Reliability Function of Variable-length Block Codes with Feedback under Cost Constraints

Barış Nakiboḡlu

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology

April 10, 2006

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. Lower bounds are found on expected decoding time for a given message set size, average error probability, and cost constraint. It is also shown that there exist coding strategies that reaches these bounds in certain asymptotic sense. Consequently the reliability function is found as a function of rate and power constraint for variable decoding time DMC's with perfect feedback under a cost constraint. The results in this work generalize Burnashev's results,[1] to the cost constrained case, however as oppose to the unconstrained case, except few cases, concavity of the reliability function is strict.

1 Introduction

The effect of feedback in communication has been studied from the early days of information theory. We will start with a brief overview of those studies. The results depend heavily on the model used and the constraints imposed on the system.

All of the results considered here are '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. In a fixed-length block-coding scheme with feedback duration of the codewords will be fixed but the symbols at each time will depend on previous channel outputs as well as the message. In ‘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.[1] In all of the following results feedback is assumed to be perfect: infinite capacity, error free, delay free. However it is evident in many of the cases that the assumption of infinite capacity and instantaneous feedback is not necessary for the corresponding results.

When we look at the maximum achievable rate, i.e., or minimum expected time to send a ‘large amount of information’ with arbitrarily small error probability, i.e., channel capacity, feedback does not yield an improvement. For block coding with feedback, Shannon [12], showed that channel capacity does not increase with feedback in stationary discrete memoryless channels (DMC). Although it is not stated explicitly in [12], one can generalize this result to the ‘generalized block-coding’ case, using the weak law of large numbers.

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 [3], 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 [5], 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 discrete time additive white Gaussian noise channel (AWGNC) is given by Pinsker [8]. He showed that the sphere packing exponent is still an upper bound on error exponent even with feedback if decoding time is constant and the total amount of energy in each codeword is the average power constraint times the block length.

A first relaxation would be having a block code with a constraint on the expected energy. Schalkwijk and Kailath [11], [10], 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 [7], proved that error probability can be made to decay n-fold exponentially. In fact no lower bound to error

---

[1] In other words the receiver decides when to make a decision about the transmitted message.

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

[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.

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), give us the corresponding relaxation for DMC. In contrast to the case of channel capacity, where Shannon's result in [12] can be extended to the generalized block-coding 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 [1]. 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 [9] a 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 with feedback under cost constraints on DMCs, which does not have have a non-trivial zero transition probability.

For DMC's which have a zero transition probability, zero-error capacity will be the appropriate performance measure. Shannon showed in [12] 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 [12] 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. Burnashev, [1], extended this result to the generalized block codes. Furthermore he showed that if the channel has one or more zero probability transitions then zero-error capacity is equal to the channel capacity for generalized block codes. We will show that Burnashev's, results extends to the case with cost constraints.

We will start with a description of channel and constraints in the second section. In the third section we will propose a family of coding schemes for DMC's with feedback. In the fourth and fifth section we will investigate the performance of DMC's under cost constraints, and derive lower bounds on the expected decoding time. In the sixth section we will extend the conventional definitions of the channel capacity, zero-error channel capacity and reliability function to the generalized block codes with feedback and derive the expression for reliability function using the bounds we obtained in section three and five.

---

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

2 Model And Notation

We start by describing the forward and feedback channel models to be considered here, and then we continue with the depiction of possible coding and decoding algorithms. After that we will define the cost constraints for variable decoding time systems and explain why cost constraints of this form are appropriate. Finally we will mention some trivial extensions of the known results to the cost constraint case.[5]

2.1 Channel Model

The forward channel is assumed to be a discrete memoryless channel (DMC) with input alphabet $\{1, \ldots, |\mathcal{X}|\}$ and output alphabet $\{1, \ldots, |\mathcal{Y}|\}$. The input and output at time n are denoted X_n, Y_n respectively, and X^n, Y^n denote the n-tuples $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ respectively. In addition for each letter $k \in \mathcal{X}$, there is a transmission cost ρ_k.

The feedback channel is a noiseless DMC with an arbitrarily large alphabet, i.e., $\mathbf{P}\left[Z'_n \middle| Z_n\right] = 1 \quad \forall n$ where Z_n and Z'_n denote the input and the output of the feedback channel at time n, respectively. Thus we will use Z_n to describe both the feedback symbol send by the receiver and the feedback symbol received by the transmitter.

The symbol Z_n sent from the receiver at time n can depend on Y^n and is received without error at the transmitter after X_n is sent and before X_{n+1} is sent. Z^n denotes $Z_1, \ldots, Z_n$.

Because of the memorylessness assumption,

$$\mathbf{P}\left[Y_n \middle| X^n, Y^{n-1}, Z^{n-1}\right] = \mathbf{P}\left[Y_n \middle| X_n\right]$$

Thus the forward DMC is defined by the $|\mathcal{X}|$ by $|\mathcal{Y}|$ transition matrix $\{P_{kl}\}$ where, for each time n, $\mathbf{P}\left[Y_n = l \middle| X_n = k\right] = P_{kl}$. If all of the elements of column of the probability transition matrix is zero, corresponding output symbol is not reachable from any input symbol, then we can safely discard that symbol, column, from the model. Thus we will assume that there are no non-reachable output symbols from now on.

The existence of a zero transition probability, is a crucial factor on characterizing the error probability of generalized block codes. Although our main focus will be the channels which does not have zero-transition probabilities, we will mention corresponding results for channels with zero transition probabilities as well.

---

[5] We will be considering discrete time systems only.

2.2 Coding Algorithm

The transmitter is given the message θ, which is drawn from the message set $\mathcal{M} = \{1, \ldots, M\}$ according to a uniform probability distribution at time 0. Until 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.[6]

In general a coding scheme is an assignment of messages to the probability measures on input letters. The probability mass function used in these assignments can only depend on the observation of the transmitter, i.e., output of the feedback channel, and the random variables generated at the transmitter.

When we are finding bounds on best possible performance we need to consider the set of all possible coding algorithms, $\mathfrak{C}$. However when we have perfect feedback we can consider a much restricted set of coding algorithms, $\mathfrak{C}_R$, and argue that any coding algorithm outside this set can be replaced by an algorithm which is in $\mathfrak{C}_R$, without sacrificing the performance, in any sense.

$\mathfrak{C}_R$ is the set of coding schemes which are deterministic assignments of messages to the input letters, at each time depending on feedback.

$$X_n(m) = \mathcal{C}_n(m, Z^{n-1}) \qquad \forall Z^{n-1} \tag{1}$$

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

$$X_n = \mathcal{C}_n(\theta, Z^{n-1}) \qquad \forall Z^{n-1}, \forall \theta \tag{2}$$

In order to understand why $\mathfrak{C}_R$ is sufficient under the perfect feedback assumption, note that a coding algorithm can be outside of $\mathfrak{C}_R$ but in $\mathfrak{C}$ if and only if it has one or more experiments that is done at the transmitter which determines the assignment of one or more of the messages. However any experiment done at time n at the transmitter, can be done at time $n-1$ at the receiver[7] and its results can be sent to the transmitter through the feedback channel via Z_{n-1}. Note that disregarding that knowledge at the receiver can only degrade the performance, and will be equivalent to the original coding scheme. Consequently for any coding algorithm outside of $\mathfrak{C}_R$ there is at least one coding algorithm in $\mathfrak{C}_R$, which has a performance at least as good as the one outside $\mathfrak{C}_R$.

---

[6] The only work that considers non-block algorithms with feedback was by Horstein, [6], until recently. Anant Sahai has some new analysis for fixed delay codes.

[7] We might be required to do the M separate random experiments, one for each of the messages.

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

$$\mathcal{F}_n = \sigma(Y^n, Z^n) = \sigma(Z^n)$$

It is worth mentioning that because of the infinite feedback link rate assumption we need not to include any random variable other then Z^n, because anything observed at receiver can be send to the transmitter. The σ-field describing the whole system, $\mathcal{G}_n$ will include θ as one of its generators.[8]

$$\mathcal{G}_n = \sigma(Z^n, \theta)$$

The specific element of $\mathcal{F}_n$ observed at time n, is denoted by $\mathfrak{f}_n$. The filtration formed by the sequence of $\mathcal{F}_n$'s is denoted by $\mathcal{F}$

2.3 Decoding Criteria

A decoding criteria is a decision rule about continuing the coding or decoding the message depending on the receivers observation. In other words it is a Markov stopping time with respect to the filtration $\mathcal{F}$, together with a decoded message at each $\mathfrak{f}_n$ that stops the coding. Consequently $\mathcal{F}$ can be partitioned into $M+1$ disjoint sets $\chi_1, \chi_2, \ldots, \chi_{M+1}$. In first M of them corresponding message will be decoded. In the last one, χ_{M+1}, decoding will never occur.

The probability of error of the i^{th} message is $\mathbf{P}\left[e \middle| \theta = i\right] = 1 - \mathbf{P}\left[\chi_i \middle| \theta = i\right]$. Consequently the probability of error of the code, P_e is given by

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

The expected transmission time is

$$\mathbf{E}\left[\tau\right] = \frac{1}{M} \sum_{i=1}^{M} \mathbf{E}\left[\tau \middle| \theta = i\right]$$

At each time n, depending on the realization $\mathfrak{f}_n$ of σ-field $\mathcal{F}_n$, we will have an a posteriori probability distribution $p_i(\mathfrak{f}_n)$ on $\mathcal{M}$. Consequently the conditional entropy

---

[8] In the most general case where feedback is not necessarily error free, $\mathcal{F}_n$ 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.

of the message, given $\mathcal{F}_n = \mathfrak{f}_n$, will be a random variable measurable in $\mathcal{F}_n$, given by the expression;

$$\mathcal{H}_{\mathfrak{f}_n} = H(\theta \mid \mathcal{F}_n = \mathfrak{f}_n) = -\sum_{i=1}^{M} p_i(\mathfrak{f}_n) \ln p_i(\mathfrak{f}_n)$$

2.4 Cost Constraints For Generalized Block Codes

In order to make use of the variable nature of the decoding time we will relax our cost constraint from a fixed total energy constraint for every channel realization to an expected energy constraint. We will put restrictions on the expected energy that is used for sending one message in terms of the expected duration of the transmission.[9]

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 $\mathcal{S}_n$, i.e. $\mathcal{S}_n$, is measurable in $\mathcal{G}_n$ but not in $\mathcal{F}_n$. Nevertheless the constraint on $\mathcal{S}_n$ is in terms of the expected value of $\mathcal{S}_n$, $\mathbf{E}\left[\mathcal{S}_\tau \middle| \mathcal{F}_0\right]$, which is measurable in the filtration $\mathcal{F}$. The random variable $\mathcal{S}_n(i)$ be the cost for the codeword that corresponds to the i^{th} message up to and including time n. Since at time n the part of the codeword up to time n is known at the receiver for each message, receiver can calculate $\mathcal{S}_n(i)$, without knowing the actual message, θ. Thus each $\mathcal{S}_n(i)$ is 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. Similarly we can calculate the expected energy of sending a message, $\mathbf{E}\left[\mathcal{S}_\tau \middle| \mathcal{F}_0\right]$, where τ is the decoding time. The cost constraint, $\mathcal{P}$, will restrict the set of transmission strategies to the ones satisfying

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

Before starting the analysis we will discuss some facts about the cost constraint. Note that $\mathcal{P}$ is a meaningful constraint iff $\mathcal{P} \in (\rho_{min}, \rho_{max})$. If $\mathcal{P} \ge \rho_{max}$ then the cost

---

[9] This allows an uneven distribution of the expected energy and its time average between the different decoding paths. Thus some of the possible decoding paths might have an average power is larger then the average constraint we had.

constraint does not introduce any restriction on the coding algorithms or decoding rules that can be used. Thus this case will be equivalent to the unconstrained case, which is studied by Burnashev, [1]. If $\mathcal{P} = \rho_{min}$, i.e. $\mathbf{E}[\mathcal{S}_\tau] \leq \rho_{min}\mathbf{E}[\tau]$, then only the letters with $\rho_i = \rho_{min}$ can be used so this again reduces to the unconstrained case but with a smaller set of input letters. Finally the constraint is not satisfiable if $\mathcal{P} < \rho_{min}$.

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

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

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 $\rho_{min} = 0$.

3 Asymptotically Optimal Coding Algorithm

We will first focus on proposing a coding scheme for channels that does not have transitions with zero probability. After obtaining upper bounds on the error probability of those codes, we will briefly describe a family of codes with zero error probability for channels with zero probability transitions. Claimed optimality of the families for corresponding cases will be apparent only after the discussion of converse (lower bounds on expected decoding times) and channel parameters like reliability function, zero error capacity.

3.1 Channels with no Zero Transition Probability

Before going into details of the algorithm we will revisit two well known problems, and restate widely known results about them in a way that will be handy for our treatment.

First problem is coding. Conventional coding theorem[10] can easily be reformulated as follows.

---

[10] Theorem 7.3.2 in [4] for example.

Lemma 1 *For any DMC, for every cost constraint $\mathcal{P} \geq 0$, and every $\delta > 0$, there exist a $\epsilon(\delta) > 0$ such that for large enough block length ℓ, one can find a code with $M \geq e^{\ell(\mathbf{C}(\mathcal{P})-\delta)}$ codewords, such that each code word satisfies $\mathcal{S}_\ell \leq \mathcal{P}\ell$ and has a probability of error $P_e \leq e^{-\ell\epsilon(\delta)}$.*

The function $\mathbf{C}(\mathcal{P})$ is given by

$$\mathbf{C}(\mathcal{P}) = \max_{\substack{\varphi \\ \sum_{k=1}^{K} \varphi_k \rho_k \leq \mathcal{P}}} \sum_{k=1,l=1}^{K,L} \varphi_k P_{kl} \ln \frac{P_{kl}}{\sum_{j=1}^{K} \varphi_j P_{jl}}$$

The second problem we will visit is binary hypothesis testing. The transmitter will transmit one of the two a codewords of length ℓ depending on the message, θ, it is given; and receiver will make a guess, $\hat{\theta}$, depending on its observation,Y^ℓ. There are only two possible messages A and R, and the error probabilities of them, are not necessarily of equal importance. It is shown in [2], for any codeword pair $(\mathbf{x}^{\mathcal{A}}, \mathbf{x}^{\mathcal{R}})$and corresponding $q_i(l) = \mathbf{P}\left[Y_i = l \middle| \theta = R\right]$, $p_i(l) = \mathbf{P}\left[Y_i = l \middle| \theta = A\right]$ there exist a decoding rule that solely depends on log-likelihood ratio, $\ln \frac{\mathbf{P}\left[Y^\ell \middle| \theta = A\right]}{\mathbf{P}\left[Y^\ell \middle| \theta = R\right]}$, such that

$$P_{A|R} = \mathbf{P}\left[\hat{\theta} = A \middle| \theta = R\right] \leq e^{-\sum_i D(\xi_i^s || q_i)}$$

$$P_{R|A} = \mathbf{P}\left[\hat{\theta} = R \middle| \theta = A\right] \leq e^{-\sum_i D(\xi_i^s || p_i)}$$

where $\xi_i^s(l) = \frac{(p_i(l))^s (q_i(l))^{1-s}}{\sum_j (p_i(j))^s (q_i(j))^{1-s}}$.

Evidently

$$\lim_{s \to 1} D(\xi_i^s || q_i) = D(p_i || q_i) \quad \text{and} \quad \lim_{s \to 1} D(\xi_i^s || p_i) = 0$$

In addition if $s < 1$, $\min_{p,q;p \neq q} D(\xi^s || p) > 0$. Thus one can write the above result as follows $\forall \delta > 0$, $\exists \epsilon(\delta) > 0$ such that, for any code word pair $(\mathcal{A}_i, \mathcal{R}_i)$ such that $\mathcal{A}_i \neq \mathcal{R}_i \quad \forall i$, there exists a threshold decoding such that

$$P_{A|R} \leq e^{-\sum_i (D(p_i || q_i) - \delta)} \quad \text{and} \quad P_{R|A} \leq e^{-\ell\epsilon(\delta)}$$

In our case not only the errors but also the costs of the codewords will be of different importance. We will require that the cost of the codeword associated with $\mathcal{A}$ satisfy, $\sum_i \rho_{\mathbf{x}_i^{\mathcal{A}}} \leq \ell\mathcal{P}$.

Lemma 2 *For any DMC which does not have zero probability transitions, for any constraint $\mathcal{P} \geq 0$, for every $\delta > 0$ for large enough ℓ there exist two code word $\mathbf{x}^{\mathcal{A}}$ and $\mathbf{x}^{\mathcal{R}}$such that*

$$P_{A|R} \leq e^{-\ell(\mathbf{D}(\mathcal{P})-\delta)} \quad \textit{and} \quad P_{R|A} \leq e^{-\ell\epsilon(\delta)} \tag{3}$$

and $\sum_i \rho_{\mathbf{x}_i^{\mathcal{A}}} \leq \ell\mathcal{P}$

The function $\mathbf{D}(\mathcal{P})$ is given by

$$\mathbf{D}(\mathcal{P}) = \max_{\substack{\varphi \\ \sum_{k=1}^{K} \varphi_k \rho_k \leq \mathcal{P}}} \sum_k \varphi_k \mathbf{D}_k \qquad \mathbf{D}_i = \max_j \sum_{l=1}^{L} P_{il} \log \frac{P_{il}}{P_{jl}}$$

$\mathbf{D}(\mathcal{P})$ is the maximum of a linear function over linear constraints. The maximization can be most easily visualized by Figure 1, which shows that $\mathcal{D}(\mathcal{P})$ is piecewise linear, non-decreasing, and concave in its region of definition, $\mathcal{P} \geq 0$.

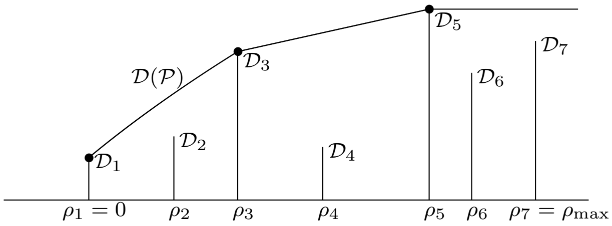

Figure 1: Calculation of $\mathcal{D}(\mathcal{P})$ as a function of the divergences $\mathcal{D}_k$ associated with individual inputs k. For convenience, the inputs are ordered in terms of cost.

Now we are ready describe the error and erasure code which will be the atomic element in the generalized block coding. For a given (η, β) pair and block length ℓ we will use a block code of length $\eta\ell$, rate $\mathbf{C}\left(\frac{\beta}{\eta}\mathcal{P}\right) - \delta$, whose average power is less then or equal to $\frac{\beta}{\eta}\mathcal{P}$. Using the feedback-link transmitter instantly learn the temporary guess of the receiver about the message. Then the transmitter will accept, $\mathcal{A}$, or reject, $\mathcal{R}$, the temporary guess using a two-codeword binary-hypothesis testing code, of length $(1-\eta)\ell$ of power $\frac{1-\beta}{1-\eta}$ whose $P_{A|R}$ exponent is $\mathbf{D}\left(\mathcal{P}\frac{1-\beta}{1-\eta}\right) - \delta$.

Using above lemmas one can easily show that, for any (η, β) pair and for any $\delta > 0$ for large enough block-length ℓ there exist an error and erasure code whose number of messages, $M \geq e^{\ell\eta\left(\mathbf{C}\left(\frac{\beta}{\eta}\mathcal{P}\right)-\delta\right)}$, probability of error $P_{e\ell} \leq e^{-\ell\left[(1-\eta)\left(\mathbf{D}\left(\frac{1-\beta}{1-\eta}\mathcal{P}\right)-\delta\right)+\eta\epsilon(\delta)\right]}$, probability of retransmission $P_{r\ell} \leq e^{-\ell\eta\epsilon(\delta)} + e^{-\ell(1-\eta)\epsilon(\delta)}$ and expected energy $\mathbf{E}[\mathcal{S}_\ell] \leq \ell(\mathcal{P} + \rho_{max} e^{-\ell\eta\epsilon(\delta)})$.

We will use this error and erasure code again and again until the message is decoded without erasure. One can write the error probability, expected energy, and expected decoding time of this generalized block code in terms of the erasure probability as follows,

$$P_e = \frac{1}{1-P_{r\ell}} P_{e\ell} \quad ; \quad \mathbf{E}[\tau] = \frac{1}{1-P_{r\ell}} \ell \quad ; \quad \mathbf{E}[\mathcal{S}_\tau] = \frac{1}{1-P_{r\ell}} \mathbf{E}[\mathcal{S}_\ell]$$

Consequently for any cost constraint $\mathcal{P}$, and (η, β) pair, for any $\delta' > 0$, there exist an $\epsilon'(\delta') > 0$ for large enough $\mathbf{E}[\tau]$, there exist a generalized block code such that

$$M \geq e^{\mathbf{E}[\tau]\eta\left(\mathbf{C}\left(\frac{\beta}{\eta}\mathcal{P}\right) - \delta'\right)}$$

$$P_{e\ell} \leq e^{-\mathbf{E}[\tau](1-\eta)\left(\mathbf{D}\left(\frac{1-\beta}{1-\eta}\mathcal{P}\right) - \delta'\right)}$$

$$\mathbf{E}[\mathcal{S}_\tau] \leq \mathbf{E}[\tau]\left(\mathcal{P} + \rho_{max} e^{-\mathbf{E}[\tau]\epsilon'(\delta')}\right)$$

Note that $\mathcal{E}_{\mathbf{c}}(x) = \lim_{y \to x^+} \frac{y}{\mathbf{C}(y)}$ is an increasing function of x from $[0, \infty)$ to $[\mathcal{E}_{\mathbf{c}}(0), \infty)$. In addition it is strictly increasing possible except and interval[11] of the form $[0, x_0]$ on which $\mathcal{E}_{\mathbf{c}}(x) = \mathcal{E}_{\mathbf{c}}(0)$. Evidently $\mathcal{E}_{\mathbf{c}}(\cdot)$ has an inverse if we restrict its domain to be $[x_0, \infty)$, we will denote this function by ${\mathcal{E}_{\mathbf{c}}}^{-1}(\cdot)$.

Then for any $(\mathcal{P}, R)$ pair such that $\mathcal{P} \geq 0$ and $0 \leq R \leq \mathbf{C}(\mathcal{P})$, for any $\eta \geq \eta^*(\mathcal{P}, R) = \frac{\mathcal{P}}{{\mathcal{E}_{\mathbf{c}}}^{-1}\left(\frac{\mathcal{P}}{R}\right)}$, there exist a $\beta(\mathcal{P}, R, \eta) \leq 1$ such that $\eta \mathbf{C}\left(\frac{\beta(\mathcal{P},R,\eta)}{\eta}\mathcal{P}\right) \geq R$.

$\mathbf{C}^{-1}(\cdot)$ is defined on the interval $[\mathbf{C}(0), \mathbf{C}]$. But we will extend its domain to $[0, \mathbf{C}]$ by defining $\mathbf{C}^{-1}(x) = 0$, $\forall x \in [0, \mathbf{C}(0))$. Evidently with this extended definition $\beta(\mathcal{P}, R, \eta) = \frac{\eta}{\mathcal{P}} \mathbf{C}^{-1}\left(\frac{R}{\eta}\right)$.

If we plug in these values in the expression we have found, and use the scheduling technique, which will be described in appendix B we get the following result.

Theorem 1 *For any DMC which does not have zero probability transition, with an error free feedback channel of rate* $\mathbf{C}$ *(or higher) of finite delay* $\mathbf{T}$, $\forall \mathcal{P} \geq 0$, *for all rates* $R \leq \mathbf{C}(\mathcal{P})$, *and for all* $\eta \in \left(\frac{\mathcal{P}}{{\mathcal{E}_{\mathbf{c}}}^{-1}\left(\frac{\mathcal{P}}{R}\right)}, 1\right)$ *and* $\forall \delta > 0$, *there exist* $\epsilon(\delta) > 0$ *such that for large enough* $\mathbf{E}[\tau]$ *there exist a generalized block-coding scheme such that*

---

[11] These facts are proved in the appendix C

$$M \geq e^{(R-\delta)\mathbf{E}[\tau]} \quad P_e \leq e^{-(E(\mathcal{P},R,\eta)-\delta)\mathbf{E}[\tau]} \quad \mathbf{E}\left[\mathcal{S}_\tau\right] \leq \mathbf{E}\left[\tau\right]\left(\mathcal{P} + \rho_{max} e^{-\mathbf{E}[\tau]\epsilon(\delta)}\right) \tag{4}$$

where $E(\mathcal{P},R,\eta) = (1-\eta)\mathbf{D}\left(\frac{\mathcal{P}-\eta\mathbf{C}^{-1}\left(\frac{R}{\eta}\right)}{1-\eta}\right)$

3.2 Channels with Zero Transition Probability

For a DMC with a zero transition probability, there are two input symbol, r, a and one output letter s, such that $P_{rs} = 0$ and $P_{as} = \mathfrak{q} > 0$. For a hypothesis testing problem if we choose $\mathbf{x}^{\mathcal{A}} = aaa\ldots a$ and $\mathbf{x}^{\mathcal{R}} = rrr\ldots r$ and if we decode $\mathcal{A}$ only if it observes at least on s, on the channel output,

$$P_{A|R} = 0 \qquad \text{and} \qquad P_{R|A} = (1-\mathfrak{q})^{\ell} \tag{5}$$

where ℓ is the length of the codewords.

The coding algorithm will be almost identical to the one we used previously, except two facts. First one is that the hypothesis testing in the second phase will be done according to the above algorithm, rather then the ones described previously. Second difference is that the duration of the second phase will approximately equal to the logarithm of the duration of the trial. Consequently the amount of time and energy we spend in the second phase increases only logarithmically with the block-length of the error and erasure code.

Thus for a DMC with one or more zero probability transitions, for any $\delta > 0$ for large enough block-length ℓ there exist an error and erasure code whose number of messages, $M \geq e^{(\ell-\ln\ell)(\mathbf{C}(\mathcal{P})-\delta)}$, probability of error $P_e = 0$, probability of retransmission $P_{r\ell} \leq e^{-(\ell-\ln\ell)\epsilon(\delta)} + \ell^{\ln(1-\mathfrak{q})}$ and expected energy $\mathbf{E}\left[\mathcal{S}_\ell\right] \leq \ell\mathcal{P} + \rho_{max}\ln\ell$.

Using this error and erasure code repetitively as described previously, one can obtain a variable decoding time block-code. After little bit of algebraic manipulation one can get the following theorem.

Theorem 2 *For any DMC which has at least one zero probability transition and an error free feedback channel of rate* $\mathbf{C}$ *(or higher) of finite delay* $\mathbf{T}$*, for any* $\mathcal{P} \geq 0$ *and for any* $\delta > 0$ *for large enough* $\mathbf{E}\left[\tau\right]$*, there exists coding algorithm with decoding criteria such that*

$$M \geq e^{\mathbf{E}[\tau](\mathbf{C}(\mathcal{P})-\delta)} \qquad P_e = 0 \qquad \mathbf{E}\left[\mathcal{S}_\tau\right] \leq \mathcal{P}\mathbf{E}\left[\tau\right] + 2\rho_{max}\ln\mathbf{E}\left[\tau\right]$$

4 Basic Lemmas

In this section we will present basic tools we will use in the proof of the converse result. Proofs of these lemmas can be found in the appendix A.

4.1 Generalized Fano's Inequality

Lemma 3 (Generalized Fano Inequality for variable decoding time systems) *For any coding algorithm and decoding rule such that* $\mathbf{P}\left[\tau < \infty\right] = 1$,

$$\mathbf{E}\left[\mathcal{H}_{\mathfrak{f}_\tau}\right] \leq \mathfrak{h}(P_e) + P_e \ln(M-1) \tag{6}$$

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

The proof of this lemma only requires the forward channel to be stationary and memoryless. Thus it can be used in the AWGNC case also.

In proving non-existence (converse) results, the Fano inequality is generally used as a lower bound on error probability P_e in terms of the conditional entropy, $\mathbf{E}\left[\mathcal{H}_{\mathfrak{f}_\tau}\right]$. Our approach will be a little bit different; we will use P_e to find an upper bound on $\mathbf{E}\left[\mathcal{H}_{\mathfrak{f}_\tau}\right]$. 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 $\mathbf{E}\left[X|\,Y\right] < A$ means that for every value y of the random variable Y, $f(y) = \mathbf{E}\left[X|\,y\right] < A$. Equivalently $f(Y) = \mathbf{E}\left[X|\,Y\right] < A$. The following lemmas about the change of entropy can best be understood with this interpretation.

4.2 Bounds On Change of Entropy and Cost

Consider the stochastic sequence $V_n^{\mathcal{P}}$, which is measurable in $\mathcal{F}_n$

$$V_n^{\mathcal{P}} = \mathcal{H}_{\mathfrak{f}_n} + \gamma_{\mathbf{C}}^{\mathcal{P}} \mathbf{E}\left[\mathcal{S}_n|\,\mathcal{F}_n\right] \tag{7}$$

where $\gamma_{\mathbf{C}}^{\mathcal{P}}$ is the value of the Lagrange multiplier for the cost constraint in the maximization of mutual information given in equation (19), in appendix A.

Lemma 4 *$\forall n \geq 0$, and $\forall \mathcal{P} \geq \rho_{min}$;*

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

Let us define a stochastic sequence, $W_n^{\mathcal{P}}$ measurable in $\mathcal{F}_n$ as

$$W_n^{\mathcal{P}} = \ln \mathcal{H}_{\mathfrak{f}_n} + \gamma_{\mathbf{D}}^{\mathcal{P}} \mathbf{E}\left[\mathcal{S}_n \middle| \mathcal{F}_n\right] \tag{8}$$

where $\gamma_{\mathbf{D}}^{\mathcal{P}}$ is the value of the Lagrange multiplier for the cost constraint in the maximization given in equation (19), in appendix A. Evidently $W_n^{\mathcal{P}}$ is a more accurate tool to handle small values of entropy.

Lemma 5 *$\forall n \geq 0$, and $\forall \mathcal{P} \geq \rho_{min}$;*

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

Finally we can establish a bound on the maximum decrease in logarithm of entropy as follows.

Lemma 6 *For any $n \geq 0$, $Y_{n+1} = l$*

$$|\ln \mathcal{H}_{\mathfrak{f}_n} - \ln \mathcal{H}_{\mathfrak{f}_{n+1}}| \leq \max_{i,k} \ln \frac{P_{kl}}{P_{il}} \leq \max_{i,k,l} \ln \frac{P_{kl}}{P_{il}} = \mathbf{F} \tag{9}$$

4.3 Measuring Time with Submartingales

We will use the following lemma, to obtain lower bounds on the expected value of stopping times, in terms of the expected values of stopped stochastic sequences.

Lemma 7 *Let the stochastic sequence $\{\Gamma_i\}$ be measurable in the filtration $\mathcal{F}$ and assume that for some K and R*

$$\begin{aligned} \mathbf{E}\left[\, |\Gamma_n - \Gamma_{n+1}| \, \middle| \mathcal{F}_n\right] &< K \qquad \forall n \\ \mathbf{E}\left[\Gamma_n - \Gamma_{n+1} \middle| \mathcal{F}_n\right] &\leq R \qquad \forall n \end{aligned}$$

If τ_i and τ_f are stopping times with respect to the filtration $\mathcal{F}$ such that

$$\mathbf{E}\left[\tau_f\right] < \infty \qquad \& \qquad \tau_i \leq \tau_f$$

Then

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

The following corollary immediately follows from lemmas 4,5, 6 and 7.

Corallary 1 *For any pair of stopping times* (τ_i, τ_f)*, and any coding algorithm; if*

$$\tau_f \geq \tau_i \quad \textit{and} \quad \mathbf{E}\left[\mathcal{S}_{\tau_f} - \mathcal{S}_{\tau_i}\right] \leq \mathcal{P}\mathbf{E}\left[\tau_f - \tau_i\right]$$

then

$$\mathbf{E}\left[\mathcal{H}_{\mathfrak{f}_{\tau_i}} - \mathcal{H}_{\mathfrak{f}_{\tau_f}}\right] \leq \mathbf{C}(\mathcal{P})\mathbf{E}\left[\tau_f - \tau_i\right] \quad \textit{and} \quad \mathbf{E}\left[\ln \mathcal{H}_{\mathfrak{f}_{\tau_i}} - \ln \mathcal{H}_{\mathfrak{f}_{\tau_f}}\right] \leq \mathbf{D}(\mathcal{P})\mathbf{E}\left[\tau_f - \tau_i\right]$$

Proof:
Since $0 \leq H_n \leq \ln M$, we have $|V_n^{\mathcal{P}} - V_{n+1}^{\mathcal{P}}| \leq \ln M + \gamma_{\mathbf{C}}^{\mathcal{P}} \rho_{\max}$. Also since $|\ln H_n - \ln H_{n+1}| \leq \mathbf{F}$ we have $|W_n^{\mathcal{P}} - W_{n+1}^{\mathcal{P}}| \leq \mathbf{F} + \gamma_{\mathbf{D}}^{\mathcal{P}} \rho_{\max}$.

Thus we can use Lemmas 4, 5 together with 7. Inserting our assumption $\mathbf{E}\left[\mathcal{S}_{\tau_f} - \mathcal{S}_{\tau_i}\right] \leq \mathcal{P}\mathbf{E}\left[\tau_f - \tau_i\right]$ we get the claimed results.

QED

The bounds asserted by this lemma are tight provided that we use them in ‘appropriate’ intervals. Indeed the proof of the converse will simply use these bounds with an appropriate intermediate stopping time.

5 Lower Bound For The Expected decoding Time

In this section we will derive lower bounds on expected decoding time for a given value of cost constraint, $\mathcal{P}$, number of possible messages, M, and probability of error, P_e. These will later used to obtain upper bounds on channel capacity , zero-error channel capacity and the reliability function.

Theorem 3 *For any DMC channel with feedback; if a generalized block-code with* $M \geq 2$*,* $P_e \geq 0$ *and* $\mathcal{P} \geq 0$*, satisfies* $\mathbf{E}\left[\mathcal{S}_\tau\right] \leq \mathcal{P}\mathbf{E}\left[\tau\right]$ *then the expected value of decoding time will satisfy*

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

Proof:
Note that $\mathcal{H}_{\mathfrak{f}_0} = \ln M$, and because of the generalized Fano's inequality $\mathbf{E}\left[\mathcal{H}_{\mathfrak{f}_\tau}\right] \leq \mathfrak{h}(P_e) + P_e \ln(M-1)$. Applying the corollary 1, for $\tau_i = 0$, $\tau_f = \tau$ we get the claimed in equality.

QED

Note that this bound is valid for any code on any DMC. However if we focus on the channels which does not have a zero transition probability, it is possible to find a tighter bound given by the following theorem.

Theorem 4 *For any DMC with feedback which does not have a zero transition probability; if a generalized block code with* $M > 2$*,* $P_e > 0$*, and* $\mathcal{P} \geq 0$*, satifies satisfies*[12] $\mathbf{E}[\mathcal{S}_\tau] \leq \mathcal{P}\mathbf{E}[\tau]$ *then the expected number of observations* $\mathbf{E}[\tau]$ *satisfies the inequality*

$$\mathbf{E}[\tau] \geq \min_{\substack{0 \leq \beta \leq 1 \\ 0 < \eta < 1}} \max \left\{ \frac{\mathcal{V}_1}{\eta \mathbf{C}\left(\mathcal{P}\frac{\beta}{\eta}\right)}, \frac{\mathcal{V}_2}{(1-\eta)\mathbf{D}\left(\mathcal{P}\frac{1-\beta}{1-\eta}\right)} \right\}$$

where

$$\mathcal{V}_1 = \ln M \left(1 - P_e(\ln M - \ln P_e + 1) - \frac{1}{\ln M}\right)$$
$$\mathcal{V}_2 = -\ln P_e \left(1 - \frac{\mathbf{F}}{-\ln P_e} - \frac{\ln(\ln M - \ln P_e + 1)}{-\ln P_e}\right)$$

The functions $\frac{1}{\eta \mathbf{C}\left(\mathcal{P}\frac{\beta}{\eta}\right)}$ and $\frac{1}{(1-\eta)\mathbf{D}\left(\mathcal{P}\frac{1-\beta}{1-\eta}\right)}$ are convex in the pair (η, β). So we have a convex optimization problem provided that $\mathcal{V}_1$ and $\mathcal{V}_2$ are non-negative.

When proving converse results it is important to have constraints as weak as possible. Accordingly we have not put a constraint on the expected energy at each decoding point. Our constraint is on the expected amount of energy that is spent for sending 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:
If $\mathbf{E}[\tau | \mathcal{F}_0] = \infty$ then the theorem holds trivially; therefore we will assume $\mathbf{E}[\tau | \mathcal{F}_0] < \infty$, henceforth.

The main problem in the proof is to find an intermediate Markov stopping time τ_1 which will divide the message transmission interval into two disjoint phases such that the duration of each can be tightly lower bounded by the corollary. Consider

---

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

the stopping time $t_1 = \min\{n | \mathcal{H}_{\mathfrak{f}_n} \leq 1\}$ in filtration $\mathcal{F}$. Clearly $\tau_1 = \min(\tau, t_1)$ is also a Markov stopping time, and $0 \leq \tau_1 \leq \tau$ with probability one.

For any coding algorithm define the corresponding η and β as

$$\eta = \frac{\mathbf{E}[\tau_1]}{\mathbf{E}[\tau]} \qquad \beta = \frac{\mathbf{E}[\mathcal{S}_{\tau_1}]}{\mathbf{E}[\mathcal{S}_\tau]}$$

Then using the corollary in $(0, \tau_1)$ and (τ_1, τ) we get

$$\mathbf{E}[\tau_1] \geq \frac{\mathbf{E}\left[\mathcal{H}_{\mathfrak{f}_0} - \mathcal{H}_{\mathfrak{f}_{\tau_1}}\right]}{\mathbf{C}\left(\mathcal{P}\frac{\beta}{\eta}\right)} \quad \text{and} \quad \mathbf{E}[\tau - \tau_1] \geq \frac{\mathbf{E}\left[\ln \mathcal{H}_{\mathfrak{f}_{\tau_1}} - \ln \mathcal{H}_{\mathfrak{f}_\tau}\right]}{\mathbf{D}\left(\mathcal{P}\frac{1-\beta}{1-\eta}\right)}$$

Using the fact $\frac{\mathbf{E}[\tau_1]}{\mathbf{E}[\tau]} = \eta$

$$\mathbf{E}[\tau] \geq \max\left(\frac{1}{\eta}\frac{\mathbf{E}\left[\mathcal{H}_{\mathfrak{f}_0} - \mathcal{H}_{\mathfrak{f}_{\tau_1}}\right]}{\mathbf{C}\left(\mathcal{P}\frac{\beta}{\eta}\right)}, \frac{1}{1-\eta}\frac{\mathbf{E}\left[\ln \mathcal{H}_{\mathfrak{f}_{\tau_1}} - \ln \mathcal{H}_{\mathfrak{f}_\tau}\right]}{\mathbf{D}\left(\mathcal{P}\frac{1-\beta}{1-\eta}\right)}\right)$$

This is a lower bound on $\mathbf{E}[\tau]$, in terms of (η, β), and $\mathcal{P}$. However the pair (η, β) depends on the coding-scheme, we need to minimize the bound over (η, β) to remove this dependence.

$$\mathbf{E}[\tau] \geq \min_{\substack{0 \leq \beta \leq 1 \\ 0 < \eta < 1}} \max\left(\frac{1}{\eta}\frac{\mathbf{E}\left[\mathcal{H}_{\mathfrak{f}_0} - \mathcal{H}_{\mathfrak{f}_{\tau_1}}\right]}{\mathbf{C}\left(\mathcal{P}\frac{\beta}{\eta}\right)}, \frac{1}{1-\eta}\frac{\mathbf{E}\left[\ln \mathcal{H}_{\mathfrak{f}_{\tau_1}} - \ln \mathcal{H}_{\mathfrak{f}_\tau}\right]}{\mathbf{D}\left(\mathcal{P}\frac{1-\beta}{1-\eta}\right)}\right)$$

Finally we need to prove that $\mathbf{E}\left[\mathcal{H}_{\mathfrak{f}_0} - \mathcal{H}_{\mathfrak{f}_{\tau_1}}\right] \geq \mathcal{V}_1$ and $\mathbf{E}\left[\ln \mathcal{H}_{\mathfrak{f}_{\tau_1}} - \ln \mathcal{H}_{\mathfrak{f}_\tau}\right] \geq \mathcal{V}_2$.

$$\mathbf{E}\left[\mathcal{H}_{\mathfrak{f}_{\tau_1}}\right] = \mathbf{E}\left[\mathcal{H}_{\mathfrak{f}_{t_1}} \mathbb{I}_{\{\tau > t_1\}}\right] + \mathbf{E}\left[\mathcal{H}_{\mathfrak{f}_\tau} \mathbb{I}_{\{\tau \leq t_1\}}\right]$$

Since $\mathcal{H}_{\mathfrak{f}_n} \leq \ln M$

$$\begin{aligned}\mathbf{E}\left[\mathcal{H}_{\mathfrak{f}_{\tau_1}}\right] &\leq \mathbf{E}\left[1 \mathbb{I}_{\{\tau > t_1\}}\right] + \mathbf{E}\left[(\ln M) \mathbb{I}_{\{\tau \leq t_1\}}\right] \\ &= 1\mathbf{P}[\tau > t_1] + (\ln M)\mathbf{P}[\tau \leq t_1]\end{aligned}$$

Using $\mathbf{P}[\tau \leq t_1] \leq \mathbf{P}[\mathcal{H}_{\mathfrak{f}_\tau} > 1]$ and Markov in eqaulity $\mathbf{P}[\mathcal{H}_{\mathfrak{f}_\tau} > 1] \leq \mathbf{E}[\mathcal{H}_{\mathfrak{f}_\tau}]$ we get

$$\mathbf{E}\left[\mathcal{H}_{\mathfrak{f}_{\tau_1}}\right] \leq 1 + \ln M \mathbf{E}[\mathcal{H}_{\mathfrak{f}_\tau}]$$

Using Fano's in equality we get

$$\mathbf{E}\left[\mathcal{H}_{\mathfrak{f}_0} - \mathcal{H}_{\mathfrak{f}_{\tau_1}}\right] \geq \mathcal{V}_1$$

Using Lemma 6 and the definition of τ_1, lower bound $\mathbf{E}\left[\ln \mathcal{H}_{\mathfrak{f}_{\tau_1}}\right]$

$$\mathbf{E}\left[\ln \mathcal{H}_{\mathfrak{f}_{\tau_1}}\right] \geq -\mathbf{F}$$

Using Jensen's in equality, the concavity of $\log(\cdot)$ function, and Fano's inequality, we can upper bound $\mathbf{E}\left[\ln \mathcal{H}_{\mathfrak{f}_{\tau}}\right]$ as

$$\mathbf{E}\left[\ln \mathcal{H}_{\mathfrak{f}_{\tau}}\right] \leq \ln \mathbf{E}\left[\mathcal{H}_{\mathfrak{f}_{\tau}}\right] \leq \ln P_e + \ln(\ln M - \ln P_e + 1)$$

Consequently

$$\mathbf{E}\left[\ln \mathcal{H}_{\mathfrak{f}_{\tau_1}} - \ln \mathcal{H}_{\mathfrak{f}_{\tau}}\right] \geq \mathcal{V}_2 \tag{10}$$

QED

6 Reliability Function and Zero-Error Capacity

In this section we will extend the conventional definitions of channel parameters to the generalized block-codes, and use our analysis in sections 3 and 5 to obtain the reliability function and zero-error capacity of the DMCs, with feedback.

6.1 Definitions

A code is a coding algorithm, and a decoding criteria, for a given number of messages M, and probability of error P_e. All of the macroscopic properties of the code can be summarized by by the quadruple $(M, P_e, \mathbf{E}[\tau], \mathbf{E}[\mathcal{S}_\tau])$.

Definition 1 *A sequence of generalized block codes, $\mathcal{Q}$, is a 'valid sequences' iff* $\limsup_{i\to\infty} {P_e}^i = 0$

Definition 2 *A valid sequence, $\mathcal{Q}$, will called 'zero-error sequences' iff there exist $\forall k$, $\exists i \geq k$ such that ${P_e}^i = 0$. We will denote the zero error valid sequences with subscript zero, e.g. $\mathcal{Q}_0$.*

Definition 3 *$\mathfrak{Q}_{\mathcal{P}}$ is the set of valid sequences $\mathcal{Q}$ satisfying the cost constraint $\mathcal{P}$, i.e.* $\limsup_{i\to\infty} \frac{\mathbf{E}[\mathcal{S}_\tau]^i}{\mathbf{E}[\tau]^i} \leq \mathcal{P}$

Definition 4 *The Rate and error exponent of a valid sequence $\mathcal{Q}$ will be given by*

$$R_{\mathcal{Q}} = \liminf_{i\to\infty} \frac{\ln M^i}{\mathbf{E}\left[\tau\right]^i} \quad \textit{and} \quad E_{\mathcal{Q}} = \liminf_{i\to\infty} \frac{-\ln {P_e}^i}{\mathbf{E}\left[\tau\right]^i}$$

It is worth mentioning that although rate is a well defined quantity for all valid sequences, error exponent is only defined for valid sequences which are not zero-error sequences.

An alternative, relatively robust, definition of channel parameters can be given using valid sequences.

Definition 5 *The cost constraint channel capacity*[13] *for generalized block codes with feedback is given by* $\mathbf{C}[\mathcal{P}] = \sup_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{P}}} R_{\mathcal{Q}}$

Definition 6 *The cost constraint zero error capacity is* $\mathbf{C}_0[\mathcal{P}] = \sup_{\mathcal{Q}_0\in\mathfrak{Q}_{\mathcal{P}}} R_{\mathcal{Q}_0}$

Definition 7 *The reliability function of a channel under cost constraint is*
$E[\mathcal{P},R] = \sup_{\substack{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{P}} \\ R_{\mathcal{Q}}\geq R}} E_{\mathcal{Q}}$

6.2 Derivation of Functional Values

As a result of theorem 1 and 2 for any DMC's we have $\mathbf{C}[\mathcal{P}] \geq \mathbf{C}(\mathcal{P})$. Furthermore for DMCs with zero probability transitions we have $\mathbf{C}_0[\mathcal{P}] \geq \mathbf{C}(\mathcal{P})$ and for DMCs which does not have a zero probability transition we have $E[\mathcal{P},R] \geq \sup_{\eta\in(\eta^*(\mathcal{P},R),1)} E(\mathcal{P},R,\eta)$, where $\eta^*(\mathcal{P},R) = \frac{\mathcal{P}}{{\mathcal{E}_{\mathbf{c}}}^{-1}\left(\frac{\mathcal{P}}{R}\right)}$.

We will continue with giving upper bounds to $\mathbf{C}[\mathcal{P}]$, $\mathbf{C}_0[\mathcal{P}]$ and $E[\mathcal{P},R]$. First note that as result of theorem 3 for any DMC, $\mathbf{C}[\mathcal{P}] \leq \mathbf{C}(\mathcal{P})$. Consequently $\mathbf{C}[\mathcal{P}] = \mathbf{C}(\mathcal{P})$ for all channels and since $\mathbf{C}_0[\mathcal{P}] \leq \mathbf{C}[\mathcal{P}]$, $\mathbf{C}_0[\mathcal{P}] = \mathbf{C}(\mathcal{P})$ for channels with zero probability transitions. Our final task will be proving $E[\mathcal{P},R] \leq \sup_{\eta\in(\eta^*(\mathcal{P},R),1)} E(\mathcal{P},R,\eta)$

---

[13] Cost constraint channel capacity, $\mathbf{C}[\mathcal{P}]$ is defined in terms of 'achievable performance', it should not be confused with, function $\mathbf{C}(\mathcal{P})$ which is merely the supremum of mutual information under cost constraint.

First note that as a result of lemma 6 and 3, one can not have zero error code/sequence on DMCs which does not have zero probability transitions. Thus error exponent is properly defined for all valid sequences for these channels. Remember that

$$\mathcal{V}_1 = \ln M \left(1 - P_e(\ln M - \ln P_e + 1) - \frac{1}{\ln M}\right)$$
$$\mathcal{V}_2 = -\ln P_e \left(1 - \frac{\mathbf{F}}{-\ln P_e} - \frac{\ln(\ln M - \ln P_e + 1)}{-\ln P_e}\right)$$

For any $\mathcal{Q}$ with $E_{\mathcal{Q}} > 0$

$$\liminf_{i\to\infty} \frac{\mathcal{V}_1^i}{\mathbf{E}\left[\tau\right]^i} = R_{\mathcal{Q}} \qquad \liminf_{i\to\infty} \frac{\mathcal{V}_2^i}{\mathbf{E}\left[\tau\right]^i} = E_{\mathcal{Q}}$$

In addition since both $\mathbf{C}(\cdot)$ and $\mathbf{D}(\cdot)$ are continuous and increasing function for any $\mathcal{Q} \in \mathfrak{Q}_{\mathcal{P}}$ we have

$$\limsup_{i\to\infty} \mathbf{C}\left(\frac{\beta}{\eta}\frac{\mathbf{E}\left[\mathcal{S}_\tau\right]^i}{\mathbf{E}\left[\tau\right]^i}\right) = \mathbf{C}\left(\frac{\beta}{\eta}\mathcal{P}\right) \qquad \limsup_{i\to\infty} \mathbf{D}\left(\frac{1-\beta}{1-\eta}\frac{\mathbf{E}\left[\mathcal{S}_\tau\right]^i}{\mathbf{E}\left[\tau\right]^i}\right) = \mathbf{D}\left(\frac{1-\beta}{1-\eta}\mathcal{P}\right)$$

Thus as a result of theorem 4 any $\mathcal{Q} \in \mathfrak{Q}_{\mathcal{P}}$ with $E_{\mathcal{Q}} > 0$ will satisfy

$$\inf_{0<\eta<1, 0\leq\beta\leq 1} \max\left\{\frac{R_{\mathcal{Q}}}{\eta\mathbf{C}\left(\mathcal{P}\frac{\beta}{\eta}\right)}, \frac{E_{\mathcal{Q}}}{(1-\eta)\mathbf{D}\left(\mathcal{P}\frac{1-\beta}{1-\eta}\right)}\right\} \leq 1$$

Evidently this is a joint constraint on possible $(R_{\mathcal{Q}}, E_{\mathcal{Q}})$ pairs. We will try to obtain a constraint on $E_{\mathcal{Q}}$ in terms of $R_{\mathcal{Q}}$ from it. For an $(R_{\mathcal{Q}}, E_{\mathcal{Q}})$ pair to be achievable there must exist an (η, β) pair such that,[14]

$$R_{\mathcal{Q}} \leq \eta\mathbf{C}\left(\mathcal{P}\frac{\beta}{\eta}\right) \quad \text{and} \quad E_{\mathcal{Q}} \leq (1-\eta)\mathbf{D}\left(\mathcal{P}\frac{1-\beta}{1-\eta}\right)$$

We can find a lower bounds on the η values for a given $R_{\mathcal{Q}}$

$$R_{\mathcal{Q}} \leq \eta\mathbf{C}\left(\mathcal{P}\frac{\beta}{\eta}\right) \leq \eta\mathbf{C}\left(\frac{\mathcal{P}}{\eta}\right)$$

---

[14] Indeed we must have a sequence of (η, β) pairs, and state the equations in terms of the limits, but it is conceivable that result will be exactly same as the result of our treatment here.

Thus

$$\frac{\frac{\mathcal{P}}{\eta}}{\mathbf{C}\left(\frac{\mathcal{P}}{\eta}\right)} \leq \frac{\mathcal{P}}{R_{\mathcal{Q}}} \Rightarrow \mathcal{E}_{\mathbf{c}}\left(\frac{\mathcal{P}}{\eta}\right) \leq \frac{\mathcal{P}}{R_{\mathcal{Q}}} \tag{11}$$

Noting that $\mathcal{E}_{\mathbf{c}}(\cdot)$ is a increasing function on $[0, \infty)$ and strictly increasing function on[15] $[\mathcal{P}_0, \infty)$.

$$\frac{\mathcal{P}}{\eta} \leq {\mathcal{E}_{\mathbf{c}}}^{-1}\left(\frac{\mathcal{P}}{R_{\mathcal{Q}}}\right) \Rightarrow \eta \geq \eta^*(R_{\mathcal{Q}}, \mathcal{P}) = \frac{\mathcal{P}}{{\mathcal{E}_{\mathbf{c}}}^{-1}\left(\frac{\mathcal{P}}{R_{\mathcal{Q}}}\right)}$$

For any given $(\mathcal{P}, R_{\mathcal{Q}})$ pair such that $R_{\mathcal{Q}} \leq \mathbf{C}(\mathcal{P})$ and $\eta \leq \eta^*(\mathcal{P}, R_{\mathcal{Q}})$ only the values of $\beta \geq \beta^*(\mathcal{P}, R_{\mathcal{Q}}, \eta) = \frac{\eta}{\mathcal{P}}\mathbf{C}^{-1}\left(\frac{R_{\mathcal{Q}}}{\eta}\right)$, will satisfy $R_{\mathcal{Q}} \leq \eta\mathbf{C}\left(\mathcal{P}\frac{\beta}{\eta}\right)$. Considering $E_{\mathcal{Q}} \leq (1-\eta)\mathbf{D}\left(\mathcal{P}\frac{1-\eta}{1-\beta}\right)$ we get

$$E_{\mathcal{Q}} \leq \sup_{\eta \in \left(\frac{\mathcal{P}}{{\mathcal{E}_{\mathbf{c}}}^{-1}\left(\frac{\mathcal{P}}{R_{\mathcal{Q}}}\right)} 1\right)} (1-\eta)\mathbf{D}\left(\frac{\mathcal{P} - \eta\mathbf{C}^{-1}\left(\frac{R_{\mathcal{Q}}}{\eta}\right)}{1-\eta}\right) \tag{12}$$

Above relation holds trivially for valid sequences with a zero error exponent. Thus for any sequence of rate $R_{\mathcal{Q}} \geq R$, which satisfy the power constraint $\mathcal{P}$, $E_{\mathcal{Q}} \leq E(\mathcal{P}, R)$. Consequently the reliability function of the channel satisfies $E[\mathcal{P}, R] \leq E(\mathcal{P}, R)$. Together with the lower bound we previously derive on $E[\mathcal{P}, R]$

$$E(\mathcal{P}, R) = \sup_{\eta \in \left(\frac{\mathcal{P}}{{\mathcal{E}_{\mathbf{c}}}^{-1}\left(\frac{\mathcal{P}}{R_{\mathcal{Q}}}\right)} 1\right)} E(\mathcal{P}, R, \eta)$$

where

$$E(\mathcal{P}, R, \eta) = (1-\eta)\mathbf{D}\left(\frac{\mathcal{P} - \eta\mathbf{C}^{-1}\left(\frac{R}{\eta}\right)}{1-\eta}\right) \qquad \forall \eta \in \left(\frac{\mathcal{P}}{{\mathcal{E}_{\mathbf{c}}}^{-1}\left(\frac{\mathcal{P}}{R}\right)}, 1\right)$$

It is worth mentioning at this point that if we had imposed a cost constraint on every single member of the sequence $\mathcal{Q}$, rather then the limit, we would have a discontinuity on $E[\mathcal{P}, R]$ for some channels, at $\mathcal{P} = 0$. Reason is that we can not use any symbol of non-zero cost ever, if we impose zero-cost constraint in individual

---

[15] $\mathcal{P}_0 = \arg \inf_{\mathcal{E}_{\mathbf{c}}(\mathcal{P}) \geq \mathcal{E}_{\mathbf{c}}(0)} \mathcal{P}$

codes. However there is no problem of this sort for cost constraints $\mathcal{P}$ greater then zero, both definitions yields to the same result.

Also it should be noted that on some boundary values directly putting in the values or $\mathcal{P}$ and R might not be possible, i.e., might give us undefined values, like the case of $\mathcal{P} > 0$ and $R = 0$. But all of such cases can be solved by thinking of them as the limit of other cases that one can calculate the exponent with the given expression.

6.3 Concavity of $E(\mathcal{P}, R)$

We will prove that the $E(\mathcal{P}, P, \eta)$ is concave function,i.e., it has convex domain and linear interpolation of values of the function between any two points in the domain under estimates the function. Then this will immediately give us the concavity of $E[\mathcal{P}, R] = E(\mathcal{P}, R)$ as the supremum of $E(\mathcal{P}, P, \eta)$ over its domain.

Note that

$$\begin{array}{c} R < \mathbf{C}(\mathcal{P}) \\ \eta \in \left(\frac{\mathcal{P}}{\mathcal{E}_{\mathbf{c}}{}^{-1}\left(\frac{\mathcal{P}}{R}\right)}, 1 \right) \end{array} \iff \begin{array}{c} 0 < \eta < 1 \\ R < \mathbf{C}(\mathcal{P}) \\ R < \eta \mathbf{C}\left(\frac{\mathcal{P}}{\eta}\right) \end{array}$$

Lets assume that $(\mathcal{P}_1, R_1, \eta_1)$ and $(\mathcal{P}_2, R_2, \eta_2)$ is in the domain, then, we will prove that for any $\alpha \in [0, 1]$, $(\mathcal{P}_\alpha, R_\alpha, \eta_\alpha)$ is also in the domain, where

$$\mathcal{P}_\alpha = \alpha\mathcal{P}_1 + (1-\alpha)\mathcal{P}_2 \qquad R_\alpha = \alpha R_1 + (1-\alpha) R_2 \qquad \eta_\alpha = \alpha\eta_1 + (1-\alpha)\eta_2$$

1. $$\begin{array}{c} 0 < \eta_1 < 1 \\ 0 < \eta_2 < 1 \end{array} \Rightarrow 0 < \eta_\alpha < 1$$

2. $$\begin{array}{c} R_1 < \mathbf{C}(\mathcal{P}_1) \\ R_2 < \mathbf{C}(\mathcal{P}_2) \end{array} \Rightarrow R_\alpha < \alpha\mathbf{C}(\mathcal{P}_1) + (1-\alpha)\mathbf{C}(\mathcal{P}_2)$$

   Because of the concavity of the $\mathbf{C}(\cdot)$

   $$\alpha\mathbf{C}(\mathcal{P}_1) + (1-\alpha)\mathbf{C}(\mathcal{P}_2) < \mathbf{C}(\alpha\mathcal{P}_1 + (1-\alpha)\mathcal{P}_2) \Rightarrow R_\alpha < \mathbf{C}(\mathcal{P}_\alpha)$$

3.

$$\begin{array}{l} R_1 < \eta_1 \mathbf{C}\left(\frac{\mathcal{P}_1}{\eta_1}\right) \\ R_2 < \eta_2 \mathbf{C}\left(\frac{\mathcal{P}_2}{\eta_2}\right) \end{array} \Rightarrow R_\alpha < \alpha\eta_1 \mathbf{C}\left(\frac{\mathcal{P}_1}{\eta_1}\right) + (1-\alpha)\mathbf{C}\left(\frac{\mathcal{P}_2}{\eta_2}\right)$$

$$\begin{aligned} R_\alpha &< \eta_\alpha \left(\frac{\alpha\eta_1}{\eta_\alpha}\mathbf{C}\left(\frac{\mathcal{P}_1}{\eta_1}\right) + \frac{(1-\alpha)\eta_2}{\eta_\alpha}\mathbf{C}\left(\frac{\mathcal{P}_2}{\eta_2}\right)\right) \\ &\leq \eta_\alpha \mathbf{C}\left(\frac{\alpha\eta_1}{\eta_\alpha}\frac{\mathcal{P}_1}{\eta_1} + \frac{(1-\alpha)\eta_2}{\eta_\alpha}\frac{\mathcal{P}_2}{\eta_2}\right) \\ &= \eta_\alpha \mathbf{C}\left(\frac{\alpha\mathcal{P}_1 + (1-\alpha)\mathcal{P}_2}{\eta_\alpha}\right) \\ &= \eta_\alpha \mathbf{C}\left(\frac{\mathcal{P}_\alpha}{\eta_\alpha}\right) \end{aligned}$$

Thus any linear combination of the points in the domain will also be point in the domain, i.e., domain of the function $E(\mathcal{P}, R, \eta)$ is convex. We will continue with proving

$$\alpha E(\mathcal{P}_1, R_1, \eta_1) + (1-\alpha)E(\mathcal{P}_2, R_2, \eta_2) \leq E(\mathcal{P}_\alpha, R_\alpha, \eta_\alpha)$$

$$\begin{aligned} &\alpha E(\mathcal{P}_1, R_1, \eta_1) + (1-\alpha)E(\mathcal{P}_2, R_2, \eta_2) \\ &\qquad = \alpha(1-\eta_1)\mathbf{D}\left(\frac{\mathcal{P}_1 - \eta_1\mathbf{C}^{-1}\left(\frac{R_1}{\eta_1}\right)}{1-\eta_1}\right) + (1-\alpha)(1-\eta_2)\mathbf{D}\left(\frac{\mathcal{P}_2 - \eta_2\mathbf{C}^{-1}\left(\frac{R_2}{\eta_2}\right)}{1-\eta_2}\right) \\ &\qquad \leq (1-\eta_\alpha)\mathbf{D}\left(\frac{\alpha\left(\mathcal{P}_1 - \eta_1\mathbf{C}^{-1}\left(\frac{R_1}{\eta_1}\right)\right) + (1-\alpha)\left(\mathcal{P}_2 - \eta_2\mathbf{C}^{-1}\left(\frac{R_2}{\eta_2}\right)\right)}{1-\eta_\alpha}\right) \\ &\qquad = \eta_\alpha\mathbf{D}\left(\frac{\mathcal{P}_\alpha - \left[\alpha\eta_1\mathbf{C}^{-1}\left(\frac{R_1}{\eta_1}\right) + (1-\alpha)\eta_2\mathbf{C}^{-1}\left(\frac{R_2}{\eta_2}\right)\right]}{1-\eta_\alpha}\right) \\ &\qquad \leq \eta_\alpha\mathbf{D}\left(\frac{\mathcal{P}_\alpha - \eta_\alpha\mathbf{C}^{-1}\left(\frac{\alpha R_1 + (1-\alpha)R_2}{\eta_\alpha}\right)}{1-\eta_\alpha}\right) \\ &\qquad = \eta_\alpha\mathbf{D}\left(\frac{\mathcal{P}_\alpha - \eta_\alpha\mathbf{C}^{-1}\left(\frac{R_\alpha}{\eta_\alpha}\right)}{1-\eta_\alpha}\right) \\ &\qquad = E(\mathcal{P}_\alpha, R_\alpha, \eta_\alpha) \end{aligned}$$

Consequently $E(\mathcal{P}, R, \eta)$ is jointly concave in all of its arguments, consequently $E(\mathcal{P}, R)$ which is just the supremum of it is also a convex function. A typical reliability function curve, for a given $\mathcal{P}$ would be

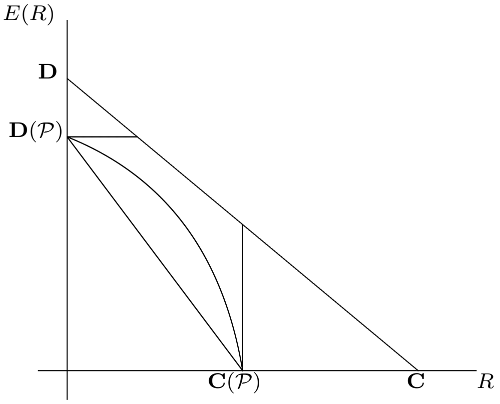

Figure 2: A typical $E(\mathcal{P}, R)$ curve for fixed $\mathcal{P}$

7 Conclusion

We have considered generalized block codes on a DMC with perfect feedback. The conventional additive cost constraint is relaxed in a certain sense, to be compatible with the variable nature of decoding time. Instead of imposing a constraint on expected energy for all decoding instances as in the case of fixed decoding time block codes, we only impose a constraint on the expected energy of one message transmission, in terms of expected decoding time. i.e.,

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

where $\mathbf{E}\left[\mathcal{S}_\tau\right]$ is the expected value of the energy spent in one trial.[16] This approach will allow an uneven average power distribution on different decoding points.

For a DMC that has non-trivial zero transition probabilities, Burnashev, [1] showed that zero error capacity, with generalized block coding, is equal to the channel capacity without feedback, **C**. We extended this result to the cost constraint case as follows. $\forall \mathcal{P} \geq 0$ under the cost constraint $\mathbf{E}\left[\mathcal{S}_\tau\right] \leq \mathcal{P}\mathbf{E}\left[\tau\right]$, zero-error capacity, $\mathbf{C}_0[\mathcal{P}]$ for generalized block codes is equal to $\mathbf{C}(\mathcal{P})$.

---

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

For a DMC that does not have any zero transition probabilities, Burnashev,[1], 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 without cost constraint.

We have generalized his results to the cost constrained case. The reliability function of generalized block coding schemes, $\forall \mathcal{P} \geq 0$ under cost constraint, $\mathbf{E}\left[\mathcal{S}_\tau\right] \leq \mathcal{P}\mathbf{E}\left[\tau\right]$ on a DMC with perfect feedback is a concave function of $(R, \mathcal{P})$ pair given by;

$$E[\mathcal{P}, R] = \sup_{\eta \in \left(\frac{\mathcal{P}}{\mathcal{E}_{\mathbf{c}}{}^{-1}\left(\frac{\mathcal{P}}{R}\right)}, 1\right)} (1-\eta)\mathbf{D}\left(\frac{\mathcal{P} - \eta C^{-1}\left(\frac{R}{\eta}\right)}{1-\eta}\right) \tag{13}$$

Also it is shown that a two phase coding scheme very similar to the one proposed by Yamamoto and Itoh, [15], is asymptotically optimal in both of the cases.

A Proof of Lemmas in 4

A.1 Proof of Lemma 3

Proof:
Decoding time τ is Markov stopping time in the filtration $\mathcal{F}$. The conditional entropy of the messages given the observation $\mathcal{F}_\tau = \mathfrak{f}_\tau$ is

$$\mathcal{H}_{\mathfrak{f}_\tau} = \sum_{n=0}^{\infty} \mathcal{H}_{\mathfrak{f}_n} \mathbb{I}_{\{\tau = n\}}$$

where $\mathbb{I}_{\{q\}}$ is the indicator function for the event q. Since $\mathcal{H}_{\mathfrak{f}_n}$ is a bounded random variable and $\mathbf{P}\left[\tau < \infty\right] = 1$, $\mathbf{E}\left[\mathcal{H}_{\mathfrak{f}_\tau}\right]$ can be written as

$$\mathbf{E}\left[\mathcal{H}_{\mathfrak{f}_\tau}\right] = \lim_{N \to \infty} \sum_{n=0}^{N} \mathbf{E}\left[\mathcal{H}_{\mathfrak{f}_n} \middle| \, \tau = n\right] \mathbf{P}\left[\tau = n\right] \tag{14}$$

One can use the conventional Fano inequality to get

$$\mathcal{H}_{\mathfrak{f}_n} \leq \mathfrak{h}(P_e(\mathfrak{f}_n)) + P_e(\mathfrak{f}_n) \ln(M-1) \tag{15}$$

where $P_e(\mathfrak{f}_n)$ is the probability of error of any detector we can use at the point $\mathcal{F}_n = \mathfrak{f}_n$.

We can write probability of error as

$$P_e = \mathbf{E}\left[P_e(\mathfrak{f}_n)\mathbb{I}_{\{\mathcal{F}_\tau = \mathfrak{f}_n\}}\right] = \sum_{\mathfrak{f}_\tau} P_e(\mathfrak{f}_\tau)\mathbf{P}\left[\mathcal{F}_\tau = \mathfrak{f}_\tau\right]$$

Using equations (14), (15), (16), together with the concavity of the binary entropy function, $\mathfrak{h}(x) = -x\ln x - (1-x)\ln(1-x)$ we get equation (6).

QED

A.2 Proof of Lemma 4

Proof:

$$\begin{aligned}
\mathbf{E}\left[V_n^{\mathcal{P}}\middle|\,\mathcal{F}_n\right] &= \mathbf{E}\left[H(\theta|\mathcal{F}_n = \mathfrak{f}_n) + \gamma_{\mathbf{C}}^{\mathcal{P}}\mathcal{S}_n|\,\mathcal{F}_n\right] \\
\mathbf{E}\left[V_{n+1}^{\mathcal{P}}\middle|\,\mathcal{F}_n\right] &= \mathbf{E}\left[H(\theta|\mathcal{F}_n = \mathfrak{f}_n, Y_{n+1} = l_0) + \gamma_{\mathbf{C}}^{\mathcal{P}}(\mathcal{S}_n + \rho_{X_{n+1}})\middle|\,\mathcal{F}_n\right] \\
\mathbf{E}\left[V_n - V_{n+1}|\,\mathcal{F}_n\right] &= I(\theta; Y_{n+1}|\mathcal{F}_n = \mathfrak{f}_n) - \gamma_{\mathbf{C}}^{\mathcal{P}}\mathbf{E}\left[\rho_{X_{n+1}}\middle|\,\mathcal{F}_n\right]
\end{aligned}$$

Because of Markov relation $\theta \leftrightarrow X_{n+1} \leftrightarrow Y_{n+1}$ and data processing inequality we have

$$\begin{aligned}
\mathbf{E}\left[V_n - V_{n+1}|\,\mathcal{F}_n = \mathfrak{f}_n\right] &\leq I(X_{n+1}; Y_{n+1}|\mathcal{F}_n = \mathfrak{f}_n) - \gamma_{\mathbf{C}}^{\mathcal{P}}\mathbf{E}\left[\rho_{X_{n+1}}\middle|\,\mathcal{F}_n\right] \\
&= \mathcal{L}^{\mathbf{C}}(\mathcal{P}, \varphi, \gamma_{\mathbf{C}}^{\mathcal{P}}) - \gamma_{\mathbf{C}}^{\mathcal{P}}\mathcal{P}
\end{aligned}$$

Using the fact fact $\max_\varphi \mathcal{L}^{\mathbf{C}}(\mathcal{P}, \varphi, \gamma_{\mathbf{C}}^{\mathcal{P}}) = \mathbf{C}(\mathcal{P})$ we get

$$\mathbf{E}\left[V_n - V_{n+1}|\,\mathcal{F}_n = \mathfrak{f}_n\right] \leq \mathbf{C}(\mathcal{P}) - \gamma_{\mathbf{C}}^{\mathcal{P}}\mathcal{P}$$

QED

A.3 Proof of Lemma 5

Proof:
Let us introduce the short hand

$$\begin{aligned} p(i) &= p_i(\mathfrak{f}_n) & p(i|l) &= \mathbf{P}\left[\theta = i \middle|\, Y_{n+1} = l, \mathcal{F}_n = \mathfrak{f}_n\right] \\ \varphi(k|i) &= \mathbf{P}\left[X_{n+1} = k \middle|\, \mathcal{F}_n = \mathfrak{f}_n, \theta = i\right] & \varphi(k) &= \mathbf{P}\left[X_{n+1} = k \middle|\, \mathcal{F}_n = \mathfrak{f}_n\right] \\ \psi(l|i) &= \mathbf{P}\left[Y_{n+1} = l \middle|\, \mathcal{F}_n = \mathfrak{f}_n, \theta = i\right] & \psi(l) &= \mathbf{P}\left[Y_{n+1} = l \middle|\, \mathcal{F}_n = \mathfrak{f}_n\right] \end{aligned}$$

Then

$$\mathbf{E}\left[\ln \mathcal{H}_{\mathfrak{f}_n} - \ln \mathcal{H}_{\mathfrak{f}_{n+1}} \middle|\, \mathcal{F}_n = \mathfrak{f}_n\right] = \sum_{l=1}^{L} \psi(l) \ln \frac{-\sum_{i=1}^{M} p(i) \ln p(i)}{-\sum_{i=1}^{M} p(i|l) \ln p(i|l)}$$

Using log-sum inequality

$$\mathbf{E}\left[\ln \mathcal{H}_{\mathfrak{f}_n} - \ln \mathcal{H}_{\mathfrak{f}_{n+1}} \middle|\, \mathcal{F}_n = \mathfrak{f}_n\right] \leq \sum_i \frac{-p(i) \log p(i)}{\sum_j -p(j) \log p(j)} \sum_l \psi(l) \ln \frac{-p(i) \ln p(i)}{-p(i|l) \ln p(i|l)}$$

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

$$\begin{aligned} \sum_l \psi(l) \ln \frac{-p(i) \ln p(i)}{-p(i|l) \ln p(i|l)} &= \sum_l \psi(l) \ln \left(\frac{\psi(l)}{\psi(l|i)} \frac{\ln 1/p(i)}{\ln \frac{\psi(l)}{p(i)\psi(l|i)}} \right) \\ &= \sum_l \left(p(i)\psi(l|i) + (1 - p(i))\psi(l|\bar{i})\right) \ln \left(\frac{\psi(l)}{\psi(l|i)} \frac{\ln 1/p(i)}{\ln \frac{\psi(l)}{p(i)\psi(l|i)}} \right) \\ &= p(i) \sum_l \psi(l|i) \ln \frac{\psi(l|i)}{\psi(l|\bar{i})} + (1 - p(i)) \sum_l \psi(l|\bar{i}) \ln \frac{\psi(l|\bar{i})}{\psi(l|i)} \\ &+ p(i) \sum_l \psi(l|i) \ln \left(\frac{\psi(l)}{\psi(l|i)} \frac{\psi(l|\bar{i})}{\psi(l|i)} \frac{\ln 1/p(i)}{\ln \frac{\psi(l)}{p(i)\psi(l|i)}} \right) \\ &+ (1 - p(i)) \sum_l \psi(l|\bar{i}) \ln \left(\frac{\psi(l)}{\psi(l|\bar{i})} \frac{\ln 1/p(i)}{\ln \frac{\psi(l)}{p(i)\psi(l|i)}} \right) \end{aligned}$$

We will prove that the last two terms are in deed less then zero.

$$\begin{aligned}\sum_l \psi(l|i) \ln \left(\frac{\psi(l)}{\psi(l|i)} \frac{\psi(l|\bar{i})}{\psi(l|i)} \frac{\ln 1/p(i)}{\ln \frac{\psi(l)}{p(i)\psi(l|i)}} \right) &= -\sum_l \psi(l|i) \ln \frac{\psi(l|i)}{\psi(l)} \\ &\quad - \sum_l \psi(l|i) \ln \frac{\psi(l|i)}{\psi(l|\bar{i})} \frac{\ln \left(1 + \frac{1-p(i)}{p(i)} \frac{\psi(l|\bar{i})}{\psi(l|i)}\right)}{\ln \left(1 + \frac{1-p(i)}{p(i)}\right)} \\ &\leq 0\end{aligned}$$

Where we used the fact $\log\left(\frac{1}{x}\log(1+x)\right)$ is a convex function and Kullback Leibler divergence is always non-negative in the last step

$$\begin{aligned}\sum_l \psi(l|\bar{i}) \ln \left(\frac{\psi(l)}{\psi(l|\bar{i})} \frac{\ln 1/p(i)}{\ln \frac{\psi(l)}{p(i)\psi(l|i)}} \right) &\leq -\sum_l \psi(l|\bar{i}) \ln \frac{\psi(l|\bar{i})}{\psi(l)} \\ &\quad - \sum_l \psi(l|\bar{i}) \ln \frac{\ln \left(1 + \frac{1-p(i)}{p(i)} \frac{\psi(l|\bar{i})}{\psi(l|i)}\right)}{\ln \left(1 + \frac{1-p(i)}{p(i)}\right)} \\ &\leq 0\end{aligned}$$

Where we used the fact $\log\left(\log(1+\frac{1}{x})\right)$ is a convex function and Kullback Leibler divergence is always non-negative in the last step

Consequently we have

$$\sum_l \psi(l) \ln \frac{-p(i)\ln p(i)}{-p(i|l)\ln p(i|l)} \leq p(i) \sum_l \psi(l|i) \ln \frac{\psi(l|i)}{\psi(l|\bar{i})} + (1-p(i)) \sum_l \psi(l|\bar{i}) \ln \frac{\psi(l|\bar{i})}{\psi(l|i)}$$

Note that because of convexity of Kullback Leibler Divergence and definition of $\mathbf{D}_i$ we have

$$p(i) D(\psi(\cdot|i)||\psi(\cdot|\bar{i})) + (1-p(i)) D(\psi(\cdot|\bar{i})||\psi(\cdot|i)) \leq \sum_k \varphi(k) \mathbf{D}_k$$

Thus

$$\mathbf{E}\left[\ln \mathcal{H}_{\mathfrak{f}_n} - \ln \mathcal{H}_{\mathfrak{f}_{n+1}} \middle| \mathcal{F}_n\right] \leq \sum_k \varphi(k) \mathbf{D}_k$$

Noting that $W_n^{\mathcal{P}} = \ln \mathcal{H}_{\mathfrak{f}_n} + \gamma_{\mathbf{D}}^{\mathcal{P}} \mathbf{E}\left[\mathcal{S}_n \middle| \mathcal{F}_n\right]$

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

where we have used the fact $\max_\varphi \mathcal{L}^{\mathbf{D}}(\mathcal{P}, \varphi, \gamma_{\mathbf{D}}^{\mathcal{P}}) = \mathbf{D}(\mathcal{P})$ in the last step.

QED

A.4 Proof of Lemma 6

Proof:

$$\ln \mathcal{H}_{\mathfrak{f}_n} - \ln \mathcal{H}_{\mathfrak{f}_{n+1}} = \ln \frac{-\sum_{i=1}^{M} p(i) \ln p(i)}{-\sum_{i=1}^{M} p(i|l) \ln p(i|l)}$$

Evidently

$$\min_i \frac{-p(i) \ln p(i)}{-p(i|l) \ln p(i|l)} \leq \frac{\mathcal{H}_{\mathfrak{f}_n}}{\mathcal{H}_{\mathfrak{f}_{n+1}}} \leq \max_i \frac{-p(i) \ln p(i)}{-p(i|l) \ln p(i|l)} \tag{16}$$

Note that $\psi(l) = p(i)\psi(l|i) + (1 - p(i))\psi(l|\bar{i})$

$$\frac{-p(i) \ln p(i)}{-p(i|l) \ln p(i|l)} = \frac{\psi(l)}{\psi(l|i)} \frac{\ln 1/p(i)}{\ln\left(1 + \frac{1-p(i)}{p(i)} \frac{\psi(l|\bar{i})}{\psi(l|i)}\right)}$$

If $\psi(l|\bar{i}) \geq \psi(l|i)$ then

$$\frac{\psi(l)}{\psi(l|\bar{i})} \leq \frac{\psi(l)}{\psi(l|i)} \frac{\ln 1/p(i)}{\ln\left(1 + \frac{1-p(i)}{p(i)} \frac{\psi(l|\bar{i})}{\psi(l|i)}\right)} \leq \frac{\psi(l)}{\psi(l|i)}$$

If $\psi(l|\bar{i}) \leq \psi(l|i)$ then

$$\frac{\psi(l)}{\psi(l|i)} \leq \frac{\psi(l)}{\psi(l|i)} \frac{\ln 1/p(i)}{\ln\left(1 + \frac{1-p(i)}{p(i)} \frac{\psi(l|\bar{i})}{\psi(l|i)}\right)} \leq \frac{\psi(l)}{\psi(l|\bar{i})}$$

Consequently

$$\min_{i,k} \ln \frac{P_{kl}}{P_{il}} \leq \ln \frac{-p(i)\ln p(i)}{-p(i|l)\ln p(i|l)} \leq \max_{i,k} \ln \frac{P_{kl}}{P_{il}}$$

Thus if $Y_{n+1} = l$

$$|\ln \mathcal{H}_{\mathfrak{f}_n} - \ln \mathcal{H}_{\mathfrak{f}_{n+1}}| \leq \max_{i,k} \ln \frac{P_{kl}}{P_{il}}$$

QED

A.5 Proof of Lemma 7

Proof:
Consider the following stochastic process

$$\begin{aligned} \xi_n &= \Gamma_n + Rn - [\Gamma_{n\wedge\tau_i} + R(n \wedge \tau_i)] \\ &= [\Gamma_n - \Gamma_{\tau_i} + R(n - \tau_i)]\mathbb{I}_{\{\tau_i \leq n\}} \end{aligned}$$

Note that

$$\begin{aligned} \xi_{n+1} - \xi_n &= [\Gamma_{n+1} - \Gamma_{\tau_i} + R(n+1-\tau_i)](\mathbb{I}_{\{\tau_i \leq n\}} + \mathbb{I}_{\{\tau_i = n+1\}}) - [\Gamma_n - \Gamma_{\tau_i} + R(n-\tau_i)]\mathbb{I}_{\{\tau_i \leq n\}} \\ &= [\Gamma_{n+1} - \Gamma_n + Rn]\mathbb{I}_{\{\tau_i \leq n\}} + [\Gamma_{n+1} - \Gamma_{\tau_i} + R(n+1-\tau_i)]\mathbb{I}_{\{\tau_i = n+1\}} \\ &= [\Gamma_{n+1} - \Gamma_n + Rn]\mathbb{I}_{\{\tau_i \leq n\}} \qquad (17) \end{aligned}$$

Thus $\mathbf{E}\left[|\xi_n - \xi_{n+1}|\right| \mathcal{F}_n] < K + |R|$ and $\mathbf{E}\left[\xi_n - \xi_{n+1}\right| \mathcal{F}_n] \leq 0$ Since $\mathbf{E}\left[\tau_f\right] < \infty$ the conditions of theorem 2 in [13] p459, will hold for ξ_n and τ_f. Thus

$$\mathbf{E}\left[\xi_0 - \xi_{\tau_f}\middle| \mathcal{F}_0\right] \leq 0$$

Using the fact $\mathbf{P}\left[\tau_f \geq \tau_i\right] = 1$ and $\tau_i \geq 0$ we can extend this result for τ_i and τ_f

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

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

QED

B 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 $\ln|\mathcal{Y}|$ should be enough. Instead of proving this result we will prove a stronger result.[17]

Let us assume the feedback channel is not perfect, in the sense that it is error free but it has only a finite capacity, which is equal to the capacity of the forward channel[18] $\mathbf{C}$. Furthermore assume that the feedback channel has a constant delay, $\mathbf{T}$. Under this assumptions the decisions of the receiver will not be revealed to the transmitter instantly. Accordingly we will end the phases when these decisions are revealed to the transmitter rather than the instances they are made. We will call the interval, between the transmission of the first symbol of a trial and the decoding of the temporary estimate of the decoder by transmitter as the first phase. The rest of the time until the instance that the transmitter learns the decision of decoder about permanent decoding or retransmission will be called second phase.

Sending back the estimate will take an appreciable amount of time, thus we want to start sending back the estimate as soon as possible. Accordingly we will send the message in parts, so that receiver can decode the parts and send them back while receiving the new parts.

Let $M = N_s^{N_r}$. We can use the coding theorem to argue that, for every $\mathcal{P} \geq 0$ and every $\delta > 0$, there exists a $\epsilon(\delta) > 0$ such that for large enough ℓ_s, one can find block codes such that

$$N_s \geq e^{\ell_s(\mathbf{C}(\mathcal{P})-\delta)} \quad \text{and} \quad P_e \leq e^{-\epsilon(\delta)\ell_s}$$

We can use the same code N_r times to send the over all message, while sending back the estimate about its parts as they are decoded. Then one can argue that the duration of the first phase ℓ_1 will be

$$\ell_1 = (N_r + 1)\ell_s + \mathbf{T}$$

---

[17] This result is also a simple consequence of the work of Şimşek and Sahahi, [9].

[18] Note that $\mathbf{C} \geq \mathbf{C}(\mathcal{P})$.

After little bit of algebra these expressions can be put in to the following form. For any $\mathcal{P} \geq 0$ and $\delta > 0$ there exists $\epsilon(\delta) > 0$ such that for large enough ℓ_1 a coding algorithm and a scheduling scheme can be found such that

$$M \geq e^{\ell_1(\mathbf{C}(\mathcal{P})-\delta)} \quad \text{and} \quad P_e \leq e^{-\ell_1 \epsilon(\delta)}$$

Similarly one can write the duration of the second phase in terms of delay, $\mathbf{T}$ and the block-length of the hypothesis testing code ℓ_h as

$$\ell_2 = \ell_h + \frac{\ln 2}{\mathbf{C}} + \mathbf{T}$$

After obtaining this expression it is evident that we can do exactly the same calculations we have done before to reach the achievability result we get even under finite delay, and limited feedback capacity.

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. 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 the rate of the first phase, $\mathbf{C}\left(\frac{\beta}{\eta}\mathcal{P}\right)$. This rate will be different for different R's.

C Properties of $\mathbf{C}(\mathcal{P})$ and $\mathbf{D}(\mathcal{P})$

The functions $\mathbf{C}(\mathcal{P})$ and $\mathbf{D}(\mathcal{P})$ are defined as

$$\mathbf{C}(\mathcal{P}) = \max_{\substack{\varphi \\ \sum_{k=1}^{K} \varphi_k \rho_k \leq \mathcal{P}}} \mathcal{I}_{\mathfrak{P}}(\varphi) \quad \& \quad \mathbf{D}(\mathcal{P}) = \max_{\substack{\varphi \\ \sum_{k=1}^{K} \varphi_k \rho_k \leq \mathcal{P}}} \sum_k \varphi_k \mathbf{D}_k$$

where

$$\mathcal{I}_{\mathfrak{P}}(\varphi) = \sum_{k=1,l=1}^{K,L} \varphi_k P_{kl} \ln \frac{P_{kl}}{\sum_{j=1}^{K} \varphi_j P_{jl}} \quad \& \quad \mathbf{D}_i = \max_i \sum_{l=1}^{L} P_{il} \log \frac{P_{il}}{P_{jl}}$$

We can write the cost constraints as minimizations, as max-min problems as follows.

$$\mathbf{C}(\mathcal{P}) = \max_{\varphi} \min_{\gamma_{\mathbf{C}} \geq 0} \mathcal{L}^{\mathbf{C}}(\mathcal{P}, \varphi, \gamma_{\mathbf{C}}) \quad \& \quad \mathbf{D}(\mathcal{P}) = \max_{\varphi} \min_{\gamma_{\mathbf{D}} \geq 0} \mathcal{L}^{\mathbf{D}}(\mathcal{P}, \varphi, \gamma_{\mathbf{D}})$$

where

$$\mathcal{L}^{\mathbf{C}}(\mathcal{P}, \varphi, \gamma_{\mathbf{C}}) = \mathcal{I}_{\mathfrak{P}}(\varphi) + \gamma_{\mathbf{C}}(\mathcal{P} - \sum_k \varphi_k \rho_k) \quad \& \quad \mathcal{L}^{\mathbf{D}}(\mathcal{P}, \varphi, \gamma_{\mathbf{D}}) = \sum_k \rho_k \mathbf{D}_k + \gamma_{\mathbf{D}}(\mathcal{P} - \sum_k \varphi_k \rho_k)$$

It is known[19] that $\mathcal{L}$ has a saddle point if φ has a compact set constraint, and there exists a φ^0 with a corresponding R such that $\{\gamma | \mathcal{L}(\varphi^0, \gamma) < R\}$ 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

$$\mathbf{C}(\mathcal{P}) = \min_{\gamma_{\mathbf{C}} \geq 0} \max_{\varphi} \mathcal{L}^{\mathbf{C}}(\mathcal{P}, \varphi, \gamma_{\mathbf{C}}) \quad \& \quad \mathbf{D}(\mathcal{P}) = \min_{\gamma_{\mathbf{D}} \geq 0} \max_{\varphi} \mathcal{L}^{\mathbf{D}}(\mathcal{P}, \varphi, \gamma_{\mathbf{D}})$$

Thus

$$\mathbf{C}(\mathcal{P}) = \max_{\varphi} \mathcal{L}^{\mathbf{C}}(\mathcal{P}, \varphi, \gamma_{\mathbf{C}}^{\mathcal{P}}) \quad \& \quad \mathbf{D}(\mathcal{P}) = \max_{\varphi} \mathcal{L}^{\mathbf{D}}(\mathcal{P}, \varphi, \gamma_{\mathbf{D}}^{\mathcal{P}}) \tag{18}$$

where

$$\gamma_{\mathbf{C}}^{\mathcal{P}} = \arg\min \max_{\varphi} \mathcal{L}^{\mathbf{C}}(\mathcal{P}, \varphi, \gamma_{\mathbf{C}}) \quad \& \quad \gamma_{\mathbf{D}}^{\mathcal{P}} = \arg\min \max_{\varphi} \mathcal{L}^{\mathbf{D}}(\mathcal{P}, \varphi, \gamma_{\mathbf{D}}) \tag{19}$$

Since $\mathcal{L}(\mathcal{P}, \phi, \gamma)$ is linear in $\mathcal{P}$ for any given γ, $\mathbf{C}(\mathcal{P})$ and $\mathbf{D}(\mathcal{P})$ will also be concave in $\mathcal{P}$. Thus for all values of $\mathcal{P}$ that results in a $\mathbf{C}(\mathcal{P}) < \mathbf{C}$; $\mathbf{C}(\mathcal{P})$ will be strictly increasing function. Similarly all values of $\mathcal{P}$ that results in a $\mathbf{D}(\mathcal{P}) < \mathbf{D}$; $\mathbf{D}(\mathcal{P})$ will be strictly increasing function.

Lemma 8 *For any DMC with associated cost on input letters;*

- $\mathbf{C}(\mathcal{P})$ *and* $\mathbf{D}(\mathcal{P})$ *are concave, nondecreasing, and non-negative. Furthermore* $\mathbf{C}(\mathcal{P})$ *and* $\mathbf{D}(\mathcal{P})$ *are strictly increasing for values below* $\mathbf{C}$ *and* $\mathbf{D}$ *respectively.*
- $\mathbf{D}(\mathcal{P})$ *is also piecewise linear and positive.*

---

[19] Bertsekas [14] Saddle Point Theorem, Proposition 2.6.9 (case 2), pp151.

- $\frac{1}{\mathbf{C}(\mathcal{P})}$ *and* $\frac{1}{\mathbf{D}(\mathcal{P})}$ *are convex, non-increasing, positive. They are strictly decreasing and strictly concave for values above* $\frac{1}{\mathbf{C}}$ *and* $\frac{1}{\mathbf{D}}$ *respectively.*
- $\mathcal{E}_{\mathbf{d}}(\mathcal{P}) = \frac{\mathcal{P}}{\mathbf{D}(\mathcal{P})}$ *is strictly increasing, non-negative, everywhere.*
- $\mathcal{E}_{\mathbf{c}}(\mathcal{P}) = \frac{\mathcal{P}}{\mathbf{C}(\mathcal{P})}$ *is strictly increasing on the interval* $[\mathcal{P}_0, \infty)$*, and constant on* $[0, \mathcal{P}_0]$*, where* $\mathcal{P}_0 = \inf_{\mathcal{E}_{\mathbf{c}}(x) > \mathcal{E}_{\mathbf{c}}(0)} x$

Proof:
Note that first three assertions have already been proved, except $\mathbf{D}(\mathcal{P})$ being piecewise linear which is obvious. For the last two we need to consider a specific class of functions. Namely non-negative, non-decreasing, concave function $f(x)$ of x. For any such function $f(x)$, because of the concavity $\forall x_0$, $\exists\ d_{x_0}$ such that

$$f(x) \leq f(x_0) + d_{x_0}(x - x_0) \qquad \forall x$$

Applying the identity for $x = 0$ we can find a bound on d_{x_0} as follows

$$d_{x_0} \leq \frac{f(x_0) - f(0)}{x_0}$$

Let $x_1 > x_0$

$$\begin{aligned} \frac{f(x_1)}{x_1} &\leq \frac{f(x_0) + d_{x_0}(x_1 - x_0)}{x_1} \\ &\leq \frac{f(x_0) + \frac{f(x_0)-f(0)}{x_0}(x_1 - x_0)}{x_1} \\ &= \frac{f(x_0)}{x_0} - f(0)\frac{x_1 - x_0}{x_0 x_1} \end{aligned}$$

Consequently $\frac{x}{f(x)}$ is a non-decreasing function.

It is strictly-increasing everywhere if $f(0) > 0$. Since $\mathbf{D}(0) > 0$, $\mathcal{E}_{\mathbf{d}}(\mathcal{P})$ is a strictly increasing function.

In addition $\frac{x}{f(x)}$ is strictly increasing on any point x, such that $d_x < \frac{f(x)-f(0)}{x}$. Thus $\mathcal{E}_{\mathbf{c}}(\mathcal{P})$ is strictly increasing if $\mathbf{C}(\cdot)$ is strictly concave at $\mathcal{P} = 0$ or $\mathbf{C}(0) > 0$. If $\mathbf{C}(0) = 0$ and if $\mathbf{C}(\cdot)$ in not strictly concave at $\mathcal{P} = 0$, $\mathbf{C}(\mathcal{P})$ will be linear on an on interval $[0, \mathcal{P}_0]$, where $\mathcal{P}_0 = \inf_{\mathcal{E}_{\mathbf{c}}(\mathcal{P}) > \mathcal{E}_{\mathbf{c}}(0)} \mathcal{P}$.

Although $\mathbf{C}(\mathcal{P})$ is strictly concave in most of the examples, there are channels such that $\mathbf{C}(\cdot)$ is linear at least for an interval of the form $[0, \mathcal{P}_0]$.

QED

References

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

[2] E.R. Berlekamp C.E. Shannon, R.G. Gallager. Lower bounds to error probability for coding on discrete memoryless channels. *Information and Control*, 10, No. 1:65–103, 1967.

[3] 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.

[4] Robert G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, Inc., New York, NY, USA, 1968.

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

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

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

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

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

[15] 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.