0 \leq < \mathcal{P}_A, \mathcal{P}_B \leq \rho_{max}} \max\{\mathcal{V}_1, \mathcal{V}_2\}$$

where

$$\mathcal{V}_1 = \frac{\ln M - P_e(\ln P_e + \ln M + 1)\ln M - 1}{\mathbf{C}(\mathcal{P}_A)} - \frac{\ln P_e + \mathbf{F} + \ln(\ln M - \ln P_e + 1)}{\mathbf{D}(\mathcal{P}_B)}$$
$$\mathcal{V}_2 = \frac{\mathcal{P}_A}{\mathcal{P}} \frac{\ln M - P_e(\ln P_e + \ln M + 1)\ln M - 1}{\mathbf{C}(\mathcal{P}_A)} - \frac{\mathcal{P}_B}{\mathcal{P}} \frac{\ln P_e + \mathbf{F} + \ln(\ln M - \ln P_e + 1)}{\mathbf{D}(\mathcal{P}_B)}$$

When proving converse results it is important to have constraints as weak as possible. Accordingly we have not even put a constraint on the expected energy at each decoding time, our constraint is on the expected amount of energy that is spent for one message. Being more specific, under this constraint it is possible to have decoding points, at which expected energy spent up until that time, is much higher then the product of power constraint and time.

Proof:

[1]Remember our reasoning that leads to the assumption $\rho_{min} = 0$.

Note that if $\mathbf{E}[\tau|\mathcal{F}_0] = \infty$ then the theorem holds trivially. Therefore we will assume $\mathbf{E}[\tau|\mathcal{F}_0] < \infty$.

Using the generalized Fano inequality, together with the Markov inequality we have already shown that, for any $\delta > 0$,

$$\mathbf{E}[H_\tau] \geq \delta \mathbf{P}[H_\tau > \delta] \qquad \mathbf{P}[H_\tau > \delta] \leq P_e \frac{\ln P_e + \ln M + 1}{\delta} \tag{8.1}$$

In filtration $\mathcal{F}$, define a stopping time t_δ by

$$t_\delta = \min\{n|H_n \leq \delta\}$$

Clearly $\tau_\delta = \tau \wedge t_\delta$ is also a stopping time, and $\tau_\delta \leq \tau$ for all realizations of $\mathcal{F}$.

We will start by lower bounding $\mathbf{E}[\tau_\delta]$. Let us find an upper bound on $\mathbf{E}[H_{\tau_\delta}|\mathcal{F}_0]$ first.

$$\begin{aligned}
\mathbf{E}[H_{\tau_\delta}|\mathcal{F}_0] &= \mathbf{E}\left[H_{t_\delta}\mathbb{I}_{\{\tau \geq t_\delta\}} + H_\tau \mathbb{I}_{\{\tau < t_\delta\}}\middle|\mathcal{F}_0\right] \\
&= \mathbf{E}\left[H_{t_\delta}\mathbb{I}_{\{\tau \geq t_\delta\}}\middle|\mathcal{F}_0\right] + \mathbf{E}\left[H_\tau \mathbb{I}_{\{\tau < t_\delta\}}\middle|\mathcal{F}_0\right] \\
&\leq \mathbf{E}\left[\delta \mathbb{I}_{\{\tau \geq t_\delta\}}\middle|\mathcal{F}_0\right] + \mathbf{E}\left[\ln M \mathbb{I}_{\{\tau < t_\delta\}}\middle|\mathcal{F}_0\right] \\
&= \delta \mathbf{P}[\tau \geq t_\delta|\mathcal{F}_0] + \ln M \mathbf{P}[\tau < t_\delta|\mathcal{F}_0] \\
&\leq \delta + \ln M \mathbf{P}[H_\tau > \delta|\mathcal{F}_0]
\end{aligned}$$

Using equation (8.1) and setting $\delta = 1$

$$\mathbf{E}[H_{\tau_1}|\mathcal{F}_0] \leq 1 + P_e(\ln P_e + \ln M + 1)\ln M \tag{8.2}$$

Consider the following stochastic sequence

$$\xi_n = H_n + \gamma_{\mathcal{P}_A}\mathbf{E}[\mathcal{S}_n|\mathcal{F}_n]$$

Note that as a result of Lemma 8

$$\mathbf{E}\left[\xi_n - \xi_{n+1} | \mathcal{F}_n\right] \le \mathbf{C}(\mathcal{P}_A) - \gamma_{\mathcal{P}_A}\mathcal{P}_A$$

Note that $|\xi_n| \le \gamma_{\mathcal{P}_A}\rho_{\max}n + \ln M$ and by our initial assumption $\mathbf{E}\left[\tau | \mathcal{F}_0\right] < \infty$. Thus we can apply the Lemma 5, part 3, for ξ_n,τ_1 and τ

$$(\mathbf{C}(\mathcal{P}_A) - \gamma_{\mathcal{P}_A}\mathcal{P}_A)\mathbf{E}\left[\tau_1 | \mathcal{F}_0\right] \ge \mathbf{E}\left[\ln M - (H_{\tau_1} + \gamma_{\mathcal{P}_A}\mathcal{S}_{\tau_1}) | \mathcal{F}_0\right]$$

$$\mathbf{C}(\mathcal{P}_A)\mathbf{E}\left[\tau_1 | \mathcal{F}_0\right] \ge \ln M - P_e(\ln P_e + \ln M + 1)\ln M - 1 + \gamma_{\mathcal{P}_A}\mathbf{E}\left[\mathcal{P}_A\tau_1 - \mathcal{S}_{\tau_1} | \mathcal{F}_0\right] \quad (8.3)$$

Now we will find a lower bound on $\mathbf{E}\left[\tau - \tau_1 | \mathcal{F}_0\right]$. As a result of Lemma 4 and the definition of τ_δ, $H_{\tau_1} \ge e^{-\mathbf{F}}$, so $\ln H_{\tau_1} \ge -\mathbf{F}$, and thus $\mathbf{E}\left[\ln H_{\tau_1} | \mathcal{F}_0\right] \ge -\mathbf{F}$.

Because of the generalized Fano inequality we have

$$\mathbf{E}\left[H_\tau | \mathcal{F}_0\right] \le P_e(\ln M - \ln P_e + 1)$$

Taking the logarithm of both sides we get

$$\ln \mathbf{E}\left[H_\tau | \mathcal{F}_0\right] \le \ln P_e + \ln(\ln M - \ln P_e + 1)$$

Using Jensen's inequality and the concavity of $\ln(\cdot)$ we get

$$\mathbf{E}\left[\ln H_\tau | \mathcal{F}_0\right] \le \ln \mathbf{E}\left[H_\tau | \mathcal{F}_0\right] \le \ln P_e + \ln(\ln M - \ln P_e + 1) \quad (8.4)$$

Consider the stochastic sequence, ν_n given by

$$\nu_n = \ln H_n + \gamma_{\mathcal{P}_B}\mathbf{E}\left[\mathcal{S}_n | \mathcal{F}_n\right]$$

As a result of Lemma 9,

$$\mathbf{E}\left[\nu_n - \nu_{n+1} | \mathcal{F}_n\right] \le \mathbf{D}(\mathcal{P}_B) - \gamma_{\mathcal{P}_B}\mathcal{P}_B$$

As a result of Lemma 4 $|\nu_n| < (\gamma_{\mathcal{P}_B}\rho_{max} + \mathbf{F})n + \ln\ln M$, also by assumption $\mathbf{E}\left[\tau|\,\mathcal{F}_0\right] < \infty$. Then using lemma 5, part 5, for ν_n, τ_1 and τ we get

$$(\mathbf{D}(\mathcal{P}_B) - \gamma_{\mathcal{P}_B}\mathcal{P}_B)\mathbf{E}\left[\tau - \tau_1|\,\mathcal{F}_0\right] \geq \mathbf{E}\left[(\ln H_{\tau_1} + \gamma_{\mathcal{P}_B}\mathcal{S}_{\tau_1}) - (\ln H_\tau - \gamma_{\mathcal{P}_B}\mathcal{S}_\tau)|\,\mathcal{F}_0\right]$$

$$\mathbf{D}(\mathcal{P}_B)\mathbf{E}\left[\tau - \tau_1|\,\mathcal{F}_0\right] \geq -\mathbf{F} - \ln P_e - \ln(\ln M - \ln P_e + 1) + \gamma_{\mathcal{P}_B}\mathbf{E}\left[\mathcal{P}_B(\tau - \tau_1) - (\mathcal{S}_\tau - \mathcal{S}_{\tau_1})|\,\mathcal{F}_0\right] \tag{8.5}$$

It is worth mentioning that equations (8.3) and (8.5) are both valid for any value of $\mathcal{P}_A \geq 0$ and $\mathcal{P}_B \geq 0$. The cost constraint is

$$\mathbf{E}\left[\mathcal{S}_\tau|\,\mathcal{F}_0\right] \leq \mathcal{P}\mathbf{E}\left[\tau|\,\mathcal{F}_0\right]$$

Noting evident facts

$$\mathbf{E}\left[\tau|\,\mathcal{F}_0\right] = \mathbf{E}\left[\tau_1|\,\mathcal{F}_0\right] + \mathbf{E}\left[\tau - \tau_1|\,\mathcal{F}_0\right] \qquad \mathbf{E}\left[\mathcal{S}_\tau|\,\mathcal{F}_0\right] = \mathbf{E}\left[\mathcal{S}_{\tau_1}|\,\mathcal{F}_0\right] + \mathbf{E}\left[\mathcal{S}_\tau - \mathcal{S}_{\tau_1}|\,\mathcal{F}_0\right]$$

There must exist a $(\mathcal{P}_A, \mathcal{P}_B)$, $0 \leq \mathcal{P}_A, \mathcal{P}_B \leq \rho_{max}$ satisfying the following three relations;

$$\mathbf{E}\left[\mathcal{S}_{\tau_1}|\,\mathcal{F}_0\right] \leq \mathcal{P}_A\mathbf{E}\left[\tau_1|\,\mathcal{F}_0\right]$$

$$\mathbf{E}\left[\mathcal{S}_\tau - \mathcal{S}_{\tau_1}|\,\mathcal{F}_0\right] \leq \mathcal{P}_B\mathbf{E}\left[\tau - \tau_1|\,\mathcal{F}_0\right]$$

$$\mathbf{E}\left[\mathcal{S}_\tau|\,\mathcal{F}_0\right] \leq \mathcal{P}_A\mathbf{E}\left[\tau_1|\,\mathcal{F}_0\right] + \mathcal{P}_B\mathbf{E}\left[\tau - \tau_1|\,\mathcal{F}_0\right] = \mathcal{P}\mathbf{E}\left[\tau|\,\mathcal{F}_0\right] \tag{8.6}$$

Inserting these in equation (8.3) and equation(8.5) we get

$$\mathbf{C}(\mathcal{P}_A)\mathbf{E}\left[\tau_1|\,\mathcal{F}_0\right] \geq \ln M - P_e(\ln P_e + \ln M + 1)\ln M - 1$$

$$\mathbf{D}(\mathcal{P}_B)\mathbf{E}\left[\tau - \tau_1|\,\mathcal{F}_0\right] \geq -\mathbf{F} - \ln P_e - \ln(\ln M - \ln P_e + 1)$$

Thus we know that $\mathbf{E}[\tau|\mathcal{F}_0]$ should satisfy,

$$\mathbf{E}[\tau|\mathcal{F}_0] \geq \frac{\ln M - P_e(\ln P_e + \ln M + 1)\ln M - 1}{\mathbf{C}(\mathcal{P}_A)} - \frac{\ln P_e + \mathbf{F} + \ln(\ln M - \ln P_e + 1)}{\mathbf{D}(\mathcal{P}_B)} \tag{8.7}$$

Also as a result of equation(8.6), we have

$$\mathbf{E}[\tau|\mathcal{F}_0] \geq \frac{\mathcal{P}_A}{\mathcal{P}} \frac{\ln M - P_e(\ln P_e + \ln M + 1)\ln M - 1}{\mathbf{C}(\mathcal{P}_A)} - \frac{\mathcal{P}_B}{\mathcal{P}} \frac{\ln P_e + \mathbf{F} + \ln(\ln M - \ln P_e + 1)}{\mathbf{D}(\mathcal{P}_B)} \tag{8.8}$$

We can conclude that there exist an $(\mathcal{P}_A, \mathcal{P}_B)$ pair such that equations (8.7) and (8.8) are both satisfied, which immediately leads to the assertion of the theorem.

QED

Note that as a result of theorem

$$1 \geq \min_{0 \leq \mathcal{P}_A, \mathcal{P}_B \in \leq \mathcal{P}} \max \left\{ \frac{\mathcal{V}_1}{\mathbf{E}[\tau|\mathcal{F}_0]}, \frac{\mathcal{V}_2}{\mathbf{E}[\tau|\mathcal{F}_0]} \right\}$$

Thus there exist a $(\mathcal{P}_A, \mathcal{P}_B)$ pair such that

$$1 \geq \frac{\mathcal{V}_1}{\mathbf{E}[\tau|\mathcal{F}_0]} \qquad \text{and} \qquad 1 \geq \frac{\mathcal{V}_2}{\mathbf{E}[\tau|\mathcal{F}_0]}$$

This is equivalent to saying that an $(\mathcal{P}_A, \mathcal{P}_B)$ pair exists such that

$$\frac{-\ln P_e - \mathbf{F} - \ln(\ln M - \ln P_e + 1)}{\mathbf{E}[\tau|\mathcal{F}_0]} \leq \mathbf{D}(\mathcal{P}_B) \left(1 - \frac{1}{\mathbf{C}(\mathcal{P}_A)} \frac{\ln M - P_e(\ln P_e + \ln M + 1)\ln M - 1}{\mathbf{E}[\tau|\mathcal{F}_0]}\right)$$
$$\frac{-\ln P_e - \mathbf{F} - \ln(\ln M - \ln P_e + 1)}{\mathbf{E}[\tau|\mathcal{F}_0]} \leq \frac{\mathbf{D}(\mathcal{P}_B)\mathcal{P}}{\mathcal{P}_B} \left(1 - \frac{\mathcal{P}_A}{\mathcal{P}\mathbf{C}(\mathcal{P}_A)} \frac{\ln M - P_e(\ln P_e + \ln M + 1)\ln M - 1}{\mathbf{E}[\tau|\mathcal{F}_0]}\right)$$

In order to have a sequence of coding/decoding schemes at a rate R, we need to have, $\liminf_{P_e \to 0} \frac{\ln M}{\mathbf{E}[\tau|\mathcal{F}_0]} = R$. Using this fact and little bit of algebra here and there we can conclude that

$$E(R) \leq \limsup_{P_e \to 0} \frac{-\ln P_e}{\mathbf{E}[\tau|\mathcal{F}_0]} \leq \max_{\mathcal{P}_A, \mathcal{P}_B} \min \left\{ \mathbf{D}(\mathcal{P}_B)\left(1 - \frac{R}{\mathbf{C}(\mathcal{P}_A)}\right), \mathbf{D}(\mathcal{P}_B)\frac{\mathcal{P}}{\mathcal{P}_B}\left(1 - \frac{\mathcal{P}_A}{\mathcal{P}} \frac{R}{\mathbf{C}(\mathcal{P}_A)}\right) \right\} \tag{8.9}$$

We will discuss some properties of the solution of this maximization problem after proving the achievability.

Chapter 9

Achievability With Cost Constraint

We will use a scheme based on the one proposed by Yamamoto and Ito in [17]. We will prove that for all $0 \leq \mathcal{P}_A, \mathcal{P}_B \leq \rho_{max}$ a coding algorithm exists satisfying the cost constraint $\mathbf{E}\left[\mathcal{S}_\tau | \mathcal{F}_0\right] \leq \mathcal{P}\mathbf{E}\left[\tau | \mathcal{F}_0\right]$ and having an expected decoding time upper bounded by a function whose main terms are same as the main terms of $\max\{\mathcal{V}_1, \mathcal{V}_2\}$. Thus the result of the maximization problem stated in the previous section is indeed asymptotically achievable. Using this we will be able to specify what the reliability function is for generalized block code under cost constraint with perfect feedback.

Our scheme might have an auxiliary waiting phase in which the transmitter will just send a zero cost input letter.[1] This phase will ensure that the average power constraint is satisfied for the given reference average power pair $(\mathcal{P}_A, \mathcal{P}_B)$.

Phase1:

Similar to the case without cost constraint, the coding theorem states that, $\forall \mathcal{P}_A \geq 0$, $\forall P_{e1} > 0$, and $\epsilon > 0$ and $\forall M$ there exists a codebook of length l_1,

$$l_1 \leq \frac{\ln M}{C(\mathcal{P}_A)(1-\epsilon)} + \mathbf{\Delta}_1(\mathcal{P}_A, P_{e1}, \epsilon) \tag{9.1}$$

whose codewords have costs $\mathcal{S}_1 \leq \mathcal{P}_A l_1$, where $\mathbf{\Delta}_1(P_{e1}, \epsilon, \mathcal{P}_A)$ is a function that is determined by the transition probabilities and $\mathcal{S}_1$ is the amount of energy needed for

[1] Remember our discussion about ρ_{min} and resulting assumption that $\rho_{min} = 0$.

any $\theta = [1, \dots, M]$. Thus we also have

$$\mathcal{S}_1 \leq \mathcal{P}_A \frac{\ln M}{C(\mathcal{P}_A)(1-\epsilon)} + \mathcal{P}_A \mathbf{\Delta}_1(\mathcal{P}_A, P_{e1}, \epsilon) \tag{9.2}$$

Similar to the case without cost constraint we have an estimate at the end of this phase.

Phase 2:

Let $\phi_{\mathcal{P}_B}$ be the input distribution that satisfies

$$\sum_{i=1}^{K} \phi_{\mathcal{P}_B}(i)\mathbf{D}_i = \mathbf{D}(\mathcal{P}_B) \qquad \sum_{i=1}^{K} \phi_{\mathcal{P}_B}(i)\rho_i \leq \mathcal{P}_B$$

Then we will define a subset of the input letters $\Xi_{\mathcal{P}_B}$ as

$$\Xi_{\mathcal{P}_B} = \{i | \phi_{\mathcal{P}_B}(i) > 0\}$$

and we will have $|\Xi_{\mathcal{P}_B}|$ sub detections. If all of these detections result in an acceptance then we will decide that the estimate we had as a result of the first phase is true, else we will ask for a retransmission. The $A \to R$ and $R \to A$ probabilities of the i^{th} input letter will be denoted by $P_{AR}(i)$, and $P_{RA}(i)$ respectively. Then the rule on overall acceptance and rejection will lead to

$$P_{AR} \leq \sum_{i \in \Xi_{\mathcal{P}_B}} P_{AR}(i) \qquad P_{RA} = \prod_{i \in \Xi_{\mathcal{P}_B}} P_{RA}(i)$$

For a given upper bounds for P_{AR} and P_{RA}, we will chose the corresponding upper bounds on $P_{AR}(i)$ and $P_{RA}(i)$ as follows,

$$P_{RA}(i) \leq P_{RA}^{\frac{\phi_{\mathcal{P}_B}(i)\mathbf{D}_i}{\mathbf{D}(\mathcal{P}_B)}} \qquad P_{AR}(i) \leq \frac{P_{AR}}{|\Xi_{\mathcal{P}_B}|} \tag{9.3}$$

Using lemma 7 we can find an upper bound on minimum duration of each sub detection problem for equation (9.3), as we did for the case without cost constraints. For each element $i \in \Xi_{\mathcal{P}_B}$ we will send the input letter i itself for acceptance and the letter $\kappa(i)$ for rejection where

$$\mathbf{D}_i = \max_j \mathbf{D}(P_{il}||P_{jl}) = \mathbf{D}(P_{il}||P_{\kappa(i)l})$$

$\forall i \in \Xi_{\mathcal{P}_B}$ using equation (5.15), $\forall P_{RA}(i) > 0$, $\forall P_{AR}(i) > 0$ we can upper bound $l_2(i)$, as follows $\forall \epsilon > 0$,

$$l_2(i) < \frac{-\ln P_{RA}(i)}{\mathbf{D}_i(1-\epsilon)} + \mathbf{\Delta}_2^i(P_{AR}(i), \epsilon)$$

Using the equation (9.3), $\forall \epsilon > 0$, $\forall P_{RA} > 0$, $\forall P_{AR} > 0, \forall \mathcal{P}_B \geq 0$

$$l_2(i) < \phi_{\mathcal{P}_B}(i) \frac{-\ln P_{RA}}{\mathbf{D}(\mathcal{P}_B)(1-\epsilon)} + \mathbf{\Delta}_2^i(P_{AR}, \epsilon)$$

Using $l_2 = \sum_{i \in \Xi_{\mathcal{P}_B}} l_2(i)$ we can calculate an upper bound on l_2 and energies required for acceptance and rejection for a (P_{RA}, P_{AR}) pair,

$$l_2 < \frac{-\ln P_{RA}}{\mathbf{D}(\mathcal{P}_B)(1-\epsilon)} + \mathbf{\Delta}_2(\mathcal{P}_B, P_{AR}, \epsilon) \qquad \forall \epsilon > 0, \forall P_{RA} > 0, \forall P_{AR} > 0, \forall \mathcal{P}_B \geq 0 \quad (9.4)$$

The acceptance and the rejection costs will then be given respectively by

$$\mathcal{S}_{2A} = \sum_{i \in \Xi_{\mathcal{P}_B}} l_2(i)\rho_i \leq -\mathcal{P}_B \frac{\ln P_{RA}}{\mathbf{D}(\mathcal{P}_B)(1-\epsilon)} + \mathbf{\Delta}_2^*(\mathcal{P}_B, P_{AR}, \epsilon)$$

$$\mathcal{S}_{2R} = \sum_{i \in \Xi_{\mathcal{P}_B}} l_2(i)\rho_{\kappa(i)} \leq \sum_{i \in \Xi_{\mathcal{P}_B}} l_2(i)\rho_{max} \leq \rho_{max} l_2$$

Similar to *Phase 1*, we can calculate the energy spent in this phase, but only as an expected value over the two cases about the estimate, i.e., the case where the estimate is true and the case where it is wrong.

$$\mathbf{E}[\mathcal{S}_2] = P_{e1}\mathcal{S}_{2R} + (1 - P_{e1})\mathcal{S}_{2A}$$

$$\mathbf{E}[\mathcal{S}_2] \leq \mathcal{P}_B \frac{\ln P_{RA}}{\mathbf{D}(\mathcal{P}_B)(1-\epsilon)} + \mathcal{P}_B \mathbf{\Delta}_2^*(\mathcal{P}_B, P_{AR}, \epsilon) + \rho_{max} l_2 P_{e1} \quad (9.5)$$

Considering the four possible outcomes of a trial and the rules governing the overall process we can write expressions for retransmission probability P_η and error probability P_e

$$P_\eta = P_{e1}(1-P_{RA}) + (1-P_{e1})P_{AR} \qquad P_e = \frac{P_{e1}P_{RA}}{1-P_\eta}$$

As in the case without cost constraints we will make substitutions for P_{e1} and P_{AR}

$$P_{e1} = \frac{S}{\rho_{max}}\frac{\delta}{2} \qquad P_{AR} = \frac{\delta}{2}$$

Assuming[2] also $\delta \leq 2/3$

$$P_\eta \leq \delta \qquad P_e \leq \frac{\delta}{2(1-\delta)}P_{RA} \leq P_{RA}$$

Then we can argue that there exist a function $\boldsymbol{\Delta}(\mathcal{P}, \mathcal{P}_A, \mathcal{P}_B, \delta, \epsilon)$ determined by transition probabilities and costs, such that following relations hold for ant $S > 0$.

$$l_1 + l_2 \leq \left(\frac{\ln M}{C(\mathcal{P}_A)} - \frac{\ln P_e}{\mathbf{D}(\mathcal{P}_B)}\right)\frac{1}{1-\epsilon} + \boldsymbol{\Delta}(\mathcal{P}, \mathcal{P}_A, \mathcal{P}_B, \delta, \epsilon) \tag{9.6}$$

$$\mathbf{E}\left[\mathcal{S}_1 + \mathcal{S}_2\right] \leq \left(\mathcal{P}_A\frac{\ln M}{\mathbf{C}(\mathcal{P}_A)} + \mathcal{P}_B\frac{\ln P_e}{\mathbf{D}(\mathcal{P}_B)}\right)\frac{1}{1-\epsilon} + \mathcal{P}\boldsymbol{\Delta}(\mathcal{P}, \mathcal{P}_A, \mathcal{P}_B, \delta, \epsilon) + \mathcal{P}l_2\frac{\delta}{2} \tag{9.7}$$

Phase 3

We will let l_3 be

$$l_3 = \left(\max\left\{\frac{\ln M}{C(\mathcal{P}_A)} - \frac{\ln P_e}{\mathbf{D}(\mathcal{P}_B)}, \mathcal{P}_A\frac{\ln M}{\mathbf{C}(\mathcal{P}_A)} + \mathcal{P}_B\frac{\ln P_e}{\mathbf{D}(\mathcal{P}_B)}\right\} + \boldsymbol{\Delta}\right)(1+\delta/2) - l_1 - l_2$$

Note that because of equation (9.6) and equation (9.7), the condition on l_3, being positive, is satisfied. Thus

$$l_0 = \max\left\{\frac{\ln M}{C(\mathcal{P}_A)} - \frac{\ln P_e}{\mathbf{D}(\mathcal{P}_B)}, \mathcal{P}_A\frac{\ln M}{\mathbf{C}(\mathcal{P}_A)} + \mathcal{P}_B\frac{\ln P_e}{\mathbf{D}(\mathcal{P}_B)}\right\}(1+\delta/2) + \boldsymbol{\Delta}(1+\delta/2)$$

$$\mathcal{S}_0 \leq \mathcal{P}l_0$$

[2]Remember the discussion about $\mathcal{P}$, i.e., if $\mathcal{P} \geq \rho_{max}$ we have a trivial cost constraint which is equivalent to having none. Thus we investigate the case for $\mathcal{P} < \rho_{max}$. Also it can be seen easily by a little algebra that the theorem holds for $\mathcal{P} \geq \rho_{max}$.

Note that

$$\mathbf{E}[\tau] = \frac{l_0}{1-P_\eta} \leq \frac{l_0}{1-\delta} \qquad \mathbf{E}[\mathcal{S}_\tau] = \frac{\mathcal{S}_0}{1-P_\eta}$$

Thus $\forall \mathcal{P} > 0$, $\forall \epsilon > 0$ and $\forall \delta > 0$ there exists a coding algorithm which satisfies $\mathbf{E}[\mathcal{S}_\tau] \leq \mathcal{P}\mathbf{E}[\tau]$ such that

$$\mathbf{E}[\tau] \leq \frac{1+\delta/2}{(1-\delta)(1-\epsilon)} \max\left\{\left(\frac{\ln M}{C(\mathcal{P}_A)} - \frac{\ln P_e}{\mathbf{D}(\mathcal{P}_B)}\right), \left(\frac{\mathcal{P}_A}{\mathcal{P}}\frac{\ln M}{\mathbf{C}(\mathcal{P}_A)} + \frac{\mathcal{P}_B}{\mathcal{P}}\frac{\ln P_e}{\mathbf{D}(\mathcal{P}_B)}\right)\right\} + \mathbf{\Delta}(1+\delta/2)$$

where $\mathbf{\Delta} = \mathbf{\Delta}(\mathcal{P}, \mathcal{P}_A, \mathcal{P}_B, \delta, \epsilon)$ is a function determined by transition probabilities and input letter costs.

Going through similar analysis as we did for case with out cost constraint we can easily extend this discussion to the finite delay feedback systems, with restricted feedback channel capacity. As a result we get the following theorem

Theorem 6. *For any DMC with cost constraints and an error free feedback channel of rate* $\mathbf{C}$ *or higher of finite delay* $\mathbf{T}$, $\forall S > 0$, $\forall \delta > 0$, *and for all* $(\mathcal{P}_A, \mathcal{P}_B)$ *such that* $0 \leq \mathcal{P}_A, \mathcal{P}_B \leq \rho_{max}$, *there exist a coding scheme such that for all* $M > 1$ *and* $\forall P_e > 0$ *the expected decoding time* $\mathbf{E}[\tau]$ *is upper bounded as follows*

$$\mathbf{E}[\tau] \leq \frac{1}{1-\delta} \max\left\{\left(\frac{\ln M}{C(\mathcal{P}_A)} - \frac{\ln P_e}{\mathbf{D}(\mathcal{P}_B)}\right), \left(\frac{\mathcal{P}_A}{\mathcal{P}}\frac{\ln M}{\mathbf{C}(\mathcal{P}_A)} + \frac{\mathcal{P}_B}{\mathcal{P}}\frac{\ln P_e}{\mathbf{D}(\mathcal{P}_B)}\right)\right\} + \mathbf{\Delta}(\mathcal{P}, \mathcal{P}_A\mathcal{P}_B, \mathbf{T}, \delta) \tag{9.8}$$

where $\mathbf{\Delta}$ *is a function determined by transition probabilities and costs of input letters.*

This result together with the converse and previous discussions about $\mathcal{P}$ give us the expression for the reliability function for any $\mathcal{P} > 0$ as follows:

$$E(R) = \max_{\mathcal{P}_A, \mathcal{P}_B} \min\{\mathcal{W}_1(\mathcal{P}_A, \mathcal{P}_B)\mathcal{W}_2(\mathcal{P}_A, \mathcal{P}_B)\} \tag{9.9}$$

where

$$\mathcal{W}_1(\mathcal{P}_A, \mathcal{P}_B) = \mathbf{D}(\mathcal{P}_B)\left(1 - \frac{R}{\mathbf{C}(\mathcal{P}_A)}\right)$$
$$\mathcal{W}_2(\mathcal{P}_A, \mathcal{P}_B) = \mathbf{D}(\mathcal{P}_B)\frac{\mathcal{P}}{\mathcal{P}_B}\left(1 - \frac{\mathcal{P}_A}{\mathcal{P}}\frac{R}{\mathbf{C}(\mathcal{P}_A)}\right)$$

Even without knowing anything specific about the channel or the constraint we can describe the region where the reliability function lies, and some other properties as follows.

If we look at the operating point for an even distribution of power which is equal to the average power constraint, i.e., $(\mathcal{P}_A, \mathcal{P}_B) = (\mathcal{P}, \mathcal{P})$, we get a lower bound on the reliability function.

$$\begin{aligned} E(R) &= \max_{\mathcal{P}_A, \mathcal{P}_B} \min \{\mathcal{W}_1(\mathcal{P}_A, \mathcal{P}_B) \mathcal{W}_2(\mathcal{P}_A, \mathcal{P}_B)\} \\ &\geq \min \{\mathcal{W}_1(\mathcal{P}, \mathcal{P}) \mathcal{W}_2(\mathcal{P}, \mathcal{P})\} \\ &= \mathbf{D}(\mathcal{P}) \left(1 - \frac{R}{\mathbf{C}(\mathcal{P})}\right) \end{aligned}$$

Similarly the case without cost constraints is an upper bound, which can also be seen algebraically as follows

$$\begin{aligned} E(R) &= \max_{\mathcal{P}_A, \mathcal{P}_B} \min \{\mathcal{W}_1(\mathcal{P}_A, \mathcal{P}_B) \mathcal{W}_2(\mathcal{P}_A, \mathcal{P}_B)\} \\ &\leq \max_{\mathcal{P}_A, \mathcal{P}_B} \mathcal{W}_1(\mathcal{P}_A, \mathcal{P}_B) \\ &\leq \mathcal{W}_1(\rho_{max}, \rho_{max}) \\ &= \mathbf{D} \left(1 - \frac{R}{\mathbf{C}}\right) \end{aligned}$$

We know that $P_e > 0$ if $R > \mathbf{C}$. Similar fact will follow the converse theorem that will be proved int the next section for the cost constraint case, i.e., $P_e > 0$ if $R > \mathbf{C}(\mathcal{P})$. Thus

$$E(R) = 0 \quad \forall R > \mathbf{C}(\mathcal{P})$$

Note that

$$\mathcal{W}_1(\mathcal{P}_A, \mathcal{P}_B) = \mathbf{D}(\mathcal{P}_B) \qquad \mathcal{W}_2(\mathcal{P}_A, \mathcal{P}_B) = \mathbf{D}(\mathcal{P}_B) \frac{\mathcal{P}}{\mathcal{P}_B}$$

Using the concavity and positivity of $\mathbf{D}(\mathcal{P})$ function we can conclude that

$$E(R) \leq \mathbf{D}(\mathcal{P})$$

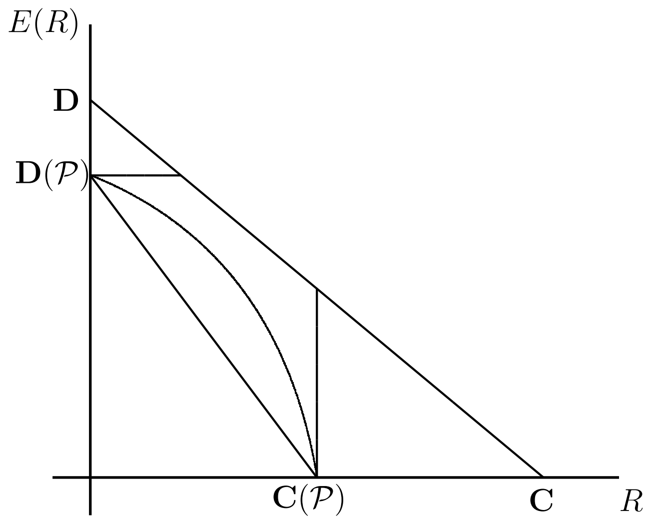

Figure 9-1: Reliability Function Of a DMC with cost constraint.

Thus the reliability function should lie in the region that whose boundaries are given by four identities we get. The typical shape of the region together with the reliability function itself is as follows.

Chapter 10

Zero Error Capacity

Our last topic is to consider the case when one or more transition probabilities are zero, i.e., one or more $\mathbf{D}_i$'s are infinity. First we will show that the cost constraint capacity with perfect feedback, is equal to the cost constraint capacity without feedback. This will also imply that zero error capacity is at most $\mathbf{C}(\mathcal{P})$. Then we will propose a method based on the Yamamoto Itoh scheme that reaches the cost-constrained capacity, $\mathbf{C}(\mathcal{P})$, with zero error probability, for all $\mathcal{P} > 0$. The case $\mathcal{P} = 0$ is investigated later.

Theorem 7. *For any DMC channel with feedback, under the cost constraint $\mathcal{P}$, for any $P_e \geq 0$, the expected value of decoding time will satisfy*

$$\mathbf{E}\left[\tau \middle| \mathcal{F}_0\right] \geq \frac{\ln M - \mathfrak{h}(P_e) - P_e \ln(M-1)}{C(\mathcal{P})}$$

Note that this bound on expected decoding time is valid for any DMC. But it is not as tight as Theorem 5 which is valid only for channels without zero transition probabilities.

Proof:

Consider the the stochastic process

$$\xi_n = H_n + \mathcal{S}_n + n(\mathbf{C}(\mathcal{P}) - \gamma_{\mathcal{P}}\mathcal{P})$$

We have already shown that $|\xi_n| < Kn$ and ξ_n is a submartingale. Using lemma 5, part 4, for ξ_n and τ we get

$$(\mathbf{C}(\mathcal{P}) - \gamma_{\mathcal{P}}\mathcal{P})\mathbf{E}\left[\tau|\,\mathcal{F}_0\right] \geq \mathbf{E}\left[\xi_0 - \xi_\tau|\,\mathcal{F}_0\right]$$

$$\mathbf{C}(\mathcal{P})\mathbf{E}\left[\tau|\,\mathcal{F}_0\right] \geq \mathbf{E}\left[\ln M - H_\tau|\,\mathcal{F}_0\right] + \gamma_{\mathcal{P}}\mathbf{E}\left[S\tau - \mathcal{S}_\tau|\,\mathcal{F}_0\right]$$

Considering the condition $\mathbf{E}\left[S\tau|\,\mathcal{F}_0\right] \geq \mathbf{E}\left[\mathcal{S}_\tau|\,\mathcal{F}_0\right]$ this leads to

$$\mathbf{E}\left[\tau|\,\mathcal{F}_0\right] \geq \frac{\mathbf{E}\left[\ln M - H_\tau|\,\mathcal{F}_0\right]}{\mathbf{C}(\mathcal{P})}$$

Using the generalized Fano inequality we get the required relation.

QED

As a result of the coding theorem without feedback, for any cost constraint $\mathcal{P} \geq 0$, $\forall P_{e1} > 0$ and $\epsilon > 0$ there exists a block code of length l_1,

$$l_1 \leq \frac{\ln M}{C(\mathcal{P})(1-\epsilon)} + \mathbf{\Delta}(\mathcal{P}, P_{e1}, \epsilon)$$

whose code words satisfy $\mathcal{S}_1 \leq \mathcal{P} l_1$.

In the second phase we we will either accept or reject the estimate of the receiver, for a duration of l_2. Let $X = r$ be the input letter which has a zero transition probability, i.e. say $P_{ru} = 0$ and let $X = a$ be an input letter whose transition probability to u is non zero $P_{au} = (1-q) \neq 0$. We will use $X = r$ for rejection and $X = a$ for acceptance. If we observe a $Y = u$ we would decide that the estimate is true else will try once more. This will guarantee that $P_{RA} = 0$, thus $P_e = 0$ also. The expression for retransmission probability, P_η will be

$$P_\eta = P_{e1} + (1 - P_{e1})q^{l_2}$$

If we let $P_{e1} = \frac{\delta}{2}$ and $q^{l_2} = \frac{\delta}{2}$ then $l_2 = \frac{\ln(\delta/2)}{\ln q}$ then $P_\eta < \delta$

$$l_1 + l_2 \leq \frac{\ln M}{C(\mathcal{P})(1-\epsilon)} + \mathbf{\Delta}(\mathcal{P}, \delta/2, \epsilon) + \frac{\ln(\delta/2)}{\ln q}$$

$$\mathcal{S}_1 + \mathcal{S}_2 \leq \mathcal{P}\frac{\ln M}{C(\mathcal{P})(1-\epsilon)} + \mathcal{P}\mathbf{\Delta}(\mathcal{P}, \delta/2, \epsilon) + \rho_{max}\frac{\ln(\delta/2)}{\ln q}$$

Then we will have a *phase 3*, of duration

$$l_3 = \frac{\rho_{max}}{S}\frac{\ln(\delta/2)}{\ln q} - l_2$$

It is evident at the end of *phase 3*, the energy spend, $\mathcal{S}_0 = \mathcal{S}_1 + \mathcal{S}_2 + \mathcal{S}_3$, satisfies the average power constraint for all messages and realization, i.e.,

$$\mathcal{S}_0 \leq \mathcal{P} l_0$$

Also l_0 is bounded as follows

$$l_0 \leq \frac{\ln M}{\mathbf{C}(\mathcal{P})(1-\epsilon)} + \mathbf{\Delta}(\mathcal{P}, \epsilon)$$

Using the fact

$$\mathbf{E}[\tau] = \frac{l_0}{1-P_\eta} \qquad \text{and} \qquad \mathbf{E}[\mathcal{S}_\tau] = \frac{\mathcal{S}_0}{1-P_\eta}$$

Thus there exist a coding algorithm that satisfies the cost constraint such that

$$\mathbf{E}[\tau] \leq \frac{1}{1-P_\eta}\frac{\ln M}{\mathbf{C}(\mathcal{P})(1-\epsilon)} + \mathbf{\Delta}(\mathcal{P}, \epsilon, \delta)$$

Theorem 8. *For any DMC with cost constraints and an error free feedback channel of rate* $\mathbf{C}$ *or higher of finite delay* $\mathbf{T}$*, if there exists a vanishing transition probability* $\forall \mathcal{P} > 0$,$\forall \delta > 0$,*there exists a coding algorithm such that*

$$\mathbf{E}[\tau] = \frac{1}{1-\delta}\frac{\ln M}{\mathbf{C}(\mathcal{P})} + \mathbf{\Delta}(\mathcal{P}, \mathbf{T}, \delta) \qquad \textit{and} \qquad \mathbf{E}[\mathcal{S}_\tau] \leq \mathcal{P}\mathbf{E}[\tau]$$

with zero-error probability.

Indeed this together with the converse says that, for any $\mathcal{P} > 0$ zero-error capacity under cost constraint, $\mathbf{E}\left[\mathcal{S}_\tau\right] \leq \mathcal{P}\mathbf{E}\left[\tau\right]$, is equal to cost constraint capacity $\mathbf{C}(S)$, if there exist on zero transition probability.

For $\mathcal{P} = 0$ case if there is only one letter with $\rho_i = 0$, then $\mathbf{C}(0) = 0$. Thus zero-error capacity is also zero under cost constraint $\mathcal{P} = 0$ is also zero.

If there exist more than one input letters the $\mathbf{C}(0) > 0$, but we are restricted to set of input letters which has zero cost. If this restricted set has a zero transition probability[1] then zero error capacity will be equal to $\mathbf{C}(0)$. Indeed as mentioned previously if we allow a 'vanishing' relaxation in the cost constraint, we will not have this 'discontinuity' of the problem at $\mathcal{P} = 0$. Zero error capacity at zero cost will be just cost constraint capacity at zero cost, $\mathbf{C}(0)$.

[1]Here we are considering a restricted set of input letters, i.e., input letters which has zero cost, and a restricted set of output letters, output letters that has a non-zero transition probability from at least one of the element's of the restricted input letters.

Chapter 11

Conclusion

We have considered generalized block codes on a DMC with perfect feedback. Since the decoding time is not fixed we can not use conventional definitions of rate and error exponent. Instead we have used the following by replacing the block length with its expectation

$$R = \frac{\ln M}{\mathbf{E}[\tau]} \qquad E(R) = -\frac{\ln P_e}{\mathbf{E}[\tau]}$$

For a DMC that does not have zero transition probabilities Burnashev,[2], showed that the reliability function is just a straight line, of the form

$$E(R) = \mathbf{D}\left(1 - \frac{R}{\mathbf{C}}\right)$$

for the case with out cost constraint.

We have generalized his results to the cost constrained case. As it is done for rate and error exponent, a conventional additive cost constraint is relaxed in a certain sense, to be compatible with the variable nature of decoding time. In order to have a more comprehensive set for admissible coding/decoding algorithms, instead of imposing a constraint on expected energy for all decoding instances, we only impose a constraint on the expected energy of one message transmission, in terms of expected decoding time. i.e.,

$$\mathbf{E}[\mathcal{S}_\tau] = \mathcal{P}\mathbf{E}[\tau]$$

where $\mathbf{E}[\mathcal{S}_\tau]$ is the expected value of the energy spent in one trial.[1] This definition/approach will allow time an uneven average power distribution on different decoding points. We have shown that the reliability function of generalized block coding schemes, $\forall \mathcal{P} > 0$ under cost constraint, $\mathbf{E}[\mathcal{S}_\tau] = \mathcal{P}\mathbf{E}[\tau]$ on a DMC with perfect feedback is given by;

$$E(R) = \max_{\mathcal{P}_A, \mathcal{P}_B} \min \left\{ \mathbf{D}(\mathcal{P}_B) \left(1 - \frac{R}{\mathbf{C}(\mathcal{P}_A)}\right), \mathbf{D}(\mathcal{P}_B) \frac{\mathcal{P}}{\mathcal{P}_B} \left(1 - \frac{\mathcal{P}_A}{\mathcal{P}} \frac{R}{\mathbf{C}(\mathcal{P}_A)}\right) \right\} \quad (11.1)$$

Also it is shown that with the coding scheme proposed by Yamamoto and Itoh, [17], this reliability function is reached. We will use a code reaching cost-constraint capacity $\mathbf{C}(\mathcal{P}_A^*)$ and a simple sequence of binary signallings where $\mathcal{P}_A^*$ is the optimal value of $\mathcal{P}_A$ for the maximization problem given.

For a DMC that has zero transition probabilities Burnashev showed that zero error capacity, with generalized block coding, is equal to the channel capacity without feedback, $\mathbf{C}$. We extended this result to the cost constraint case as follows. $\forall \mathcal{P} > 0$ under the cost constraint $\mathbf{E}[\mathcal{S}_\tau] = \mathcal{P}\mathbf{E}[\tau]$, zero-error capacity with generalized block codes is equal to the cost constraint capacity without feedback, $\mathbf{C}(\mathcal{P})$.

Finally for the $\mathcal{P} = 0$ case we conclude that both the zero-error rate problem and the error exponent problem are equivalent to the corresponding problems without cost constraint for a restricted set of input letters, namely the set of input letters having a zero cost. This discontinuity of the problem is a result of the stringency of the form of the cost constraint, i.e., $\mathbf{E}[\mathcal{S}_\tau] \leq \mathcal{P}\mathbf{E}[\tau]$. If we allow a 'vanishing' additional cost than our result for $\mathcal{P} = 0$, will just be the limits of the results for $\mathcal{P} > 0$ case, as expected.

[1] With the assumption $\rho_{min} = 0$, validity of which has already been discussed.

Appendix A

Proof of Lemma 5

Proof:

1. *$(\nu_n, \mathcal{F}_n)$ is a submartingale and $|\nu_n| < K'n \quad \forall n$:*

   Because of equation(3.11), $|\nu_n| \leq K'n$ for $K' = K + |R|$, so $\mathbf{E}[\nu_n] < \infty \quad \forall n$. Because of equation(3.12) $\mathbf{E}[\nu_{n+1} | \mathcal{F}_n] \geq \nu_n$; thus $(\nu_n, \mathcal{F}_n)$ is a submartingale.

2. *$\xi_n = \nu_n - \nu_{n \wedge \tau_i}$ is a submartingale and $|\xi_n| < K''n \quad \forall n$*

   Note that one can write $\nu_{n \wedge \tau_i}$ as

   $$\nu_{n \wedge \tau_i} = \nu_{\tau_i} \mathbb{I}_{\{\tau_i \leq n\}} + \nu_n \mathbb{I}_{\{\tau_i > n\}}$$

   Since τ_i is a stopping time with respect to the filtration $\mathcal{F}$, $\mathbb{I}_{\{\tau_i \leq n\}}$ is a measurable random variable in $\mathcal{F}_n$. Thus the random variables $\nu_{n \wedge \tau_i}$ and ξ_n are also measurable in $\mathcal{F}_n$.

   $$\begin{aligned} \xi_n &= \nu_n - \nu_{\tau_i} \mathbb{I}_{\{\tau_i \leq n\}} - \nu_n \mathbb{I}_{\{\tau_i > n\}} \\ &= (\nu_n - \nu_{\tau_i}) \mathbb{I}_{\{\tau_i \leq n\}} \end{aligned}$$

Now we can write the expression for the expected value of ξ_{n+1}

$$\begin{aligned}\mathbf{E}\left[\xi_{n+1}\middle|\mathcal{F}_n\right] &= \mathbf{E}\left[(\nu_{n+1}-\nu_{\tau_i})\mathbb{I}_{\{\tau_i\le n+1\}}\middle|\mathcal{F}_n\right]\\ &= \mathbf{E}\left[(\nu_{n+1}-\nu_{\tau_i})(\mathbb{I}_{\{\tau_i\le n\}}+\mathbb{I}_{\{\tau_i=n+1\}})\middle|\mathcal{F}_n\right]\end{aligned}$$

If we add and subtract ξ_n within the expectation,

$$\begin{aligned}\mathbf{E}\left[\xi_{n+1}\middle|\mathcal{F}_n\right] &= \mathbf{E}\left[\xi_n-(\nu_n-\nu_{\tau_i})\mathbb{I}_{\{\tau_i\le n\}}+(\nu_{n+1}-\nu_{\tau_i})(\mathbb{I}_{\{\tau_i\le n\}}+\mathbb{I}_{\{\tau_i=n+1\}})\middle|\mathcal{F}_n\right]\\ &= \xi_n+\mathbf{E}\left[(\nu_{n+1}-\nu_n)\mathbb{I}_{\{\tau_i\le n\}}\middle|\mathcal{F}_n\right]+\mathbf{E}\left[(\nu_{n+1}-\nu_{\tau_i})\mathbb{I}_{\{\tau_i=n+1\}}\middle|\mathcal{F}_n\right]\\ &= \xi_n+\mathbf{E}\left[\nu_{n+1}-\nu_n\middle|\mathcal{F}_n\right]\mathbb{I}_{\{\tau_i\le n\}}+\mathbf{E}\left[(\nu_{n+1}-\nu_{n+1})\mathbb{I}_{\{\tau_i=n+1\}}\middle|\mathcal{F}_n\right]\end{aligned}$$

Using submartingale property on ν_n and using the fact that the last term is 0.

$$\mathbf{E}\left[\xi_{n+1}\middle|\mathcal{F}_n\right]\ge\xi_n$$

It is evident that $|\nu_n|<K'n$ implies, $|\xi_n|<K''n$ for some K''. Thus $\mathbf{E}\left[\xi_n\right]<\infty$ and $(\xi_n,\mathcal{F}_n)$ is a submartingale.

3. $R\mathbf{E}\left[\tau_i\middle|\mathcal{F}_0\right]\ge\mathbf{E}\left[\Gamma_0-\Gamma_{\tau_i}\middle|\mathcal{F}_0\right]$:

   Note that since $\mathbf{E}\left[\tau_i\right]<\infty$ and $|\nu_n|\le K'n\quad\forall n$, the conditions of theorem 6 in [6] p250, will hold for ν_n and τ_i. Thus

$$\begin{aligned}\mathbf{E}\left[\nu_{\tau_i}\middle|\mathcal{F}_0\right]&\ge\mathbf{E}\left[\nu_0\middle|\mathcal{F}_0\right]\\ \mathbf{E}\left[R\tau+\Gamma_{\tau_i}\middle|\mathcal{F}_0\right]&\ge\nu_0\\ R\mathbf{E}\left[\tau_i\middle|\mathcal{F}_0\right]&\ge\mathbf{E}\left[\Gamma_0-\Gamma_{\tau_i}\middle|\mathcal{F}_0\right]\end{aligned}$$

4. $R\mathbf{E}\left[\tau_f\middle|\mathcal{F}_0\right]\ge\mathbf{E}\left[\Gamma_0-\Gamma_{\tau_f}\middle|\mathcal{F}_0\right]$:

   This is the same as part 3 replacing τ_i with τ_f.

5. $R\mathbf{E}\left[\tau_f-\tau_i\middle|\mathcal{F}_0\right]\ge\mathbf{E}\left[\Gamma_{\tau_i}-\Gamma_{\tau_f}\middle|\mathcal{F}_0\right]$:

Note that since $\mathbf{E}[\tau_f] < \infty$ and $|\xi_n| \le K''n \quad \forall n$, the conditions of theorem 6 in [6] p250, will hold for ξ_n and τ_f. Thus

$$\mathbf{E}\left[\xi_{\tau_f} \middle| \mathcal{F}_0\right] \ge \mathbf{E}\left[\xi_0 \middle| \mathcal{F}_0\right]$$

Using the definition of ξ_n

$$\mathbf{E}\left[\nu_{\tau_f} - \nu_{\tau_f \wedge \tau_i} \middle| \mathcal{F}_0\right] \ge 0$$

Using the fact $\tau_f \ge \tau_i$

$$\mathbf{E}\left[\nu_{\tau_f} - \nu_{\tau_i} \middle| \mathcal{F}_0\right] \ge 0$$

As a result of definition of ν_n

$$\mathbf{E}\left[(R\tau_f + \Gamma_{\tau_f}) - (R\tau_i + \Gamma_{\tau_i}) \middle| \mathcal{F}_0\right] \ge 0$$

$$R\mathbf{E}\left[\tau_f - \tau_i \middle| \mathcal{F}_0\right] \ge \mathbf{E}\left[\Gamma_{\tau_i} - \Gamma_{\tau_f} \middle| \mathcal{F}_0\right]$$

QED

Appendix B

Alternative Proof Of Lemma 2

Proof:

We will first introduce a short hand that will also be used in the next proof.

$$f_i = p_i(\mathfrak{f}_n) \qquad f_i(l) = \mathbf{P}\left[\theta = i \middle| Y_{n+1} = l, \mathcal{F}_n = \mathfrak{f}_n\right]$$
$$w(k|i) = \mathbf{P}\left[X_{n+1} = k \middle| \mathcal{F}_n = \mathfrak{f}_n, \theta = i\right] \qquad p(l|i) = \mathbf{P}\left[Y_{n+1} = l \middle| \theta = i, \mathcal{F}_n = \mathfrak{f}_n\right]$$

Using the probabilistic relation of channel input and channel output

$$p(l|i) = \sum_{k=1}^{K} w(k|i) P_{kl} \qquad p(l) = \sum_{i=1}^{M} p_i(\mathfrak{f}_n) p(l|i) \qquad f_i p(l|i) = p(l) f_i(l)$$

Using the shorthand and above relations

$$\begin{aligned}
\mathbf{E}\left[H_n - H_{n+1} \middle| \mathcal{F}_n = \mathfrak{f}_n\right] &= \mathbf{E}\left[H_n \middle| \mathcal{F}_n = \mathfrak{f}_n\right] - \mathbf{E}\left[H_{n+1} \middle| \mathcal{F}_n = \mathfrak{f}_n\right] \\
&= H_n - \mathbf{E}\left[H_{n+1} \middle| \mathcal{F}_n = \mathfrak{f}_n\right] \\
&= -\sum_{i=1}^{M} f_i \ln f_i + \sum_{l=1}^{L} p(l) \sum_{i=1}^{M} f_i(l) \ln f_i(l) \\
&= \sum_{i=1}^{M} \sum_{l=1}^{L} f_i p(l|i) \ln \frac{f_i(l)}{f_i} \\
&= \sum_{i=1}^{M} \sum_{l=1}^{L} f_i p(l|i) \ln \frac{p(l|i)}{p(l)} \\
&= I(\theta; Y_{n+1} | \mathcal{F}_n = \mathfrak{f}_n) \leq \mathbf{C}
\end{aligned}$$

Or being more straightforward and doing the algebra for the last inequality

$$\begin{aligned}
\mathbf{E}\left[H_n - H_{n+1} \middle| \mathcal{F}_n = \mathfrak{f}_n\right] &= \min_{\psi} \sum_{i=1}^{M} \sum_{l=1}^{L} f_i p(l|i) \ln \frac{p(l|i)}{\psi_l} \\
&= \min_{\psi} \sum_{i=1}^{M} \sum_{l=1}^{L} f_i \sum_{k=1}^{K} w(k|i) P_{kl} \ln \frac{\sum_{k'=1}^{K} w(k'|i) P_{k'l}}{\psi_l \sum_{k'=1}^{K} w(k'|i)} \\
&\leq \min_{\psi} \sum_{l=1}^{L} \sum_{i=1}^{M} \sum_{k=1}^{K} f_i w(k|i) P_{kl} \ln \frac{w(k|i) P_{kl}}{\psi_l w(k|i)} \\
&= \min_{\psi} \sum_{l=1}^{L} \sum_{i=1}^{M} \sum_{k=1}^{K} f_i w(k|i) P_{kl} \ln \frac{P_{kl}}{\psi_l}
\end{aligned}$$

$$\leq \min_{\psi} \max_{k} \sum_{l=1}^{L} P_{kl} \ln \frac{P_{kl}}{\psi_l} = \mathbf{C} \tag{B.1}$$

The first equality is just writing mutual information as a minimization over output distributions which can be found in [1], the first inequality is a result of log-sum inequality, and the second inequality is just because a weighted sum of some quantity should be less then the maximum of weighted ones.

QED

Bibliography

[1] R. Blahut. Computation of channel capacity and rate-distortion functions. *Information Theory, IEEE Transactions on*, Vol. 18, Iss 4:460–473, 1972.

[2] M. V. Burnashev. Data transmission over a discrete channel with feedback, random transmission time. *Problemy Perdachi Informatsii*, 12, No. 4:10–30, 1976.

[3] M. V. Burnashev and K. Sh. Zigangirov. One problem of observation control. *Problemy Perdachi Informatsii*, 11, No. 3:44–52, 1975.

[4] Imre Csiszár and János G. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, Inc., Orlando, FL, USA, 1982.

[5] R. L. Dobrushin. An asymptotic bound for the probability error of information transmission through a channel without memory using the feedback. *Problemy Kibernetiki*, vol 8:161–168, 1962.

[6] Robert G. Gallager. *Title Discrete Stochastic Processes*. Kluwer Academic Publishers, USA, 1995.

[7] E. A. Haroutunian. A lower bound of the probability of error for channels with feedback. *Problemy Peredachi Informatsii*, vol 13:36–44, 1977.

[8] M. Horstein. Sequential transmission using noiseless feedback. *IEEE Transactions on Information Theory*, Vol.9 Iss 3:136–143, 1963.

[9] A. Kramer. Improving communication reliability by use of an intermittent feedback channel. *Information Theory, IEEE Transactions on*, Vol.15, Iss.1:52–60, 1969.

[10] M. S. Pinsker. The probability of error in block transmission in a memoryless gaussian channel with feedback. *Problemy Perdachi Informatsii*, 4(4):1–4, 1968.

[11] A. Sahai and T. Şimşek. On the variable-delay reliability function of discrete memoryless channels with access to noisy feedback. In *ITW2004, San Antonio, Texas, October 24 29, 2004*, 2004.

[12] J. P. M. Schalkwijk. A coding scheme for additive noise channels with feedback–ii: Band-limited signals. *Information Theory, IEEE Transactions on*, Vol.12, Iss.2:183–189, 1966.

[13] J. P. M. Schalkwijk and M. Barron. Sequential signalling under a peak power constraint. *Information Theory, IEEE Transactions on*, Vol.17, Iss.3:278–282, 1971.

[14] J. P. M. Schalkwijk and T. Kailath. A coding scheme for additive noise channels with feedback–i: No bandwidth constraint. *Information Theory, IEEE Transactions on*, Vol.12, Iss.2:172–182, 1966.

[15] C. Shannon. The zero error capacity of a noisy channel. *Information Theory, IEEE Transactions on*, Vol. 2, Iss 3:8–19, 1956.

[16] Dimitri P. Bertsekas (with Angelia Nedic and Asuman E. Özdağlar). *Convex Analysis and Optimization.* Athena Scientific, New York, NY, USA, 2003.

[17] H. Yamamoto and K. Itoh. Asymptotic performance of a modified schalkwijk-barron scheme for channels with noiseless feedback. *IEEE Transactions on Information Theory*, Vol.25, Iss.6:729–733, 1979.