Error Exponents for Variable-Length Block Codes With Feedback and Cost Constraints

Barış Nakiboğlu and Robert G. Gallager, *Life Fellow, IEEE*

***Abstract*—Variable-length block-coding schemes are investigated for discrete memoryless channels with ideal feedback under cost constraints. Upper and lower bounds are found for the minimum achievable probability of decoding error $P_{e,\min}$ as a function of constraints $R, \mathcal{P}$, and $\overline{\tau}$ on the transmission rate, average cost, and average block length, respectively. For given R and $\mathcal{P}$, the lower and upper bounds to the exponent $-(\ln P_{e,\min})/\overline{\tau}$ are asymptotically equal as $\overline{\tau} \to \infty$. The resulting reliability function, $\lim_{\overline{\tau}\to\infty}(-\ln P_{e,\min})/\overline{\tau}$, as a function of R and $\mathcal{P}$, is concave in the pair $(R, \mathcal{P})$ and generalizes the linear reliability function of Burnashev to include cost constraints. The results are generalized to a class of discrete-time memoryless channels with arbitrary alphabets, including additive Gaussian noise channels with amplitude and power constraints.**

***Index Terms*—Block codes, cost constraints, feedback, memoryless channels, variable-length communication.**

I. Introduction

THE information-theoretic effect of feedback in communication that has been studied since Shannon [15] showed in 1956 that feedback cannot increase the capacity $\mathcal{C}$ of a discrete memoryless channel (DMC). At about the same time, Elias [4] and Chang [14] gave examples showing that feedback could greatly simplify error correction at rates below capacity.

This paper, as well as much of the existing literature on feedback communication, is restricted to block coding, i.e., coding in which messages are transmitted sequentially and each message is completely decoded and released to the destination before transmission of the next message begins. Non-block codes, with overlapping messages, raise a somewhat different and complimentary set of conceptual issues, as discussed in an excellent paper by Sahai [10].

Block coding for feedback communication can be further separated into fixed-length and variable-length coding. The codewords in a fixed-length block code all have the same length, but, due to the feedback, the symbols in each codeword can depend on previous channel outputs as well as the choice of transmitted message. For variable-length block codes, the decoding time can also depend dynamically on the previously received symbols. We assume that the feedback is ideal, meaning that it is noiseless, instantaneous, and of unlimited capacity. Thus, we can assume that all information available at the receiver is also available at the transmitter, and consequently, the transmitter can determine when the receiver decodes each message. The assumption of ideal feedback is unrealistic, of course, but we feel a thorough understanding of this case will play a major role in studying the far more complex problem of nonideal feedback.

A widely used quality criterion for fixed-length block codes of a given rate is the error exponent, $\frac{-\ln P_e}{\tau}$, where P_e is the probability of decoding error and τ is the block length. Dobrushin [3] showed that the sphere-packing exponent (the well-known upper bound to the error exponent without feedback) is also an upper bound for fixed-length block coding with feedback on symmetric DMCs. It has been long conjectured that this is also true for nonsymmetric DMCs, but the current best upper bound, by Haroutunian [6], is larger than the sphere-packing bound in the nonsymmetric case.

Variable-length block coding allows the decoding to be delayed under unusually severe noise, thus usually providing a dramatic increase in error exponent. As motivated in the discussion following Theorem 1, the error exponent for a variable-length block code is defined as $\frac{-\ln P_e}{\overline{\tau}}$ where $\overline{\tau}$ is the expected block length. Similarly, the rate is defined as $\frac{\ln M}{\overline{\tau}}$ where M is the size of the message set. Since successive messages require independent and identically distributed message transmission times, this rate (converted to base 2) is the long-term rate at which message bits can be transferred to the receiver.

The reliability function $E(R)$, for a class of coding schemes on a given channel, is defined as the asymptotic maximum achievable exponent, as $\overline{\tau} \to \infty$, for codes of rate greater than or equal to R. Burnashev [1] developed upper and lower bounds to $E(R)$ for variable-length block codes on DMCs with ideal feedback. For DMCs in which all outputs can be reached with positive probability from all inputs, Burnashev's upper and lower bounds to $E(R)$ are equal. The resulting function $E(R)$ is linear, going from a positive constant at $R = 0$ to 0 at $R = \mathcal{C}$.

For DMCs in which at least one output can be reached from only a proper subset of inputs, Burnashev implicitly showed that $P_e = 0$ is asymptotically achievable for all $R < \mathcal{C}$. This means that $\mathcal{C}$ is the zero-error capacity of variable-length block codes for such DMCs. Thus, the zero-error capacity for variable-length block codes with feedback can be strictly larger than that for fixed-length codes with feedback, which in turn can be strictly larger than zero-error capacity without feedback.

The main objective of this paper is to generalize Burnashev's results to DMCs subject to a cost criterion. That is, a nonnegative cost $\rho_k \geq 0$ is associated with each letter k of the channel input alphabet, $\{0, 1, \ldots, |\mathcal{X}|-1\}$. It is assumed[1] that $\rho_k = 0$ for at least one choice of k. The energy in a

Manuscript received December 18, 2006; revised October 11, 2007. This work was supported in part by the National Science Foundation under Grants ANI-0335256 and by DARPA ITMANET program.

The authors are with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: nakib@mit.edu; gallager@mit.edu).

Communicated by G. Kramer, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2007.915913

[1]The assumption that the minimum cost symbol has cost 0 causes no loss of generality, since otherwise the minimum cost could be trivially subtracted from all symbol costs.

codeword $X_1, X_2, \ldots, X_\tau$, where X_n is transmitted at time n, $1 \leq n \leq \tau$ and τ is the decoding time, is defined to be $S_\tau = \rho_{X_1} + \cdots + \rho_{X_\tau}$. As explained more fully later, a variable-length block code is defined to satisfy an average cost (power) constraint $\mathcal{P} \geq 0$ if $\boldsymbol{E}[S_\tau] \leq \boldsymbol{E}[\tau]\mathcal{P}$. We will find the corresponding reliability function for all $\mathcal{P} \geq 0$. For all DMCs whose transition probabilities are all positive, this reliability function is a concave function of $(R, \mathcal{P})$. If zero transition probabilities exist, then zero error probability can be achieved at all rates below the cost constrained capacity.

Our interest in cost criteria for DMCs is motivated by the desire to separate the effect of cost constraints from that of infinite alphabet size, thus allowing a better understanding of channels such as additive Gaussian noise where these effects are combined. Pinsker [9] considered fixed-length codes for the discrete-time additive white Gaussian noise channel (AWGNC) with feedback. He showed that the sphere-packing exponent upper-bounds the error exponent if the energy of each codeword has a fixed upper bound, independent of the noise sample values. Schalkwijk [13] considered the same model but allowed the codeword energy to depend on the noise, subject to an average energy constraint. He developed a simple algorithm for which the error probability decays as a twofold exponential of the block length (and thus also of the energy). This was an extension of joint work with Kailath [12] where the infinite bandwidth limit of the problem was considered. Kramer [7] later showed that the error probability could be made to decay n-fold exponentially for any n for the infinite bandwidth case, and it is not hard to slightly strengthen these results for both finite and infinite bandwidth.

In the following section, we consider a class of variable-length block codes for DMCs with feedback and cost constraints. These generalize the Yamamoto and Itoh [18] codes to allow for cost constraints. We lower-bound the achievable error exponent for these codes as a function of constraints $R, \mathcal{P}$, and $\overline{\tau}$ on rate, average cost, and average block length, respectively.

In Section III, we consider all possible variable-length block codes and derive a lower bound on $\overline{\tau}$ as a function of power constraint $\mathcal{P}$, average error probability P_e, and message-set size M. This is then converted into an upper bound on the error exponent over all codes of given $R, \mathcal{P}$, and $\overline{\tau}$. We show that as $\overline{\tau} \to \infty$, this upper bound coincides with the lower bound of Section II, thus determining the reliability function in the presence of a cost constraint.

In Section IV, the results are generalized to a broader class of discrete-time memoryless channels that includes AWGNCs with both power and amplitude constraints.

II. Achievability: Asymptotically Optimum Codes

A. Forward and Feedback Channel Models and Cost Constraint

The forward channel is assumed to be a DMC of positive capacity with input alphabet $\mathcal{X} = \{0, \ldots, |\mathcal{X}|-1\}$ and output alphabet $\mathcal{Y} = \{0, \ldots, |\mathcal{Y}|-1\}$. The input and output at time n are denoted by X_n and Y_n; the n-tuples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are denoted by X^n and Y^n. The feedback channel is *ideal* in the sense that it is discrete and noiseless with an arbitrarily large alphabet size $|\mathcal{Z}|$ (although $|\mathcal{Z}| = |\mathcal{Y}|$ is sufficient). 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 and before X_{n+1} is sent. Z^n denotes $Z_1, \ldots, Z_n$.

The forward DMC is defined by the $|\mathcal{X}|$ by $|\mathcal{Y}|$ transition matrix $\{P_{kj}\}$ where, for each time n, $P_{kj} = \boldsymbol{P}[Y_n = j | X_n = k]$. The channel is memoryless in the sense that

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

For each input letter $k \in \mathcal{X}$, there is a nonnegative transmission cost $\rho_k \geq 0$ and at least one ρ_k is zero. The cost S_τ of transmitting a codeword of length τ is the sum of the costs of the τ symbols in the codeword. A cost constraint $\mathcal{P}$ means that $\boldsymbol{E}[\mathcal{S}_\tau] \leq \mathcal{P}\boldsymbol{E}[\tau]$. We usually refer to $\mathcal{P}$ as a *power constraint* and to S_τ as *energy*. With this definition of power constraint, $\mathcal{P}$ can be seen to upper-bound the long-term time-average cost per symbol over a long string of independent successive message transmissions.

B. Fixed-Length Block Codes With Error-or-Erasure Decoding

We begin with the slightly simpler problem of finding fixed-length block codes for an error-or-erasure decoder, i.e., a decoder which can either decode the message or produce an erasure symbol. The objective will be to minimize (or approximately minimize) the error probability while making the erasure probability small but potentially much larger than the error probability. In the following subsection, this error-and-erasure scheme will be converted into a variable-length block-coding scheme by retransmitting the erased messages.

Consider a code of fixed-length ℓ containing two phases of length ℓ_1 and ℓ_2, respectively. The first phase uses a power constraint $\mathcal{P}_1$ and the second $\mathcal{P}_2$. This provides an overall power constraint $\mathcal{P}$ where $\ell_1\mathcal{P}_1 + \ell_2\mathcal{P}_2 = \ell\mathcal{P}$. Define η as ℓ_1/ℓ, so that this power constraint becomes

$$\mathcal{P} = \eta\mathcal{P}_1 + (1-\eta)\mathcal{P}_2. \tag{1}$$

Phase 1 consists of a conventional block code without feedback, operating incrementally close to the capacity $\mathcal{C}(\mathcal{P}_1)$ of the channel subject to constraint $\mathcal{P}_1$

$$\mathcal{C}(\mathcal{P}_1) \triangleq \max_{\phi:\, \sum_k \phi_k \rho_k \leq \mathcal{P}_1} \sum_{k,j} \phi_k P_{kj} \ln \frac{P_{kj}}{\sum_m \phi_m P_{mj}}. \tag{2}$$

Here and throughout, ϕ is assumed to be a probability assignment, i.e., $\phi_k \geq 0$ for each k and $\sum_k \phi_k = 1$. The conventional coding theorem for a constrained DMC with fixed block length and no feedback is as follows:[2] for any $\delta_1 > 0$, there is an $\epsilon_1(\delta_1) > 0$ such that, for all large enough ℓ_1, codes of block length ℓ_1 exist with $M \geq e^{\ell_1[\mathcal{C}(\mathcal{P}_1)-\delta_1]}$ codewords, each of energy at most $\ell_1\mathcal{P}_1$ and each with error probability upper-bounded by

$$P_{e1} \leq \exp -\ell_1\epsilon_1(\delta_1). \tag{3}$$

Using such a code in phase 1, the decoder makes a tentative decision at the end of phase 1. The transmitter (knowing the decision

[2]See, for example, [5, Theorem 7.3.2].

via feedback) then sends a binary codeword $\boldsymbol{x}_A$ for "accept" and $\boldsymbol{x}_R$ for "reject" in phase 2. Let P_{RA} be the probability that the receiver decodes $\boldsymbol{x}_A$ given that $\boldsymbol{x}_R$ is sent. Similarly, P_{AR} is the probability of decoding $\boldsymbol{x}_R$ given $\boldsymbol{x}_A$.

If $\boldsymbol{x}_A$ is decoded, the receiver gives its tentative decision from phase 1 to the user and the overall probability of error $\tilde{P}_e$ satisfies $\tilde{P}_e \leq P_{RA}$. If $\boldsymbol{x}_R$ is decoded, an erasure is released and the probability of erasure $\tilde{P}_r$ satisfies $\tilde{P}_r \leq P_{AR} + P_{e1}$. Assume for now that the power constraint may be violated by an incrementally small amount. Thus, we choose $\boldsymbol{x}_A$ to satisfy the constraint, and choose $\boldsymbol{x}_R$ arbitrarily since it is rarely used. We bound $-\ln P_{RA}$ by the divergence between the output distribution conditional on $\boldsymbol{x}_A$ and the output distribution conditional on $\boldsymbol{x}_R$.

To be more explicit, define the maximum single-letter divergence for the input letter k as

$$\mathcal{D}_k \triangleq \max_m \sum_j P_{kj} \ln \frac{P_{kj}}{P_{mj}}. \tag{4}$$

Note that if $P_{mj} = 0$ for some channel transition, then $D_k = \infty$ for each k such that $P_{kj} > 0$. We will see in Section II-E that this leads to error-free codes at rates below capacity. In the following subsection, we consider only channels for which $P_{kj} > 0$ for all $k \in \mathcal{X}$ and $j \in \mathcal{Y}$.

1) Error-and-Erasure Decoding With All $P_{kj} > 0$: Assume that $P_{kj} > 0$ for all $k \in \mathcal{X}$ and $j \in \mathcal{Y}$ and, for each $k \in \mathcal{X}$, let m_k be an input letter m maximizing $\sum_j P_{kj} \ln \frac{P_{kj}}{P_{mj}}$. If $\boldsymbol{x}_A$ contains $\phi_k \ell_2$ occurrences of letter k and $\boldsymbol{x}_r$ is chosen to contain the letter m_k whenever $\boldsymbol{x}_A$ contains k, then the following minor variation of Stein's lemma results:[3] for any $\delta_2 > 0$, there is an $\epsilon_2(\delta_2) > 0$ such that

$$P_{RA} \leq \exp\left[\sum_k -\ell_2 \phi_k \mathcal{D}_k + \ell_2 \delta_2\right] \tag{5}$$

$$P_{AR} \leq \exp\left[-\ell_2 \epsilon_2(\delta_2)\right]. \tag{6}$$

From (5), we want to choose $\boldsymbol{x}_A$ to maximize $\sum_k \mathcal{D}_k \phi_k$ subject to the power constraint. Thus, for a power constraint $\mathcal{P}_2$ in phase 2, define $\mathcal{D}(\mathcal{P}_2)$ as

$$\mathcal{D}(\mathcal{P}_2) \triangleq \max_{\phi:\, \sum_k \phi_k \rho_k \leq \mathcal{P}_2} \sum_k \mathcal{D}_k \phi_k. \tag{7}$$

The function $\mathcal{D}(\mathcal{P})$ in (7) is the maximum of a linear function of ϕ over linear constraints. As illustrated in Fig. 1, $\mathcal{D}(\mathcal{P})$ is piecewise linear, nondecreasing, and concave in its domain of definition $\mathcal{P} \geq 0$.

Choosing the phase 2 codewords $\boldsymbol{x}_A$ and $\boldsymbol{x}_R$ according to this maximization, (5) becomes

$$P_{RA} \leq \exp\left[-\ell_2 \mathcal{D}(\mathcal{P}_2) + \ell_2 \delta_2\right]. \tag{8}$$

The power constraint $\mathcal{P}_2$ is then satisfied by $\boldsymbol{x}_A$. The power in $\boldsymbol{x}_R$ (whose probability of usage vanishes exponentially with ℓ_1) can be upper-bounded by $\rho_{\max}$. The preceding results are summarized in the following lemma.

[3]This can be derived, for example, by starting with Theorem 5 in [16] and specializing to the case of asymptotically small s.

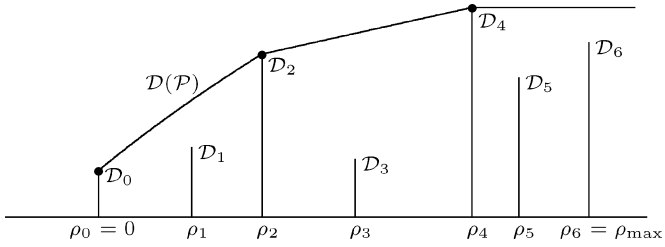

Fig. 1. The function $\mathcal{D}(\mathcal{P})$ for a channel satisfying $P_{kj} > 0$ for all $k \in \mathcal{X}$ and $j \in \mathcal{Y}$. The maximum single-letter divergences $\mathcal{D}_k$ are also shown. For convenience, the inputs are ordered in terms of cost. For any given $\mathcal{P}$, $\mathcal{D}(\mathcal{P})$ can be achieved with at most two positive ϕ_k.

Lemma 1: Assume ideal feedback for a DMC with all $P_{kj} > 0$. Then for all $\mathcal{P}_1 \geq 0, \mathcal{P}_2 \geq 0, \eta \in (0,1), \delta_1 > 0, \delta_2 > 0$, and all sufficiently large ℓ, there is an error-and-erasure code with $M \geq \exp\{\eta\ell[\mathcal{C}(\mathcal{P}_1) - \delta_1]\}$ messages such that, for each message, $\theta \in \mathcal{M} = \{1, 2 \ldots M\}$, the probability of error $\tilde{P}_e(\theta)$, the probability of erasure $\tilde{P}_r(\theta)$, and the expected energy $\boldsymbol{E}[\mathcal{S}(\theta)]$ satisfy

$$\tilde{P}_e(\theta) \leq \exp\{-(1-\eta)\ell[\mathcal{D}(\mathcal{P}_2) - \delta_2)]\} \tag{9}$$

$$\tilde{P}_r(\theta) \leq e^{-\eta\ell\epsilon_1(\delta_1)} + e^{-(1-\eta)\ell\epsilon_2(\delta_2)} \tag{10}$$

$$\boldsymbol{E}[\mathcal{S}(\theta)] \leq \ell[\eta\mathcal{P}_1 + (1-\eta)\mathcal{P}_2 + \rho_{\max} e^{-\eta\ell\epsilon_1(\delta_1)}]. \tag{11}$$

C. Variable-Length Block Codes; All $P_{kj} > 0$

The above error-or-erasure code can form the basis of a variable-length block code with ideal feedback. As in Yamamoto and Itoh [18], the transmitter observes each erasure via the feedback and repeats the original message until a message, not necessarily correct, is accepted. For simplicity, we assume that when a message is repeated, the receiver ignores the previously received symbols and uses the same decoding algorithm as before. Since an error then occurs independently after each repetition of the fixed-length codeword, the overall error probability satisfies

$$P_e \leq \frac{1}{1 - \tilde{P}_r} \exp\{-(1-\eta)\ell[\mathcal{D}(\mathcal{P}_2) - \delta_2]\}.$$

The duration τ of a block is ℓ times the number of error-or-erasure tries until acceptance, so $\boldsymbol{E}[\tau] = \ell/(1 - \tilde{P}_r)$. The coefficient $1/(1 - \tilde{P}_r)$ goes to 1 with increasing ℓ and thus can be absorbed into the arbitrary term δ_2 for sufficiently large ℓ. Similarly, ℓ can be replaced with $\overline{\tau} = \boldsymbol{E}[\tau]$, yielding $P_e \leq \exp\{-\overline{\tau}(1-\eta)[\mathcal{D}(\mathcal{P}_2) - \delta_2]\}$.

In the same way, the expected energy $\boldsymbol{E}[\mathcal{S}_\tau]$ over the entire transmission satisfies $\boldsymbol{E}[\mathcal{S}_\tau] \leq \boldsymbol{E}[\mathcal{S}]/(1 - \tilde{P}_r)$. Finally, using (11), the average power for each codeword is

$$\frac{\boldsymbol{E}[\mathcal{S}_\tau]}{\boldsymbol{E}[\tau]} \leq \eta\mathcal{P}_1 + (1-\eta)\mathcal{P}_2 + \rho_{\max} e^{-\eta\ell\epsilon_1(\delta_1)}.$$

The following lemma summarizes these results.

Lemma 2: Assume ideal feedback for a DMC with all $P_{kj} > 0$. Then for all $\mathcal{P}_1 \geq 0$, $\mathcal{P}_2 \geq 0$, $\eta \in (0,1)$, $\delta > 0$, and for all sufficiently large ℓ, there is a variable-length block code of

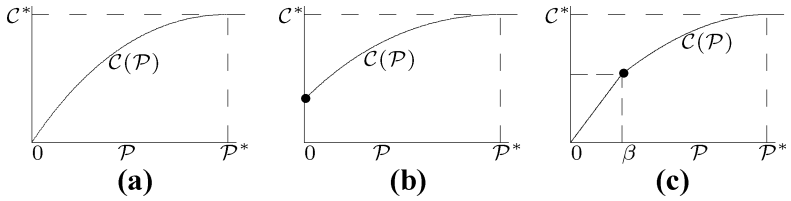

Fig. 2. Typical capacity functions. Parts (a) and (b) illustrate that $\mathcal{C}(0)$ can be either 0 or positive. Part (c) illustrates an important special case where $\mathcal{C}(x)$ is linear from 0 to $\beta > 0$ where β is defined as the largest x for which $\mathcal{C}(y)/y = \mathcal{C}(x)/x$ for all $y \in (0, x]$.

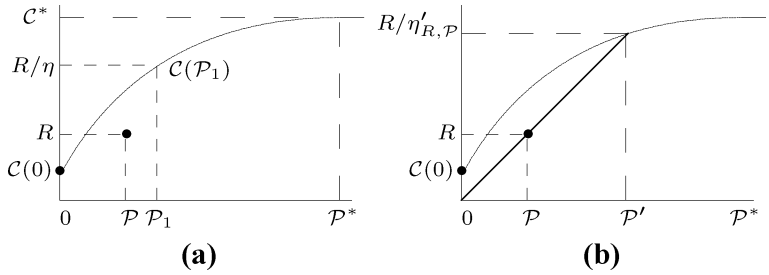

Fig. 3. Part (a) gives a graphic construction for $\mathcal{P}_1$ from R and η. Part (b) illustrates the value of η at which $\frac{R}{\eta} = \mathcal{C}(\frac{\mathcal{P}}{\eta})$.

expected length $\overline{\tau}$, $\ell \leq \overline{\tau} < \ell+1$ with $M \geq \exp[\overline{\tau}(R-\delta)]$ messages such that for each message, $\theta \in \mathcal{M} = \{1, 2 \ldots M\}$, the probability of error $P_e(\theta)$, and the expected energy $\boldsymbol{E}[\mathcal{S}_\tau(\theta)]$ satisfy

$$P_e(\theta) \leq \exp\{-\overline{\tau}[(1-\eta)\mathcal{D}(\mathcal{P}_2) - \delta]\} \tag{12}$$

$$\boldsymbol{E}[\mathcal{S}_\tau(\theta)] \leq (\eta\mathcal{P}_1 + (1-\eta)\mathcal{P}_2 + \delta)\overline{\tau}. \tag{13}$$

D. Optimization of the Bound; All $P_{kj} > 0$

Lemma 2 can be interpreted as providing a nominal rate of transmission $R = \eta\mathcal{C}(\mathcal{P}_1)$, a nominal power constraint $\mathcal{P} = \eta\mathcal{P}_1 + (1-\eta)\mathcal{P}_2$, and a nominal exponent of error probability $(1-\eta)\mathcal{D}(\mathcal{P}_2)$. We have demonstrated the existence of variable-length block codes for which the actual average rate, power, and exponent approach these values arbitrarily closely as $\overline{\tau}$ becomes large.

For any given $\mathcal{P}$ and R satisfying $0 < R < \mathcal{C}(\mathcal{P})$, we now maximize the exponent $(1-\eta)\mathcal{D}(\mathcal{P}_2)$ subject to the constraints

$$\eta\mathcal{C}(\mathcal{P}_1) = R; \quad \mathcal{P}_1 \geq 0, \tag{14}$$

$$\eta\mathcal{P}_1 + (1-\eta)\mathcal{P}_2 = \mathcal{P}; \quad \mathcal{P}_1 \geq 0,\ \mathcal{P}_2 \geq 0,\ 0 < \eta < 1. \tag{15}$$

It will be shown that there is a certain interval of values of η for which (14) and (15) have solutions, and that $\mathcal{P}_1$ and $\mathcal{P}_2$ are essentially uniquely defined as a function of η in that interval. Thus, the maximization will reduce to a maximization of a function of one variable over an interval.

As can be seen from (2) and visualized in Fig. 2, the function $\mathcal{C}(\mathcal{P})$ is nonnegative, concave, continuous, and nondecreasing in its domain $\mathcal{P} \geq 0$. It is strictly increasing for $0 \leq \mathcal{P} \leq \mathcal{P}^*$ where $\mathcal{P}^*$ is the smallest $\mathcal{P}$ for which $\mathcal{C}(\mathcal{P}) = \mathcal{C}^*$, the unconstrained channel capacity.

For the case where $\mathcal{P} > 0$, Fig. 3 illustrates how to construct $\mathcal{P}_1$ to satisfy (14) for a given R, $\mathcal{P}$, and η. Note that $\mathcal{P}_1$ is uniquely specified if $\mathcal{C}(0) \leq R/\eta < \mathcal{C}^*$. If $R/\eta = \mathcal{C}^*$, then $\mathcal{P}_1 \geq \mathcal{P}^*$. Choosing $\mathcal{P}_1 > \mathcal{P}^*$ for a given η simply reduces $\mathcal{P}_2$, and thus reduces the exponent. Thus, we restrict attention in this case to $\mathcal{P}_1 = \mathcal{P}^*$ and the now unique solution for $\mathcal{P}_1$, here denoted $\mathcal{P}_1(\eta)$, is given by

$$\mathcal{P}_1(\eta) = \mathcal{C}^{-1}(R/\eta), \qquad \text{for } \frac{R}{\mathcal{C}^*} \leq \eta \leq \frac{R}{\mathcal{C}(0)} \tag{16}$$

where $\mathcal{C}^{-1}(y)$ over $y \in [\mathcal{C}(0), \mathcal{C}^*]$ is the inverse of $\mathcal{C}(x)$ over $x \in [0, \mathcal{P}^*]$. This inverse exists since $\mathcal{C}(\cdot)$ is strictly increasing over this interval. Next note from (15) that $\eta\mathcal{P}_1$ must be less than or equal to $\mathcal{P}$. This constraint is illustrated in the second part of Fig. 3. A straight line is constructed passing through the origin and the point $(\mathcal{P}, R)$. For each η, a horizontal line at height R/η intersects this line at $\mathcal{P}/\eta$. For the choice of η in the figure, the capacity curve lies to the left of the straight line at height R/η, so $\mathcal{P}_1 < \mathcal{P}/\eta$, thus ensuring that $\mathcal{P}_2 > 0$.

The above straight line is linearly increasing and the capacity curve is bounded, so they must intersect at some positive power, rate pair, say $(\mathcal{P}', R')$. Define[4] $\eta'_{R,\mathcal{P}}$ in terms of $\mathcal{P}'$ as

$$\eta'_{R,\mathcal{P}} \triangleq \frac{\mathcal{P}}{\mathcal{P}'} = \frac{R}{\mathcal{C}(\mathcal{P}')}. \tag{17}$$

Since the capacity curve and the straight line cross at $(\mathcal{P}', R')$, we see that for $\eta < \eta'_{R,\mathcal{P}}$ (i.e., $R/\eta > R' = \mathcal{C}(\mathcal{P}')$) a straight line at height R/η passes through the straight line before the capacity curve (if it passes through the capacity curve at all). If it does pass through the capacity curve, then $\mathcal{P}/\eta < \mathcal{P}_1$, so that (15) has no solution. If it does not pass through the capacity curve, then (14) has no solution.

If $\eta \geq \eta'_{R,\mathcal{P}}$, on the other hand, $\mathcal{P}_1$ is uniquely given by (16), provided that $\eta \in \left[\frac{R}{\mathcal{C}^*}, \frac{R}{\mathcal{C}(0)}\right]$. One can see that $\eta \geq \frac{R}{\mathcal{C}^*}$ is implied by $\eta \geq \eta'_{R,\mathcal{P}}$, by considering the definition of $\eta'_{R,\mathcal{P}}$ given in (17).

[4]The point $\eta'_{R,\mathcal{P}}$ expressed implicitly in (17) can be expressed explicitly as

$$\eta'_{R,\mathcal{P}} = \frac{\mathcal{P}}{\Gamma^{-1}(R/\mathcal{P})}$$

where $\Gamma^{-1}(\cdot)$ is the inverse of the function $\Gamma(x) = \mathcal{C}(x)/x$ taken over the domain $x \geq \beta$, where β is the largest x for which $\Gamma(x) = \Gamma(0)$.

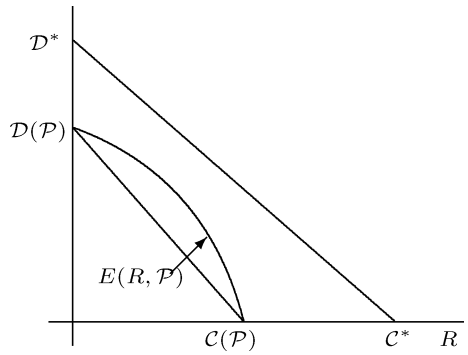

Fig. 4. A typical $E(R,\mathcal{P})$ curve. The figure illustrates that $E(R,\mathcal{P})$, as a function of R for fixed $\mathcal{P}$, is concave, decreasing, and bounded.

In summary, we have shown that (14) and (15) have a unique solution for $(\mathcal{P}_1,\mathcal{P}_2)$ for all η values in the interval

$$\mathcal{I}_{R,\mathcal{P}} \triangleq \left[\eta'_{R,\mathcal{P}},\ \min\left(1,\frac{R}{\mathcal{C}(0)}\right)\right]. \tag{18}$$

and otherwise no solutions exist.

Finally, we must show that $\mathcal{I}_{R,\mathcal{P}}$ is not empty. Since by assumption $R<\mathcal{C}(\mathcal{P})$, it follows that $R'>R$ and, thus, $\eta'_{R,\mathcal{P}}<1$. Also, from (17), $\eta'_{R,\mathcal{P}}<R/\mathcal{C}(0)$, which shows that $\mathcal{I}_{R,\mathcal{P}}$ is nonempty.

The previous results assumed $\mathcal{P}>0$. For $\mathcal{P}=0$, (14) is satisfied only by $\eta=R/\mathcal{C}(0)$, and if one extends the previous definition of $\eta'_{R,\mathcal{P}}$ to be $R/\mathcal{C}(0)$ in this case, then the above results apply to this case also (although $\mathcal{I}_{R,\mathcal{P}}$ shrinks to the single point $\eta'_{R,\mathcal{P}}$).

Using (15), the nominal exponent then becomes

$$E(R,\mathcal{P},\eta)=(1-\eta)\mathcal{D}\left(\frac{\mathcal{P}-\eta\mathcal{C}^{-1}(R/\eta)}{1-\eta}\right). \tag{19}$$

The following lemma, proven in the Appendix, shows that $E(R,\mathcal{P},\eta)$ is concave.

Lemma 3: The set of points $(R,\mathcal{P},\eta)$ such that $0<R<\mathcal{C}(\mathcal{P})$ and $\eta\in\mathcal{I}_{R,\mathcal{P}}$ is convex. The function $E(R,\mathcal{P},\eta)$ is concave over this domain.

We next maximize the exponent $E(R,\mathcal{P},\eta)$ over $\eta\in\mathcal{I}_{R,\mathcal{P}}$

$$E(R,\mathcal{P})\triangleq \sup_{\eta\in\mathcal{I}_{R,\mathcal{P}}}(1-\eta)\mathcal{D}\left(\frac{\mathcal{P}-\eta\mathcal{C}^{-1}(R/\eta)}{1-\eta}\right). \tag{20}$$

This is simply a concave maximization over an interval. The resulting function $E(R,\mathcal{P})$ is then also concave as a function of $(R,\mathcal{P})$, and thus also as a function of R for any given $\mathcal{P}$. This is illustrated in Fig. 4. It can be shown that $E(R,\mathcal{P})$ is strictly decreasing in R from $\mathcal{D}(\mathcal{P})$ at $R=0$.

One can extend the definition of $E(R,\mathcal{P})$ to $R=\mathcal{C}(\mathcal{P})$, for any $\mathcal{P}$ as

$$E(\mathcal{C}(\mathcal{P}),\mathcal{P})\triangleq \lim_{\delta\to 0^+}E(\mathcal{C}(\mathcal{P})-\delta,\mathcal{P}). \tag{21}$$

The following theorem results from using $E(R,\mathcal{P})$ in Lemma 2.

Theorem 1: Assume ideal feedback for a DMC with all $P_{kj}>0$. Then for all $\mathcal{P}\geq 0$, $R\leq\mathcal{C}(\mathcal{P})$, $\delta>0$, and all sufficiently large integer ℓ, there is a variable-length block code of expected length $\overline{\tau}$, $\ell\leq\overline{\tau}<\ell+1$ with $M\geq\exp[\overline{\tau}(R-\delta)]$ messages such that for each message $\theta\in\mathcal{M}=\{1,2\ldots M\}$, the probability of error $P_e(\theta)$, and the expected energy $\boldsymbol{E}[\mathcal{S}_\tau(\theta)]$ satisfy

$$P_e(\theta)\leq\exp\{-\overline{\tau}[E(R,\mathcal{P})-\delta]\} \tag{22}$$

$$\boldsymbol{E}[\mathcal{S}_\tau(\theta)]\leq\left(\mathcal{P}+\rho_{\max}e^{-\overline{\tau}\epsilon(\delta)}\right)\overline{\tau} \tag{23}$$

where $\epsilon(\delta)>0$ for each $\delta>0$. Furthermore, the probability that the codeword length exceeds ℓ is at most δ.

Theorem 1 shows that the exponent $E(R,\mathcal{P})$ can be asymptotically achieved by this particular class of variable-length block codes. The converse in the next section will show that no variable-length block code can do better asymptotically, i.e., that $E(R,\mathcal{P})$ is the reliability function for constrained variable-length block codes.

Theorem 1 also shows that these codes are almost fixed-length block codes, deviating from fixed length only with arbitrarily small probability. It is also possible to analyze the queueing delay for this class of codes. Note that if the source bits arrive equally spaced in time, then, even for a fixed-length block code, bits are delayed waiting for the next block and additionally delayed waiting for the block to be received and decoded. The additional delay, for variable-length block codes, is the queueing delay of waiting blocks while earlier blocks are retransmitted. At any $R<\mathcal{C}(\mathcal{P})$, the probability of retransmission decreases exponentially (albeit with a small exponent) with $\overline{\tau}$, so it is not surprising that the expected additional delay due to retransmissions goes to 0 with increasing $\overline{\tau}$. We have shown that this indeed happens.

For $0<R<\mathcal{C}(\mathcal{P})$ and $\mathcal{P}>0$, (23) can be simplified by absorbing the term $\rho_{\max}e^{-\overline{\tau}\epsilon(\delta)}$ into the δ of (22). This cannot be done for $\mathcal{P}=0$ since the constraint $\boldsymbol{E}[\mathcal{S}_\tau(\theta)]\leq\mathcal{P}\overline{\tau}$ for all θ reduces to the unconstrained case where only zero-cost inputs are used. In (23), on the other hand, we are using a reject message of positive power with asymptotically vanishing probability, with increasing $\overline{\tau}$.

The requirement of ideal feedback can be relaxed to that of a noiseless feedback link of capacity $\mathcal{C}_{\text{fb}}\geq\mathcal{C}$ and finite delay T by using a modification of the error-and-erasure scheme first suggested by Şimşek and Sahai in [11] for unconstrained channels. For phase 1, the message is divided into equal length submessages which are separately encoded at a rate close to capacity and sent one after the other. A temporary decision about each submessage is made at the receiver and sent reliably to the transmitter with a delay equal to T plus the submessage transmission time. In phase 2, the entire message is rejected if any submessage was in error and otherwise it is accepted. A single bit of feedback is required for phase 2, and it can be shown that the various delays become amortized over the entire message transmission as $M\to\infty$.

1) Channels for Which $E(R,\mathcal{P})>0$ for $R=\mathcal{C}(\mathcal{P})$: In certain cases $E(\mathcal{C}(\mathcal{P}),\mathcal{P})$ as defined in (21) is strictly positive. We start with a simple example of this phenomenon and then delineate the cases where it is possible.

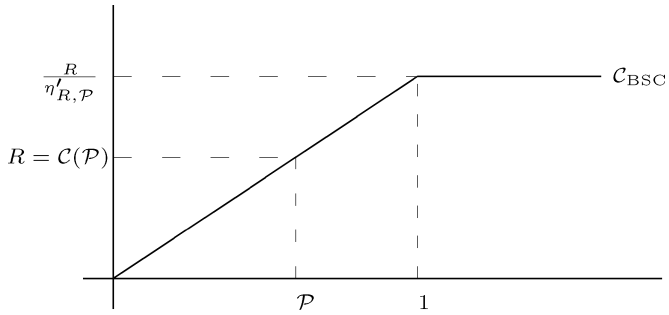

Fig. 5. $\mathcal{C}(\mathcal{P})$ versus $\mathcal{P}$ for a BSC with a zero-cost noise symbol and resulting value of $\eta'_{R,\mathcal{P}}$ at $R = \mathcal{C}(\mathcal{P})$.

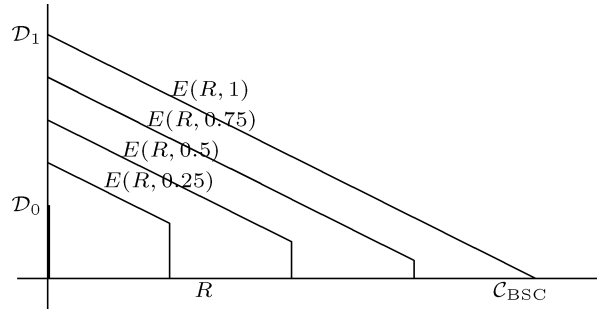

Fig. 6. $E(R,\mathcal{P})$ for a BSC with a zero-cost noise symbol. For $\mathcal{P} < 1$, the exponent decreases linearly to a positive value at capacity.

Example 1: BSC With Extra Free Symbol: Consider a binary-symmetric channel (BSC) in which each input symbol has unit cost. There is an additional cost-free symbol that is completely noisy. That is, the transition probabilities and costs are as follows:

$$P_{kj} = \begin{bmatrix} 1/2 & 1/2 \\ \alpha & 1-\alpha \\ 1-\alpha & \alpha \end{bmatrix} \qquad \rho_k = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \tag{24}$$

where $0 < \alpha < 1/2$. Let $\mathcal{C}_{\rm BSC} = \ln 2 - \mathfrak{h}(\alpha)$ where $\mathfrak{h}(\alpha) = -(1-\alpha)\ln(1-\alpha) - \alpha \ln \alpha$ is the binary entropy. As illustrated in Fig. 5

$$\mathcal{C}(\mathcal{P}) = \begin{cases} \mathcal{P}\mathcal{C}_{\rm BSC}, & \text{for } 0 \le \mathcal{P} \le 1 \\ \mathcal{C}_{\rm BSC}, & \text{for } \mathcal{P} > 1. \end{cases}$$

Assume a power constraint $\mathcal{P} < 1$ and choose the rate to be at capacity $R = \mathcal{C}(\mathcal{P}) = \mathcal{P}\mathcal{C}_{\rm BSC}$. It can be seen from Fig. 5 that $\eta'_{\mathcal{C}(\mathcal{P}),\mathcal{P}} = \mathcal{C}(\mathcal{P})/\mathcal{C}_{\rm BSC} = \mathcal{P}$. Thus, the power $\mathcal{P}_1$ in phase 1 can be chosen anywhere between $\mathcal{P}$ and 1. However it is chosen, all the available energy will be used in phase 1 and $\mathcal{P}_2$ will be 0. However, $\mathcal{D}_0 > 0$ since the divergence of the free symbol with either of the BSC symbols is positive. Thus, as intuition might suggest, the maximum exponent results from maximizing the interval available for phase 2, i.e., $\eta'_{\mathcal{C}(\mathcal{P}),\mathcal{P}} = \mathcal{P}$. Thus

$$E(\mathcal{C}(\mathcal{P}),\mathcal{P}) = (1-\mathcal{P})\mathcal{D}_0.$$

This ability to transmit at channel capacity (subject to the ϵ's and δ's) with a positive exponent is quite surprising, but arises from the ability to choose $\mathcal{P}_1 = 1$ throughout phase 1, thus using the BSC in that phase and transmitting at the BSC capacity. In other words, since, $\mathcal{C}(\mathcal{P})$ is linear with $\mathcal{P}$ for $\mathcal{P} \le 1$, one can transmit the packet faster in phase 1 by using greater power without increasing the overall energy required for the transmission. The exponent is maximized by choosing $\mathcal{P}_1 = 1$, saving a fraction $1 - \mathcal{P}$ of time for phase 2.

More generally, any DMC for which β, as defined in Fig. 2 is greater than 0 has the property that if $\mathcal{P} < \beta$, then there is a positive exponent at $R = \mathcal{C}(\mathcal{P})$. To see this, note that $\eta'_{\mathcal{C}(\mathcal{P}),\mathcal{P}} = \mathcal{C}(\mathcal{P})/\mathcal{C}(\beta) < 1$. Thus, by choosing $\mathcal{P}_1 = \beta$, the resulting exponent is $E(\mathcal{C}(\mathcal{P}),\mathcal{P}) = (1 - \mathcal{P}/\beta)\mathcal{D}_0$. It can be seen from Fig. 2 that when $\mathcal{P} > \beta$, and $R = \mathcal{C}(\mathcal{P})$, then $\eta'_{R,\mathcal{P}} = 1$, and there is no time left for phase 2.

For the example of a BSC with a free symbol, $E(R,\mathcal{P})$ can also be calculated for any $R < \mathcal{C}(\mathcal{P})$ and $\mathcal{P} \le 1$. It can be seen that $\eta'_{R,\mathcal{P}} = R/\mathcal{C}_{\rm BSC}$, and with a little thought it can be seen that this value of η, corresponding to $\mathcal{P}_1 = 1$, maximizes $E(R,\mathcal{P},\eta)$. Thus, as illustrated in Fig. 6,

$$E(R,\mathcal{P}) = (1-\eta)\mathcal{D}\left(\frac{\mathcal{P}-\eta}{1-\eta}\right), \qquad \text{where } \eta = \frac{R}{\mathcal{C}_{\rm BSC}}.$$

From (7), $\mathcal{D}(x) = \mathcal{D}_0 + (\mathcal{D}_1 - \mathcal{D}_0)x$ for $0 \le x \le 1$, so

$$E(R,\mathcal{P}) = (1-\eta)\mathcal{D}_0 + (\mathcal{P}-\eta)(\mathcal{D}_1 - \mathcal{D}_0), \quad \text{where } \eta = \frac{R}{\mathcal{C}_{\rm BSC}}.$$

2) Alternative Approaches to Finding $E(R,\mathcal{P})$: The reliability function $E(R,\mathcal{P})$ is expressed in (20) as an optimization over $E(R,\mathcal{P},\eta)$ and as such involves calculating $\mathcal{C}(\mathcal{P}_1)$ and $\mathcal{D}(\mathcal{P}_2)$ as subproblems. An alternative that might be more convenient numerically is to express $E(R,\mathcal{P})$ directly as a concave optimization over the input probabilities in phases 1 and 2 subject to the constraints corresponding to a given R and $\mathcal{P}$.

Another alternative, which is more interesting conceptually, is to investigate how the phase 1 and phase 2 powers must be related. Consider the equivalent problem of finding the minimal power $\mathcal{P}$ required for a given rate R and exponent E. We will derive a necessary condition for $\mathcal{P}_1 > 0$, $\mathcal{P}_2 > 0$, and $0 < \eta < 1$ to achieve this minimum power. First consider the special case in which $\mathcal{C}(\mathcal{P})$ is continuously differentiable for $\mathcal{P} > 0$ and let $A_1 = \eta\mathcal{P}_1$ be the phase 1 power amortized over both phases. The partial derivative of A_1 with respect to η for a given $R = \eta\mathcal{C}(A_1/\eta)$ is then

$$\left.\frac{\partial A_1}{\partial \eta}\right|_R = -\frac{\partial[\eta\mathcal{C}(A_1/\eta)]/\partial\eta}{\partial[\eta\mathcal{C}(A_1/\eta)]/\partial A_1} = \mathcal{P}_1 - \frac{\mathcal{C}(\mathcal{P}_1)}{\mathcal{C}'(\mathcal{P}_1)}. \tag{25}$$

Geometrically, this is the horizontal axis intercept of the tangent to $\mathcal{C}(\cdot)$ at $\mathcal{P}_1$.

In the general case, $\mathcal{C}(\mathcal{P}_1)$ can have slope discontinuities at particular values of $\mathcal{P}_1$; because of these discontinuities, the left and right derivatives and the corresponding tangents and intercepts becomes different from each other (see Fig. 7).

In the same way, let $A_2 = (1-\eta)\mathcal{P}_2$. Then, holding the exponent E fixed

$$\begin{aligned} \left.\frac{\partial A_2}{\partial \eta}\right|_E &= -\frac{\partial[(1-\eta)\mathcal{D}(A_2/[(1-\eta)]/\partial\eta}{\partial[[(1-\eta)\mathcal{D}(A_2/[(1-\eta))]/\partial A_2} \\ &= -\mathcal{P}_2 + \frac{\mathcal{D}(\mathcal{P}_2)}{\mathcal{D}'(\mathcal{P}_2)}. \end{aligned} \tag{26}$$

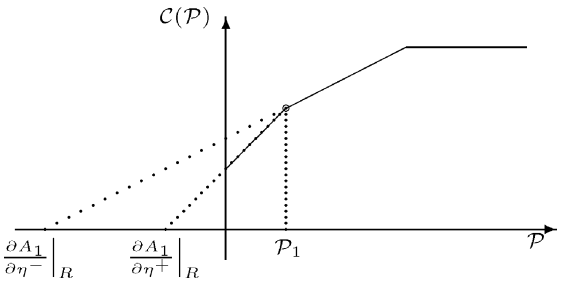

Fig. 7. $\frac{\partial A_1}{\partial \eta^-}|_R$ is the derivative corresponding to negative change in η (positive change in $\mathcal{P}_1$) and $\frac{\partial A_1}{\partial \eta^+}|_R$ is the derivative corresponding to positive change in η (negative change in $\mathcal{P}_1$).

This is the negative of the horizontal axis intercept of the tangent to $\mathcal{D}(\cdot)$ at $\mathcal{P}_2$. At points of slope discontinuity in $\mathcal{D}(\cdot)$, this must be replaced with

$$\left.\frac{\partial A_2}{\partial \eta^-}\right|_E = -\mathcal{P}_2 + \frac{\mathcal{D}(\mathcal{P}_2)}{\left.\frac{d\mathcal{D}(x)}{dx^-}\right|_{x=\mathcal{P}_2}} \tag{27}$$

$$\left.\frac{\partial A_2}{\partial \eta^+}\right|_E = -\mathcal{P}_2 + \frac{\mathcal{D}(\mathcal{P}_2)}{\left.\frac{d\mathcal{D}(x)}{dx^+}\right|_{x=\mathcal{P}_2}}. \tag{28}$$

Finally, the overall power constraint is $\mathcal{P} = A_1 + A_2$, so

$$\left.\frac{\partial \mathcal{P}}{\partial \eta^+}\right|_{R,E} = \left.\frac{\partial A_1}{\partial \eta^+}\right|_R + \left.\frac{\partial A_2}{\partial \eta^+}\right|_E \tag{29}$$

$$\left.\frac{\partial \mathcal{P}}{\partial \eta^-}\right|_{R,E} = \left.\frac{\partial A_1}{\partial \eta^-}\right|_R + \left.\frac{\partial A_2}{\partial \eta^-}\right|_E. \tag{30}$$

For $\mathcal{P}_1, \mathcal{P}_2$, and η to minimize $\mathcal{P}$ for fixed R, E, it is necessary that $\frac{\partial \mathcal{P}}{\partial \eta^+}|_{R,E} \geq 0$ (i.e., that an incremental increase in η does not reduce $\mathcal{P}$) and that $\frac{\partial \mathcal{P}}{\partial \eta^-}|_{R,E} \leq 0$ (i.e., that an incremental decrease in η does not reduce $\mathcal{P}$). Geometrically, what this says is that the horizontal intercept of the tangent to $\mathcal{C}(\cdot)$ at $\mathcal{P}_1$, which in general is the interval $[\frac{\partial A_1}{\partial \eta^-}|_R, \frac{\partial A_1}{\partial \eta^+}|_R]$, must overlap with the horizontal intercept of the tangent to $\mathcal{D}(\cdot)$ at $\mathcal{P}_2$, i.e., with the interval $[-\frac{\partial A_2}{\partial \eta^+}|_E, -\frac{\partial A_2}{\partial \eta^-}|_E]$. Note that these intervals reduce to single points in the absence of slope discontinuities in $\mathcal{C}(\mathcal{P}_1)$ or $\mathcal{D}(\mathcal{P}_2)$.

It is surprising that these conditions do not involve η. The following example shows how these conditions can be used.

Example 2: Combined Four-Input Symmetric Channel, BSC, and Free Symbol: Consider the following DMC with seven input letters and four output letters

$$P_{kj} = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ \delta & \delta & 1/2-\delta & 1/2-\delta \\ 1/2-\delta & 1/2-\delta & \delta & \delta \\ 1-3\epsilon & \epsilon & \epsilon & \epsilon \\ \epsilon & 1-3\epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & 1-3\epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & 1-3\epsilon \end{bmatrix}$$

$$\rho_k = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 4 \\ 4 \\ 4 \\ 4 \end{bmatrix}.$$

where $\epsilon = 1/75$, $\delta = 1/100$.

$\mathcal{C}(\mathcal{P})$ is piecewise linear for the same reason as in the previous example; $\mathcal{C}(\mathcal{P})$ and $\mathcal{D}(\mathcal{P})$ are given in Fig. 8.

The above necessary conditions on $\mathcal{P}_1$ and $\mathcal{P}_2$ imply

$$\begin{aligned} \mathcal{P}_1 = 1 &\Leftrightarrow 1 > \mathcal{P}_2 > 0 \\ 4 > \mathcal{P}_1 > 1 &\Leftrightarrow \mathcal{P}_2 = 1 \\ \mathcal{P}_1 = 4 &\Leftrightarrow \mathcal{P}_2 > 1. \end{aligned} \tag{31}$$

Using these conditions and the set constraint, we can calculate $E(R, \mathcal{P})$ for any given $\mathcal{P}$; the solutions for $\mathcal{P} = 1$ and $\mathcal{P} = 5$ are given in Fig. 9.

E. Zero-Error Capacity; Channels With at Least One $P_{kj} = 0$

The form of $E(R, \mathcal{P})$ relies heavily on the assumption that $P_{kj} > 0$ for all k, j. To see why, assume $P_{mj} = 0$ for some m, j. We assume throughout, without loss of generality, that for each output j, $P_{kj} > 0$ for at least one input k. Thus, with $P_{mj} = 0$ and $P_{kj} > 0$, $\mathcal{D}_k = \infty$. Suppose that the "accept" codeword of Section II uses all k's, the "reject" message all m's, and that the receiver decodes "accept" only if it receives one or more j's. In this case, no errors can ever occur for the corresponding variable-length block code.

Asymptotically, phase 2 can occupy a negligible portion of the block, say $\ln \ell$ of ℓ symbols. Then for any $\delta > 0$, and all large enough block lengths ℓ, an error-and-erasure code exists with $M \geq e^{(\ell - \ln \ell)(\mathcal{C}(\mathcal{P}) - \delta)}$, $\tilde{P}_e = 0$, and $\hat{P}_r \leq e^{-(\ell - \ln \ell)\epsilon(\delta)} + \ell^{\ln(1-P_{kj})}$, and expected energy $\boldsymbol{E}[\mathcal{S}_\ell] \leq \ell \mathcal{P} + \rho_{\max} \ln \ell$. After a little analysis the following theorem results.

Theorem 2: Assume ideal feedback for a DMC with at least one $P_{mj} = 0$. Then for all $0 < R \leq \mathcal{C}(\mathcal{P})$, all positive δ, and all sufficiently large $\overline{\tau}$, there is a variable-length block code satisfying

$$M \geq e^{\overline{\tau}(R-\delta)}, \qquad P_e = 0, \qquad \boldsymbol{E}[\mathcal{S}_\tau] \leq \mathcal{P}\overline{\tau} + 2\rho_{\max} \ln \overline{\tau}.$$

III. The Converse: Relating $\overline{\tau}$ and P_e

We have established an upper bound on P_e for given rate R, power $\mathcal{P}$, and expected block length $\overline{\tau}$ by developing and analyzing a particular class of variable-length block codes. Here we develop a lower bound to P_e which, for large enough $\overline{\tau}$, is valid for all variable-length block codes. The lower bound uses the idea of a two-phase analysis, but, as will be seen, this does not restrict the encoding or decoding. We start by finding a lower bound on the expected time spent in the first phase $\boldsymbol{E}[\tau_1]$ and a related lower bound on the expected time spent in the second phase $\boldsymbol{E}[\tau - \tau_1]$.

The analysis is a simplification and generalization of Burnashev [1] and is based on the evolution at each time n of the con-

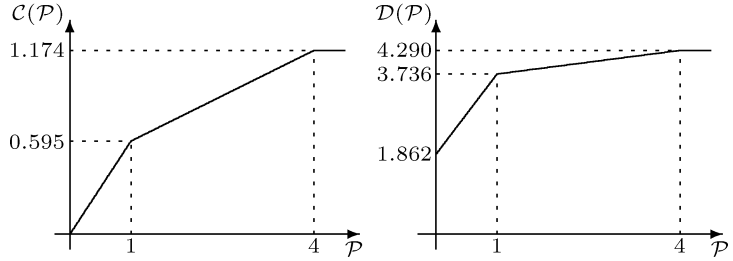

Fig. 8. Capacity and divergence functions, to different scales.

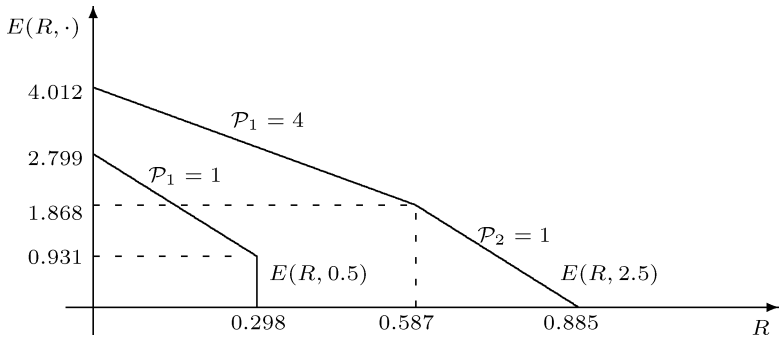

Fig. 9. Reliability function for $\mathcal{P} = 0.5$ and $\mathcal{P} = 2.5$; each straight-line segment is characterized by either constant $\mathcal{P}_1$ or constant $\mathcal{P}_2$ according to (31).

ditional message entropy, conditioned on the observations at the receiver. The first phase is the interval until this conditional entropy drops from $\ln M$ to some fixed intermediate value, taken here to be 1. The second phase is the remaining time interval until the actual decoding time. Fano's inequality is used to link the conditional entropy at the decoding time to the error probability. In the first phase, we analyze a stochastic sequence related to the decrease in conditional entropy at each instant n, and in the second phase, we analyze a stochastic sequence related to the decrease in the logarithm of the conditional entropy.

Establishing this lower bound to P_e is more involved than the upper bound to P_e, since the lower bound must apply to *all* variable-length block codes. We start with a more precise definition of variable-length block codes. After that we bound the expected change of conditional entropy and its logarithm, first in one time unit and second between two stopping times. Then these are used to lower-bound the probability of error. The resulting upper bound on the reliability function agrees with the lower bound in Section II.

A. Mathematical Preliminaries and Fano's Inequality

In a variable-length block code, the transmitter is assumed to initially receive one of M equiprobable messages from the set $\mathcal{M} = \{1, \ldots, M\}$. It transmits successive channel symbols about that message, say message θ, until the receiver makes a decision and releases the decoded message to the user. The time of this decision is a random variable denoted by τ. We assume throughout that $\boldsymbol{E}[\tau] = \overline{\tau} < \infty$, since otherwise our exponential lower bound on error probability in terms of $\overline{\tau}$ trivially holds.

Given noiseless feedback, we can restrict our attention to encoding algorithms in which each input symbol X_n is a deterministic function of message and feedback.[5]

$$X_n = X_n(\theta, Z^{n-1}) \qquad \forall Z^{n-1}, \forall \theta. \tag{32}$$

The entire observation of the receiver up to time n, including Y^n and any additional random choices, can be summarized by the σ-field $\mathcal{F}_n$ generated by these random variables. The nested sequence of $\mathcal{F}_n$'s is called a filtration $\mathcal{F}$.

At each time n, depending on the realization $\mathfrak{f}_n$ of σ-field $\mathcal{F}_n$, the receiver has an *a posteriori* probability $p_{\mathfrak{f}_n}(i)$ for each i in $\mathcal{M}$. The corresponding conditional entropy of the message, given $\mathcal{F}_n$, is a random variable $\mathcal{H}_{\mathcal{F}_n}$, measurable in $\mathcal{F}_n$. Its sample value for any realization $\mathfrak{f}_n \in \mathcal{F}_n$, is given by

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

A decoding algorithm includes a decision rule about continuing or stopping the communication, depending on the observations up to that time, i.e., a Markov stopping time (see, for example, [17, p. 476]) with respect to the filtration $\mathcal{F}$. The message is also decoded at this stopping time. In order to define the various random variables at all times $n \geq 1$, rather than only times up to the stopping time, we will assume that $X_n(\theta, Z^{n-1})$ is equal to some given zero-cost symbol for all $n > \tau$ and all θ. Thus, $S_n = S_\tau$ for all $n > \tau$. Thus, if a variable-length block code (henceforth simply called a code) satisfies a cost constraint $\boldsymbol{E}[S_\tau] < \mathcal{P}\overline{\tau}$, then $\boldsymbol{E}[S_\tau] < \infty$ and $\boldsymbol{E}[S_n] < \infty$ for all n.

[5]This still permits the receiver to send some random choices for codewords to the transmitter. There is no obvious advantage to such randomization, but it is easy to include the possibility. Random choices at the transmitter provide no added generality since those choices (for all possible θ) could be made earlier at the receiver with no loss of performance.

Fano's inequality can be applied for each $\mathfrak{f}_\tau$ to upper-bound the conditional entropy $\mathcal{H}_{\mathfrak{f}_\tau}$ in terms of the error probability of the decoding at $\mathfrak{f}_\tau$. Taking the expectation[6] of these inequalities over $\mathfrak{f}_\tau \in \mathcal{F}_\tau$, and using the concavity of the binary entropy $\mathfrak{h}(x)$, the expected value $\boldsymbol{E}[\mathcal{H}_{\mathcal{F}_\tau}]$ of the entropy at the decoding time can be upper-bounded by

$$\begin{aligned}\boldsymbol{E}[\mathcal{H}_{\mathcal{F}_\tau}] &\le \mathfrak{h}(P_e) + P_e \ln(M-1) \\ &\le P_e(\ln M - \ln P_e + 1). \end{aligned} \tag{33}$$

This suggests that the conditional entropy is usually very small at the decoding time, motivating a focus on how fast the logarithm of the entropy changes in the second phase of the analysis below.

B. Bounds on the Change of Conditional Entropy

For any DMC, any code, and any $\mathcal{P} \ge 0$, define a sequence of random variables $\{\boldsymbol{V}_n; n \ge 0\}$ as

$$\boldsymbol{V}_n \triangleq \mathcal{H}_{\mathcal{F}_n} + n\mathcal{C}(\mathcal{P}) + \gamma_{\mathcal{C}}^{\mathcal{P}}(\boldsymbol{E}[\mathcal{S}_n|\,\mathcal{F}_n] - n\mathcal{P}) \tag{34}$$

where $\gamma_{\mathcal{C}}^{\mathcal{P}} \ge 0$ is the Lagrange multiplier for the cost constraint in the maximization of $\mathcal{C}(\mathcal{P})$ over input probability distributions in (2). This stochastic sequence will be used to bound the entropy and energy changes in phase 1.

Lemma 4: For any DMC, any code, and any $\mathcal{P} \ge 0$, the sequence $\{\boldsymbol{V}_n; n \ge 0\}$ is a submartingale,[7] i.e.,

$$\boldsymbol{E}[|\boldsymbol{V}_n|] < \infty \text{ and } \boldsymbol{E}[\boldsymbol{V}_{n+1}|\,\mathcal{F}_n] \ge \boldsymbol{V}_n, \quad \text{for all } n \ge 0.$$

Note that $\boldsymbol{E}[\boldsymbol{V}_{n+1}|\,\mathcal{F}_n]$ is a random variable whose sample values are determined by the particular $\mathfrak{f}_n \in \mathcal{F}_n$ realized. In that respect, inequality in Lemma 4 is between two random variables.

This lemma applies to all codes, whether or not they have a cost constraint equal to the $\mathcal{P}$ in the definition of $\boldsymbol{V}_n$. Proofs of Lemmas 4 to 8 are given in the Appendix .

The following two lemmas develop another submartingale based on the log entropy.

Lemma 5: For any DMC with all $P_{kj} > 0$, any code, and any $n \ge 0$

$$\boldsymbol{E}\left[\ln \mathcal{H}_{\mathcal{F}_n} - \ln \mathcal{H}_{\mathcal{F}_{n+1}} \big|\, \mathcal{F}_n\right] \le \boldsymbol{E}\left[\mathcal{D}_{X_{n+1}} \big|\, \mathcal{F}_n\right]. \tag{35}$$

Another stochastic sequence, $\{\boldsymbol{W}_n; n \ge 0\}$ is now defined that combines the changes in log entropy and cost.

$$\boldsymbol{W}_n \triangleq \ln \mathcal{H}_{\mathcal{F}_n} + n\mathcal{D}(\mathcal{P}) + \gamma_{\mathcal{D}}^{\mathcal{P}}(\boldsymbol{E}[\mathcal{S}_n|\,\mathcal{F}_n] - n\mathcal{P}) \tag{36}$$

where $\gamma_{\mathcal{D}}^{\mathcal{P}} \ge 0$ is the Lagrange multiplier for the cost constraint in the maximization of $\mathcal{D}(\mathcal{P})$ over input probabilities in (7).

[6]The facts that $|\mathcal{H}_{\mathcal{F}_\tau}| \le \ln M$ and $|P_e| \le 1$, combined with Lebesgue's dominated convergence theorem, [17, p. 187], allow us to interchange the limit and expectation here.

[7]See, for example, [17, Ch. VII] for a treatment of submartingales.

Lemma 6: For any DMC with all $P_{kj} > 0$, for any code, and for any $\mathcal{P} \ge 0$, the sequence $\{\boldsymbol{W}_n; n \ge 0\}$ is a submartingale, i.e.,

$$\boldsymbol{E}[|\boldsymbol{W}_n|] < \infty \text{ and } \boldsymbol{E}[\boldsymbol{W}_{n+1}|\,\mathcal{F}_n] \ge \boldsymbol{W}_n, \quad \text{for all } n \ge 0.$$

C. Measuring Time With Submartingales

The following lemmas are used to lower-bound the expected value of certain stopping times and, consequently, the durations of phases 1 and 2 in terms of $\{\boldsymbol{V}_n; n \ge 0\}$ and $\{\boldsymbol{W}_n; n \ge 0\}$, respectively.

Lemma 7: For any DMC, any $\mathcal{P} \ge 0$, and any code, if a stopping time τ_1 satisfies

$$\boldsymbol{E}[\tau_1] < \infty \quad \text{and} \quad \boldsymbol{E}[\mathcal{S}_{\tau_1}] \le \mathcal{P}\boldsymbol{E}[\tau_1]$$

then

$$\mathcal{C}(\mathcal{P})\boldsymbol{E}[\tau_1] \ge \boldsymbol{E}\left[\mathcal{H}_{\mathcal{F}_0} - \mathcal{H}_{\mathcal{F}_{\tau_1}}\right]. \tag{37}$$

Lemma 8: For any DMC with all $P_{kj} > 0$, any $\mathcal{P} \ge 0$, and any code, if a pair of stopping times $(\tau_1 \le \tau_2)$ satisfies

$$\boldsymbol{E}[\tau_2] < \infty \quad \text{and} \quad \boldsymbol{E}[\mathcal{S}_{\tau_2} - \mathcal{S}_{\tau_1}] \le \mathcal{P}\boldsymbol{E}[\tau_2 - \tau_1]$$

then

$$\mathcal{D}(\mathcal{P})\boldsymbol{E}[\tau_2 - \tau_1] \ge \boldsymbol{E}\left[\ln \mathcal{H}_{\mathcal{F}_{\tau_1}} - \ln \mathcal{H}_{\mathcal{F}_{\tau_2}}\right]. \tag{38}$$

The bounds asserted by these lemmas are tight in the sense that when they are used with the stopping times to be specified later, they will show that $E(R, \mathcal{P})$ in (20) is an upper bound on the reliability function.

D. Lower-Bounding $\overline{\tau}$

We now derive lower bounds on the expected decoding time for any variable-length block code with M equiprobable messages, subject to a given cost constraint $\mathcal{P}$ and a required probability of error P_e. The first result is simply an explicit statement of the well-known impossibility of transmitting reliably at rates above $\mathcal{C}(\mathcal{P})$.

Theorem 3: For any DMC, any code with $M \ge 2$ equiprobable messages, any $\mathcal{P} \ge 0$, and any required error probability $P_e \ge 0$ and cost constraint $\mathcal{P}$, the expected decoding time satisfies

$$\boldsymbol{E}[\tau] \ge \frac{\ln M - P_e(\ln M - \ln P_e + 1)}{\mathcal{C}(\mathcal{P})}. \tag{39}$$

Proof: From (33), $\boldsymbol{E}[\mathcal{H}_{\mathcal{F}_\tau}] \le P_e(\ln M - \ln P_e + 1)$. Thus, since $\mathcal{H}_{\mathcal{F}_0} = \ln M$, (39) results from (37) with $\tau_1 = \tau$. **QED**

This result is valid both for the case where all $P_{kj} > 0$ and the zero-error case where some $P_{kj} = 0$. In the zero-error case, we already know that $P_e = 0$ is asymptotically achievable for $R < \mathcal{C}(\mathcal{P})$, so our remaining task is to show that $E(R, \mathcal{P})$ is an

upper bound as well as a lower bound to the reliability function in the case where $R < \mathcal{C}(\mathcal{P})$ and all $P_{kj} > 0$.

E. Lower Bounding $\overline{\tau}$ for DMCs With All $P_{kj} > 0$

The main issue in this lower bound is finding an intermediate Markov stopping time τ_1 which will divide the message transmission interval into two disjoint phases[8] such that the duration of each can be lower-bounded tightly by Lemmas 7 and 8, respectively. Consider the stopping time $t_1 = \min\{n \mid \mathcal{H}_{\mathcal{F}_n} \leq 1\}$ in filtration $\mathcal{F}$. This does not quite work as an intermediate stopping time, since a variable-length code could in principle occasionally decode before the conditional entropy, $\mathcal{H}_{\mathcal{F}_n}$ drops below 1. Instead, we use $\tau_1 = \min(\tau, t_1)$ to define the end of the first phase. This is also a Markov stopping time, and $0 \leq \tau_1 \leq \tau$, so this is a well-defined intermediate time for all codes.

We now apply Lemma 7 to τ_1. Let $\boldsymbol{E}[\mathcal{S}_{\tau_1}]$ be the expected energy used by any given code in this first phase and let $\mathcal{P}_1 = \frac{\boldsymbol{E}[\mathcal{S}_{\tau_1}]}{\boldsymbol{E}[\tau_1]}$. Then (37) becomes

$$\boldsymbol{E}\left[\mathcal{H}_{\mathcal{F}_0} - \mathcal{H}_{\mathcal{F}_{\tau_1}}\right] \leq \mathcal{C}(\mathcal{P}_1)\boldsymbol{E}[\tau_1]. \tag{40}$$

We next find a lower bound to $\boldsymbol{E}\left[-\mathcal{H}_{\mathcal{F}_{\tau_1}}\right]$. By definition of t_1, $\mathcal{H}_{\mathcal{F}_{t_1}} \leq 1$, but $\mathcal{H}_{\mathcal{F}_{\tau_1}}$ might be greater than 1 if $\mathcal{H}_{\mathcal{F}_\tau} > 1$. Thus, we can upper-bound $\boldsymbol{E}\left[\mathcal{H}_{\mathcal{F}_{\tau_1}}\right]$ by

$$\begin{aligned}
\boldsymbol{E}\left[\mathcal{H}_{\mathcal{F}_{\tau_1}}\right] &\leq 1 + \boldsymbol{P}[\mathcal{H}_{\mathcal{F}_\tau} > 1]\,\boldsymbol{E}\left[\mathcal{H}_{\mathcal{F}_{\tau_1}} \mid \mathcal{H}_{\mathcal{F}_{\tau_1}} > 1\right] \\
&\leq 1 + \boldsymbol{P}[\mathcal{H}_{\mathcal{F}_\tau} > 1] \ln M \qquad (41)\\
&\leq 1 + \boldsymbol{E}[\mathcal{H}_{\mathcal{F}_\tau}] \ln M \qquad (42)\\
&\leq 1 + P_e(\ln M - \ln P_e + 1) \ln M, \qquad (43)
\end{aligned}$$

where in (41) we upper-bounded $\boldsymbol{E}\left[\mathcal{H}_{\mathcal{F}_{\tau_1}} \middle| \mathcal{H}_{\mathcal{F}_\tau} > 1\right]$ by $\ln M$, the maximum entropy for any ensemble of M elements. We used the Markov inequality in (42) and then (33) in (43).

Since the messages are *a priori* equiprobable, $\mathcal{H}_{\mathcal{F}_0} = \ln M$, so substituting this and (43) into (40)

$$\boldsymbol{E}[\tau_1] \geq \frac{\ln M\left[1 - P_e(\ln M - \ln P_e + 1) - \frac{1}{\ln M}\right]}{\mathcal{C}(\mathcal{P}_1)}. \tag{44}$$

As shown later, the term in brackets essentially approaches 1 as $P_e \to 0$ and, thus, $\boldsymbol{E}[\tau_1]$ is approximately lower-bounded by $(\ln M)/\boldsymbol{E}[\mathcal{C}(\mathcal{P}_1)]$.

Next we find the expected time $\boldsymbol{E}[\tau - \tau_1]$ spent in phase 2. Here we use (38) from Lemma 8, with the initial time τ_i chosen to be τ_1 and the final time τ_f chosen to be τ. Let $\boldsymbol{E}[\mathcal{S}_\tau - \mathcal{S}_{\tau_1}]$ be the expected energy used by the given code in this second phase and let $\mathcal{P}_2 = \frac{\boldsymbol{E}[\mathcal{S}_\tau - \mathcal{S}_{\tau_1}]}{\boldsymbol{E}[\tau - \tau_1]}$. Then

$$\boldsymbol{E}\left[\ln \mathcal{H}_{\mathcal{F}_{\tau_1}} - \ln \mathcal{H}_{\mathcal{F}_\tau}\right] \leq \mathcal{D}(\mathcal{P}_2)\boldsymbol{E}[\tau - \tau_1]. \tag{45}$$

[8]There is a nice intuitive relation between these two phases used in the converse and the two phases used in the variable-length block codes of Section II-C, since in each case the first phase deals with a large sea of messages and the second deals essentially with a binary hypothesis. When an error-and-erasure codeword is repeated, however, phase 1 as defined here could end during any one of those repetitions.

We lower-bound $\boldsymbol{E}\left[\ln \mathcal{H}_{\mathcal{F}_{\tau_1}} - \ln \mathcal{H}_{\mathcal{F}_\tau}\right]$ by upper-bounding $\boldsymbol{E}[\ln \mathcal{H}_{\mathcal{F}_\tau}]$ and lower-bounding $\boldsymbol{E}\left[\ln \mathcal{H}_{\mathcal{F}_{\tau_1}}\right]$. By Jensen's inequality, $\boldsymbol{E}[\ln \mathcal{H}_{\mathcal{F}_\tau}] \leq \ln \boldsymbol{E}[\mathcal{H}_{\mathcal{F}_\tau}]$, so from (33)

$$\boldsymbol{E}[\ln \mathcal{H}_{\mathcal{F}_\tau}] \leq \ln[P_e(\ln M - \ln P_e + 1)]. \tag{46}$$

To lower-bound $\boldsymbol{E}\left[\ln \mathcal{H}_{\mathcal{F}_{\tau_1}}\right]$, we use the following lemma.

Lemma 9: For any DMC with all $P_{kj} > 0$, any code, and any $n \geq 0$

$$\ln \mathcal{H}_{\mathcal{F}_n} - \ln \mathcal{H}_{\mathcal{F}_{n+1}} \leq \max_{k,m,j} \ln \frac{P_{kj}}{P_{mj}} \quad \triangleq \quad \mathsf{F}. \tag{47}$$

Since $\mathcal{H}_{\mathcal{F}_{\tau_1 - 1}} > 1$, i.e., $\ln \mathcal{H}_{\mathcal{F}_{\tau_1 - 1}} > 0$, the lemma implies that $\ln \mathcal{H}_{\mathcal{F}_{\tau_1}} \geq -\mathsf{F}$.

Substituting this and (46) into (45)

$$\boldsymbol{E}[\tau - \tau_1] \geq \frac{-\ln P_e - \mathsf{F} - \ln[\ln M - \ln P_e + 1]}{\mathcal{D}(\mathcal{P}_2)}. \tag{48}$$

As shown later, the numerator is essentially $(-\ln P_e)$ in the limit of small P_e. Now we can find a lower bound on $(-\ln P_e)/\overline{\tau}$, for codes of rate $(\ln M)/\overline{\tau}$, using the above result.

Theorem 4: Assume a DMC with all $P_{kj} > 0$. Let $\mathcal{P} \geq 0$, $0 \leq R \leq \mathcal{C}(\mathcal{P})$, and $\delta > 0$ be arbitrary. Then, for all sufficiently large $\overline{\tau}$, all variable-length block codes with

- expected energy $\boldsymbol{E}[\mathcal{S}_\tau] \leq \mathcal{P}\overline{\tau} + \delta$
- $M \geq \exp[\overline{\tau}(R + \delta)]$ equiprobable messages

must satisfy

$$P_e \geq \exp\{-\overline{\tau}[E(R, \mathcal{P}) + \delta]\}. \tag{49}$$

We now give an intuitive justification of the theorem; a proof is given in the Appendix. Leaving out the "negligible" terms, (44) and (48) are

$$\overline{\tau}_1 \geq \frac{\ln(M)}{\mathcal{C}(\mathcal{P}_1)}; \qquad \overline{\tau} - \overline{\tau}_1 \geq \frac{-\ln(P_e)}{\mathcal{D}(\mathcal{P}_2)}.$$

Defining $\eta = \frac{\overline{\tau}_1}{\overline{\tau}}$ for the given code and rearranging terms

$$\frac{\ln(M)}{\overline{\tau}} \leq \eta\mathcal{C}(\mathcal{P}_1); \qquad \frac{-\ln(P_e)}{\overline{\tau}} \leq (1 - \eta)\mathcal{D}(\mathcal{P}_2). \tag{50}$$

In any code that allocates its time and power between the two phases as $\overline{\tau}_1, \mathcal{P}_1$ and $\overline{\tau} - \overline{\tau}_1, \mathcal{P}_2$, (50) is the converse of Lemma 2. The exponent $E(R, \mathcal{P})$ is the result of optimizing over these parameters for given $R, \mathcal{P}$. The proof in the Appendix treats the neglected quantities and this optimization carefully.

IV. Extension to Other Memoryless Channels

The channel model of Sections II and III assumes finite input and output alphabets, but, as will be seen, the analysis is more general, and with some added assumptions, continues to hold with minor changes such as replacing sums and max's with integrals and sup's. A later paper by Burnashev [2], extends his results for the DMC to a more general class of discrete-time memoryless channels. The extension below generalizes his class of channels and adds arbitrary finite but not necessarily bounded cost assignments.

A. Assumptions About the Channel Model

The channel input and output alphabets can be countable or uncountably infinite and will be denoted by $\mathcal{X}$ and $\mathcal{Y}$, respectively. Each element $x \in \mathcal{X}$ has an associated cost ρ_x and, as before, we assume that the infimum of these costs is equal to zero.

Each input $x \in \mathcal{X}$ will have an associated probability measure ϑ_x governing the output conditional on input x; this replaces the transition matrix $\{P_{kj}\}$ for the DMC case. We will assume that there exists a probability measure ν, with respect to which all ϑ_x are absolutely continuous, i.e.,

$$\forall x \in \mathcal{X}, \qquad \nu \gg \vartheta_x.$$

Indeed, without this ν one can hardly begin to analyze such a memoryless channel. For each $x \in \mathcal{X}$, let ψ_x be the Radon–Nikodym derivative of ϑ_x with respect to ν

$$\psi_x = \frac{\vartheta_x}{\nu}. \tag{51}$$

If $\mathcal{X}$ and $\mathcal{Y}$ are each the set of real numbers and if probability densities exist, then $\psi_x(y)$ can be taken as the probability density of y conditional on x.

Our previous definitions can be extended by replacing sums with integrals, max's with sup's, etc.

$$\mathcal{C}(\mathcal{P}) \triangleq \sup_{\mu:\int_{\mathcal{X}} \rho_x d\mu \le \mathcal{P}} \int_{\mathcal{X}} d\mu \int_{\mathcal{Y}} \psi_x \ln \frac{\psi_x}{\psi_\mu} d\nu \tag{52}$$

where $\psi_\mu = \int_{\mathcal{X}} \psi_x d\mu$ is the unconditional probability measure on $\mathcal{Y}$ corresponding to the measure μ on $\mathcal{X}$. Similarly

$$\mathcal{D}_x \triangleq \sup_{\alpha \in \mathcal{X}} \int_{\mathcal{Y}} \psi_x \ln \frac{\psi_x}{\psi_\alpha} d\nu, \tag{53}$$

$$\mathcal{D}(\mathcal{P}) \triangleq \sup_{\mu:\int_{\mathcal{X}} \rho_x d\mu \le \mathcal{P}} \int_{\mathcal{X}} \mathcal{D}_x d\mu. \tag{54}$$

The following assumption, which clearly includes DMCs with all $P_{kj} > 0$, will ensure that $\mathcal{D}(\mathcal{P})$ and $\mathcal{C}(\mathcal{P})$ are finite for all $\mathcal{P} \ge 0$.

Assumption 1: The discrete time memoryless channel satisfies the following:

- $\forall x \in \mathcal{X}$, $\mathcal{D}_x < \infty$ and $\rho_x < \infty$,
- $\forall \mathcal{P} > 0$, $\Lambda(\mathcal{P}) \triangleq \sup_{x:\rho_x \le \mathcal{P}} \mathcal{D}_x < \infty$,
- $\limsup_{\mathcal{P}\to\infty} \frac{\Lambda(\mathcal{P})}{\mathcal{P}} < \infty$.

With the above assumptions replacing the DMC assumptions, equivalents of Lemma 1, Lemma 2, and Theorem 1 can be proven. The essential difference is that the symbol costs are not universally bounded by a constant like $\rho_{\max}$, as assumed earlier for the reject codeword. Consequently, the statements and proofs must be modified slightly, but the essence of the propositions will be the same. As an example, consider Lemma 1.

In establishing Lemma 1 for the DMC case, the use of letter k in phase 2 for the accept codeword $\boldsymbol{x}_a$ was matched, in the reject codeword $\boldsymbol{x}_r$, by the letter m_k that yielded the maximum divergence $\mathcal{D}_k$ with k. Here the use of letter x in $\boldsymbol{x}_a$ will be matched in $\boldsymbol{x}_r$ by some $g_\delta(x)$ that comes within a given δ of $\mathcal{D}_x$ in (53) and has a finite cost $\rho_{g_\delta(x)}$. For a cost constraint $\mathcal{P}_2$ in phase 2 here, the supremum $\mathcal{D}(\mathcal{P}_2)$ from (54) can be approached within δ by a linear combination of at most two input letters, say x_1 and x_2, so that

$$\alpha \mathcal{D}_{x_1} + (1-\alpha)\mathcal{D}_{x_2} \ge \mathcal{D}(\mathcal{P}_2) - \delta \tag{55}$$

$$\alpha \rho_{x_1} + (1-\alpha)\rho_{x_2} \le \mathcal{P}_2. \tag{56}$$

Using such a linear combination for $\boldsymbol{x}_a$, and using the matching $g_\delta(x_1)$ and $g_\delta(x_2)$ for $\boldsymbol{x}_r$, we see that $\boldsymbol{x}_a$ meets the cost constraint $\mathcal{P}_2$, the divergence per letter between $\boldsymbol{x}_a$ and $\boldsymbol{x}_r$ is at least $\mathcal{D}(\mathcal{P}_2) - 3\delta$, and the cost per letter of $\boldsymbol{x}_r$ is bounded by $\max(\rho_{g_\delta(x_1)}, \rho_{g_\delta(x_2)})$. With these modifications, Lemma 1 follows as before.

Lemma 2 and Theorem 1 $\rho_{\max}$ will then follow as before by replacing $\rho_{\max}$ by $\max(\rho_{g_\delta(x_1)}, \rho_{g_\delta(x_2)})$. The optimization problem for $\mathcal{C}(\mathcal{P})$ and $\mathcal{D}(\mathcal{P})$ is then the same as before, with the caveat that $\mathcal{C}(\mathcal{P})$, $\mathcal{D}(\mathcal{P})$, $\mathcal{P}$ can all be unbounded. However, the value $\eta'_{R,\mathcal{P}}$ can still be found in exactly the same way as in the DMC case, using the intersection of the $(x, \mathcal{C}(x))$ curve and the straight line passing through the origin and $(\mathcal{P}, R)$. If this intersect does not exist it will mean that all values of $\eta \ge 0$ are permissible. In short, the function $E(R, \mathcal{P})$ can be found using the same optimization problem and solution techniques in terms of the functions $\mathcal{C}(\mathcal{P})$ and $\mathcal{D}(\mathcal{P})$.

Proceeding on to the converse, it can be seen that the proofs of Lemmas 4 to 8 all hold under Assumption 1. In verifying these proofs, however, one must assume that all codes have finite expected energy; this is tacitly assumed in Theorem 4 since it is assumed throughout that $\overline{\tau} < \infty$. Lemma 9 does not hold in all cases, and in particular does not hold even for the amplitude-limited AWGNC. The following additional assumption will hold in many cases where Lemma 9 does not hold and will enable Theorem 4 to be proven.

Assumption 2: The discrete-time memoryless channel has an associated function $\xi(\cdot)$ such that:

- for any coding and any n

$$\boldsymbol{E}\left[\left[\ln \frac{\mathcal{H}_{\mathcal{F}_n}}{\mathcal{H}_{\mathcal{F}_{n+1}}}\right]_{(a)} \middle| \mathcal{F}_n\right] \le \xi(a)\left(1 + \boldsymbol{E}\left[\mathcal{S}_{n+1} - \mathcal{S}_n \middle| \mathcal{F}_n\right]\right);$$

- $\lim_{a\to\infty} \xi(a) = 0$;

where $[x]_{(a)} = x \mathbb{I}_{\{x \ge a\}}$.

Theorem 4 is proved for stationary memoryless channels satisfying Assumptions 1 and 2 in the Appendix.

B. Discussion of Extended Channel Models

It is natural at this point to ask what kinds of channels satisfy Assumptions 1 and 2. A partial answer comes from considering the class of channels without cost constraints considered by Burnashev in [2]. He shows that any channel satisfying the following conditions has an error exponent given by $E = (1 - R/\mathcal{C})\mathcal{D}$.

- $\mathcal{D} = \sup_{\alpha,\beta\in\mathcal{X}} \int \psi_\alpha \ln\left(\frac{\psi_\alpha}{\psi_\beta}\right) d\nu < \infty$.
- $\phi(a) = \sup_{\alpha,\beta\in\mathcal{X}} \int \psi_\alpha \ln\left(\frac{\psi_\alpha}{\psi_\beta}\right) \mathbb{I}_{\left\{\ln\left(\frac{\psi_\alpha}{\psi_\beta}\right) > a\right\}} d\nu < \infty$, and $\lim_{a\to\infty} \phi(a) = 0$.
- At least one of the following is satisfied:
  - The channel is an additive noise channel whose input alphabet is a closed interval on the real line and whose noise has a unimodal density.

— $\exists K > 0$ such that

$$\sup_{\alpha,\beta\in\mathcal{X}} \int \psi_\alpha \left(\ln \frac{\psi_\alpha}{\psi_\beta}\right)^{(1+K)} d\nu < \infty.$$

His first assumption, $\sup_{x\in\mathcal{X}} \mathcal{D}_x < \infty$, implies that the channel satisfies our Assumption 1 for all nonnegative finite-cost assignments. He shows that the other assumptions imply that a function $\xi(a)$, $a \geq 0$ exists such that $\lim_{a\to\infty} \xi(a) = 0$ and such that for all codes

$$\boldsymbol{E}\left[\left[\ln \mathcal{H}_{\mathcal{F}_n} - \ln \mathcal{H}_{\mathcal{F}_{n+1}}]_{(a)}\right| \mathcal{F}_n\right] \leq \xi(a). \tag{57}$$

This implies that the channel satisfies our Assumption 2 for all nonnegative finite-cost assignments. Thus, his assumptions imply our assumptions for all cost assignments and, consequently, for all cost assignments, the corresponding $E(R,\mathcal{P})$ exists and is the reliability function.

We next give an example of a channel that does not satisfy the conditions given in [2] and does not have a finite reliability function without a cost constraint, but does satisfy our conditions and has a finite cost constrained reliability function $E(R,\mathcal{P})$.

Let $\mathcal{X}$ and $\mathcal{Y}$ be the set of nonnegative integers and assume the cost function $\rho_x = x^2$, i.e., the cost of each input letter is equal to the square of the value of the corresponding real number. Let the transition probability P_{xy} be

$$P_{xy} = \frac{1}{3}\left(\frac{1}{2^y} + \delta[x-y]\right)$$

where $\delta[\cdot]$ is 1 when its argument is 0, and 0 elsewhere.

This channel can be proved to satisfy Assumptions 1 and 2, and thus its error exponent is given by $E(R,\mathcal{P})$. On the other hand, it does not satisfy the necessary conditions of [2] and the reliability function is unbounded for any rate below capacity if there is no cost constraint.

A major reason for this investigation of cost constraints with feedback has been to achieve a better understanding of why error probability can be made to decrease faster than exponentially with constraint length for the AWGNC with feedback. It is easy to see that $\mathcal{D}_x$ in (53) is infinite for all x, and thus codes of the Yamamoto and Itoh type have unbounded exponents for the AWGNC. If an amplitude constraint A is imposed at the channel input, then it is equally easy to see that $\mathcal{D}_x$ is finite for all x. The supremum in (53) becomes a maximum at $\alpha = \pm A$, with the sign opposite to that of x. In fact, assuming unit variance noise

$$\mathcal{D}(\mathcal{P}) = \frac{(A+\sqrt{\mathcal{P}})^2}{2}, \qquad \text{for } \mathcal{P} \leq A^2.$$

After the optimization procedure of Section II-D, an exponent $E_A(R,\mathcal{P})$ can be calculated which is rapidly increasing in A but finite for all $A < \infty$. This exponent is an upper bound to that in the fixed length case.

The insight to be gained by this is that the faster-than-exponential decay of error probability for the classical AWGNC is due to the ability to transmit at unbounded amplitudes when errors are immanent. It is important to note that having variable (and unbounded) block length subject to a mean is not a substitute for unbounded amplitude. The variable length increases the exponent, but cannot make it unbounded.

V. Conclusion

Theorems 1 and 4 specify the reliability function for the class of variable-length block codes for DMCs with cost constraints, all $P_{kj} > 0$, and ideal[9] feedback. The results are extended to a more general class of discrete-time memoryless channels satisfying Assumptions 1 and 2 of Section IV. AWGNCs with amplitude and power constraints provide examples satisfying these assumptions. Theorem 2 shows that zero error probability is achievable at all rates up to the cost constrained capacity and moreover is achievable by a very simple scheme if the channel has one or more zero probability transitions.

The rate and the error exponent are specified in terms of the expected block length. By looking at a long sequence of successive message transmissions, it is evident from the law of large numbers that the rate corresponds to the average number of message bits transmitted per unit time. In the same way, the cost constraint is satisfied as an average over both time and channel behavior. The theorems then say essentially that the probability of error $P_{e,\min}$ for the best variable-length block code of given $R,\mathcal{P}$, and $\overline{\tau}$ satisfies $\frac{-\ln P_{e,\min}}{\overline{\tau}} \to E(R,\mathcal{P})$ as $\overline{\tau}\to\infty$.

Mathematically, these theorems are quite similar to the conventional non-feedback block-coding results except for the following differences: first, the reliability function is known for all rates rather than rates sufficiently close to capacity; second, the reliability function is concave (and sometimes positive at capacity); and third, the reliability function is given in terms of expected rather than actual block length. The first two differences have been discussed in detail in the previous sections.

In order to understand the role of the expected block length on the exponent, look at the coding scheme used for achievability. The expected block length $\overline{\tau}$ is close to the fixed block length of the underlying error-and-erasure code and it is this code that determines $E(R,\mathcal{P})$. In other words, the variable-length feature is essential for the small error probability, but τ is constant with high probability.

One might think that a variable-length block code has many system disadvantages over a fixed-length code, but this is not really true (except for time-sensitive systems) since variable-length protocols are almost invariably used at all higher layers. As discussed in Section II, it can be shown that the expected additional queuing delay introduced by those codes can be made to approach 0 with increasing $\overline{\tau}$.

Appendix

A. Proof of Lemma 3 (Concavity of $E(R,\mathcal{P},\eta)$)

For any given DMC, let Ω be the set of triples $(R,\mathcal{P},\eta)$ for which $0 < R < \mathcal{C}(\mathcal{P})$ and $\eta \in \mathcal{I}_{R,\mathcal{P}}$.

$$\Omega = \{(R,\mathcal{P},\eta) : \mathcal{P} \geq 0, 0 < R < \mathcal{C}(\mathcal{P}), \eta \in \mathcal{I}_{R,\mathcal{P}}\}. \tag{58}$$

First we show that Ω is a convex set, and then we show that $E(R,\mathcal{P},\eta)$ is a concave function over the domain Ω.

[9]Indeed, as argued previously, noiseless feedback of rate $\mathcal{C}$ or higher with bounded delay is enough.

Assume that $(R_a, \mathcal{P}_a, \eta_a)$ and $(R_b, \mathcal{P}_b, \eta_b)$ are arbitrary points of Ω. We show that Ω is a convex set by showing that for any $\alpha \in (0,1)$, the point $(R_\alpha, \mathcal{P}_\alpha, \eta_\alpha)$ given by

$$\mathcal{P}_\alpha = \alpha\mathcal{P}_a + (1-\alpha)\mathcal{P}_b \tag{59}$$
$$R_\alpha = \alpha R_a + (1-\alpha)R_b \tag{60}$$
$$\eta_\alpha = \alpha\eta_a + (1-\alpha)\eta_b \tag{61}$$

is also in the set Ω.

R_α is clearly positive, and using the concavity of $\mathcal{C}(\cdot)$, we get

$$\begin{aligned} R_\alpha &< \alpha\mathcal{C}(\mathcal{P}_a) + (1-\alpha)\mathcal{C}(\mathcal{P}_b) \\ &\le \mathcal{C}(\alpha\mathcal{P}_a + (1-\alpha)\mathcal{P}_b) \\ &= \mathcal{C}(\mathcal{P}_\alpha). \end{aligned} \tag{62}$$

We must also show that $\eta_\alpha \in \mathcal{I}_{R_\alpha,\mathcal{P}_\alpha}$. It suffices to show that $\eta_\alpha \ge \eta'_{R_\alpha,\mathcal{P}_\alpha}$, $\eta_\alpha < 1$, and $\eta_\alpha \le R_\alpha/C(0)$. The latter two conditions are obvious, so we will show only that $\eta_\alpha \ge \eta'_{R_\alpha,\mathcal{P}_\alpha}$. As discussed in Section II-D, the condition $\eta \ge \eta'_{R,\mathcal{P}}$ is equivalent to $R \le \eta\mathcal{C}(\mathcal{P}/\eta)$.

To show that first note that

$$R_a \le \eta_a \mathcal{C}\left(\frac{\mathcal{P}_a}{\eta_a}\right) \quad \text{and} \quad R_b \le \eta_b \mathcal{C}\left(\frac{\mathcal{P}_b}{\eta_b}\right)$$

which implies that

$$R_\alpha \le \alpha\eta_a \mathcal{C}\left(\frac{\mathcal{P}_a}{\eta_a}\right) + (1-\alpha)\mathcal{C}\left(\frac{\mathcal{P}_b}{\eta_b}\right).$$

Thus

$$\begin{aligned} R_\alpha &\le \eta_\alpha \left(\frac{\alpha\eta_a}{\eta_\alpha}\mathcal{C}\left(\frac{\mathcal{P}_a}{\eta_a}\right) + \frac{(1-\alpha)\eta_b}{\eta_\alpha}\mathcal{C}\left(\frac{\mathcal{P}_b}{\eta_b}\right)\right) \\ &\le \eta_\alpha \mathcal{C}\left(\frac{\alpha\eta_a}{\eta_\alpha}\frac{\mathcal{P}_a}{\eta_a} + \frac{(1-\alpha)\eta_b}{\eta_\alpha}\frac{\mathcal{P}_b}{\eta_b}\right) \\ &= \eta_\alpha \mathcal{C}\left(\frac{\alpha\mathcal{P}_a + (1-\alpha)\mathcal{P}_b}{\eta_\alpha}\right) \\ &= \eta_\alpha \mathcal{C}\left(\frac{\mathcal{P}_\alpha}{\eta_\alpha}\right). \end{aligned}$$

Consequently, Ω is a convex region. We next show that $E(R,\mathcal{P},\eta)$ is concave over Ω. That is, given points $(R_a,\mathcal{P}_a,\eta_a), (R_b,\mathcal{P}_b,\eta_b)$, and $(R_\alpha,\mathcal{P}_\alpha,\eta_\alpha)$ in Ω, we will show that

$$\alpha E(R_a,\mathcal{P}_a,\eta_a) + (1-\alpha)E(R_b,\mathcal{P}_b,\eta_b) \le E(R_\alpha,\mathcal{P}_\alpha,\eta_\alpha). \tag{63}$$

Let us start with the left-hand side of the inequality (63)

$$\begin{aligned} &\alpha E(R_a,\mathcal{P}_a,\eta_a) + (1-\alpha)E(R_b,\mathcal{P}_b,\eta_b) \\ &\quad = \alpha(1-\eta_a)\mathcal{D}(\mathcal{P}_{a2}) + (1-\alpha)(1-\eta_b)\mathcal{D}(\mathcal{P}_{b2}) \\ &\quad \le (1-\eta_\alpha)\mathcal{D}\left(\frac{\alpha(1-\eta_a)\mathcal{P}_{a2}}{1-\eta_\alpha} + \frac{(1-\alpha)(1-\eta_b)\mathcal{P}_{b2}}{1-\eta_\alpha}\right) \end{aligned} \tag{64}$$

where we have used the concavity of $\mathcal{D}(\cdot)$ together with (61) in the last step.

$$\begin{aligned} &\frac{\alpha(1-\eta_a)\mathcal{P}_{a2}}{1-\eta_\alpha} + \frac{(1-\alpha)(1-\eta_b)\mathcal{P}_{b2}}{1-\eta_\alpha} \\ &= \frac{\alpha\left(\mathcal{P}_a - \eta_a\mathcal{C}^{-1}\left(\frac{R_a}{\eta_a}\right)\right)}{1-\eta_\alpha} + \frac{(1-\alpha)\left(\mathcal{P}_b - \eta_b\mathcal{C}^{-1}\left(\frac{R_b}{\eta_b}\right)\right)}{1-\eta_\alpha} \\ &= \frac{\mathcal{P}_\alpha - \left[\alpha\eta_a\mathcal{C}^{-1}\left(\frac{R_a}{\eta_a}\right) + (1-\alpha)\eta_b\mathcal{C}^{-1}\left(\frac{R_b}{\eta_b}\right)\right]}{1-\eta_\alpha} \\ &\le \frac{\mathcal{P}_\alpha - \eta_\alpha\mathcal{C}^{-1}\left(\frac{\alpha R_a + (1-\alpha)R_b}{\eta_\alpha}\right)}{1-\eta_\alpha}. \end{aligned} \tag{65}$$

The inequality above follows the convexity of $\mathcal{C}^{-1}(\cdot)$. Using the fact that $\mathcal{D}(\cdot)$ is nondecreasing together with the inequalities given in (64) and (65) we get the inequality (63) **QED**

B. Proof of Lemma 4

In the following proofs, we will use the following shorthand notation:

$$\begin{aligned} p(i) &= \boldsymbol{P}[\theta = i \,|\, \mathcal{F}_n] & p(i|j) &= \boldsymbol{P}[\theta = i \,|\, \mathcal{F}_n, Y_{n+1} = j] \\ \varphi(k) &= \boldsymbol{P}[X_{n+1} = k \,|\, \mathcal{F}_n] & \varphi(k|i) &= \boldsymbol{P}[X_{n+1} = k \,|\, \mathcal{F}_n, \theta = i] \\ \psi(j) &= \boldsymbol{P}[Y_{n+1} = j \,|\, \mathcal{F}_n] & \psi(j|i) &= \boldsymbol{P}[Y_{n+1} = j \,|\, \mathcal{F}_n, \theta = i]. \end{aligned}$$

Each of these quantities are random variables whose sample values for any given $\mathfrak{f}_n \in \mathcal{F}_n$ are probabilities or conditional probabilities of messages, channel inputs, or channel outputs. The reader can then interpret the following arguments as holding for each sample value $\mathfrak{f}_n \in \mathcal{F}_n$ and thus for the random variables themselves.

We will first prove that $\boldsymbol{E}[|\boldsymbol{V}_n|] < \infty$. Recall

$$\boldsymbol{V}_n = \mathcal{H}_{\mathcal{F}_n} + n\mathcal{C}(\mathcal{P}) + \gamma_{\mathcal{C}}^{\mathcal{P}}\left(\boldsymbol{E}[\mathcal{S}_n \,|\, \mathcal{F}_n] - \mathcal{P}n\right). \tag{66}$$

Since $\gamma_{\mathcal{C}}^{\mathcal{P}} \ge 0$, $\mathcal{C}(\mathcal{P}) \ge 0$, and $\boldsymbol{E}[\mathcal{S}_n \,|\, \mathcal{F}_n] \ge 0$

$$|\boldsymbol{V}_n| \le |\mathcal{H}_{\mathcal{F}_n}| + n\mathcal{C}(\mathcal{P}) + \gamma_{\mathcal{C}}^{\mathcal{P}}\left(\boldsymbol{E}[\mathcal{S}_n \,|\, \mathcal{F}_n] + \mathcal{P}n\right). \tag{67}$$

Note that $|\mathcal{H}_{\mathcal{F}_n}| \le \ln M$, thus

$$\boldsymbol{E}[|\boldsymbol{V}_n|] \le \ln M + n\mathcal{C}(\mathcal{P}) + \gamma_{\mathcal{C}}^{\mathcal{P}}\left(\boldsymbol{E}[\mathcal{S}_n] + \mathcal{P}n\right). \tag{68}$$

In addition for any finite-energy code[10] $\boldsymbol{E}[\mathcal{S}_n] < \infty$. Consequently, $\boldsymbol{E}[|\boldsymbol{V}_n|] < \infty$.

We next prove that $\boldsymbol{E}[\boldsymbol{V}_{n+1} \,|\, \mathcal{F}_n] \ge \boldsymbol{V}_n$. Following the definition of $\boldsymbol{V}_n$

$$\begin{aligned} \boldsymbol{E}[\boldsymbol{V}_{n+1} \,|\, \mathcal{F}_n] = \boldsymbol{E}\left[\mathcal{H}_{\mathcal{F}_{n+1}} \,\middle|\, \mathcal{F}_n\right] + (n+1)\left(\mathcal{C}(\mathcal{P}) - \gamma_{\mathcal{C}}^{\mathcal{P}}\mathcal{P}\right) \\ + \gamma_{\mathcal{C}}^{\mathcal{P}}\boldsymbol{E}\left[\mathcal{S}_n + \rho_{X_{n+1}} \,\middle|\, \mathcal{F}_n\right]. \end{aligned} \tag{69}$$

Note that

$$\begin{aligned} \mathcal{H}_{\mathcal{F}_n} - \boldsymbol{E}\left[\mathcal{H}_{\mathcal{F}_{n+1}} \,\middle|\, \mathcal{F}_n\right] &= \sum_{\theta\in\mathcal{M}} p(i)\log\frac{1}{p(i)} \\ &\quad - \sum_{j\in\mathcal{Y}}\psi(j)\sum_{\theta\in\mathcal{M}} p(i|j)\log\frac{1}{p(i|j)} \\ &= \sum_{\theta\in\mathcal{M}, j\in\mathcal{Y}} \psi(j)p(i|j)\log\frac{p(i|j)}{p(i)} \\ &\triangleq I(\theta; Y_{n+1}|\mathcal{F}_n). \end{aligned} \tag{70}$$

Note in the notation above about the conditional mutual information is different from the usual information theory conven-

[10]The convention for extending the encoding algorithms beyond decoding time is assigning all of the codewords to the same zero-cost symbol.

tion, where "$I(\theta;Y_{n+1}|\mathcal{F}_n)$" denotes an expectation over $\mathcal{F}_n$ as well as $\mathcal{M}$ and $\mathcal{Y}$, i.e., $\bar{\boldsymbol{E}}[I(\theta;Y_{n+1}|\mathcal{F}_n)]$.

Thus, (66) and (69) lead to

$$\boldsymbol{E}[\boldsymbol{V}_{n+1}|\mathcal{F}_n] = \boldsymbol{V}_n - I(\theta;Y_{n+1}|\mathcal{F}_n) + \mathcal{C}(\mathcal{P}) - \gamma_{\mathcal{C}}^{\mathcal{P}}\mathcal{P} + \gamma_{\mathcal{C}}^{\mathcal{P}}\boldsymbol{E}\left[\rho_{X_{n+1}}\middle|\mathcal{F}_n\right].$$

Because of the Markov relation $\theta \leftrightarrow X_{n+1} \leftrightarrow Y_{n+1}$ which holds for all $\mathfrak{f}_n$ combined with the data processing inequality, we have

$$\boldsymbol{E}[\boldsymbol{V}_{n+1}|\mathcal{F}_n] \geq \boldsymbol{V}_n - I(X_{n+1};Y_{n+1}|\mathcal{F}_n) + \mathcal{C}(\mathcal{P}) - \gamma_{\mathcal{C}}^{\mathcal{P}}\mathcal{P} + \gamma_{\mathcal{C}}^{\mathcal{P}}\boldsymbol{E}\left[\rho_{X_{n+1}}\middle|\mathcal{F}_n\right]. \quad (71)$$

Note that

$$\mathcal{C}(\mathcal{P}) - \gamma_{\mathcal{C}}^{\mathcal{P}}\mathcal{P} = \max_{\varphi}\left(\boldsymbol{\mathcal{I}}(\varphi) - \gamma_{\mathcal{C}}^{\mathcal{P}}\sum_k \varphi(k)\rho_k\right) \quad (72)$$

where

$$\boldsymbol{\mathcal{I}}(\varphi) = \sum_{k,j}\varphi(k)P_{kj}\ln\frac{P_{kj}}{\sum_{k'}\varphi(k')P_{k'j}}$$

is the mutual information between the input and output symbols when the distribution of the input symbol is φ. Thus

$$\bar{\boldsymbol{E}}[\boldsymbol{V}_{n+1}|\mathcal{F}_n] \geq \boldsymbol{V}_n \quad (73)$$

and the stochastic sequence $\{\boldsymbol{V}_n, n\}$ is a submartingale. **QED**

C. Proof of Lemma 5

First note that for all j, log-sum inequality implies

$$\begin{aligned}\ln\frac{\mathcal{H}_{\mathcal{F}_n}}{\mathcal{H}_{\mathcal{F}_{n+1}}} &= \ln\frac{\sum_i p(i)\ln\frac{1}{p(i)}}{\sum_i p(i|j)\ln\frac{1}{p(i|j)}} \\ &= \frac{\sum_i p(i)\ln\frac{1}{p(i)}}{\mathcal{H}_{\mathcal{F}_n}}\ln\frac{\sum_i p(i)\ln\frac{1}{p(i)}}{\sum_i p(i|j)\ln\frac{1}{p(i|j)}} \\ &\leq \frac{1}{\mathcal{H}_{\mathcal{F}_n}}\sum_i p(i)\ln\frac{1}{p(i)}\ln\frac{p(i)\ln\frac{1}{p(i)}}{p(i|j)\ln\frac{1}{p(i|j)}}. \quad (74)\end{aligned}$$

Consequently

$$\begin{aligned}&\bar{\boldsymbol{E}}\left[\ln\mathcal{H}_{\mathcal{F}_n} - \ln\mathcal{H}_{\mathcal{F}_{n+1}}\middle|\mathcal{F}_n = \mathfrak{f}_n\right] \\ &\quad\leq \sum_j \psi(j)\frac{1}{\mathcal{H}_{\mathcal{F}_n}}\sum_i\left(p(i)\ln\frac{1}{p(i)}\right)\ln\frac{p(i)\ln\frac{1}{p(i)}}{p(i|j)\ln\frac{1}{p(i|j)}} \\ &\quad= \sum_i \frac{p(i)\ln\frac{1}{p(i)}}{\mathcal{H}_{\mathcal{F}_n}}\sum_j\psi(j)\ln\frac{p(i)\ln\frac{1}{p(i)}}{p(i|j)\ln\frac{1}{p(i|j)}}. \quad (75)\end{aligned}$$

Using $\frac{p(i)}{p(i|j)} = \frac{\psi(j)}{\psi(j|i)}$ and defining

$$\psi(j|\bar{i}) \triangleq \boldsymbol{P}[Y_{n+1}{=}j|\mathcal{F}_n{=}\mathfrak{f}_n, \theta{\neq}i]$$

we get

$$\begin{aligned}&\sum_j\psi(j)\ln\frac{p(i)\ln p(i)}{p(i|j)\ln p(i|j)} \\ &\quad= \sum_j\psi(j)\ln\left(\frac{\psi(j)}{\psi(j|i)}\frac{\ln p(i)}{\ln p(i|j)}\right) \\ &\quad= \sum_j[p(i)\psi(j|i) + (1{-}p(i))\psi(j|\bar{i})]\ln\left(\frac{\psi(j)}{\psi(j|i)}\ \frac{\ln p(i)}{\ln p(i|j)}\right) \\ &\quad\leq p(i)\sum_j\psi(j|i)\ln\frac{\psi(j|i)}{\psi(j|\bar{i})} + (1{-}p(i))\sum_j\psi(j|\bar{i})\ln\frac{\psi(j|\bar{i})}{\psi(j|i)}. \quad (76)\end{aligned}$$

In order to verify the inequality in (76), denote the right side minus the left side as A, and note that by substitution

$$\begin{aligned}A &= p(i)\sum_j\psi(j|i)\ln\frac{\psi(j|i)}{\psi(j)} + (1{-}p(i))\sum_j\psi(j|\bar{i})\ln\frac{\psi(j|\bar{i})}{\psi(j)} \\ &\quad + p(i)\sum_j\psi(j|i)\ln\left(\frac{\psi(j|i)}{\psi(j|\bar{i})}\frac{\ln p(i|j)}{\ln p(i)}\right) \\ &\quad + (1{-}p(i))\sum_j\psi(j|\bar{i})\ln\left(\frac{\ln p(i|j)}{\ln p(i)}\right).\end{aligned}$$

The first two terms above are divergences, and thus nonnegative. The third term can be rewritten as below and is shown to be nonnegative by applying Jensen's inequality to the function $\ln\left(\frac{1}{x}\ln(1+\alpha x)\right)$ which is convex for any $\alpha > 0$.

$$\begin{aligned}&\sum_j\psi(j|i)\ln\left(\frac{\psi(j|i)}{\psi(j|\bar{i})}\frac{\ln p(i|j)}{\ln p(i)}\right) \\ &\quad= \sum_j\psi(j|i)\ln\frac{\psi(j|i)}{\psi(j|\bar{i})}\frac{\ln\left(1+\frac{1-p(i)}{p(i)}\frac{\psi(j|\bar{i})}{\psi(j|i)}\right)}{\ln\left(1+\frac{1-p(i)}{p(i)}\right)} \quad (77) \\ &\quad\geq 0. \quad (78)\end{aligned}$$

Similarly, the fourth term can be rewritten as below and is shown to be nonnegative by applying Jensen's inequality to the convex function $\ln\left(\ln(1+\frac{1}{\alpha x})\right)$ for $\alpha > 0$.

$$\begin{aligned}&\sum_j\psi(j|\bar{i})\ln\left(\frac{\ln p(i|j)}{\ln p(i)}\right) \\ &\quad= \sum_j\psi(j|\bar{i})\ln\frac{\ln\left(1+\frac{1-p(i)}{p(i)}\frac{\psi(j|\bar{i})}{\psi(j|i)}\right)}{\ln\left(1+\frac{1-p(i)}{p(i)}\right)} \quad (79) \\ &\quad\geq 0. \quad (80)\end{aligned}$$

This verifies (76). The final term in (76) can be upper-bounded by

$$\begin{aligned}&\sum_j\psi(j|\bar{i})\ln\frac{\psi(j|\bar{i})}{\psi(j|i)} \\ &\quad= \sum_j\left[\sum_k\varphi(k|\bar{i})P_{kj}\right]\ln\frac{\sum_k\varphi(k|\bar{i})P_{kj}}{\psi(j|i)\sum_k\varphi(k|\bar{i})}\end{aligned}$$

$$\stackrel{(a)}{\leq} \sum_j \sum_k \varphi(k|\bar{i}) P_{kj} \ln \frac{P_{kj}}{\psi(j|i)} \tag{81}$$

$$\stackrel{(b)}{\leq} \sum_k \varphi(k|\bar{i}) \mathcal{D}_k. \tag{82}$$

where (a) uses the log sum inequality over k for each j and (b) follows from the definition of $\mathcal{D}_k$.

By a similar argument on the first term in (76)

$$\sum_j \psi(j|i) \ln \frac{\psi(j|i)}{\psi(j|\bar{i})} \leq \sum_k \varphi(k|i) \mathcal{D}_k. \tag{83}$$

Substituting the above two inequalities into (76)

$$\sum_j \psi(j) \ln \frac{p(i) \ln p(i)}{p(i|j) \ln p(i|j)} \leq \sum_k \varphi(k) \mathcal{D}_k. \tag{84}$$

Thus

$$\begin{aligned} \boldsymbol{E}\left[\ln \frac{\mathcal{H}_{\mathcal{F}_n}}{\mathcal{H}_{\mathcal{F}_{n+1}}} \middle| \mathcal{F}_n = \mathfrak{f}_n \right] &\leq \sum_k \varphi(k) \mathcal{D}_k \\ &= \boldsymbol{E}\left[\mathcal{D}_{X_{n+1}} \middle| \mathcal{F}_n = \mathfrak{f}_n \right] \end{aligned} \tag{85}$$

which is equivalent to (35). **QED**

D. Proof of Lemma 6

We will first prove that $\boldsymbol{E}[\boldsymbol{W}_{n+1}|\mathcal{F}_n] \geq \boldsymbol{W}_n$. Recall from (36) that

$$\boldsymbol{W}_n = \ln \mathcal{H}_{\mathcal{F}_n} + n\mathcal{D}(\mathcal{P}) + \gamma_{\mathcal{D}}^{\mathcal{P}} (\boldsymbol{E}[\mathcal{S}_n|\mathcal{F}_n] - n\mathcal{P}) \tag{86}$$

where $\gamma_{\mathcal{D}}^{\mathcal{P}} \geq 0$ is the Lagrange multiplier for the cost constraint in the maximization of $\mathcal{D}(\mathcal{P})$ over input probabilities in (7). Consequently

$$\mathcal{D}(\mathcal{P}) = \max_{\phi} \sum_k \phi_k \mathcal{D}_k + \gamma_{\mathcal{D}}^{\mathcal{P}} \left(\sum_k \phi_k \rho_k - \mathcal{P} \right). \tag{87}$$

Using first Lemma 5 and then (87) we get

$$\begin{aligned} \boldsymbol{E}[\boldsymbol{W}_{n+1}|\mathcal{F}_n] &\geq \boldsymbol{W}_n + \mathcal{D}(\mathcal{P}) - \boldsymbol{E}\left[\mathcal{D}_{X_{n+1}} \middle| \mathcal{F}_n\right] \\ &\quad - \gamma_{\mathcal{D}}^{\mathcal{P}} \left(\boldsymbol{E}\left[\rho_{X_{n+1}} \middle| \mathcal{F}_n\right] - \mathcal{P} \right) \\ &\geq \boldsymbol{W}_n. \end{aligned} \tag{88}$$

We next prove that $\boldsymbol{E}[|\boldsymbol{W}_n|] < \infty$. Using the definition of $\boldsymbol{W}_n$ and the fact that $\gamma_{\mathcal{D}}^{\mathcal{P}} \geq 0$, $\mathcal{D}(\mathcal{P}) \geq 0$, and $\boldsymbol{E}[\mathcal{S}_n|\mathcal{F}_n] \geq 0$

$$|\boldsymbol{W}_n| \leq |\ln \mathcal{H}_{\mathcal{F}_n}| + n\mathcal{D}(\mathcal{P}) + \gamma_{\mathcal{D}}^{\mathcal{P}} (\boldsymbol{E}[\mathcal{S}_n|\mathcal{F}_n] + \mathcal{P}n). \tag{89}$$

Note that since $\mathcal{H}_{\mathcal{F}_n} \leq \mathcal{H}_{\mathcal{F}_0} = \ln M$

$$\boldsymbol{E}[|\boldsymbol{W}_n|] \leq \ln M + \boldsymbol{E}\left[\ln \frac{\mathcal{H}_{\mathcal{F}_0}}{\mathcal{H}_{\mathcal{F}_n}} \right] + n\mathcal{D}(\mathcal{P}) + \gamma_{\mathcal{D}}^{\mathcal{P}} (\boldsymbol{E}[\mathcal{S}_n] + \mathcal{P}n).$$

Since for any finite energy code $\boldsymbol{E}[\mathcal{S}_n] < \infty$, proving that $\boldsymbol{E}\left[\ln \frac{\mathcal{H}_{\mathcal{F}_0}}{\mathcal{H}_{\mathcal{F}_n}}\right] < \infty$, will establish $\boldsymbol{E}[|\boldsymbol{W}_n|] < \infty$.

Note that $\forall n < \infty$

$$\boldsymbol{E}\left[\ln \frac{\mathcal{H}_{\mathcal{F}_0}}{\mathcal{H}_{\mathcal{F}_n}} \right] = \boldsymbol{E}\left[\sum_{k=1}^{n} \ln \frac{\mathcal{H}_{\mathcal{F}_{k-1}}}{\mathcal{H}_{\mathcal{F}_k}} \right]. \tag{90}$$

Using the concavity of $\mathcal{D}(\mathcal{P})$ together with (85) we get

$$\boldsymbol{E}\left[\sum_{k=1}^{n} \ln \frac{\mathcal{H}_{\mathcal{F}_{k-1}}}{\mathcal{H}_{\mathcal{F}_k}} \right] \leq n\mathcal{D}\left(\frac{\boldsymbol{E}[\mathcal{S}_n]}{n} \right). \tag{91}$$

Recalling that $\boldsymbol{E}[\mathcal{S}_n] < \infty$, will prove $\boldsymbol{E}\left[\ln \frac{\mathcal{H}_{\mathcal{F}_0}}{\mathcal{H}_{\mathcal{F}_n}}\right] < \infty$ and thus, $\boldsymbol{E}\left[|W_n^{\mathcal{P}}|\right] < \infty$. **QED**

E. Proof of Lemma 7

By the definition of $\boldsymbol{V}_n$

$$\boldsymbol{V}_{\tau_i} = \mathcal{H}_{\mathcal{F}_{\tau_i}} + \tau_i \mathcal{C}(\mathcal{P}) + \gamma_{\mathcal{C}}^{\mathcal{P}} \left(\boldsymbol{E}[\mathcal{S}_{\tau_i}|\mathcal{F}_{\tau_i}] - \mathcal{P}\tau_i \right). \tag{92}$$

Since the expected value of each term on the right side exists

$$\begin{aligned} \boldsymbol{E}\left[\boldsymbol{V}_{\tau_i}^{\mathcal{P}}\right] &= \boldsymbol{E}\left[\mathcal{H}_{\mathcal{F}_{\tau_i}}\right] + \boldsymbol{E}[\tau_i] \mathcal{C}(\mathcal{P}) + \gamma_{\mathcal{C}}^{\mathcal{P}} \boldsymbol{E}[\mathcal{S}_{\tau_i} - \mathcal{P}\tau_i] \\ &\leq \boldsymbol{E}\left[\mathcal{H}_{\mathcal{F}_{\tau_i}}\right] + \boldsymbol{E}[\tau_i] \mathcal{C}(\mathcal{P}) \end{aligned} \tag{93}$$

where we have used $\gamma_{\mathcal{C}}^{\mathcal{P}} \geq 0$ along with the hypothesis of the lemma that $\boldsymbol{E}[\mathcal{S}_{\tau_i}] \leq \mathcal{P}\boldsymbol{E}[\tau_i]$. Since $\boldsymbol{E}\left[V_0^{\mathcal{P}}\right] = \boldsymbol{E}[\mathcal{H}_{\mathcal{F}_0}] = \ln M$, the result of the lemma, i.e., $\mathcal{C}(\mathcal{P})\boldsymbol{E}[\tau_i] \geq \boldsymbol{E}[\mathcal{H}_{\mathcal{F}_0}] - \boldsymbol{E}\left[\mathcal{H}_{\mathcal{F}_{\tau_i}}\right]$ will then hold if

$$\boldsymbol{E}[\boldsymbol{V}_{\tau_i}] \geq \boldsymbol{E}[\boldsymbol{V}_0] \tag{94}$$

holds. Doob's theorem (see [17, p. 485]) states that a submartingale $\boldsymbol{V}_n$ satisfies (94) if it satisfies the following two conditions:

$$\boldsymbol{E}[|\boldsymbol{V}_{\tau_i}|] < \infty \quad \text{and} \quad \lim_{n\to\infty} \boldsymbol{E}\left[|\boldsymbol{V}_n| \mathbb{I}_{\{\tau_i \geq n\}}\right] = 0. \tag{95}$$

The first condition follows from modifying (92) to bound $|\boldsymbol{V}_{\tau_i}^{\mathcal{P}}|$.

$$|\boldsymbol{V}_{\tau_i}^{\mathcal{P}}| \leq \mathcal{H}_{\mathcal{F}_{\tau_i}} + \tau_i \mathcal{C}(\mathcal{P}) + \gamma_{\mathcal{C}}^{\mathcal{P}} \left(\boldsymbol{E}[\mathcal{S}_{\tau_i}|\mathcal{F}_{\tau_i}] + \mathcal{P}\tau_i \right). \tag{96}$$

To establish the second condition, let

$$\begin{aligned} \xi_n &= |\boldsymbol{V}_n| \mathbb{I}_{\{\tau_i \geq n\}} \\ &= |\mathcal{H}_{\mathcal{F}_n} + n\mathcal{C}(\mathcal{P}) + \gamma_{\mathcal{C}}^{\mathcal{P}} \boldsymbol{E}[\mathcal{S}_n|\mathcal{F}_n] - \gamma_{\mathcal{C}}^{\mathcal{P}} \mathcal{P}n| \mathbb{I}_{\{\tau_i \geq n\}} \\ &\leq \left[\mathcal{H}_{\mathcal{F}_n} + n\mathcal{C}(\mathcal{P}) + \gamma_{\mathcal{C}}^{\mathcal{P}} \boldsymbol{E}[\mathcal{S}_n|\mathcal{F}_n] + \gamma_{\mathcal{C}}^{\mathcal{P}} \mathcal{P}n \right] \mathbb{I}_{\{\tau_i \geq n\}}. \end{aligned} \tag{97}$$

We want to find a random variable ζ of finite expectation that upper-bounds ξ_n for each n; the troublesome term here is $\boldsymbol{E}[\mathcal{S}_n|\mathcal{F}_n]$. Let $\mathcal{S}_n(m)$ (which is measurable in $\mathcal{F}_n$) denote the cost of the codeword corresponding to the message m at time n. The following very weak bound is sufficient for our purposes:

$$\boldsymbol{E}[\mathcal{S}_n|\mathcal{F}_n] = \sum_{m=1}^{M} \boldsymbol{P}[\theta{=}m|\mathcal{F}_n] \, \mathcal{S}_n(m) \leq \sum_{m=1}^{M} \mathcal{S}_n(m). \tag{98}$$

Substituting (98) into (97)

$$\xi_n \leq \left[\ln M + n\mathcal{C}(\mathcal{P}) + \gamma_{\mathcal{C}}^{\mathcal{P}} \sum_{m=1}^{M} \mathcal{S}_n(m) + \gamma_{\mathcal{C}}^{\mathcal{P}} \mathcal{P}n \right] \mathbb{I}_{\{\tau_i \geq n\}}$$

$$\begin{aligned}&\overset{(a)}{\leq}\left[\ln M+\tau_i\mathcal{C}(\mathcal{P})+\gamma_{\mathcal{C}}^{\mathcal{P}}\sum_{m=1}^{M}\mathcal{S}_{\tau_i}(m)+\gamma_{\mathcal{C}}^{\mathcal{P}}\mathcal{P}\tau_i\right]\mathbb{I}_{\{\tau_i\geq n\}}\\&\leq\left[\ln M+\tau_i\mathcal{C}(\mathcal{P})+\gamma_{\mathcal{C}}^{\mathcal{P}}\sum_{m=1}^{M}\mathcal{S}_{\tau_i}(m)+\gamma_{\mathcal{C}}^{\mathcal{P}}\mathcal{P}\tau_i\right]\\&\leq\left[\ln M+\tau_i\mathcal{C}(\mathcal{P})+\gamma_{\mathcal{C}}^{\mathcal{P}}M\boldsymbol{E}\left[\mathcal{S}_{\tau_i}\right]+\gamma_{\mathcal{C}}^{\mathcal{P}}\mathcal{P}\tau_i\right]\triangleq\zeta.\end{aligned}\quad(99)$$

In (a) we have used the fact that indicator function is zero if $\tau_i < n$; and $\mathcal{S}_{\tau_i}(m)\geq\mathcal{S}_n(m)$ if $\tau_i > n$. Note that $0\leq\xi_n\leq\zeta$ for all n. Since $\boldsymbol{E}[\tau_i]<\infty$ and $\boldsymbol{E}[\mathcal{S}_{\tau_i}]\leq\mathcal{P}\boldsymbol{E}[\tau_i]<\infty$ it follows that $\boldsymbol{E}[\zeta]<\infty$. Since $\lim_{n\to\infty}\boldsymbol{P}[\xi_n=0]=1$, Lebesgue's dominated convergence theorem (see [17, p. 187]) shows that $\lim_{n\to\infty}\boldsymbol{E}[\xi_n]=0$. **QED**

F. Proof of Lemma 8

Lemma 6 showed that the sequence

$$\boldsymbol{W}_n=\ln\mathcal{H}_{\mathcal{F}_n}+n\mathcal{D}(\mathcal{P})+\gamma_{\mathcal{D}}^{\mathcal{P}}(\boldsymbol{E}[\mathcal{S}_n|\,\mathcal{F}_n]-n\mathcal{P})\quad(100)$$

for $n\geq 0$ is a submartingale, and we will use Doob's theorem to prove the lemma. In particular, for two stopping times, $\tau_i\leq\tau_f$, Doob's theorem says that if, for both $s=i$ and $s=f$

$$\boldsymbol{E}[|\boldsymbol{W}_{\tau_s}|]<\infty\quad\text{and}\quad\lim_{n\to\infty}\boldsymbol{E}\left[|\boldsymbol{W}_n|\mathbb{I}_{\{\tau_s\geq n\}}\right]=0\quad(101)$$

then $\boldsymbol{E}[\boldsymbol{W}_{\tau_i}]$ and $\boldsymbol{E}\left[\boldsymbol{W}_{\tau_f}\right]$ exist and satisfy

$$\boldsymbol{E}\left[\boldsymbol{W}_{\tau_f}\right]\geq\boldsymbol{E}[\boldsymbol{W}_{\tau_i}].\quad(102)$$

For the moment, assume that the condition of (101) is satisfied. Then substituting the definition of $\boldsymbol{W}_n$ for $n=\tau_i$ and $n=\tau_f$ into (102)

$$\begin{aligned}\boldsymbol{E}\left[\ln\mathcal{H}_{\mathcal{F}_{\tau_f}}\ln\mathcal{H}_{\mathcal{F}_{\tau_i}}\right]+\mathcal{D}(\mathcal{P})\boldsymbol{E}[\tau_f-\tau_i]\\+\gamma_{\mathcal{D}}^{\mathcal{P}}\left(\boldsymbol{E}\left[\mathcal{S}_{\tau_f}-\mathcal{S}_{\tau_i}\right]-\mathcal{P}\boldsymbol{E}[\tau_f-\tau_i]\right)\geq 0.\end{aligned}\quad(103)$$

Inserting the assumption $\boldsymbol{E}\left[\mathcal{S}_{\tau_f}-\mathcal{S}_{\tau_i}\right]\leq\mathcal{P}\boldsymbol{E}[\tau_f-\tau_i]$

$$\boldsymbol{E}\left[\ln\frac{\mathcal{H}_{\mathcal{F}_{\tau_f}}}{\mathcal{H}_{\mathcal{F}_{\tau_i}}}\right]+\mathcal{D}(\mathcal{P})\boldsymbol{E}[\tau_f-\tau_i]\geq 0.\quad(104)$$

This is equivalent to the result of the lemma, so we need only establish the conditions in (101) to complete the proof. For the first part, we can modify (100) to bound $|\boldsymbol{W}_{\tau_s}|$ as

$$|\boldsymbol{W}_{\tau_s}|\leq|\ln\mathcal{H}_{\mathcal{F}_{\tau_s}}|+\tau_s\mathcal{D}(\mathcal{P})+\gamma_{\mathcal{D}}^{\mathcal{P}}(\boldsymbol{E}[\mathcal{S}_{\tau_s}|\,\mathcal{F}_{\tau_s}]+\tau_s\mathcal{P}).\quad(105)$$

All but the first of these terms clearly have finite expectations, so the first part of (101) reduces to proving that $\boldsymbol{E}\left[|\ln\mathcal{H}_{\mathcal{F}_{\tau_s}}|\right]<\infty$. Since $\mathcal{H}_{\mathcal{F}_0}=\ln M$

$$\begin{aligned}|\ln\mathcal{H}_{\mathcal{F}_{\tau_s}}|&=\left|\sum_{n=0}^{\tau_s-1}\ln\frac{\mathcal{H}_{\mathcal{F}_{n+1}}}{\mathcal{H}_{\mathcal{F}_n}}+\ln\ln M\right|\\&\leq\sum_{n=1}^{\tau_s-1}\left|\ln\frac{\mathcal{H}_{\mathcal{F}_n}}{\mathcal{H}_{\mathcal{F}_{n+1}}}\right|+|\ln\ln M|.\end{aligned}\quad(106)$$

Now using Lemma 5, we have

$$\begin{aligned}\boldsymbol{E}\left[\ln\frac{\mathcal{H}_{\mathcal{F}_n}}{\mathcal{H}_{\mathcal{F}_{n+1}}}\middle|\,\mathcal{F}_n\right]&\leq\boldsymbol{E}\left[\mathcal{D}_{X_{n+1}}\middle|\,\mathcal{F}_n\right]&(107)\\&\leq\mathcal{D}\left(\boldsymbol{E}\left[\mathcal{S}_{n+1}-\mathcal{S}_n\middle|\,\mathcal{F}_n\right]\right)&(108)\end{aligned}$$

where the second inequality follows from the concavity of the function $\mathcal{D}(\cdot)$.

For any random variable v, we have $v=v\mathbb{I}_{\{v\geq 0\}}+v\mathbb{I}_{\{v<0\}}$ and $|v|=v\mathbb{I}_{\{v\geq 0\}}-v\mathbb{I}_{\{v<0\}}$. Combining these, $|v|=v-2v\mathbb{I}_{\{v<0\}}$. Applying this to the random variable $v=\ln\mathcal{H}_{\mathcal{F}_n}-\ln\mathcal{H}_{\mathcal{F}_{n+1}}$ and using the (108) we get

$$\begin{aligned}\boldsymbol{E}\left[\left|\ln\frac{\mathcal{H}_{\mathcal{F}_n}}{\mathcal{H}_{\mathcal{F}_{n+1}}}\right|\middle|\,\mathcal{F}_n\right]&\leq\mathcal{D}\left(\boldsymbol{E}\left[\mathcal{S}_{n+1}-\mathcal{S}_n\middle|\,\mathcal{F}_n\right]\right)\\&\quad-2\boldsymbol{E}\left[\ln\frac{\mathcal{H}_{\mathcal{F}_n}}{\mathcal{H}_{\mathcal{F}_{n+1}}}\mathbb{I}_{\{\mathcal{H}_{\mathcal{F}_n}<\mathcal{H}_{\mathcal{F}_{n+1}}\}}\middle|\,\mathcal{F}_n\right].\end{aligned}$$

The last term above can be bounded as

$$\begin{aligned}&\boldsymbol{E}\left[\ln\frac{\mathcal{H}_{\mathcal{F}_{n+1}}}{\mathcal{H}_{\mathcal{F}_n}}\mathbb{I}_{\{\mathcal{H}_{\mathcal{F}_n}\leq\mathcal{H}_{\mathcal{F}_{n+1}}\}}\middle|\,\mathcal{F}_n\right]\\&=\boldsymbol{E}\left[\frac{\mathcal{H}_{\mathcal{F}_{n+1}}}{\mathcal{H}_{\mathcal{F}_n}}\left[-\frac{\mathcal{H}_{\mathcal{F}_n}}{\mathcal{H}_{\mathcal{F}_{n+1}}}\ln\frac{\mathcal{H}_{\mathcal{F}_n}}{\mathcal{H}_{\mathcal{F}_{n+1}}}\right]\mathbb{I}_{\{\mathcal{H}_{\mathcal{F}_n}\leq\mathcal{H}_{\mathcal{F}_{n+1}}\}}\middle|\,\mathcal{F}_n\right]\\&\overset{(a)}{\leq}e\boldsymbol{E}\left[\frac{\mathcal{H}_{\mathcal{F}_{n+1}}}{\mathcal{H}_{\mathcal{F}_n}}\mathbb{I}_{\{\mathcal{H}_{\mathcal{F}_n}\leq\mathcal{H}_{\mathcal{F}_{n+1}}\}}\middle|\,\mathcal{F}_n\right]\\&\leq e\boldsymbol{E}\left[\frac{\mathcal{H}_{\mathcal{F}_{n+1}}}{\mathcal{H}_{\mathcal{F}_n}}\middle|\,\mathcal{F}_n\right]\\&\overset{(b)}{\leq}e\end{aligned}\quad(109)$$

where in (a) we have used the fact that $-x\ln x\leq e$ for all $x>0$ and in (b) we used $\boldsymbol{E}\left[\mathcal{H}_{\mathcal{F}_{n+1}}\middle|\,\mathcal{F}_n\right]\leq\mathcal{H}_{\mathcal{F}_n}$. Thus

$$\begin{aligned}&\boldsymbol{E}\left[\left|\ln\mathcal{H}_{\mathcal{F}_n}-\ln\mathcal{H}_{\mathcal{F}_{n+1}}\right|\middle|\,\mathcal{F}_n\right]\\&\leq\mathcal{D}\left(\boldsymbol{E}\left[\mathcal{S}_{n+1}-\mathcal{S}_n\middle|\,\mathcal{F}_n\right]\right)+2e\\&\leq\mathcal{D}(0)+2e+\mathcal{D}'(0)\,\boldsymbol{E}\left[\mathcal{S}_{n+1}-\mathcal{S}_n\middle|\,\mathcal{F}_n\right]\end{aligned}\quad(110)$$

where $\mathcal{D}'(0)=\frac{d}{d\mathcal{P}}\mathcal{D}(\mathcal{P})\big|_{\mathcal{P}=0}$. Substituting this into the expectation of (106)

$$\begin{aligned}\boldsymbol{E}\left[|\ln\mathcal{H}_{\mathcal{F}_{\tau_s}}|\right]\leq\mathcal{D}(0)\boldsymbol{E}[\tau_s]+\mathcal{D}'(0)\boldsymbol{E}[\mathcal{S}_{\tau_s}]\\+2e\boldsymbol{E}[\tau_s]+|\ln\ln M|.\end{aligned}\quad(111)$$

This is finite, verifying the first part of the condition in (101).

Finally, we must verify the second part of the condition, i.e., that

$$\lim_{n\to\infty}\boldsymbol{E}\left[|\boldsymbol{W}_n|\mathbb{I}_{\{\tau_s\geq n\}}\right]=0.$$

Let ξ_n' be

$$\begin{aligned}\xi_n'&\triangleq|\boldsymbol{W}_n|\mathbb{I}_{\{\tau_s\geq n\}}\\&\leq\left[|\ln\mathcal{H}_{\mathcal{F}_n}|+n\mathcal{D}(\mathcal{P})+\gamma_{\mathcal{D}}^{\mathcal{P}}\left(\boldsymbol{E}[\mathcal{S}_n|\,\mathcal{F}_n]+\mathcal{P}n\right)\right]\mathbb{I}_{\{\tau_s\geq n\}}\\&\leq\left[\ln\frac{\ln M}{\mathcal{H}_{\mathcal{F}_n}}+|\ln\ln M|+n\mathcal{D}(\mathcal{P})\right.\\&\qquad\left.+\gamma_{\mathcal{D}}^{\mathcal{P}}\left(\boldsymbol{E}[\mathcal{S}_n|\,\mathcal{F}_n]+\mathcal{P}n\right)\right]\mathbb{I}_{\{\tau_s\geq n\}}\\&\leq\left[\sum_{n=0}^{n-1}\left|\ln\frac{\mathcal{H}_{\mathcal{F}_k}}{\mathcal{H}_{\mathcal{F}_{k+1}}}\right|+|\ln\ln M|+n\mathcal{D}(\mathcal{P})\right.\end{aligned}$$

$$+\gamma_D^{\mathcal{P}}\left(\boldsymbol{E}\left[\mathcal{S}_n\middle|\mathcal{F}_n\right]+\mathcal{P}n\right)\Big]\mathbb{I}_{\{\tau_s\geq n\}}.$$

Following the same set of steps as in (99)

$$\xi_n' \leq \sum_{n=0}^{\tau_s-1}\left|\ln\frac{\mathcal{H}_{\mathcal{F}_k}}{\mathcal{H}_{\mathcal{F}_{k+1}}}\right| + |\ln\ln M| + \tau_s\mathcal{D}(\mathcal{P}) + \gamma_D^{\mathcal{P}}\left(M\boldsymbol{E}\left[\mathcal{S}_{\tau_s}\right]+\mathcal{P}\tau_s\right) \quad (112)$$

$$\triangleq \zeta'. \quad (113)$$

Taking the expectation of both sides, using (110) together with the hypotheses $\boldsymbol{E}[\tau]<\infty$ and $\boldsymbol{E}[\mathcal{S}_\tau]<\infty$, we see that $\boldsymbol{E}[\zeta']<\infty$. Thus, using Lebesgue's dominated convergence theorem,[11] together with the fact that $\lim_{n\to\infty}\boldsymbol{P}[\xi_n'=0]=1$ implies that $\lim_{n\to\infty}\boldsymbol{E}[\xi_n']=0$. **QED**

G. Proof of Lemma 9

We use the shorthand notation introduced at the beginning of subsection B in proving the upper bound on $\ln\mathcal{H}_{\mathsf{f}_n}-\ln\mathcal{H}_{\mathsf{f}_{n+1}}$. Let $Y_{n+1}=j$.

$$\ln\mathcal{H}_{\mathsf{f}_n}=\ln\sum_{i=1}^{M}p(i)\ln\frac{1}{p(i)}$$

$$\ln\mathcal{H}_{\mathsf{f}_{n+1}}=\ln\sum_{i=1}^{M}p(i|j)\ln\frac{1}{p(i|j)}.$$

Note that

$$\frac{p(i)}{p(i|j)}=\frac{\psi(j)}{\psi(j|i)}\leq\max_{k,m}\frac{P_{kj}}{P_{mj}} \quad (114)$$

where we have used the fact that both $\psi(j)$ and $\psi(j|i)$ are in the convex hull of the set of transition probabilities P_{kj}. Using the nonnegativity of the divergence followed by (114)

$$\sum_i p(i)\ln\frac{1}{p(i|k)}\geq\sum_i p(i)\ln\frac{1}{p(i)} \quad (115)$$

$$\sum_i p(i)\ln\frac{1}{p(i|k)}\leq\left(\max_{k,m}\frac{P_{kj}}{P_{mj}}\right)\sum_i p(i|j)\ln\frac{1}{p(i|j)}. \quad (116)$$

If we include j in the maximization, we will get (47) which will be valid for all values of Y_{n+1}. **QED**

H. Proof of the Converse, Theorem 4

Theorem 4 will be proved for the discrete-time memoryless channels defined in Section IV. This includes DMCs with $P_{kj}>0$ for all k,j as a special case. The discussion in Section III-D.1 is valid except for $\ln\mathcal{H}_{\mathcal{F}_{\tau_1}}\geq-\mathsf{F}$ and the consequent inequality (48). As a substitute, we will use Assumption 2 of Section IV to show that $\boldsymbol{E}[\tau-\tau_1]$ can be lower-bounded, for each $\Delta>0$, by

$$\boldsymbol{E}[\tau-\tau_1]\geq\frac{1}{\mathcal{D}(\mathcal{P}_2)}\Big[-\ln P_e-\ln[\ln M-\ln P_e+1]-\Delta-\xi(\Delta)(1+\mathcal{P})\boldsymbol{E}[\tau]-\xi(\Delta)\delta\Big]. \quad (117)$$

Proof of (117) will be presented subsequent to the current proof.

[11] See, for example, Shiryaev, [17, p. 187].

Assume that the theorem is false. Then a sequence of codes, indexed by superscript i, exists such that the durations τ^i satisfy $\lim_{i\to\infty}\boldsymbol{E}[\tau^i]=\infty$ and each code satisfies all the conditions above but violates (49). Define $\eta^i=\frac{\boldsymbol{E}[\tau_1^i]}{\boldsymbol{E}[\tau^i]}$. Then using (44) for the ith code and dividing both sides by $\boldsymbol{E}[\tau^i]$, we get

$$\eta^i\geq\frac{(\ln M^i)\left[1-P_e^i(\ln M^i-\ln P_e^i+1)-\frac{1}{\ln M^i}\right]}{\boldsymbol{E}[\tau^i]\,\mathcal{C}(\mathcal{P}_1^i)}\geq\frac{(R+\delta)\left[1-P_e^i(\ln M^i-\ln P_e^i+1)-\frac{1}{\ln M^i}\right]}{\mathcal{C}(\mathcal{P}_1^i)}.$$

The term in brackets above approaches 1 since

$$P_e{}^i\leq\exp\{-\boldsymbol{E}[\tau^i]\,[E(R,\mathcal{P})+\delta]\}$$

and $\ln(M^i)$ lies between $\boldsymbol{E}[\tau^i](R+\delta)$ and $\boldsymbol{E}[\tau^i]\,\mathcal{C}(\mathcal{P}+\delta)$. Thus

$$\eta^i\geq\frac{R}{\mathcal{C}(\mathcal{P}_1^i)},\qquad\text{for sufficiently large } i. \quad (118)$$

Similarly, dividing both sides of (117) by $\boldsymbol{E}[\tau^i]$

$$1-\eta^i\geq\frac{1}{\boldsymbol{E}[\tau]\,\mathcal{D}(\mathcal{P}_2^i)}\Big[-\ln(P_e^i)-\ln\left[\ln\frac{Me}{P_e^i}\right]-\Delta-(1+\mathcal{P})\xi(\Delta)\boldsymbol{E}[\tau]-\xi(\Delta)\delta\Big]$$

$$\geq\frac{1}{\mathcal{D}(\mathcal{P}_2^i)}\Big[(E(R,\mathcal{P})+\delta)\left[1-\frac{\ln[\ln\frac{Me}{P_e^i}]}{\ln(1/P_e^i)}\right]-\frac{\Delta+\xi(\Delta)\delta}{\boldsymbol{E}[\tau^i]}-(1+\mathcal{P})\xi(\Delta)\Big].$$

Let Δ^* be such that $(1+\mathcal{P})\xi(\Delta^*)\leq\delta/4$. Then for sufficiently large $\boldsymbol{E}[\tau^i]$

$$1-\eta^i\geq\frac{E(R,\mathcal{P})+\delta/2}{\mathcal{D}(\mathcal{P}_2^i)}. \quad (119)$$

From (118), $\mathcal{P}_1^i\geq\mathcal{C}^{-1}(R/\eta^i)$ where the domain of the function $\mathcal{C}^{-1}(\cdot)$ is extended according to the following convention, for the channels for which $\mathcal{C}(0)>0$:

$$\mathcal{C}^{-1}(R)=\left\{\begin{array}{ll}\mathcal{P}s.t.\mathcal{C}(\mathcal{P})=R, & R>\mathcal{C}(0)\\ 0, & R\in[0,\mathcal{C}(0)]\end{array}\right\}. \quad (120)$$

Consequently, from the energy constraint $\boldsymbol{E}[\mathcal{S}_\tau]\leq\mathcal{P}\boldsymbol{E}[\tau]+\delta$, we have

$$\mathcal{P}_2^i\leq\frac{\mathcal{P}-\eta^i\mathcal{C}^{-1}\left(\frac{R}{\eta^i}\right)+\frac{\delta}{\boldsymbol{E}[\tau^i]}}{1-\eta^i}. \quad (121)$$

Thus, using (119) and (121) we get

$$E(R,P)+\delta/2\leq(1-\eta^i)\mathcal{D}\left(\frac{\mathcal{P}-\eta^i\mathcal{C}^{-1}\left(\frac{R}{\eta^i}\right)+\frac{\delta}{\boldsymbol{E}[\tau^i]}}{1-\eta^i}\right). \quad (122)$$

Recall that the function $E(R,P)$ was defined as

$$E(R,\mathcal{P})\triangleq\sup_{\eta\in\mathcal{I}_{R,\mathcal{P}}}(1-\eta)\mathcal{D}\left(\frac{\mathcal{P}-\eta\mathcal{C}^{-1}(R/\eta)}{1-\eta}\right) \quad (123)$$

where

$$\mathcal{I}_{R,\mathcal{P}} = \left[\eta'_{R,\mathcal{P}}, \min\left(1, \frac{R}{\mathcal{C}(0)}\right)\right].$$

Note that since $\eta^i \geq \frac{R}{\mathcal{C}(\mathcal{P}_1^i)}$ and $\eta^i \mathcal{P}_1^i \leq \mathcal{P} + \frac{\delta}{\boldsymbol{E}[\tau^i]}$, η^i is greater than $\eta'_{R,\mathcal{P}+\frac{\delta}{\boldsymbol{E}[\tau^i]}}$. Furthermore, if $\frac{R}{\mathcal{C}(0)} < \eta^i < 1$ than

$$(1-\eta^i)\mathcal{D}\left(\frac{\mathcal{P} - \eta^i \mathcal{C}^{-1}\left(\frac{R}{\eta^i}\right) + \frac{\delta}{\boldsymbol{E}[\tau^i]}}{1-\eta^i}\right) \leq (1 - \frac{R}{\mathcal{C}(0)})\mathcal{D}\left(\frac{\mathcal{P} + \frac{\delta}{\boldsymbol{E}[\tau^i]}}{1 - \frac{R}{\mathcal{C}(0)}}\right). \quad (124)$$

Thus, for sufficiently high $\boldsymbol{E}[\tau^i]$

$$(1-\eta^i)\mathcal{D}\left(\frac{\mathcal{P} - \eta^i \mathcal{C}^{-1}\left(\frac{R}{\eta^i}\right) + \frac{\delta}{\boldsymbol{E}[\tau^i]}}{1-\eta^i}\right) \leq E(R,\mathcal{P}) + \frac{\delta}{4}. \quad (125)$$

Observing that in equalities given in (122) and (125) leads to a contradiction concludes our proof. **QED**

I. Proof of Inequality (117)

For any channel and code, let υ_n be the random variable $\upsilon_n = \ln \mathcal{H}_{\mathcal{F}_n} - \ln \mathcal{H}_{\mathcal{F}_{n+1}}$. Assumption 2 of Section IV asserts that there is a function $\xi(\Delta)$ satisfying $\lim_{\Delta\to\infty}\xi(\Delta) = 0$ such that for all n and $\Delta \geq 0$

$$\boldsymbol{E}\left[\upsilon_n \mathbb{I}_{\{\upsilon_n \geq \Delta\}} \middle| \mathcal{F}_n\right] \leq \xi(\Delta)(1 + \boldsymbol{E}[\mathcal{S}_{n+1} - \mathcal{S}_n | \mathcal{F}_n]). \quad (126)$$

For all sample values $\mathfrak{f}_n$ of $\mathcal{F}_n$ such that $\mathcal{H}_{\mathfrak{f}_n} > 1$, we see that $\upsilon_n \geq -\ln \mathcal{H}_{\mathcal{F}_{n+1}}$. It follows from this that $\mathbb{I}_{\{\upsilon_n \geq \Delta\}} \geq \mathbb{I}_{\{-\ln \mathcal{H}_{\mathcal{F}_{n+1}} \geq \Delta\}}$ and thus for all $\Delta > 0$

$$\upsilon_n \mathbb{I}_{\{\upsilon_n \geq \Delta\}} \geq -\ln \mathcal{H}_{\mathcal{F}_{n+1}} \mathbb{I}_{\{-\ln \mathcal{H}_{\mathcal{F}_{n+1}} \geq \Delta\}}. \quad (127)$$

Substituting (127) into (126), we see that

$$\boldsymbol{E}\left[-\ln \mathcal{H}_{\mathcal{F}_{n+1}} \mathbb{I}_{\{-\ln \mathcal{H}_{\mathcal{F}_{n+1}} \geq \Delta\}} \middle| \mathcal{F}_n = \mathfrak{f}_n\right] \leq \xi(\Delta)(1 + \boldsymbol{E}[\mathcal{S}_{n+1} - \mathcal{S}_n | \mathcal{F}_n = \mathfrak{f}_n]). \quad (128)$$

holds for all $\mathfrak{f}_n$ such that $\mathcal{H}_{\mathcal{F}_n} \geq 1$.

Note that $\mathcal{H}_{\mathcal{F}_n} > 1$ holds, and thus (128) also holds, for each $\mathfrak{f}_n$ such that $n < \tau_1$. Thus, we can insert the indicator function $\mathbb{I}_{\{n<\tau_1\}}$ as follows:

$$\boldsymbol{E}\left[-\mathcal{H}_{\mathcal{F}_{n+1}} \mathbb{I}_{\{-\ln \mathcal{H}_{\mathcal{F}_{n+1}} \geq \Delta\}} \mathbb{I}_{\{n<\tau_1\}} \middle| \mathcal{F}_n\right] \leq \xi(\Delta)\boldsymbol{E}\left[(1 + \mathcal{S}_{n+1} - \mathcal{S}_n)\mathbb{I}_{\{n<\tau_1\}} \middle| \mathcal{F}_n\right]. \quad (129)$$

Taking the expectation over $\mathcal{F}_n$ we get

$$\boldsymbol{E}\left[-\ln \mathcal{H}_{\mathcal{F}_{n+1}} \mathbb{I}_{\{-\ln \mathcal{H}_{\mathcal{F}_{n+1}} \geq \Delta\}} \mathbb{I}_{\{n<\tau_1\}}\right] \leq \xi(\Delta)\boldsymbol{E}\left[(1 + \mathcal{S}_{n+1} - \mathcal{S}_n)\mathbb{I}_{\{n<\tau_1\}}\right]. \quad (130)$$

Note that

$$\mathbb{I}_{\{n<\tau_1\}} \mathbb{I}_{\{-\ln \mathcal{H}_{\mathcal{F}_{n+1}} \geq \Delta\}} = \mathbb{I}_{\{n+1=\tau_1\}} \mathbb{I}_{\{-\ln \mathcal{H}_{\mathcal{F}_{n+1}} \geq \Delta\}}. \quad (131)$$

For any $k > 0$, we next sum (130) over $0 \leq n < k$

$$\boldsymbol{E}\left[-\ln \mathcal{H}_{\mathcal{F}_{\tau_1}} \mathbb{I}_{\{-\ln \mathcal{H}_{\mathcal{F}_{\tau_1}} \geq \Delta\}} \mathbb{I}_{\{\tau_1 \leq k+1\}}\right] \leq \xi(\Delta)\boldsymbol{E}\left[\min(k, \tau_1) + \mathcal{S}_{\min(k,\tau_1)}\right]. \quad (132)$$

Using

$$|\ln \mathcal{H}_{\mathcal{F}_{\tau_1}}| \mathbb{I}_{\{-\ln \mathcal{H}_{\mathcal{F}_{\tau_1}} \geq \Delta\}} \mathbb{I}_{\{\tau_1 \leq k+1\}} \leq |\ln \mathcal{H}_{\mathcal{F}_{\tau_1}}| \mathbb{I}_{\{\mathcal{H}_{\mathcal{F}_{\tau_1}} \leq \exp(-\Delta)\}} \quad (133)$$

together with $\boldsymbol{E}\left[|\ln \mathcal{H}_{\mathcal{F}_{\tau_1}}|\right] < \infty$, $\boldsymbol{E}[\tau_1] < \infty$, and Lebesgue's dominated convergence theorem,[12] one can show that

$$\boldsymbol{E}\left[\ln \mathcal{H}_{\mathcal{F}_{\tau_1}} \mathbb{I}_{\{\mathcal{H}_{\mathcal{F}_{\tau_1}} \leq \exp(-\Delta)\}}\right] \geq -\xi(\Delta)(\boldsymbol{E}[\tau_1] + \boldsymbol{E}[\mathcal{S}_{\tau_1}]). \quad (134)$$

Consequently

$$\boldsymbol{E}\left[\ln \mathcal{H}_{\mathcal{F}_{\tau_1}}\right] \geq -\Delta - \xi(\Delta)(\boldsymbol{E}[\tau_1] + \boldsymbol{E}[\mathcal{S}_{\tau_1}]). \quad (135)$$

Now recall the inequality (46)

$$\boldsymbol{E}[\ln \mathcal{H}_{\mathcal{F}_\tau}] \leq \ln[P_e(\ln M - \ln P_e + 1)]. \quad (136)$$

Using Lemma 8, we get

$$\begin{aligned}\boldsymbol{E}[\tau - \tau_1] &\geq \boldsymbol{E}\left[\ln \mathcal{H}_{\mathcal{F}_{\tau_1}} - \ln \mathcal{H}_{\mathcal{F}_\tau}\right]\mathcal{D}(\mathcal{P}_2) \\ &\geq \frac{1}{\mathcal{D}(\mathcal{P}_2)}\Big[-\ln P_e - \ln[\ln M - \ln P_e + 1] - \Delta \\ &\quad -\xi(\Delta)(\boldsymbol{E}[\tau_1] + \boldsymbol{E}[\mathcal{S}_{\tau_1}])\Big] \\ &\geq \frac{1}{\mathcal{D}(\mathcal{P}_2)}\Big[-\ln P_e - \ln[\ln M - \ln P_e + 1] - \Delta \\ &\quad -\xi(\Delta)(1+\mathcal{P})\boldsymbol{E}[\tau] - \xi(\Delta)\delta\Big].\end{aligned}$$ **QED**

ACKNOWLEDGMENT

The authors are grateful to the reviewers of [8] and to Peter Berlin for pointing out the later paper by Burnashev [2].

REFERENCES

[1] M. V. Burnashev, "Data transmission over a discrete channel with feedback, random transmission time," *Probl. Perd. Inform.*, vol. 12, no. 4, pp. 10–30, 1976.
[2] M. V. Burnashev, "Sequential discrimination of hypotheses with control of observations," *Math. USSR Izv.*, vol. 15, no. 3, pp. 419–440, 1980.
[3] R. L. Dobrushin, "An asymptotic bound for the probability error of information transmission through a channel without memory using the feedback," *Probl. Kibern.*, vol. 8, pp. 161–168, 1962.
[4] P. Elias, Channel Capacity Without Coding MIT Res. Lab of Electronics, Cambridge, MA, 1956, Quart. Rep.
[5] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
[6] E. A. Haroutunian, "A lower bound of the probability of error for channels with feedback," *Probl. Pered. Inform.*, vol. 13, pp. 36–44, 1977.
[7] A. Kramer, "Improving communication reliability by use of an intermittent feedback channel," *IEEE Trans. Inf. Theory*, vol. IT-15, no. 1, pp. 52–60, Jan. 1969.

[12] Shiryaev, [17, p. 187].

[8] B. Nakiboğlu, R. G. Gallager, and M. Z. Win, "Error exponents for variable-length block codes with feedback and cost constraints," in *Proc. IEEE Int. Symp. Information Theory*, Seattle, WA, Jul. 2006, pp. 74–78.
[9] M. S. Pinsker, "The probability of error in block transmission in a memoryless Gaussian channel with feedback," *Probl. Perd. Inform.*, vol. 4, no. 4, pp. 1–4, 1968.
[10] A. Sahai, Why Block Length and Delay are not the Same Thing. [Online]. Available: arXiv:cs/0610138v1 [cs.IT], http://arxiv.org/abs/cs/0610138v1
[11] A. Sahai and T. Şimşek, "On the variable-delay reliability function of discrete memoryless channels with access to noisy feedback," in *Proc. IEEE Information Theory Workshop*, San Antonio, TX, Oct. 2004, pp. 336–341.
[12] J. P. M. Schalkwijk, "A coding scheme for additive noise channels with feedback–II: Band-limited signals," *IEEE Trans. Inf. Theory*, vol. IT-12, no. 2, pp. 183–189, Mar. 1966.
[13] J. P. M. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback–I: No bandwidth constraint," *IEEE Trans. Inf. Theory*, vol. IT-12, no. 2, pp. 172–182, Mar. 1966.
[14] S. Chang, "Theory of information feedback systems," *IEEE Trans. Inf. Theory*, vol. PGIT-2, no. 3, pp. 29–40, Sep. 1956.
[15] C. E. Shannon, "The zero error capacity of a noisy channel," *IEEE Trans. Inf. Theory*, vol. PGIT-2, no. 3, pp. 8–19, Sep. 1956.
[16] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels," *Inf. Contr.*, vol. 10, no. 1, pp. 65–103, 1967.
[17] A. N. Shiriaev, *Probability*. New York: Springer-Verlag, 1996.
[18] H. Yamamoto and K. Itoh, "Asymptotic performance of a modified Schalkwijk-Barron scheme for channels with noiseless feedback," *IEEE Trans. Inf. Theory*, vol. IT-25, no. 6, pp. 729–733, Nov. 1979.