Variable Length Coding Over an Unknown Channel

Aslan Tchamkerten and İ. Emre Telatar, *Member, IEEE*

***Abstract*—Burnashev in 1976 gave an exact expression for the reliability function of a discrete memoryless channel (DMC) with noiseless feedback. A coding scheme that achieves this exponent needs, in general, to know the statistics of the channel. Suppose now that the coding scheme is designed knowing only that the channel belongs to a family $\mathcal{Q}$ of DMCs. Is there a coding scheme with noiseless feedback that achieves Burnashev's exponent uniformly over $\mathcal{Q}$ at a nontrivial rate? We answer the question in the affirmative for two families of channels (binary symmetric, and Z). For these families we show that, for any given fraction, there is a feedback coding strategy such that for any member of the family: i) guarantees this fraction of its capacity as rate, and ii) guarantees the corresponding Burnashev's exponent. Therefore, for these families, in terms of delay and error probability, the knowledge of the channel becomes asymptotically irrelevant in feedback code design: there are blind schemes that perform as well as the best coding scheme designed with the foreknowledge of the channel under use. However, a converse result shows that, in general, even for families that consist of only two channels, such blind schemes do not exist.**

***Index Terms*—Burnashev's exponent, error exponent, feedback, universal channel coding, variable-length coding.**

I. Introduction

We consider communication over a stationary discrete memoryless channel (DMC) with causal, perfect feedback. The presence of feedback allows the encoder to let the transmitted symbol X_n at time n to depend upon the received symbols $Y_1, \ldots, Y_{n-1}$ as well as the message m. More subtly perhaps, the feedback allows the decoding time to depend on the received sequence. Consider, for example, communication with feedback over a binary erasure channel of a 1-bit message $m \in \{0, 1\}$ ([10, Problem 2.10] and [7]). The encoder, by sending $X_n = m$ until a nonerasure occurs, can ensure error-free communication, with a *random* decoding time $T = \inf\{n : Y_n \neq E\}$ at which the decoder declares $\hat{m} = Y_T$. It is easy to see that the expected decoding time is $1/(1 - \varepsilon)$ where ε is the erasure probability. One also observes that this strategy, when used to transmit a succession of bits, will achieve a long-term-average rate of $1/\mathbb{E}T = 1 - \varepsilon$ bits/channel use. Note that this is indeed the capacity of the binary erasure channel, and the above strategy achieves it, with perfect reliability, without the knowledge of the channel parameter ε. It would be appropriate to call this universal strategy "optimal for the class of binary erasure channels." [1]

If optimal feedback strategy for a class of channels were defined only in terms of achieving capacity, then it is not difficult to see that such strategies exist—e.g., first train to estimate the channel and then transmit for the estimated channel—for a very broad class of channels. If, however, the optimality criterion were to include a finer notion of reliability in terms of the decoding delay and error probability, then the existence of such strategies is far from clear.

In considering feedback communication over a known channel, Burnashev [2] gave an *exact expression* for the reliability function—the exponential rate of decay of the error probability with respect to the expected decoding time. In the following, we will refer to this function as the "Burnashev exponent." [2]

We choose to include the notion of reliability in our optimality criterion by asking a strategy, when used on a member of a class of channels, to "attain the Burnashev exponent" for this member. Note, however, that there is an ambiguity in this requirement: the *rate* that is achieved by a strategy depends, in general, on the member, and it is thus necessary to specify for what rate we evaluate the Burnashev exponent. Furthermore, a strategy that transmits above capacity for all channels in the class, would trivially achieve the Burnashev's exponent at this rate (as the exponent equals zero); to admit such a strategy as universal is clearly undesirable. We thus see that universality needs to be defined with some care.

In this paper, we define a suitable optimality criterion and show that, for two nontrivial classes of channels (binary symmetric and Z), there are strategies that are universal in this sense. Loosely speaking, for these families, it is possible to both attain the Burnashev's exponent and have a certain control on the rate. The control on the rate is in the form of guaranteeing that the rate stays above (or below[3]) a certain fraction of the channel

Manuscript received May 3, 2004; revised April 13, 2005. This work was supported in part by the National Competence Center in Research on Mobile Information and Communication Systems (NCCR–MICS), a center supported by the Swiss National Science Foundation under Grant 5005-67322. This work was performed while A. Tchamkerten was with the Information Theory Laboratory, School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne, Switzerland. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Chicago, IL, June/July 2004.

A. Tchamkerten is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: tcham@mit.edu).

İ. E. Telatar is with the Information Theory Laboratory (LTHI), ISC–I&C, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland (e-mail: emre.telatar@epfl.ch).

Communicated by A. Lapidoth, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2006.872974

[1]Notice that without feedback, the communication task at hand would be that of communication over a compound channel [1]; a strategy could hope at best to transmit at the capacity of the worst channel in the class.

[2]In particular, even though feedback does not increase the capacity of a memoryless channel (Shannon [17] and Csiszàr [3]), it does, in general, increase the reliability function. That the feedback strategies need not commit to a fixed block length turns out to be critical for this gain in reliability: Dobrushin [5] shows that for symmetric channels the error exponent attainable by block feedback strategies cannot exceed the sphere packing exponent.

[3]In certain cases, it might be more desirable to achieve a low error probability rather than a high communication rate. Therefore, one may want to communicate at a rate that does not exceed a certain limit while attaining a low error probability.

capacity. Such strategies, in terms of rate and reliability, thus do asymptotically as well as the best coding schemes tuned for the channel under use. Therefore, in terms of achievable rates and error exponent, the knowledge of the channel becomes irrelevant: no penalty occurs because of the channel uncertainty. However, families of channels for which universally optimal feedback strategies exist are rather specific. More precisely, we provide a converse result that shows that, even for certain simple families of channels that contain only two channels, no universally optimal coding strategy may exists.

The rest of the paper is organized as follows. In Section II, we first review a few important definitions on variable length coding schemes.

In Section III, we exhibit optimal coding strategies for the sets of binary symmetric and Z channels in a sequence of steps. In Section III-A, we exhibit a decoder that performs without knowing the statistics of the channel under use for an arbitrary class of DMCs, i.e., a universal decoder. For the average error probability over the ensemble of codes randomly generated according to a distribution P (i.e., each symbol of each codeword is chosen independently according to P), this decoder, when operating at a rate R, achieves an error exponent equal to $I(PQ)-R$, where Q is the current channel and $I(PQ)$ is the mutual information between the input and the output induced by the joint distribution $(PQ)(x,y)=P(x)Q(y|x)$. For each of the two classes of binary symmetric and Z channels, one can find a (universal) encoder that, combined with the above universal decoder, yields a coding strategy that achieves the error exponent $I(PQ)-R$ for every channel. In Section III-B, we append a second coding phase to the above universal coding strategy. The addition of this second phase augments the error exponent and makes it possible to attain Burnashev's exponent for the binary symmetric and the Z families.

In Section III-C, we set, in a general framework, the problem of finding universally optimal coding strategies for a given set of channels. We then show that for general families of channels optimal coding strategies do not necessarily exist.

In Section IV, we prove our results. In Section IV-A, we prove the claims related to Section III-A whereas Section IV-B concerns the claims of Sections III-B and III-C.

We conclude this section with notational conventions. With a slight abuse of notation, when we refer to some channel Q it is intended to be a DMC with transition probability matrix Q. The Z channel is the binary-input binary-output channel Q given by $Q(0|0)=1$ and $Q(0|1)=\varepsilon$, for some $\varepsilon\in[0,1]$. Random variables are denoted by capital letters, e.g., X, and their samples by lower case letters, e.g., x. The notation $\mathbb{E}X$ stands for the expectation of X. Given a sequence $x^n=x_1,x_2,\ldots,x_n$ we define its empirical distribution $\hat{P}_{x^n}(x)$ as

$$\frac{\sum_{j=1}^{n}\mathbb{1}_x(x_j)}{n}$$

where $\mathbb{1}_x(x_j)=1$ if $x=x_j$ and 0 if $x\neq x_j$. Given two sequences x^n and y^n, the joint empirical distribution is denoted by $\hat{P}_{x^n,y^n}(x,y)$, i.e.,

$$\hat{P}_{x^n,y^n}(x,y)\triangleq\frac{\sum_{j=1}^{n}\mathbb{1}_{(x,y)}(x_j,y_j)}{n}.$$

II. Preliminaries

Let Q be a stationary DMC with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, and with causal instantaneous noiseless feedback.

Definition 1 (Codebook (or Encoder)): Given a message set $\mathcal{M}$ of size $M\geq 1$, a codebook (or encoder) is a sequence of functions

$$f=\{f_n:\mathcal{M}\times\mathcal{Y}^{n-1}\longrightarrow\mathcal{X}\}_{n\geq 1}\,. \tag{1}$$

The symbol to be sent at time n is given by $f_n(m,y^{n-1})$. A codeword for message m is the sequence of functions $\{f_n(m,\cdot)\}_{n\geq 1}$.

Definition 2 (Random Codebook): A random codebook is a set of randomly and independently generated codewords, such that each codeword $\{f_n(m,\cdot)\}_{n\geq 1}$ is replaced by a sequence of random variables

$$X(m)\triangleq X_1(m),X_2(m),\ldots$$

drawn independently according to some probability distribution P defined over $\mathcal{X}$.

A perhaps more natural definition of a random codebook might be a codebook where the elements in the set $\{f_n(m,y^{n-1})\}$ are replaced by samples drawn independently according to some probability distribution P defined over $\mathcal{X}$. With this definition, for a given message, the nth symbol to be sent depends on the received symbols y^{n-1}. In Definition 2, the nth symbol to be sent is the same, regardless of y^{n-1}. In other words, the random codebook as defined in Definition 2 ignores feedback. However, one can easily check that our results related to random codebooks (in particular, Proposition 1) hold with both definitions.

Definition 3 (Decoder): Given a message set $\mathcal{M}$ of size $M\geq 1$, a decoder is a sequence of functions

$$\phi=\{\phi_n:\mathcal{Y}^n\longrightarrow\mathcal{M}\}_{n\geq 1} \tag{2}$$

together with a stopping time T relative to the received symbols $Y_1,Y_2,\ldots$.[4] The decoded message is $\phi_T(y^T)$.

Definition 4 (Coding Scheme and Sequence of Coding Schemes): Given a message set $\mathcal{M}$ of size $M\geq 1$, a coding scheme is a tuple $c=(f,\phi,T)$. A sequence of coding schemes indexed by the number of messages $\{c^M\}_{M\geq 1}$ is denoted by $\mathbf{S}$.

Example 1: In the language of the definitions introduced, the coding scheme in the Introduction for the binary erasure channel can be described by $\mathcal{M}=\{0,1\}$, $f_n(m,y^{n-1})=m$, $\phi_n(y^n)=y_n$ if $y_n\neq E$ and $\phi_n(y^n)=0$ if $y_n=E$, and $T=\inf\{n:Y_n\neq E\}$.

[4]An integer-valued random variable T is called a stopping time with respect to a sequence of random variables $Y_1,Y_2,\ldots$ if, for all $n\geq 1$, conditioned on $Y_1,\ldots,Y_n$, the event $\{T=n\}$ is independent of $Y_{n+1},Y_{n+2},\ldots$

Definition 5 (Rate): Given a message set of size $M \geq 1$ and a coding scheme $c = (f, \phi, T)$, the transmission rate is[5]

$$R(c, Q) \triangleq \frac{\ln M}{\mathbb{E}_Q T} \quad \text{nats per channel use} \tag{3}$$

where $\mathbb{E}_Q$ denotes the expected decision time (over uniformly chosen messages) when the channel Q is used, i.e.,

$$\mathbb{E}_Q T \triangleq \frac{1}{M} \sum_{m \in \mathcal{M}} \mathbb{E}_Q(T | \text{ message } m \text{ is sent}). \tag{4}$$

Indeed, by the law of large numbers, when a large number of messages are sent using a coding scheme c, the average rate approaches $R(c, Q)$.

The asymptotic rate for a sequence of coding schemes $\mathtt{S} = \{c^M\}_{M \geq 1}$ and a given channel Q is

$$R(\mathtt{S}, Q) \triangleq \lim_{M \to \infty} R(c^M, Q) \tag{5}$$

whenever the limit exists. Notice that we use the same "R" for two different quantities $R(c, Q)$ and $R(\mathtt{S}, Q)$. No confusion should occur since, while both functions have the same range, they are defined over different domains.

Definition 6 (Error Probability and Average Error Probability): Given a message set of size M and a coding scheme c, the average (over uniformly chosen messages) error probability is defined as

$$\mathbb{P}_Q(\mathcal{E}|c) = \frac{1}{M} \sum_{m \in \mathcal{M}} \mathbb{P}_Q \left(\phi_T(Y^T) \neq m \middle| c, \text{ message } m \text{ is sent} \right) \tag{6}$$

where the subscript $_Q$ refers to the channel over which communication is carried. Alternatively we will use the notation $\mathbb{P}_Q(\mathcal{E}|f, \phi, T)$ instead of $\mathbb{P}_Q(\mathcal{E}|c)$ to emphasize the coding scheme under consideration.

Given a decoder (ϕ, T) and a codebook randomly generated according to some distribution P,[6] the average error probability (over uniformly chosen messages) is denoted by $\mathbb{P}_Q(\mathcal{E}|\{X_n(m)\}_{m=1}^M, \phi, T)$.

In general, given a message set of finite size, finding a coding scheme that minimizes the error probability for a certain coding delay is an open question. For this reason, we shall instead consider the behavior of the error probability when the message set size tends to infinity.

Definition 7 (Error Exponent): Given a channel Q and a sequence of coding schemes

$$\mathtt{S} = \{c^M\}_{M \geq 1} = \{(f^M, \phi^M, T^M)\}_{M \geq 1}$$

such that $\mathbb{P}_Q(\mathcal{E}|c^M) \to 0$ as $M \to \infty$, the error exponent is

$$E(\mathtt{S}, Q) \triangleq \liminf_{M \to \infty} -\frac{1}{\mathbb{E}_Q T^M} \ln \mathbb{P}_Q(\mathcal{E}|c^M) \,. \tag{7}$$

[5]We use "ln" for the logarithm to the base e.

[6]See Definition 2 of a random codebook.

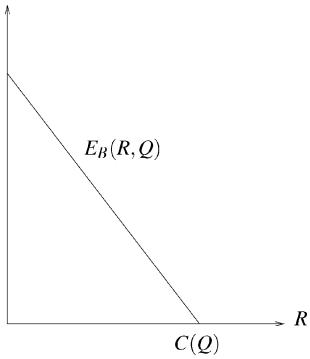

Fig. 1. For a given DMC Q with perfect feedback, the maximum achievable error exponent is given by $E_B(R, Q)$. The slope of $E_B(R, Q)$ is always equal or steeper than -1.

Remarkably, the exponential behavior of the error probability as a function of the expected decoding time of the best coding schemes is known.

Theorem (Burnashev [2]): Let Q be a DMC with input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$, and with capacity $C(Q)$. Let R be any constant in $[0, C(Q)]$. For any $\mathtt{S} = \{c^M\}_{M \geq 1}$ such that $R(\mathtt{S}, Q) = R$

$$\limsup_{M \to \infty} -\frac{1}{\mathbb{E}_Q T^M} \ln \mathbb{P}_Q(\mathcal{E}|c^M) \leq E_B(R, Q) \tag{8}$$

where

$$E_B(R, Q) \triangleq \max_{(x, x') \in \mathcal{X} \times \mathcal{X}} D(Q(\cdot|x) \| Q(\cdot|x')) \left(1 - \frac{R}{C(Q)} \right) \tag{9}$$

and where $D(Q(\cdot|x) \| Q(\cdot|x'))$ denotes the Kullback–Leibler distance between the output distributions induced by the input symbols x and x'. Moreover, there exists an $\mathtt{S}$ such that $R(\mathtt{S}, Q) = R$ and $E(\mathtt{S}, Q) = E_B(R, Q)$.

The typical shape of E_B is given in Fig. 1. In the sequel, the function E_B will be referred to as the Burnashev's exponent.

III. Results

A. A Universal Coding Scheme

Suppose we use some random codebook $\{X(m)\}_{m=1}^M$ to communicate through a channel Q that is revealed neither to the transmitter nor to the receiver. The transmitter starts sending $X_1(l), X_2(l), X_3(l), \ldots$ for some $l \in \mathcal{M}$ until a decision is made by the receiver. What is a good time to decode? Since the code has been generated according to P, we may hope to achieve rates up to $I(PQ)$ over the channel Q, and aim for a rate $I(PQ)/\alpha$ with $\alpha > 1$. But, since Q is unknown, we cannot use $I(PQ)$ directly in our decoding rule. However, one would expect that the empirical distribution of the sent codeword and the received sequence would be close to PQ, and that among all codewords the sent one would have the largest empirical mutual information with the received sequence. Hence, a reasonable candidate for the decoding instant is the first time n for which $\max_m I(\hat{P}_{X^n(m), Y^n})/\alpha \geq (\ln M)/n$. Accordingly, consider

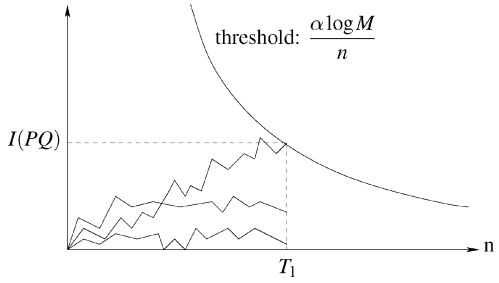

Fig. 2. This figure illustrates the first phase of one transmission cycle with $M = 3$. Each trace represents a sequence of empirical mutual informations $\{I(\hat{P}_{x^n(m),y^n})\}_{n\geq 1}$, $m = 1, 2, 3$. As soon as a trace exceeds the threshold curve $\frac{\alpha \ln M}{n}$, the decoder declares the corresponding message.

the following universal decoding time $T_1 = T_1(\alpha, M)$ defined as:

$$T_1(\alpha, M) \triangleq \inf \left\{ n \geq 1 \, : \exists \, m \in \{1, \ldots, M\} \right. \\ \left. \text{with } I(\hat{P}_{X^n(m),Y^n}) > \frac{\alpha \ln M}{n} \right\} \quad (10)$$

where $\alpha > 1$ is some fixed constant. At time T_1, the receiver declares the message m for which the empirical mutual information exceeds the threshold that defines T_1 (see Fig. 2). If multiple messages have empirical mutual informations that exceed this threshold, the receiver picks the one with the smallest index. Through feedback this decision is also known to the transmitter. This universal decoder, which we denote by $(\phi_u^M, T_1(\alpha, M))$, is an extension of the well-known maximum mutual information (MMI) decoder [12], [4]. The difference between $(\phi_u^M, T_1(\alpha, M))$ and the MMI decoder stands in that the MMI decoder is used in combination with fixed length codebooks, whereas $(\phi_u^M, T_1(\alpha, M))$ chooses the moment to decode according to the stopping time defined in (10). Another variation of the MMI decoder with variable length decision time was previously introduced by Shulman [18, Ch. 3]. The related results will be discussed after Proposition 2.

Proposition 1: Let Q be a DMC with input alphabet $\mathcal{X}$ and let P be a probability distribution over $\mathcal{X}$. Let $\alpha > 1$ and let $\mathcal{E}$ denote the decoding error event at time T_1. Then we get (11) at the bottom of the page, where

$$R = \lim_{M\to\infty} \frac{\ln M}{\mathbb{E}_P \mathbb{E}_Q T_1(\alpha, M)} = \frac{I(PQ)}{\alpha}$$

and where $\mathbb{E}_P \mathbb{E}_Q T_1(\alpha, M)$ denotes the expected decoding time averaged over the ensemble of codebooks randomly generated according to P.

From Proposition 1 we deduce that, if the transmitter and the receiver share a common source of randomness that generates independent and identically distributed (i.i.d.) samples according to some distribution P, the error exponent $I(PQ) - R$ is achievable at a rate $R = I(PQ)/\alpha$ without the transmitter and the receiver knowing the underlying channel.

Remark: One of the parameters in Proposition 1 is the input distribution P, and this might be considered as a weakness of the proposition. A question that naturally arises is the choice of this distribution when different channels in the class have different capacity achieving distributions. We don't have an answer to this question, but, for any set $\mathcal{Q}$ of binary input channels, setting P to be the Bernoulli 1/2 distribution yields $I(PQ) \geq 0.94 C(Q)$ for any element Q in $\mathcal{Q}$, where $C(Q)$ denotes the capacity of the channel Q (see [14] and [18, Chapter 5]).

The next proposition shows that for some classes of channels the error exponent $I(PQ) - R$ is universally achievable with one single sequence of (nonrandom) codebooks. In other words, in certain cases, the error exponent $I(PQ) - R$ is universally achievable without the transmitter and the receiver sharing a common source of randomness. The universal coding strategy obtained still requires only 1-bit feedback.

We denote by BSC_L and Z_L the set of BSCs and Z channels, respectively, with crossover probability $\varepsilon \in [0, L]$.

Let P be a probability distribution over $\mathcal{X}$ and Q a channel with input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$ such that $I(PQ) > 0$. Let $\mathcal{P}$ denote the set of joint distributions on $\mathcal{X} \times \mathcal{Y}$. For any $\alpha > 1$ and any integer $M \geq 1$ define

$$n_*(\alpha, P, Q, M) \triangleq \min \left\{ n \geq 1 : \right. \\ \left. \min_{V \in \mathcal{P} : I(V) \leq \frac{\alpha \ln M}{n}} D(V||PQ) \geq (\alpha - 1) \frac{\ln M}{n} \right\}. \quad (12)$$

We use the notation $a \wedge b$ to denote $\min\{a, b\}$.

Proposition 2: Let $L \in [0, 1/2)$ and let

$$n_*(\alpha, P, M) \triangleq \max_{Q \in \mathrm{BSC}_L} n_*(\alpha, P, Q, M) \, . \quad (13)$$

For any $\alpha > 1$ and any probability distribution P over $\{0, 1\}$, there exists a sequence of codebooks $\{f^M\}_{M\geq 1}$ such that, for every $Q \in \mathrm{BSC}_L$

$$\mathtt{S} = \{c^M = (f^M, \phi_u^M, T_1(\alpha, M) \wedge n_*(\alpha, P, M))\}_{M \geq 1}$$

satisfies

$$E(\mathtt{S}, Q) \geq I(PQ) - R(\mathtt{S}, Q) \quad \text{and} \\ R(\mathtt{S}, Q) = \lim_{M\to\infty} \frac{\ln M}{\mathbb{E}_Q(T_1(\alpha, M) \wedge n_*(\alpha, P, M))} \\ = \frac{I(PQ)}{\alpha} \, . \quad (14)$$

The same result holds for the family Z_L with $0 \leq L < 1$.

Remark: The decoder $(\phi_u^M, T_1(\alpha, M) \wedge n_*(\alpha, P, M))$ differs from $(\phi_u^M, T_1(\alpha, M))$ only in that the decoding time of $(\phi_u^M, T_1(\alpha, M) \wedge n_*(\alpha, P, M))$ is bounded by $n_*(\alpha, P, M)$. In particular, if no sequence of empirical mutual informations exceeds the threshold that defines $T_1(\alpha, M)$ by time $n_*(\alpha, P, M)$ the decoder $(\phi_u^M, T_1(\alpha, M) \wedge n_*(\alpha, P, M))$ declares an error

$$\liminf_{M\to\infty} -\frac{1}{\mathbb{E}_P \mathbb{E}_Q T_1(\alpha, M)} \ln \left(\mathbb{E}_P \mathbb{P}_Q(\mathcal{E} | \{X(m)\}_{m=1}^M, \phi_u^M, T_1(\alpha, M)) \right) \geq I(PQ) - R \quad (11)$$

whereas $(\phi_u^M, T_1(\alpha, M))$ does not. An intuitive meaning of $n_*(\alpha, P, Q, M)$ is provided in the remark that follows the proof of Proposition 1.

In [18] a result similar to Proposition 2 is given.

Theorem (3.1, Shulman [18]): Let $\mathcal{Q}$ be any set of DMCs defined over the same input alphabet $\mathcal{X}$. For any probability distribution P over $\mathcal{X}$, there exists a sequence of coding schemes $\mathtt{S} = \{c^M\}_{M \geq 1}$ such that

I. for any $\mu > 0$ and M large enough, $\mathbb{P}_Q(\mathcal{E}|c^M) \leq \mu$ for all $Q \in \mathcal{Q}$;
II. the asymptotic rate $R(\mathtt{S}, Q)$ equals $I(PQ)$ for all $Q \in \mathcal{Q}$.

The preceding theorem has a more general setting than Proposition 2. The theorem says that even though the channel is almost completely unknown to both the transmitter and the receiver (only the input alphabet needs to be revealed), it is possible to reliably communicate, in the sense that the error probability can be made uniformly arbitrarily small. In Proposition 2, we restricted ourselves to smaller families of channels while having a refined expression for the error probability. Also notice that in Shulman's case, the rate is governed by the input distribution P whereas in our case the asymptotic rate is set by both P and the parameter α in the definition of $T_1(\alpha, M)$. Finally, it should be emphasized that the universal coding schemes in [18, Theorem 3.1] exploit also only 1-bit of feedback.

In the next subsection, we provide a means for boosting the error exponent obtained in Proposition 2.

B. Boosting Error Exponents

We describe a two-phase coding scheme where the first phase is carried out by the universal coding scheme mentioned in Proposition 2 with decision time $T_1(\alpha, M) \wedge n_*(\alpha, P, M)$ (see (14)). In certain cases, the addition of a properly chosen second phase boosts the error exponent from $I(PQ) - R$ to Burnashev's exponent. From now on and without loss of generality, we assume that message 1 is sent.

At time $T_1(\alpha, M) \wedge n_*(\alpha, P, M)$, the receiver labels as "most probable" the message m for which the empirical mutual information exceeds the threshold that defines $T_1(\alpha, M) \wedge n_*(\alpha, P, M)$. If multiple messages have empirical mutual informations that exceed this threshold at time T_1, the receiver picks the one with the smallest index. Through feedback this decision is also known to the transmitter.

The second phase consists in performing a hypothesis test to let the decoder decide if the message m labeled "most probable" is the sent message or not. Namely, let $x(A)$ and $x(N)$ be codewords for two additional messages "Ack" and "Nack," respectively. If $m = 1$, the transmitter acknowledges the choice of the receiver by sending $x(A)$. If $m \neq 1$, the transmitter denies the receiver's decision by sending $x(N)$. If the receiver decodes the sent codeword as "Ack," the transmission of the message is complete (either correctly or incorrectly), and the transmitter starts re-emitting a new message. Otherwise, if the decoder decides on "Nack," we begin afresh and message 1 is retransmitted (see Fig. 3).[7]

[7]The idea of a two-phase transmission procedure characterized by first choosing a "most probable" message and then accepting or rejecting this choice was previously studied, e.g., in [16], [2]. Our scheme differs from the previous works mainly because it is independent of the channel under use.

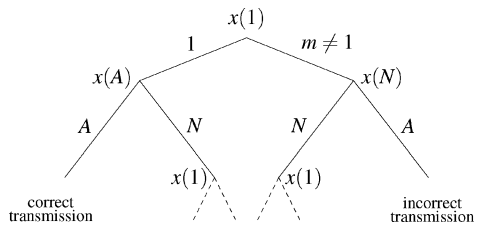

Fig. 3. The graph illustrates a two-phase transmission procedure. The vertices indicate what the transmitter sends. The edges indicate the receiver's decision. In particular, codeword $x(1)$ is correctly transmitted only if: message 1 is declared as the "most probable" codeword and $x(A)$ is correctly decoded.

The results of this subsection are obtained by studying two-phase coding strategies. Theorems 1 and 2 below are to be compared with Proposition 2.

Theorem 1: Let $L \in [0, 1/2)$. For any $\gamma \in (0, 1)$ there exist two sequences of coding schemes $\mathtt{S}'$ and $\mathtt{S}''$ such that, for every $Q \in \mathrm{BSC}_L$

$$\begin{aligned} E(\mathtt{S}', Q) &= E_B(R(\mathtt{S}', Q), Q) \quad \text{with} \\ \gamma C(Q) &\leq R(\mathtt{S}', Q) < C(Q) \end{aligned} \tag{15}$$

and

$$\begin{aligned} E(\mathtt{S}'', Q) &= E_B(R(\mathtt{S}'', Q), Q) \quad \text{with} \\ 0 &< R(\mathtt{S}'', Q) \leq \gamma C(Q). \end{aligned} \tag{16}$$

Further, there exists $\mathtt{S}''$ such that, for every $Q \in \mathrm{BSC}_L$

$$\begin{aligned} E(\mathtt{S}'', Q) &= E_B(0, Q) \quad \text{with} \\ R(\mathtt{S}'', Q) &= 0. \end{aligned} \tag{17}$$

Theorem 2: For any $L \in [0, 1)$ and $\gamma \in [0, 1)$, there exists a sequence of coding schemes $\mathtt{S}$ such that, for every $Q \in \mathrm{Z}_L$

$$E(\mathtt{S}, Q) = \infty \quad \text{with} \quad R(\mathtt{S}, Q) = \gamma C(Q)\,. \tag{18}$$

Note the difference between Theorems 1 and 2. For BSC_L and Z_L it is possible to achieve Burnashev's exponent but, for BSC_L, the constant γ allows only to bound the rate (either from above or from below), while for Z channels it allows an exact control on the rate.

We may ask if the same result for more general families of channels as for BSC_L and Z_L holds, i.e., to achieve Burnashev's exponent universally while having a certain control on the rate. An answer will be provided in the next section.

C. Optimal Feedback Strategies

We first provide a general setting for the problem of communication with feedback over an unknown channel. Then we introduce an optimality criterion for feedback strategies with respect to a family of channels. This criterion essentially asks for attaining Burnashev's error exponent, and, at the same time, it asks for having a certain control over the communication rate. In light of Theorems 1 and 2, we shall conclude that, for the binary symmetric and the Z channels families, optimal feedback strategies exist. Second, by extending a results from [19], we show that, in general, optimal feedback strategies do not exists.

The following definition introduces a main concept of this paper. We quantify the set of error exponents that can simultaneously be achieved over a given family of channels.

Definition 8 (Universally Attainable Error Exponent): Let $\mathcal{Q}$ be a set of DMCs. Let $K(Q)$ be a nonnegative function defined over $\mathcal{Q}$. Let $E(R,Q)$ be a function such that $E(R,Q) > 0$ for every $Q \in \mathcal{Q}$ and $R \in [0, K(Q))$, and such that $E(R,Q) = 0$ for every $Q \in \mathcal{Q}$ and R with $R > K(Q)$.[8] The function $E(R,Q)$ is a universally attainable error exponent over $\mathcal{Q}$ for rates in the range $[0, K(Q)]$ if, for any $Q \in \mathcal{Q}$ and any $R \in [0, K(Q))$, there exists a sequence of coding schemes $\mathtt{S}$ such that the following two conditions hold:[9]

I.

$$R(\mathtt{S},Q) \geq R \quad \text{and} \quad E(\mathtt{S},Q) \geq E(R,Q); \tag{19}$$

II. for every $W \in \mathcal{Q}$ with $K(W) > 0$

$$E(\mathtt{S},W) > 0. \tag{20}$$

Condition I of Definition 8 requires that, for a given channel Q and for any $R \in [0, K(Q))$, there exists a sequence of coding schemes $\mathtt{S}$ yielding a rate at least equal to R and a corresponding error exponent at least equal to $E(R,Q)$. By condition II, this sequence $\mathtt{S}$, if used on any channel $W \in \mathcal{Q}$ (with $K(W) > 0$), must achieve a strictly positive error exponent and therefore a rate not exceeding $K(W)$. Without condition II, the definition would have only required that for each channel there would be a good coding scheme, which does not capture the notion of universality.

For any $\mathtt{S}$ and W let

$$\Delta(\mathtt{S},W) \triangleq E_B(R(\mathtt{S},W),W) - E(\mathtt{S},W) \tag{21}$$

where $E_B(R,W)$ is defined in (9). In the case where $E(\mathtt{S},W) = \infty$ (which implies that $E_B(R(\mathtt{S},W),W) = \infty$) we set $\Delta(\mathtt{S},W) = 0$.[10] The nonnegative quantity $\Delta(\mathtt{S},W)$ compares the sequence of coding schemes $\mathtt{S}$, in terms of error exponent, with the best possible sequence of coding schemes designed for the channel W and rate $R(\mathtt{S},W)$. For any family $\mathcal{Q}$, let

$$\Delta'(\mathcal{Q}) \triangleq \inf_{\mathtt{S}:\, \forall W \in \mathcal{Q},\, E(\mathtt{S},W)>0} \sup_{W \in \mathcal{Q}} \Delta(\mathtt{S},W)\,. \tag{22}$$

A definition of an optimality criterion with respect to a set of feedback strategies and a family of channels would be to require that $\Delta'(\mathcal{Q}) = 0$. This means that there exists a sequence of $\mathtt{S}$'s each of which yields a rate not exceeding capacity and in the limit an error exponent equal to Burnashev's, on any channel in $\mathcal{Q}$. However, the existence of such a sequence gives no control on the rate achieved on the family of channels. In particular, for two different channels this rate might be negligible on one and close to capacity on the other.

[8]The function $K(Q)$ plays the role of the capacity.

[9]The same "E" is used to for two different quantities $E(R,Q)$ and $E(\mathtt{S},Q)$. No confusion should occur since, while both functions have the same range, they are defined over different domains.

[10]The definition of $\Delta(\mathtt{S},Q)$ implicitely assumes that $R(\mathtt{S},Q)$ is well defined.

A second alternative for the definition of an optimality criterion would be to introduce the quantity

$$\Delta_+(\mathcal{Q},\gamma) \triangleq \inf_{\mathtt{S}:\ \substack{\forall W,\, E(\mathtt{S},W)>0 \\ \gamma C(W) \leq R(\mathtt{S},W) < C(W)}} \sup_{W \in \mathcal{Q}} \Delta(\mathtt{S},W) \tag{23}$$

and to declare optimality if

$$\Delta''(\mathcal{Q}) \triangleq \sup_{\gamma \in [0,1)} \Delta_+(\mathcal{Q},\gamma) \tag{24}$$

equals zero. One can easily check that this second definition of optimality is stronger than the previous one in that if $\Delta''(\mathcal{Q}) = 0$, then $\Delta'(\mathcal{Q}) = 0$. Claiming $\Delta''(\mathcal{Q}) = 0$ is essentially equivalent to declare that Burnashev's exponent is universally achievable over $\mathcal{Q}$, for rates in the range $[0, C(Q)]$, while universally having a certain control on the rate through the parameter γ. However, notice that the control on the rate is only from below: we focus on coding schemes that, universally, yield a rate at least equal to $\gamma C(W)$.

We now introduce our optimality criterion that will basically require to be able to universally achieve Burnashev's exponent, and to control the rate from above and from below, on any channel in the class. Define the quantities

$$\Delta_-(\mathcal{Q},\gamma) \triangleq \inf_{\mathtt{S}:\ \substack{\forall W,\, E(\mathtt{S},W)>0 \\ 0 < R(\mathtt{S},W) \leq \gamma C(W)}} \sup_{W \in \mathcal{Q}} \Delta(\mathtt{S},W) \tag{25}$$

$$\Delta(\mathcal{Q},0) \triangleq \inf_{\mathtt{S}:\ \substack{\forall W,\, E(\mathtt{S},W)>0 \\ R(\mathtt{S},W)=0}} \sup_{W \in \mathcal{Q}} \Delta(\mathtt{S},W)\,. \tag{26}$$

For any family of DMCs $\mathcal{Q}$, we define the *diversity* $\Delta(\mathcal{Q})$ as

$$\Delta(\mathcal{Q}) \triangleq \max\left\{\Delta(\mathcal{Q},0),\ \sup_{\gamma \in (0,1)} \max\{\Delta_+(\mathcal{Q},\gamma), \Delta_-(\mathcal{Q},\gamma)\}\right\}. \tag{27}$$

We say that a family $\mathcal{Q}$ satisfies the optimality criterion if it is nondiverse, i.e., if

$$\Delta(\mathcal{Q}) = 0. \tag{28}$$

Finally, we justify the terminology "diversity" for the quantity $\Delta(\mathcal{Q})$ defined in (27). Having $\Delta(\mathcal{Q})$ large implies that there exists a pair of channels in $\mathcal{Q}$ such that Burnashev's exponent cannot be universally achieved at either sufficiently high rates or at sufficiently low rates (or both). Informally, if $\Delta(\mathcal{Q})$ is large, the family contains some "too different" channels (the family is too "diverse") in that, at some particular rate, no coding scheme can attain Burnashev's exponent on each of them.

Combining (27) with Theorems 1 and 2 we obtain the following corollaries.

Corollary 1: If $L \in [0, 1/2)$ then $\Delta(\mathrm{BSC}_L) = 0$.

Corollary 2: If $L \in [0, 1)$ then $\Delta(\mathrm{Z}_L) = 0$.

For Z_L, the optimality criterion is satisfied and, in addition, the corresponding "optimal coding schemes" have the property that they simultaneously achieve exactly any given fraction of the capacity of the channel under use.

The following theorem is an extension of the theorem presented in [19] and provides a converse to the Corollaries 1 and 2. It says that, given a pair of channels Q_1 and Q_2 that satisfies

certain conditions, Burnashev's exponent cannot be achieved simultaneously on Q_1 and Q_2 at all rates $\gamma C(Q)$ for γ below a certain threshold.

Theorem 3: Let Q_1 and Q_2 be two DMCs on $\mathcal{X} \times \mathcal{Y}$. Let

$$K(Q_i, Q_j) \triangleq \max_{(x,x') \in \mathcal{X} \times \mathcal{X}} \left[D(Q_i(\cdot|x) \| Q_i(\cdot|x')) D(Q_i(\cdot|x) \| Q_j(\cdot|x')) \right] \tag{29}$$

and assume that there exists some $\gamma \in [0, 1)$ such that

$$\begin{aligned} K(Q_1, Q_2) &< 2E_B(\gamma C(Q_1), Q_1) \text{ and} \\ K(Q_2, Q_1) &< 2E_B(\gamma C(Q_2), Q_2). \end{aligned} \tag{30}$$

For any $\gamma' \in [0, \gamma]$ and any $\mathtt{S}$ such that $R(\mathtt{S}, Q_i) = \gamma' C(Q_i)$ for $i \in \{1, 2\}$, either $E(\mathtt{S}, Q_1) < E_B(\gamma' Q_1, Q_1)$ or $E(\mathtt{S}, Q_2) < E_B(\gamma' Q_2, Q_2)$.

Example 2: Let $Q_1 = \mathrm{BSC}(\varepsilon)$ and $Q_2 = \mathrm{BSC}(1-\varepsilon)$ where $0 < \varepsilon < 1/2$. One can easily check that

$$K(Q_1, Q_2) = K(Q_2, Q_1) = E_B(0, Q_1).$$

Therefore, the condition (30) becomes

$$E_B(0, Q_1) < 2E_B(\gamma C(Q_1), Q_1)$$

from which we deduce that $\gamma < 1/2$. Hence, if we operate at a rate below half the capacity, Burnashev's error exponent cannot be simultaneously achieved on Q_1 and Q_2.

As a consequence, in general, for a given family of channels $\mathcal{Q}$, we have $\Delta(\mathcal{Q}, 0) > 0$, and hence the optimality criterion defined in (28) is not satisfied, i.e., $\Delta(\mathcal{Q}) > 0$.

A discussion on whether optimal feedback strategies exist also for more general classes of channels than the binary symmetric and the Z will be provided in Section IV-C.

IV. Analysis

This section is devoted to the proofs of the results of Section III. We would like to draw the reader's attention to the fact that Propositions 1 and 2 and Theorems 1 and 2 still hold if the feedback link has any constant delay, provided it remains noiseless. This can be easily checked from the analysis we provide in this section.

From now on, whenever we consider a channel it is assumed to have input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$.

A. Proofs of Propositions 1 and 2

We first establish five lemmas mainly using tools from the Method of Types [4].

The set of all joint types of length n defined over $\mathcal{X} \times \mathcal{Y}$ is denoted by $\mathcal{P}_n$ whereas $\mathcal{P}$ denotes the set of all joint distributions over $\mathcal{X} \times \mathcal{Y}$.

Lemma 1 gives the probability that the empirical mutual information of an incorrect codeword exceeds the threshold that defines T_1 (see (10)) at some time n, when the codebook is randomly generated according to a distribution P.

Lemma 1: Let P_X and P_Y be two distributions over $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $\{(X_i, Y_i)\}_{i \geq 1}$ be a sequence of i.i.d. pairs of random variables such that

$$\mathbb{P}(X_i = x, Y_i = y) = P_X(x) P_Y(y)$$

for all $i \geq 1$. For any real $\alpha > 1$ and integer $M \geq 1$

$$\mathbb{P}\left(I(\hat{P}_{X^n, Y^n}) > \frac{\alpha \ln M}{n} \right) \leq M^{-\alpha} (n+1)^{|\mathcal{X}||\mathcal{Y}|}. \tag{31}$$

Proof: The event $\{I(\hat{P}_{X^n,Y^n}) > \frac{\alpha \ln M}{n}\}$ is the union of all joint empirical distributions V yielding a mutual information larger than $\frac{\alpha \ln M}{n}$. From the Method of Types, the probability that $\hat{P}_{X^n,Y^n}$ equals a particular empirical joint distribution V is upper-bounded by $e^{-nD(V\|P_X P_Y)}$. A direct computation yields, for any $V \in \mathcal{P}$

$$D(V \| P_X P_Y) = I(V) + D(V_X \| P_X) + D(V_Y \| P_X) \tag{32}$$

where

$$V_X(x) \triangleq \sum_y V(x, y) \quad \text{and} \quad V_Y(y) \triangleq \sum_x V(x, y).$$

Hence, $D(V\|P_X P_Y) \geq I(V)$, and therefore,

$$\begin{aligned} \mathbb{P}\left(I(\hat{P}_{X^n,Y^n}) > \frac{\alpha \ln M}{n} \right) &\leq \sum_{V \in \mathcal{P}_n : I(V) \geq \frac{\alpha \ln M}{n}} e^{-nD(V\|P_X P_Y)} \\ &\leq \sum_{V \in \mathcal{P}_n : I(V) \geq \frac{\alpha \ln M}{n}} e^{-nI(V)} \\ &\leq M^{-\alpha} (n+1)^{|\mathcal{X}||\mathcal{Y}|} \end{aligned} \tag{33}$$

where the last inequality holds since $\mathcal{P}_n$ satisfies $|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|}$ (see, e.g., [4, Lemma 2.2]). □

We now present a technical lemma that will often be used in the sequel. It shows that the quantity $n_*(\alpha, P, Q, M)$ introduced in (12) grows logarithmically in M provided that $I(PQ) > 0$. Given two functions $f(M)$ and $g(M)$, we use the notation $f(M) = O(g(M))$ if there exist $c > 0$ and $M_0 \geq 0$ such that $|f(M)| \leq c g(M)$ for $M \geq M_0$. If $f(M) = O(g(M))$ and $g(M) = O(f(M))$ then we write $f(M) = \Theta(g(M))$. Finally, the notation $f(M) = o(g(M))$ is used if $|f(M)/g(M)| \to 0$ as $M \to \infty$.

Lemma 2: Let $\alpha > 1$, let P be a probability distribution over $\mathcal{X}$, and let Q be a channel such that $I(PQ) > 0$. The quantity

$$n_*(\alpha, P, Q, M) \triangleq \min \left\{ n \geq 1 : \min_{V \in \mathcal{P} : I(V) \leq \frac{\alpha \ln M}{n}} D(V \| PQ) \geq (\alpha - 1) \frac{\ln M}{n} \right\} \tag{34}$$

is well defined for all $M \geq 1$. Moreover

$$n_*(\alpha, P, Q, M) = \Theta(\ln M). \tag{35}$$

Proof: Fix some integer $M \geq 1$. The function $D(\cdot\|PQ)$ defined over the compact convex finite-dimensional set $\mathcal{P}$ is

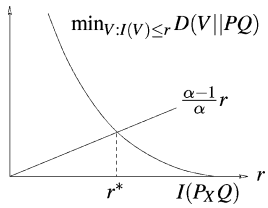

Fig. 4. $0 < r_* = r(\alpha, M, n_*) < I(PQ)$.

convex and therefore (see Luenberger [13]) continuous. Since $\{V \in \mathcal{P} : I(V) \le \frac{\alpha \ln M}{n}\}$ is compact

$$\inf_{V\in\mathcal{P}:I(V)\le\frac{\alpha\ln M}{n}} D(V||PQ) = \min_{V\in\mathcal{P}:I(V)\le\frac{\alpha\ln M}{n}} D(V||PQ)\,. \tag{36}$$

The function

$$\min_{V\in\mathcal{P}:I(V)\le\frac{\alpha\ln M}{n}} D(V||PQ)$$

is nondecreasing with n. Since $I(PQ) > 0$, and because $(\alpha-1)\frac{\ln M}{n}$ strictly decreases with n, we conclude that $n_* < \infty$ for all $M \ge 1$. Hence, n_* is well defined.

We now prove (35). Let us first rewrite (34) as

$$n_*(\alpha,P,Q,M) = \min\left\{ n \ge 1 : \min_{V\in\mathcal{P}:I(V)\le r(\alpha,M,n)} D(V||PQ) \ge \frac{\alpha-1}{\alpha} r(\alpha,M,n) \right\} \tag{37}$$

with $r(\alpha,M,n) \triangleq \frac{\alpha\ln M}{n}$.

Since $D(V||PQ) = 0$ if and only if $V = PQ$

$$\min_{V\in\mathcal{P}:I(V)\le r} D(V||PQ) > 0 \tag{38}$$

for all $r \in [0, I(PQ))$ (see Fig. 4). Therefore, since $n_* < \infty$, by defining

$$r_* \triangleq r_*(\alpha,P,Q,M) \triangleq r(\alpha,M,n_*)$$

we have $r_* \in (0, I(PQ))$ for all $M \ge 1$. Now let $\tilde{r} = \tilde{r}(\alpha,P,Q)$ be the unique solution of the equation

$$\min_{V\in\mathcal{P}:I(V)\le r} D(V||PQ) = \frac{\alpha-1}{\alpha} r\,. \tag{39}$$

The same arguments as for r_* applies and, therefore, $\tilde{r} \in (0, I(PQ))$. Let us write $r_*(M)$ for $r_*(\alpha,P,Q,M)$ since α, P, and Q are kept fixed. Using the definitions of $\tilde{r}$ and $r_*(M)$, one can easily show that, for any $M \ge 1$

$$\begin{aligned} 0 \le \tilde{r} - r_*(M) &\le \frac{\alpha\ln M}{n_*-1} - \frac{\alpha\ln M}{n_*} \\ &= \left(\frac{\alpha\ln M}{n_*}\right)^2 \frac{1}{\alpha\ln M(1-1/n_*)} \\ &= (r_*)^2 \frac{1}{\alpha\ln M - r_*} \\ &\le \frac{I(PQ)^2}{\ln M - I(PQ)} \end{aligned} \tag{40}$$

where the last inequality holds because $r_* < I(PQ)$ and $\alpha > 1$. Therefore, from (40) we conclude

$$\begin{aligned} n_* &\triangleq \frac{\alpha\ln M}{r_*} \\ &= \frac{\alpha\ln M}{\tilde{r} + o(1)} \\ &= \Theta(\ln M). \end{aligned} \tag{41}$$

□

We now want an estimate of T_1. To that aim, we consider the first time the sequence of empirical mutual informations that corresponds to the correct codeword crosses the threshold defined by the curve $\alpha \ln M/n$. Lemma 3 will show that this time has low probability to occur after n_* when the codebook is randomly generated according to a certain distribution P.

Lemma 3: Let P be a probability distribution over $\mathcal{X}$ and let Q be a channel such that $I(PQ) > 0$. Let $\{(X_i, Y_i)\}_{i\ge1}$ be a sequence of i.i.d. pairs of random variables such that

$$\mathbb{P}(X_i = x, Y_i = y) = P(x)Q(y|x).$$

For any $\alpha > 1$ and $M \ge 1$, the quantity $n_* = n_*(\alpha,P,Q,M)$ defined in (12) satisfies

$$\mathbb{P}\left(I(\hat{P}_{X^{n_*},Y^{n_*}}) \le \frac{\alpha\ln M}{n_*} \right) \le M^{-(\alpha-1)}(n_*+1)^{|\mathcal{X}||\mathcal{Y}|}. \tag{42}$$

Proof: From the Method of Types we have

$$\begin{aligned} &\mathbb{P}\left(I(\hat{P}_{X^{n_*},Y^{n_*}}) \le \frac{\alpha\ln M}{n_*} \right) \\ &\quad\le \sum_{V\in\mathcal{P}_{n_*}:I(V)\le\frac{\alpha\ln M}{n_*}} e^{-n_* D(V||PQ)} \\ &\quad\le (n_*+1)^{|\mathcal{X}||\mathcal{Y}|} e^{-n_* \min_{V\in\mathcal{P}:I(V)\le\frac{\alpha\ln M}{n_*}} D(V||PQ)} \\ &\quad= M^{-(\alpha-1)}(n_*+1)^{|\mathcal{X}||\mathcal{Y}|} \end{aligned} \tag{43}$$

where the last equality follows from the definition of n_*. □

Lemma 4 states results about T_1 in terms of its mean and its concentration around the mean as the message set size increases. Unless stated otherwise, from now on we assume, without loss of generality, that message 1 is sent.

Lemma 4: Let P be a probability distribution over $\mathcal{X}$ and let $\alpha > 1$. For any channel Q such that $I(PQ) > 0$, the ensemble of codebooks randomly generated according to P satisfies, as $M \to \infty$

I.

$$\mathbb{P}\left(T_1(\alpha,M) > \frac{\alpha\ln M}{I(PQ)}(1+o(1)) \right) \le e^{-\frac{\sqrt{\ln M}}{2I(PQ)}(1+o(1))}; \tag{44}$$

II.

$$\mathbb{P}\left(T_1(\alpha,M) \le \frac{\alpha\ln M}{I(PQ)}(1+o(1)) \right) \le e^{-\frac{\sqrt{\ln M}}{\ln|\mathcal{X}|}(1+o(1))}; \tag{45}$$

III.

$$\mathbb{E}T_1(\alpha, M) = \frac{\alpha \ln M}{I(PQ)}(1 + o(1)). \tag{46}$$

Proof: For the logic of the proof we prove the claims in the order I, III, and II.

I. Let $s \triangleq 1 + \frac{\alpha \ln M}{I(PQ)} d_1(M, P, Q)$ where

$$d_1(M, P, Q) \triangleq \frac{I(PQ)}{\min_{V \in \mathcal{P}:\, D(V||PQ) \le \frac{1}{\sqrt{\ln M}}} I(V)} . \tag{47}$$

We first show that $d_1(M, P, Q) = 1 + o(1)$ which implies that

$$s = \frac{\alpha \ln M}{I(PQ)}(1 + o(1)).$$

Since $I(V)$ is a continuous function over the closed set $\{V \in \mathcal{P} : D(V||PQ) \le \frac{1}{\sqrt{\ln M}}\}$, the minimum in the denominator of the right-hand side of (47) is well defined, and so is $d_1(M, P, Q)$. Now suppose that $V(x, y) = V_X(x)V_Y(y)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. A direct computation (as for (32)) yields

$$\begin{aligned} D(PQ||V) &= I(PQ) + D(P||V_X) + D(Q_Y||V_Y) \\ &\ge I(PQ) \end{aligned} \tag{48}$$

where

$$Q_Y(y) \triangleq \sum_{x \in \mathcal{X}} P(x)Q(y|x).$$

Then, since the set $\mathcal{P}^\pi$ of product measures in $\mathcal{P}$ is closed and $D(PQ\|\cdot)$ is continuous over $\mathcal{P}^\pi$, from (48) we have

$$\inf_{V \in \mathcal{P}^\pi} D(PQ||V) = \min_{V \in \mathcal{P}^\pi} D(PQ||V) \ge I(PQ) . \tag{49}$$

In other words, any product measure V is at least at a distance $I(PQ)$ from PQ. Hence, for large enough M, the compact set $\{V \in \mathcal{P} : D(V||PQ) \le 1/\sqrt{\ln M}\}$ contains no product measures. Therefore, for large enough M

$$\min_{V \in \mathcal{P}: D(V||PQ) \le \frac{1}{\sqrt{\ln M}}} I(V) > 0 \tag{50}$$

implying that $d_1(M, P, Q)$ is finite. This implies that $d_1(M, P, Q)$ decreases with increasing M, and since it is lower bounded by 1, we get $d_1(M, P, Q) = 1 + o(1)$.

From the definition of T_1 (see (10)) and since, without loss of generality, we assume that message 1 is sent, we have

$$\begin{aligned} &\mathbb{P}(T_1 > s) \\ &\quad \le \sum_{n \ge \lfloor s \rfloor} \mathbb{P}(T_1 = n + 1) \\ &\quad \le \sum_{n \ge \lfloor s \rfloor} \mathbb{P}\left(I(\hat{P}_{X^n(1), Y^n}) \le \frac{\alpha \ln M}{n} \right) \\ &\quad \le \sum_{n \ge \lfloor s \rfloor} (n+1)^{|\mathcal{X}||\mathcal{Y}|} e^{-n \min_{V \in \mathcal{P}: I(V) \le \frac{\alpha \ln M}{s-1}} D(V||PQ)} \end{aligned} \tag{51}$$

where $\lfloor x \rfloor$ stands for the largest integer not greater than x. Let us focus on the expression

$$\min_{V \in \mathcal{P}: I(V) \le \frac{\alpha \ln M}{s-1}} D(V||PQ).$$

If we expand d_1 in the definition of s we get

$$\frac{\alpha \ln M}{s - 1} = \min_{V \in \mathcal{P}: D(V||PQ) \le \frac{1}{\sqrt{\ln M}}} I(V) \tag{52}$$

which implies that

$$\min_{V \in \mathcal{P}: I(V) \le \frac{\alpha \ln M}{s-1}} D(V||PQ) \ge \frac{1}{\sqrt{\ln M}} . \tag{53}$$

Hence, from (51) and (53)

$$\mathbb{P}(T_1 > s) \le \sum_{n \ge s-1} (n+1)^{|\mathcal{X}||\mathcal{Y}|} e^{-n \frac{1}{\sqrt{\ln M}}} . \tag{54}$$

Since $s = 1 + \frac{\alpha \ln M}{I(PQ)} d_1(M, P, Q)$ and since

$$d_1(M, P, Q) = 1 + o(1)$$

from (54) we obtain

$$\begin{aligned} &\mathbb{P}\left(T_1 > \frac{\alpha \ln M}{I(PQ)}(1 + o(1)) \right) \\ &\quad \le \sum_{n \ge s-1} (n+1)^{|\mathcal{X}||\mathcal{Y}|} e^{-n \frac{1}{\sqrt{\ln M}}} \\ &\quad = e^{-\frac{\sqrt{\ln M}}{2I(PQ)}(1 + o(1))}. \end{aligned} \tag{55}$$

III. From (51)–(54) we deduce that

$$(n+1)\mathbb{P}(T_1 = n+1) \le (n+1)^{|\mathcal{X}||\mathcal{Y}|+1} e^{-n \frac{1}{\sqrt{\ln M}}} \tag{56}$$

for all $n \ge \lfloor s \rfloor$. Hence, from the definition of s we get

$$\mathbb{E}T_1(\alpha, M) \le \frac{\alpha \ln M}{I(PQ)}(1 + o(1)) . \tag{57}$$

Thus, in order to prove claim III it suffices to show that

$$\mathbb{E}T_1(\alpha, M) \ge \frac{\alpha \ln M}{I(PQ)}(1 + o(1)) . \tag{58}$$

First notice that from the definition of $T_1(\alpha, M)$ we have $T_1(\alpha, M) \ge \lceil \frac{\alpha \ln M}{\ln |\mathcal{X}|} \rceil$, where $\lceil x \rceil$ denotes the smallest integer not smaller than x. Let us define

$$v \triangleq \frac{\alpha \ln M}{\ln |\mathcal{X}|} \quad \text{and} \quad q \triangleq \frac{\alpha \ln M}{I(PQ)} d_2(M, P, Q) - 1$$

where

$$d_2(M, P, Q) \triangleq \frac{I(PQ)}{\max_{V \in \mathcal{P}: D(V||PQ) \le \frac{1}{\sqrt{\ln M}}} I(V)} . \tag{59}$$

Similarly as for $d_1(M, P, Q)$, we have $d_2(M, P, Q) = 1 + (o(1))$. We have

$$\begin{aligned} \mathbb{E}T_1(\alpha, M) &\ge \sum_{n=1}^{\lceil q \rceil} \mathbb{P}(T_1 \ge n) \\ &\ge q - q\mathbb{P}(\lceil v \rceil \le T_1 \le \lceil q \rceil) . \end{aligned} \tag{60}$$

From (60), and since $q = \frac{\alpha \ln M}{I(PQ)}(1 + o(1))$, in order to derive (58) it suffices to show that

$$q\mathbb{P}(\lceil v \rceil \le T_1(\alpha, M) \le \lceil q \rceil) = o(1).$$

From the definition of T_1 we get

$$\begin{aligned}&\mathbb{P}(\lceil v\rceil \le T_1 \le \lceil q\rceil)\\ &\quad=\mathbb{P}\Bigg(\exists\, m\in\{1,\dots,M\} \text{ and } j\in\{\lceil v\rceil,\dots,\lceil q\rceil\}\\ &\qquad\qquad \text{with } I(\hat{P}_{X^j(m),Y^j})>\frac{\alpha\ln M}{j}\Bigg). \end{aligned}\tag{61}$$

The union bound and Lemma 1 yield

$$\begin{aligned}&\mathbb{P}(\lceil v\rceil \le T_1 \le \lceil q\rceil)\\ &\le (M-1)\mathbb{P}\Bigg(\exists\, j\in\{\lceil v\rceil,\dots,\lceil q\rceil\}\\ &\qquad\qquad \text{with } I(\hat{P}_{X^j(2),Y^j})>\frac{\alpha\ln M}{j}\Bigg)\\ &\quad+\mathbb{P}\Bigg(\exists\, j\in\{\lceil v\rceil,\dots,\lceil q\rceil\}\\ &\qquad\qquad \text{with } I(\hat{P}_{X^j(1),Y^j})>\frac{\alpha\ln M}{j}\Bigg)\\ &\le M^{1-\alpha}(\lceil q\rceil+1)^{|\mathcal{X}||\mathcal{Y}|+1}\\ &\quad+\mathbb{P}\Bigg(\exists\, j\in\{\lceil v\rceil,\dots,\lceil q\rceil\}\\ &\qquad\qquad \text{with } I(\hat{P}_{X^j(1),Y^j})>\frac{\alpha\ln M}{j}\Bigg). \end{aligned}\tag{62}$$

An easy computation yields

$$\begin{aligned}&\mathbb{P}\Bigg(\exists j\in\{\lceil v\rceil,\dots,\lceil q\rceil\}\text{with}I(\hat{P}_{X^j(1),Y^j})>\frac{\alpha\ln M}{j}\Bigg)\\ &\quad\le(\lceil q\rceil+1)^{|\mathcal{X}||\mathcal{Y}|+1}e^{-v\min_{V\in\mathcal{P}:I(V)\ge\frac{\alpha\ln M}{q+1}}D(V\|PQ)}. \end{aligned}\tag{63}$$

Now, expanding $d_2(M,P,Q)$ in the definition of q we have

$$\frac{\alpha\ln M}{q+1}=\max_{V\in\mathcal{P}:D(V\|PQ)\le\frac{1}{\sqrt{\ln M}}} I(V)\tag{64}$$

which implies that (same as in (52)–(53))

$$\min_{V\in\mathcal{P}:I(V)\ge\frac{\alpha\ln M}{q+1}} D(V\|PQ)\ge\frac{1}{\sqrt{\ln M}}.\tag{65}$$

From (62), (63), (65), and the definitions of v and q we have

$$\mathbb{P}(\lceil v\rceil\le T_1\le\lceil q\rceil)\le e^{-\frac{\sqrt{\ln M}}{\ln|\mathcal{X}|}(1+o(1))}\tag{66}$$

and we conclude that

$$q\mathbb{P}(\lceil v\rceil\le T_1\le\lceil q\rceil)=o(1)\,.\tag{67}$$

II. Since $\mathbb{P}(T_1\ge\lceil v\rceil)=1$ and $q=\frac{\alpha\ln M}{I(PQ)}(1+o(1))$ from (66) we get

$$\mathbb{P}\left(T_1\le\frac{\alpha\ln M}{I(PQ)}(1+o(1))\right)\le e^{-\frac{\sqrt{\ln M}}{\ln|\mathcal{X}|}(1+o(1))}\,.\tag{68}$$

□

Remark: We may notice that the function $\sqrt{\ln M}$ in the definition of $d_1(M,P,Q)$ and $d_2(M,P,Q)$ in the proof of Lemma 4 can be replaced by any strictly positive function $g(M)$ such that $g(M)=o(\ln M)$.

Lemma 5 is a key lemma and Proposition 1 is an immediate consequence of it. We consider communication over a channel Q by means of a codebook randomly generated according to some distribution P, and the universal decoder $(\phi_u^M,T_1(\alpha,M))$ defined in Section III-A.

Let n be any integer such that $n\ge1$ and let $\mathcal{E}_n$ denote the event defined as: no correct decoding decision has been made during the period $[1,n]$. In particular, $\mathcal{E}_n$ includes the decoding error event of $(\phi_u^M,T_1(\alpha,M))$.

For notational convenience we shall often remove the arguments of the functions and write, for example, T_1 instead of $T_1(\alpha,M)$.

Lemma 5: Let P be a probability distribution over $\mathcal{X}$ and let Q be a channel such that $I(PQ)>0$. Let $\alpha>1$ and let $n_*=n_*(\alpha,P,Q,M)$ be defined as in (12). Then

I.

$$\begin{aligned}&\mathbb{E}_P\mathbb{P}_Q(\mathcal{E}_{n_*}|\{X(m)\}_{m=1}^M,\phi_u^M,T_1\wedge n_*)\\ &\qquad\le O\left(M^{1-\alpha}n_*^{|\mathcal{X}||\mathcal{Y}|+1}\right),\qquad \text{as } M\to\infty \end{aligned}\tag{69}$$

II.

$$\begin{aligned}&\liminf_{M\to\infty}-\frac{1}{\mathbb{E}_P\mathbb{E}T_1}\ln\mathbb{E}_P\mathbb{P}_Q(\mathcal{E}_{n_*}|\{X(m)\}_{m=1}^M,\phi_u^M,T_1\wedge n_*)\\ &\qquad\ge I(PQ)-R \end{aligned}\tag{70}$$

where $R=\lim_{M\to\infty}\frac{\ln M}{\mathbb{E}_P\mathbb{E}_QT_1}=\frac{I(PQ)}{\alpha}$.

Proof:

I. We have

$$\begin{aligned}&\mathbb{E}_P\mathbb{P}_Q(\mathcal{E}_{n_*}|\{X(m)\}_{m=1}^M,\phi_u^M,T_1\wedge n_*)\\ &\le\mathbb{P}\left(I(\hat{P}_{X^{n_*}(1),Y^{n_*}})\le\frac{\alpha\ln M}{n_*}\right)\\ &\quad+\mathbb{P}\Bigg(\exists l\in\{2,\dots,M\}\text{ and }j\in\{1,\dots,n_*\}\\ &\qquad\qquad\text{with } I(\hat{P}_{X^j(l),Y^j})\ge\frac{\alpha\ln M}{j}\Bigg). \end{aligned}\tag{71}$$

Without loss of generality, we assume that message 1 is sent, hence $\mathbb{P}(X_j(1)=x,Y_j=y)=P(x)Q(y|x)$. Lemma 3 in turn yields

$$\mathbb{P}\left(I(\hat{P}_{X^{n_*}(1),Y^{n_*}})\le\frac{\alpha\ln M}{n_*}\right)\le M^{1-\alpha}(n_*+1)^{|\mathcal{X}||\mathcal{Y}|}.\tag{72}$$

Now, for every $l\in\{2,\dots,M\}$ and $j\ge1$, we have

$$\mathbb{P}(X_j(l)=x,Y_j=y)=P(x)P_Y(y)$$

$$\mathbb{P}\left(\exists l \in \{2,\dots,M\} \text{ and } j \in \{1,\dots,n_*\} \text{ with } I(\hat{P}_{X^j(l),Y^j}) \geq \frac{\alpha \ln M}{j}\right) \leq M^{1-\alpha}(n_*+1)^{|\mathcal{X}||\mathcal{Y}|+1}. \quad (73)$$

with $P_Y(y) = \sum_x Q(y|x)P(x)$, and Lemma 1 together with the union bound gives (73) at the top of the page. Hence, from (71)–(73) we have s

$$\mathbb{E}_P\mathbb{P}_Q(\mathcal{E}_{n_*}|\{X(m)\}_{m=1}^M, \phi_u^M, T_1 \wedge n_*) \leq O\left(M^{1-\alpha} n_*^{|\mathcal{X}||\mathcal{Y}|+1}\right). \quad (74)$$

II. Let $R(M) \triangleq \frac{\ln M}{\mathbb{E}T_1(\alpha,M)}$. We readily obtain

$$\begin{aligned} M^{1-\alpha} n_*^{|\mathcal{X}||\mathcal{Y}|+1} &= e^{-(\alpha-1+o(1))\ln M} \\ &= e^{-\left(I(PQ)-\frac{I(PQ)}{\alpha}+o(1)\right)\mathbb{E}T_1(1+o(1))} \\ &= e^{-(I(PQ)-R(M)+o(1))\mathbb{E}T_1(1+o(1))} \\ &= e^{-(I(PQ)-R(M))\mathbb{E}T_1(1+o(1))}. \end{aligned} \quad (75)$$

The first equality follows from the fact that $n_* = \Theta(\ln M)$ by Lemma 2. The second and third equalities are justified by Lemma 4 claim III. Finally, combining (74) and (75) we get

$$\mathbb{E}_P\mathbb{P}_Q(\mathcal{E}_{n_*}|\{X(m)\}_{m=1}^M, \phi_u^M, T_1 \wedge n_*) \leq e^{-(I(PQ)-R(M))\mathbb{E}T_1(1+o(1))}. \quad (76)$$

Hence,

$$\begin{aligned} \liminf_{M\to\infty} &- \frac{1}{\mathbb{E}_P\mathbb{E}_Q T_1} \\ &\times \left(\ln \mathbb{E}_P\mathbb{P}_Q(\mathcal{E}_{n_*}|\{X(m)\}_{m=1}^M, \phi_u^M, T_1 \wedge n_*)\right) \\ &\geq I(PQ) - R \end{aligned} \quad (77)$$

where $R = \lim_{M\to\infty} \frac{\ln M}{\mathbb{E}_P\mathbb{E}_Q T_1} = I(PQ)/\alpha$ by Lemma 4 claim III. □

Proof of Proposition 1: Since

$$\mathbb{E}_P\mathbb{P}_Q(\mathcal{E}|\{X(m)\}_{m=1}^M, \phi_u^M, T_1) \leq \mathbb{E}_P\mathbb{P}_Q(\mathcal{E}_{n_*}|\{X(m)\}_{m=1}^M, \phi_u^M, T_1 \wedge n_*) \quad (78)$$

Proposition 1 follows from Lemma 5. □

Remark: At this point, we would like to give some intuition on $n_*(\alpha, P, Q, M)$. This quantity is introduced to find a convenient upper bound on $\mathbb{E}_P\mathbb{P}_Q(\mathcal{E}|\{X(m)\}_{m=1}^M, \phi_u^M, T_1(\alpha, M))$, the average error probability of the universal decoder $(\phi_u^M, T_1(\alpha, M))$, when the codebook is randomly generated according to some distribution P. Let $n = n(M)$ be an integer valued function growing sublinearly with M. We may upper-bound $\mathbb{P}_Q(\mathcal{E}|P, \phi_u^M, T_1(\alpha, M))$ by considering as an error the event in which $(\phi_u^M, T_1(\alpha, M))$ has not made a correct decision in the interval $[1, n]$. We denoted this event by $\mathcal{E}_n$. The event $\mathcal{E}_n$ is realized if one of the following two events happens: an incorrect decision is made in the interval $[1, n]$, or, in the interval $[1, n]$ no message has a corresponding sequence of empirical mutual informations that exceeds the threshold that defines T_1, i.e., $T_1 > n$. Denoting these two error events by A and B, respectively, we get

$$\begin{aligned} \mathbb{E}_P\mathbb{P}_Q(\mathcal{E}|\{X(m)\}_{m=1}^M, \phi_u^M, T_1) &\leq \mathbb{E}_P\mathbb{P}_Q(\mathcal{E}_n|\{X(m)\}_{m=1}^M, \phi_u^M, T_1) \\ &\leq \mathbb{P}(A) + \mathbb{P}(B). \end{aligned} \quad (79)$$

What now remains is to strike a good balance between $\mathbb{P}(A)$ and $\mathbb{P}(B)$ by choosing an appropriate n. The n_* used above is chosen such that $\mathbb{P}(A)$ and $\mathbb{P}(B)$ have the same error exponent.

From Proposition 1 we deduce that, for any channel Q, it is possible to find a codebook that, combined with the universal decoder described in Section III-A, yields a low error probability. However, in general, this does not imply the existence of a codebook that guarantees low error probability for every channel in a given family. The main point in the proof of Proposition 2 is to show that there exists a codebook that admits low error probability on all channels in BSC_L ($L \in [0, 1/2)$), and similarly for Z_L ($L \in [0, 1)$). An essential ingredient is a coupling among the channels in the families BSC_L and Z_L. This coupling is made possible because of the ordering among channels in the families.

Proof of Proposition 2: We first consider the family BSC_L where $L \in [0, 1/2)$. Pick an input distribution P over $\{0, 1\}$, a constant $\alpha > 1$, and let

$$n_*(\alpha, P, M) \triangleq \max_{Q \in \mathrm{BSC}_L} n_*(\alpha, P, Q, M). \quad (80)$$

For the moment we assume that $n_*(\alpha, P, M)$ is well defined and such that $n_*(\alpha, P, M) = \Theta(\ln M)$. We will prove this claim at the end of the proof.

Without loss of generality, we introduce a coupling between the channels in BSC_L. This coupling will be used to show the existence of a universal codebook that, combined with the universal decoder $(\phi_u^M, T_1(\alpha, M) \wedge n_*(\alpha, P, M))$, has the desired error probability and expected decision time for all channels in BSC_L.

Let $\{A_i\}_{i\geq 1}$ be an i.i.d. sequence of random variables such that A_1 is uniformly distributed within the interval $[0, L]$, and set

$$Y_i = x_i \oplus \mathbb{1}_{A_i \leq L\varepsilon} \quad (81)$$

where $\mathbb{1}_{A_i \leq L\varepsilon} = 1$ if $A_i \leq L\varepsilon$ and $\mathbb{1}_{A_i \leq L\varepsilon} = 0$ if $A_i > L\varepsilon$. We interpret Y_i as the ith output of the channel $\mathrm{BSC}(\varepsilon)$ when the input symbol is x_i (one can verify that the crossover probability of the channel described in (81) is indeed ε). The coupling introduced in (81) is such that whenever the channel $\mathrm{BSC}(\varepsilon)$ makes a crossover, all the channels $\mathrm{BSC}(\delta)$ with $\delta \in [\varepsilon, L]$ also make a crossover. Moreover, because of the coupling, at each time n the set BSC_L behaves as if there were only at most $n+1$ distinct channels. To see this, let us partition BSC_L as follows. Let $\{\mathcal{A}_i\}_{i=1}^n$ be the order statistics of $\{A_i\}_{i=1}^n$, i.e.,

$\{\mathcal{A}_i\}_{i=1}^n$ represents the same set of random variables as $\{A_i\}_{i=1}^n$ but labeled in increasing order. Then partition BSC_L as

$$\mathrm{BSC}_L = \bigcup_{\substack{l\in\mathbb{N}\\ 0\le l\le n}} b_l \tag{82}$$

where

$$b_0 \triangleq \{\mathrm{BSC}(\varepsilon)\ : \varepsilon \in [0, \mathcal{A}_1)\}$$
$$b_l \triangleq \{\mathrm{BSC}(\varepsilon)\ : \varepsilon \in [\mathcal{A}_l, \mathcal{A}_{l+1})\}$$

for $l \in \{1, \ldots, n-1\}$ and $b_n \triangleq \{\mathrm{BSC}(\varepsilon) : \varepsilon \in [\mathcal{A}_n, L]\}$. From the above partition we deduce that, at time n, given an input sequence x^n, the family BSC_L produces at most $n+1$ distinct output sequences of length n, i.e., the set BSC_L behaves as if there were only at most $n+1$ distinct channels.

We will now consider the decoding rule described in Section III-A with decision time $T_1(\alpha, M) \wedge n_*(\alpha, P, M)$ instead of $T_1(\alpha, M)$. Using a random coding argument that makes use of the coupling introduced above, we will prove that, for any M large enough, there exists a coding scheme that simultaneously over all channels in BSC_L has the desired error probability and desired expected decision time. Let us first consider the error probability that can be simultaneously achieved over BSC_L. One can check that claim I of Lemma 5 still holds when $n_*(\alpha, P, Q, M)$ is replaced by $n_*(\alpha, P, M)$ (assuming that $n_*(\alpha, P, M) = \Theta(\ln M)$). Therefore, for any $Q \in \mathrm{BSC}_L$, the average error probability over the ensemble of codes satisfies

$$\mathbb{E}_P \mathbb{P}_Q(\mathcal{E}_{n_*(\alpha,P,M)} | \{X(m)\}_{m=1}^M, \phi_u^M, T_1 \wedge n_*) \le O(M^{1-\alpha} n_*(\alpha, P, M)^{|\mathcal{X}||\mathcal{Y}|+1}). \tag{83}$$

If we apply Markov's inequality to the error probability defined over the ensemble of random codebooks generated according to P, from (83) we get

$$\mathbb{P}_P\left(\mathbb{P}_Q(\mathcal{E}_{n_*} | \{X(m)\}_{m=1}^M, \phi_u^M, T_1 \wedge n_*) > M^{1-\alpha} n_*^{3(|\mathcal{X}||\mathcal{Y}|+1)}\right) \le O\left(n_*^{-2(|\mathcal{X}||\mathcal{Y}|+1)}\right). \tag{84}$$

Now, recall that the coupling introduced among channels BSC_L is such that, at each instant n, the family BSC_L behaves as if there were at most $n+1$ distinct channels. Hence, from the union bound and (84) it follows that

$$\mathbb{P}_P\left(\bigcup_{Q\in\mathrm{BSC}_L}\left\{\mathbb{P}_Q(\mathcal{E}_{n_*} | f^M = \{X(m)\}_{m=1}^M, \phi_u^M, T_1 \wedge n_*) > M^{1-\alpha} n_*^{3(|\mathcal{X}||\mathcal{Y}|+1)}\right\}\right) \le O\left(n_*^{-(|\mathcal{X}||\mathcal{Y}|+1)}\right). \tag{85}$$

Now for the decoding time. From the union bound we get

$$\mathbb{P}\left(\bigcup_{Q\in\mathrm{BSC}_L}\left\{T_1 \wedge n_* > \frac{\alpha \ln M}{I(PQ)}(1+o(1))\right\}\right) \le (n_*+1) \max_{Q\in\mathrm{BSC}_L} \mathbb{P}\left(T_1 \wedge n_* > \frac{\alpha \ln M}{I(PQ)}(1+o(1))\right). \tag{86}$$

From Lemma 4 claim I, for every $Q \in \mathrm{BSC}_L$

$$\mathbb{P}\left(T_1 \wedge n_* > \frac{\alpha \ln M}{I(PQ)}(1+o(1))\right) \le e^{-\frac{\sqrt{\ln M}}{2I(PQ)}(1+o(1))}. \tag{87}$$

It follows that

$$\max_{Q\in\mathrm{BSC}_L} \mathbb{P}\left(T_1 \wedge n_* > \frac{\alpha \ln M}{I(PQ)}(1+o(1))\right) \le e^{-\frac{\sqrt{\ln M}}{2\min_{W\in\mathrm{BSC}_L} I(PW)}(1+o(1))}. \tag{88}$$

From (86) and (88) we have

$$\mathbb{P}\left(\bigcup_{Q\in\mathrm{BSC}_L}\left\{T_1 \wedge n_* > \frac{\alpha \ln M}{I(PQ)}(1+o(1))\right\}\right) \le (n_*+1) e^{-\frac{\sqrt{\ln M}}{2\min_{W\in\mathrm{BSC}_L} I(PW)}(1+o(1))}. \tag{89}$$

A similar argument as above together with Lemma 4 claim II yields

$$\mathbb{P}\left(\bigcup_{Q\in\mathrm{BSC}_L}\left\{T_1 \wedge n_* \le \frac{\alpha \ln M}{I(PQ)}(1+o(1))\right\}\right) \le (n_*+1) e^{-\frac{\sqrt{\ln M}}{\ln|\mathcal{X}|}(1+o(1))}. \tag{90}$$

Finally, since $n_* = n_*(\alpha, P, M)$ grows logarithmically with M, the sum of the right-hand sides of (85), (89), and (90) is smaller than 1 for M larger than some integer $M_o(\alpha, P)$, say. Hence, for every M larger than $M_o(\alpha, P)$, there exists a (nonrandom) codebook f^M such that, for every $Q \in \mathrm{BSC}_L$, the following two conditions are satisfied:

$$\mathbb{P}_Q(\mathcal{E}_{n_*} | f^M, \phi_u^M, T_1(\alpha, M) \wedge n_*) \le M^{1-\alpha} n_*^{3(|\mathcal{X}||\mathcal{Y}|+1)} \tag{91}$$

and

$$\mathbb{E}_Q(T_1(\alpha, M) \wedge n_*) = \frac{\alpha \ln M}{I(PQ)}(1+o(1)). \tag{92}$$

From (91) and (92) and a similar computation as in (75)–(77), by setting

$$\mathtt{S} = \{c^M = (f^M, \phi_u^M, T_1(\alpha, M) \wedge n_*(\alpha, P, M))\}_{M\ge 1}$$

we have

$$E(\mathtt{S}, Q) \ge I(PQ) - R(\mathtt{S}, Q) \quad \text{and}$$
$$R(\mathtt{S}, Q) = \lim_{M\to\infty} \frac{\ln M}{\mathbb{E}_Q(T_1(\alpha, M) \wedge n_*(\alpha, P, M))} = \frac{I(PQ)}{\alpha} \tag{93}$$

for every $Q \in \mathrm{BSC}_L$.

For the case where $\mathcal{Q} = \mathrm{Z}_L$ with $L \in [0,1)$, an argument similar to that for BSC_L holds. The only difference is that the coupling should be made according to

$$Y_i^{\varepsilon} = \mathbb{1}_{x_i=1}\mathbb{1}_{A_i > L\varepsilon}. \tag{94}$$

We conclude the proof of the proposition by showing, along the lines of the proof of Lemma 2, that $n_*(\alpha, P, M) = \Theta(\ln M)$. From (12) we have

$$n_*(\alpha,P,M) = \min\left\{ n \geq 1 : \min_{V\in\mathcal{P}:I(V)\leq \frac{\alpha \ln M}{n}} \min_{Q\in\mathrm{BSC}_L} D(V\|PQ) \geq (\alpha-1)\frac{\ln M}{n} \right\}. \tag{95}$$

On the one hand, the function

$$\min_{V\in\mathcal{P}:I(V)\leq \frac{\alpha \ln M}{n}} \min_{Q\in\mathrm{BSC}_L} D(V\|PQ) \tag{96}$$

is nondecreasing with n. On the other hand, $(\alpha-1)\ln M/n$ strictly decreases with n, and since $\min_{Q\in\mathrm{BSC}_L} I(PQ) > 0$, we infer that $n_*(\alpha,P,M) < \infty$. Let us define $\tilde{r} = \tilde{r}(\alpha,P)$ as the unique solution of the equation

$$\min_{V\in\mathcal{P}:I(V)\leq r} \min_{Q\in\mathrm{BSC}_L} D(V\|PQ) = \frac{\alpha-1}{\alpha} r\,. \tag{97}$$

Since $\min_{Q\in\mathrm{BSC}_L} I(PQ) > 0$, we deduce that

$$0 < \tilde{r}(\alpha,P) < \min_{Q\in\mathrm{BSC}_L} I(PQ).$$

Finally, from a reasoning similar to that concluding the proof of Lemma 2 (see from (39) onwards) we deduce that $n_*(\alpha,P,M) = \Theta(\ln M)$. □

B. Proofs of Theorems 1, 2, and 3

We start with a brief study on the two-phase coding strategy emphasizing the purposes of the first and second phase. Let us assume that communication is carried out over some channel Q and that we have a two-phase coding strategy with the following properties.

1. There is a low probability that the two-phase coding scheme makes more than one cycle, i.e., there is a low probability that, at the end of the second phase, the decoder declares "Nack." As a consequence, the average decoding time E_QT is approximatively equal to the average decoding time of the first phase, $\mathbb{E}_QT_1$, plus the average decoding time of the second phase, $\mathbb{E}_QT_2$.
2. The coding scheme used for the first phase achieves a rate R_1 close to capacity.
3. The two-message coding scheme used for the second phase is such that the probability $\mathbb{P}_Q(A|x(N))$ of declaring "Ack" while "Nack" is upper-bounded by $e^{-E_B(0,Q)\mathbb{E}_QT_2}$.

Under the above assumptions, we have that the average error probability of the two-phase coding scheme can be upper-bounded as follows:

$$\begin{aligned}
\mathbb{P}_Q(\mathcal{E}|c) &\overset{\text{a.}}{\approx} \mathbb{P}_Q(x(N))\mathbb{P}_Q(A|x(N)) \\
&\overset{\text{b.}}{\leq} e^{-E_B(0,Q)\mathbb{E}_QT_2} \\
&\overset{\text{c.}}{\approx} e^{-E_B(0,Q)\left(1-\frac{\mathbb{E}_QT_1}{\mathbb{E}_QT}\right)\mathbb{E}_QT} \\
&= e^{-E_B(0,Q))\left(1-\frac{R}{R_1}\right)\mathbb{E}_QT} \\
&\overset{\text{d.}}{\approx} e^{-E_B(0,Q)\left(1-\frac{R}{C(Q)}\right)\mathbb{E}_QT} \\
&\overset{\text{e.}}{=} e^{-E_B(R,Q)\mathbb{E}_QT}.
\end{aligned} \tag{98}$$

The approximation a. holds by property 1. If the two-phase coding scheme makes one cycle with high probability, the error probability essentially equals to the probability that a wrong message is declared "most probable" at the end of the first phase, times the probability that the receiver declares "Ack" while a "Nack" was sent.[11] Inequality b. holds by the assumption 3. The approximation c. holds because $\mathbb{E}_QT \approx \mathbb{E}_QT_1 + \mathbb{E}_QT_2$ by assumption 1. Approximation d. holds by hypothesis 2 and e. is by definition of the Burnashev's error exponent (9). Finally, note that assumption 1 requires $\mathbb{P}_Q(N|x(A))$ to be small, but not necessarily exponentially small with respect to the average coding delay of the second phase.

In order to prove Theorems 1 and 2, we will essentially show that there exist two-phase coding schemes that satisfy hypotheses 1,2, and 3 universally over BSC_L and Z_L. This will imply that Burnashev's error exponent is universally achievable over BSC_L and Z_L. In addition, we will need to prove that the rate can be controlled. For BSC_L, this means that the rate is guaranteed to be universally at least (or at most) equal to a given fraction of the channel capacity, whereas for Z_L, the control on the rate means a rate universally equal to a given fraction of the channel capacity.

Remark: From (98) we see that the role of the first phase is to carry information at a high rate while the role of the second phase is to make the probability $\mathbb{P}(A|x(N))$ as small as possible. Hence, Burnashev's exponent can be achieved with two-phase coding strategies even if the first phase has a corresponding error probability that vanishes arbitrarily slowly with increasing coding delay, i.e., the error exponent of the first phase is irrelevant. Hence, the above computation gives a simple way to prove the achievability part of Burnashev's theorem. However, the above computation hides a difficulty: finding capacity achieving coding schemes for the first phase. Therefore, the above computation gives only a conceptually simple way to reach Burnashev's exponent through a random coding argument. Indeed, the two-phase scheme proposed in [2] to prove the achievability of Burnashev's error exponent may appear very complex (at each time a complex randomized decision at both the transmitter and the receiver is required), but has the advantage that it can be implemented.

The next lemma shows that the probability of error of a two-phase scheme is approximatively $\mathbb{P}_Q(x(N))\mathbb{P}_Q(A|x(N))$. For simplicity, from now on we will often drop the subscript Q.

[11] $\mathbb{P}_Q(x(N))$ is the average error probability of the coding scheme used for the first phase.

Lemma 6: Let c be a two-phase coding scheme. If $\mathbb{P}(x(N)) + \mathbb{P}(N|x(A)) < 1$ then

$$\mathbb{P}(\mathcal{E}|c) \leq \frac{\mathbb{P}(x(N))\mathbb{P}(A|x(N))}{1 - \mathbb{P}(x(N)) - \mathbb{P}(N|x(A))}. \tag{99}$$

Proof: Let $\mathcal{E}^i$ denote the event that an error occurs at the end of the ith cycle and let N_i denote the event that the receiver declares "Nack" at the end of the ith cycle. The family of events $\{N_i\}_{i\geq 1}$ is such that $N_{i+1} \subset N_i$ and also satisfies $\mathbb{P}(N_{i+1}|N_i) = \mathbb{P}(N_1)$. Hence, we have the recursion relation $\mathbb{P}(N_{i+1}) = \mathbb{P}(N_i)\mathbb{P}(N_1)$, and therefore, $\mathbb{P}(N_{i-1}) = \mathbb{P}(N_1)^{i-1}$. It follows that

$$\begin{aligned}\mathbb{P}(\mathcal{E}^i) &= \mathbb{P}(\mathcal{E}^i, N_{i-1}) \\ &= \mathbb{P}(\mathcal{E}^i|N_{i-1})\mathbb{P}(N_{i-1}) \\ &= \mathbb{P}(A|x(N))\mathbb{P}(x(N))\mathbb{P}(N_{i-1}) \\ &= \mathbb{P}(A|x(N))\mathbb{P}(x(N))\mathbb{P}(N_1)^{i-1}\,. \end{aligned} \tag{100}$$

Now, we have

$$\begin{aligned}\mathbb{P}(N_1) &= \mathbb{P}(N|x(N))\mathbb{P}(x(N)) \\ &\quad + \mathbb{P}(N|x(A))(1 - \mathbb{P}(x(N))) \\ &\leq \mathbb{P}(x(N)) + \mathbb{P}(N|x(A)). \end{aligned} \tag{101}$$

Since $\{\mathcal{E}^i\}_{i\geq 1}$ is a family of disjoint events, (100) and (101) yield

$$\mathbb{P}(\mathcal{E}|c) = \sum_{i\geq 1} \mathbb{P}(\mathcal{E}^i) \leq \frac{\mathbb{P}(x(N))\mathbb{P}(A|x(N))}{1 - \mathbb{P}(x(N)) - \mathbb{P}(N|x(A))}. \tag{102}$$

□

Proof of Theorem 1: Pick some $L \in [0, 1/2)$ and assume that communication is carried out over a binary symmetric channel Q with crossover probability $\varepsilon \in [0, L]$. The proof is divided into a few subsections. We first introduce the coding scheme used for the first phase that we denote by c_1^M. We then propose the two-message coding scheme, denoted by c_2^M, used for the second phase of communication. In the last subsection, we show that the two-phase coding scheme has the desired properties.

a. We consider the coding scheme used for the first phase. Letting $\alpha > 1$ and P be the Bernoulli $1/2$ distribution, we deduce from Proposition 2 that there exists a sequence of coding schemes

$$\{c_1^M \triangleq (f^M, \phi_u^M, T_1(\alpha, M) \wedge n_*(\alpha, P, M))\}_{M\geq 1} \tag{103}$$

such that, for every $W \in \mathrm{BSC}_L$

$$\begin{aligned}\liminf_{M\to\infty} &\frac{1}{\mathbb{E}_W(T_1(\alpha, M)\wedge n_*(\alpha, P, M))} \ln \mathbb{P}_W(\mathcal{E}|c_1^M) \\ &\geq C(W) - \frac{C(W)}{\alpha} \end{aligned} \tag{104}$$

and

$$\lim_{M\to\infty} \frac{\ln M}{\mathbb{E}_W(T_1(\alpha, M) \wedge n_*(\alpha, P, M))} = \frac{C(W)}{\alpha}\,. \tag{105}$$

b. The messages "Ack" and "Nack" are encoded by using the all-one sequence $x(A) = 1, 1, \ldots$ and the all-zero sequence $x(N) = 0, 0, \ldots$, respectively. Suppose that $x(A)$ is being sent. The resulting output sequence $Y_1, Y_2, \ldots$ is an i.i.d. Bernoulli sequence with $\mathbb{P}(Y_1 = 1) = 1 - \varepsilon$. Define the random variables Z_i's as

$$Z_i = 1, \quad \text{if } Y_i = 1 \text{ and } \quad Z_i = -1, \quad \text{if } Y_i = 0. \tag{106}$$

Let $V_n \triangleq \sum_{i=1}^n Z_i$, $\beta > 0$, and define the stopping time

$$T_2 = T_2(\beta, M) \triangleq \inf\{n \geq 1 : |V_n| \geq \lceil \beta \ln M \rceil\}\,. \tag{107}$$

Consider the decoding rule:
- if $V_{T_2} \geq \lceil \beta \ln M \rceil$: declare "Ack,"
- if $V_{T_2} \leq -\lceil \beta \ln M \rceil$: declare "Nack."

By symmetry we have

$$\mathbb{E}_Q(T_2|x(A)) = \mathbb{E}_Q(T_2|x(N)) \triangleq \mathbb{E}_Q T_2 \tag{108}$$

and

$$\begin{aligned}\mathbb{P}_Q(A|x(N)) &= \mathbb{P}_Q(N|x(A)) \\ &= \mathbb{P}_Q(V_{T_2} = -\lceil \beta \ln M \rceil\, |x(A)). \end{aligned} \tag{109}$$

Now $\mathbb{P}_Q(V_{T_2} = -\lceil \beta \ln M \rceil\, |x(A)) \leq e^{-\lceil \beta \ln M \rceil r^*}$ where r^* is the strictly positive root of the function $\ln(\mathbb{E}e^{rZ_1})$ (see, e.g., [11, Corollary 1 p. 233]) and equals $D(\varepsilon\|1-\varepsilon)/(1-2\varepsilon)$.[12] Therefore, we have

$$\mathbb{P}_Q(V_{T_2} = -\lceil \beta \ln M \rceil\, |x(A)) \leq e^{-\frac{\lceil \beta \ln M \rceil}{1-2\varepsilon} D(\varepsilon\|1-\varepsilon)}. \tag{110}$$

From Wald's equality and (108) we have

$$\begin{aligned}\mathbb{E}_Q\left(\sum_{i=1}^{T_2} Z_i \Big| x(A)\right) &= \mathbb{E}_Q(Z_1|x(A))\mathbb{E}_Q T_2 \\ &= (1 - 2\varepsilon)\mathbb{E}_Q T_2\,. \end{aligned} \tag{111}$$

Since

$$\begin{aligned}&\mathbb{E}_Q\left(\sum_{i=1}^{T_2} Z_i \Big| x(A)\right) \\ &= \lceil \beta \ln M \rceil\, \mathbb{P}_Q(V_{T_2} = \lceil \beta \ln M \rceil\, |x(A)) \\ &\quad - \lceil \beta \ln M \rceil\, \mathbb{P}_Q(V_{T_2} = -\lceil \beta \ln M \rceil\, |x(A)) \\ &= \lceil \beta \ln M \rceil - 2\mathbb{P}_Q(V_{T_2} = -\lceil \beta \ln M \rceil\, |x(A)))\lceil \beta \ln M \rceil \end{aligned} \tag{112}$$

from (110) and (111) we get

$$\mathbb{E}_Q T_2 = \frac{\beta \ln M}{1 - 2\varepsilon}(1 + o(1))\,. \tag{113}$$

Hence, using (110) we obtain

$$\begin{aligned}\mathbb{P}_Q(N|x(A)) &\triangleq \mathbb{P}_Q(V_{T_2} = -\lceil \beta \ln M \rceil\, |x(A)) \\ &\leq e^{-D(\varepsilon\|1-\varepsilon)(1+o(1))\mathbb{E}_Q T_2}\,. \end{aligned} \tag{114}$$

For given $M \geq 1$ and $\beta > 0$, the two-message coding scheme described above will be denoted c_2^M.

[12] $D(\varepsilon\|1-\varepsilon)$ denotes the divergence $\varepsilon \ln(\varepsilon/(1-\varepsilon)) + (1-\varepsilon)\ln((1-\varepsilon)/\varepsilon)$.

c. Let c^M denote the two-phase coding scheme obtained by using c_1^M and c_2^M for the first and second phase, respectively, and let $T(\alpha,\beta,M)$ denote the decoding time of c^M. Using Lemma 6, we deduce that that, for all $Q \in \mathrm{BSC}_L$

$$\mathbb{P}_Q(\mathcal{E}|c^M) \leq e^{-D(\varepsilon\|1-\varepsilon)(1+o(1))\left(1-\frac{\mathbb{E}_Q T(\alpha,\beta,M)-\mathbb{E}_Q T_2(\beta,M)}{\mathbb{E}_Q T(\alpha,\beta,M)}\right)\mathbb{E}_Q T(\alpha,\beta,M)} \tag{115}$$

where we used the fact that $E_B(0,Q) = D(\varepsilon\|1-\varepsilon)$.

Now, one can show that (see Appendix A)

$$\mathbb{E}_Q T(\alpha,\beta,M) - \mathbb{E}_Q T_2(\beta,M) = \mathbb{E}_Q(T_1(\alpha,M) \wedge n_*(\alpha,P,M))(1+o(1)) \,. \tag{116}$$

Hence, using (113) and (105), for any $\alpha > 1$ and $\beta > 0$

$$\lim_{M\to\infty} \frac{\mathbb{E}_Q T(\alpha,\beta,M) - \mathbb{E}_Q T_2(\beta,M)}{\mathbb{E}_Q T(\alpha,\beta,M)} = \frac{\alpha R(\alpha,\beta,\varepsilon)}{C(Q)} \tag{117}$$

where

$$R(\alpha,\beta,\varepsilon) \triangleq \frac{C(Q)}{\alpha + \frac{\beta C(Q)}{1-2\varepsilon}} = \lim_{M\to\infty} \frac{\ln M}{\mathbb{E}_Q T(\alpha,\beta,M)}. \tag{118}$$

Therefore, from (115)–(118), for every $Q \in \mathrm{BSC}_L$

$$\liminf_{M\to\infty} -\frac{1}{\mathbb{E}_Q T(\alpha,\beta,M)} \ln \mathbb{P}_Q(\mathcal{E}|c^M) \geq D(\varepsilon\|1-\varepsilon)\left(1 - \frac{\alpha R(\alpha,\beta,\varepsilon)}{C(Q)}\right). \tag{119}$$

In Appendix B, we show that if (118) and (119) hold for any $\alpha > 1$, then there exists a sequence of two-phase coding schemes for which (118) and (119) also hold for $\alpha = 1$, i.e., there exists $\{c^M\}_{M\geq 1}$ such that, for every $Q \in BSC_L$

$$\liminf_{M\to\infty} -\frac{1}{\mathbb{E}_Q T(\alpha=1,\beta,M)} \ln \mathbb{P}_Q(\mathcal{E}|c^M) \geq D(\varepsilon\|1-\varepsilon)\left(1 - \frac{R(\alpha=1,\beta,\varepsilon)}{C(Q)}\right) \tag{120}$$

where

$$R(\alpha=1,\beta,Q) = \lim_{M\to\infty} \frac{\ln M}{\mathbb{E}_Q T(\alpha=1,\beta,M)} = \frac{1}{1+\frac{\beta C(Q)}{1-2\varepsilon}} C(Q). \tag{121}$$

Now, for any $Q \in \mathrm{BSC}_L$ with crossover probability ε

$$0 < \min_{W\in\mathrm{BSC}_L} C(W) \leq \frac{C(Q)}{1-2\varepsilon} \leq \frac{\ln 2}{1-2L} \,. \tag{122}$$

Choose some $\gamma \in (0,1)$. Since $\beta > 0$ is arbitrary, by setting

$$\beta = \beta(\gamma) = \frac{1-2L}{\ln 2}\left(\frac{1}{\gamma} - 1\right)$$

from (120)–(122) we conclude that there exists $\mathbb{S}'$ such that, for every $Q \in \mathrm{BSC}_L$ with crossover probability ε

$$E(\mathbb{S}',Q) \geq D(\varepsilon\|1-\varepsilon)\left(1 - \frac{R(\mathbb{S}',Q)}{C(Q)}\right) \triangleq E_B(R(\mathbb{S}',Q),Q) \tag{123}$$

where

$$\gamma C(Q) \leq R(\mathbb{S}',Q) < C(Q) \,. \tag{124}$$

Similarly, if we now set

$$\beta = \beta(\gamma) = \frac{1}{\min_{W\in\mathrm{BSC}_L} C(W)}\left(\frac{1}{\gamma} - 1\right) \tag{125}$$

we deduce from (120)–(122) that there exists $\mathbb{S}''$ such that, for every $Q \in \mathrm{BSC}_L$ with crossover probability ε

$$E(\mathbb{S}'',Q) \geq D(\varepsilon\|1-\varepsilon)\left(1 - \frac{R(\mathbb{S}'',Q)}{C(\varepsilon)}\right) \triangleq E_B(R(\mathbb{S}'',Q),Q) \tag{126}$$

where

$$0 < R(\mathbb{S}'',Q) \leq \gamma C(Q) \,. \tag{127}$$

Since Burnashev's exponent cannot be exceeded, inequalities (123) and (126) are indeed equalities.

Finally, for the case $\gamma = 0$. From (118) and (119), by fixing $\alpha > 1$ and letting $\beta = \beta(M) \to \infty$ as $M \to \infty$, on can easily check that there exists $\mathbb{S}'''$ such that, for any $Q \in \mathrm{BSC}_L$, $E(\mathbb{S}''',Q) = E_B(0,Q)$ and $R(\mathbb{S}''',Q) = 0$. □

Proof of Theorem 2: Let $L \in [0,1)$ and assume that communication is carried over some Z channel Q with crossover probability ε. We first introduce the coding scheme used for the first phase, then the two-message coding scheme for the second phase of communication. The resulting two-phase coding scheme has zero error, i.e., infinite error exponent, for rates in the range $[0, I(PQ))$, for some fixed-input distribution P. As a final step, we show that the concatenation of two two-phase coding schemes also yields error-free communication, but now for all rates in $[0, C(Q))$.

a. We consider the coding strategy for the first phase. Let P be a probability distribution over $\{0,1\}$. From Proposition 2 there exists a sequence of coding schemes

$$\{c_1^M = (f^M, \phi_u^M, T_1(\alpha,M) \wedge n_*(\alpha,P,M)\}_{M\geq 1}$$

that satisfies, for all $W \in Z_L$

$$\liminf_{M\to\infty} -\frac{1}{\mathbb{E}_W(T_1(\alpha,M) \wedge n_*(\alpha,P,M))} \ln \mathbb{P}_W(\mathcal{E}|c_1^M) \geq I(PW) - \frac{I(PW)}{\alpha} \tag{128}$$

and

$$\lim_{M\to\infty} \frac{\ln M}{\mathbb{E}_W(T_1(\alpha,M) \wedge n_*(\alpha,P,M))} = \frac{I(PW)}{\alpha}. \tag{129}$$

b. The messages "Ack" and "Nack" of the second phase are encoded by using the all-one sequence $x(A) = 1,1,\ldots$

and the all-zero sequence $x(N) = 0, 0, \ldots$. The length of the second phase depends of length of the first phase. More precisely, the length of the second phase is given by

$$T_2(k, M) \triangleq \lceil kT_1(\alpha, M) \wedge n_*(\alpha, P, M) \rceil$$

for some constant k that will be appropriately chosen later. At the end of the second phase the decoding rule is given by the following:

- if there exists $i \in [1, T_2(k, M)]$ such that $y_i = 1$: declare "Ack,"
- otherwise: declare "Nack."

It follows that

$$\mathbb{P}(A|x(N)) = 0 \,. \tag{130}$$

Now, by definition, we have $T_1(\alpha, M) \geq \left\lceil \frac{\alpha \ln M}{\ln |\mathcal{X}|} \right\rceil$. Therefore,

$$\mathbb{P}_Q(N|x(A)) \leq \varepsilon^{k \frac{\alpha \ln M}{\ln \mathcal{X}}}. \tag{131}$$

c. Choose some $\gamma \in (0, 1)$, let $\alpha \in (1, 1/\gamma)$, and let $k = 1/(\alpha\gamma) - 1$. Let c^M denotes the two-phase coding scheme obtained from a and b, and let $T(\alpha, k, M)$ denote the decision time of the two phase scheme. Since the average decoding time of the two-phase scheme is essentially equal to the average length of one cycle of the two-phase scheme (see Appendix A), we have

$$\begin{aligned} &\mathbb{E}_Q T(\alpha, k, M) \\ &\quad = (1 + k)\mathbb{E}_Q(T_1(\alpha, M) \wedge n_*(\alpha, P, M))(1 + o(1)). \end{aligned} \tag{132}$$

Letting $\mathtt{S} = \{c^M\}_{M \geq 1}$ and using (129) we get, for all $Q \in Z_L$

$$R(\mathtt{S}, Q) = \gamma I(PQ) \,. \tag{133}$$

From (130), (131), and Lemma 6 we trivially have

$$E(\mathtt{S}, Q) = \infty. \tag{134}$$

If $\gamma = 0$ it suffices to have $k = k(M)$ such that

$$k(M) \overset{M \to \infty}{\longrightarrow} \infty.$$

In this case, $R(\mathtt{S}, Q) = 0$ and $E(\mathtt{S}, Q) = \infty$.

We have shown that any rate in $[0, \gamma I(PQ))$ can be universally achieved over Z_L at Burnashev's error exponent (here infinite). In the next subsection we show that the concatenation of two two-phase coding schemes yields the same result as above,now for all rates in $[0, C(Q)]$.

d. Let $\gamma \in (0, 1)$, let P_1 be the Bernoulli $1/2$ distribution, and consider a sequence of coding schemes $\{c^{M_1}\}_{M_1 \geq 1}$ such that

$$\lim_{M_1 \to \infty} R(c^{M_1}, Q) = \gamma I(P_1 Q)$$

and

$$\liminf_{M_1 \to \infty} E(c^{M_1}, Q) = \infty.$$

Pick a particular coding scheme c^{M_1} with decoding time T^{M_1}. The transmitter starts sending a message out the M_1 messages. At time T^{M_1}, the receiver decodes the sent message (error-free) and the transmitter makes an estimate $\hat{Q}$ of the underlying Z channel according to[13]

$$I(P_1 \hat{Q}) = \frac{\ln M_1}{\gamma T^{M_1}} \tag{135}$$

then sets P_2 as being the capacity-achieving distribution of $\hat{Q}$, i.e.,

$$P_2 \triangleq \max_{P \in \nabla} I(P\hat{Q}) \tag{136}$$

where ∇ is the set of all binary distributions P such that $P(0) \in [1/e, 1/2]$.[14] At a second stage, the transmitter chooses a message out of a second message set of size M_2, and sends it using a two-phase coding scheme according to P_2. Clearly, the overall two two-phase coding scheme is error free, since at the end of each of the two coding periods no error occurs.

Let us set $M_2 = e^{M_1}$. In the rest of the proof, we show that the rate of the two two-phase scheme, that we denote by $R_{1,2}$, converges to $\gamma C(Q)$ as M_1 tends to infinity. We have

$$\begin{aligned} R_{1,2} &\triangleq \frac{\ln(M_1 \cdot M_2)}{\mathbb{E}_Q T^{M_1} + \mathbb{E}_Q T^{M_2}} \\ &= \frac{\ln M_2}{\mathbb{E}T^{M_2}} \frac{\left(1 + \frac{\ln M_1}{\ln M_2}\right)}{\left(1 + \frac{\mathbb{E}T^{M_1}}{\mathbb{E}T^{M_2}}\right)} \,. \end{aligned} \tag{137}$$

Now, since the two-phase schemes we consider have their second phase duration linear in the first phase length, Lemma 4 extends to the decoding time T^{M_1} as

$$\mathbb{P}\left(T^{M_1} > \frac{\ln M_1}{\gamma I(P_1 Q)}(1 + f(M_1))\right) \leq e^{-\frac{\sqrt{\ln M_1}}{2I(P_1 Q)}(1 + o(1))} \tag{138}$$

and

$$\mathbb{P}\left(T^{M_1} \leq \frac{\ln M_1}{\gamma I(P_1 Q)}(1 + g(M_1))\right) \leq e^{-\frac{\sqrt{\ln M_1}}{\ln |\mathcal{X}|}(1 + o(1))} \tag{139}$$

for some functions $f(M_1) = o(1)$ and $g(M_1) = o(1)$. Let us define

$$N \triangleq \left[\frac{\ln M_1}{\gamma I(P_1 Q)}(1 + f(M_1)), \frac{\ln M_1}{\gamma I(P_1 Q)}(1 + g(M_1))\right]. \tag{140}$$

From the bounds (138) and (139) we get

$$\mathbb{P}(T^{M_1} \notin N) \leq e^{-\frac{\sqrt{\ln M_0}}{2I(P_1 Q) + \ln |\mathcal{X}|}(1 + o(1))} \,. \tag{141}$$

We now derive an upper and a lower bound on $\mathbb{E}_Q(T^{M_2})$. On the one hand, we have

$$\begin{aligned} \mathbb{E}T^{M_2} \leq\ & \mathbb{P}(T^{M_1} \in N) \max_{t \in N} \mathbb{E}(T^{M^2} | T^{M_1} = t) \\ & + \mathbb{P}(T^{M_1} \notin N) \max_{t \notin N} \mathbb{E}(T^{M_2} | T^{M_2} = t) \,. \end{aligned} \tag{142}$$

From (135), (136), and (138) we deduce that

$$\max_{t \in N} \mathbb{E}(T^{M_2} | T^{M_1} = t) \leq \frac{\ln M_2}{\gamma C(Q)}(1 + o(1)) \tag{143}$$

[13]By means of feedback, this operation is performed also at the receiver.

[14]Majani and Rumsey [14] proved that, for any Z channel, the capacity-achieving distribution is such that $P(0) \in (1/e, 1/2)$.

as $M_1 \to \infty$, and

$$\max_{t \notin N} \mathbb{E}(T^{M_2}|T^{M_1}=t) \leq \frac{\ln M_2}{\gamma \min_{P\in\nabla} \min_{W\in Z_L} I(PW)}(1+o(1)) \quad (144)$$

as $M_1 \to \infty$, where $\min_{P\in\nabla} \min_{W\in \mathrm{Z}_L} I(PW) > 0$ since $L < 1$.[15] From (141)–(144) we have

$$\mathbb{E}T^{M_2} \leq \frac{\gamma \ln M_2}{C(Q)}(1+o(1))\,. \quad (145)$$

On the other hand, since

$$\mathbb{E}T^{M_2} \geq \mathbb{P}(T^{M_1} \in N) \min_{t\in N} \mathbb{E}(T^{M_2}|T^{M_1}=t) \quad (146)$$

a similar computation as above yields

$$\mathbb{E}T^{M_2} \geq \frac{\ln M_2}{\gamma C(Q)}(1+o(1)), \qquad \text{as } M_1 \to \infty. \quad (147)$$

From (145) and (147) we conclude that

$$\mathbb{E}(T^{M_2}) = \frac{\gamma \ln M_2}{C(Q)}(1+o(1)) \quad (148)$$

as $M_1 \to \infty$. Hence, from (137), (148), and the fact that $M_2 = e^{M_1}$, the two two-phase coding scheme with $M = M_1 \cdot M_2$ messages has its rate $R_{1,2}$ that converges to $\gamma C(Q)$ as M_1 tends to infinity. □

We now consider an alternative way for proving Theorems 1 and 2. Theorem 1 is proved by considering two-phase coding schemes. In particular, we used Proposition 2 that claims the existence of coding schemes, which we used for the first phase, that achieve an error exponent equal to $C(Q) - R$ with $R = C(Q)/\alpha$, uniformly over the family BSC_L, for any $\alpha > 1$. However, as follows from the discussion that precedes Lemma 6, the error exponent of the first phase coding scheme is irrelevant: what is important is to achieve capacity. Hence, let us keep the second phase of two-message communication and replace the first phase by a training-based scheme, namely, a coding scheme that first estimates the channel by means of a test sequence, then conveys information with a fixed length block codebook and a maximum-likelihood decoder tuned for the estimated channel [8], [20]. It may be easily checked that such a training-based scheme can achieve a rate $R = C(Q)/\alpha$ uniformly over BSC_L, for any $\alpha > 1$. Hence, the two-phase scheme where the first phase is a training-based scheme, achieves Burnashev's exponent at a rate that is controlled as stated in Theorem 1.

A similar argument as above holds for the family Z_L. By using a training-based scheme followed by a two-message coding scheme, it is possible to achieve Burnashev's exponent at a rate that is controlled as stated in Theorem 2. Therefore, we found a two-phase strategy that yields the same performance as the coding scheme that results from the concatenation of two two-phase schemes (see proof of Theorem 2).

The reader may ask why we did not immediately use the above arguments to prove Theorems 1 and 2, which clearly renders these proofs simpler. The reason is the following. A codebook for a channel with feedback is a set of sequences of functions $\{\{f_n(m, Y^{n-1})\}_{n\geq 1}\}_{m=1}^{M}$. As mentioned above, Proposition 2 proves the existence of coding schemes that achieve an error exponent equal to $C(Q) - R$ uniformly over the family BSC_L. A look at its proof reveals that such universal codes can be written simply as $\{\{x_n(m)\}_{n\geq 1}\}_{m=1}^{M}$ instead of $\{\{f_n(m, Y^{n-1})\}_{n\geq 1}\}_{m=1}^{M}$. In other words, the universal codebook of Proposition 2 is composed by M infinite sequences of digits, that are not functions of the received symbols, i.e., do not make use of feedback. Stated otherwise, the encoder knows, before communication starts, which symbol will be sent at any time, unless the decoder makes a decision previously. If we use training-based schemes instead, at the end of the test period, the encoder needs to have a large set of available codebooks for the different channel estimates, which is complex. For this reason, we preferred to prove Theorems 1 and 2 with a method that does not involve training-based schemes.

The proof of Theorem 3 is a straightforward extension of the theorem in [19].

Proof of Theorem 3: Consider two channels Q_1 and Q_2. Let $\mathtt{S}$ be any sequence of coding schemes yielding zero rate on Q_1 and Q_2. From the proof of the Theorem in [19] (see, in particular, the inequalities (44) and (45) therein) we deduce that either

$$E(\mathtt{S}, Q_1) \leq \frac{1}{2}K(Q_1, Q_2)$$

or

$$E(\mathtt{S}, Q_2) \leq \frac{1}{2}K(Q_2, Q_1)$$

(or both), where

$$K(Q_i, Q_j) \triangleq \max_{x,x'} \big[D(Q_i(\cdot|x)\|Q_i(\cdot|x')) + D(Q_i(\cdot|x)\|Q_j(\cdot|x'))\big]. \quad (149)$$

Since the error exponent is a nonincreasing function of the rate, if $K(Q_1, Q_2) < 2E_B(\gamma C(Q_1), Q_1)$ and $K(Q_2, Q_1) < 2E_B(\gamma C(Q_2), Q_2)$, then, for any $\gamma' \in [0, \gamma]$ and any sequence of coding schemes $\mathtt{S}$ such that $R(\mathtt{S}, Q_i) = \gamma' C(Q_i)$ for $i \in \{1, 2\}$, either

$$E(\mathtt{S}, Q_1) < E_B(\gamma' C(Q_1), Q_1)$$

or

$$E(\mathtt{S}, Q_1) < E_B(\gamma' C(Q_2), Q_2).$$ □

C. Extending the Results for BSC_L and Z_L to More General Families

A difficulty in extending Theorems 1 and 2 to a more general family of channels is due to the converse result provided by Theorem 3.

Suppose for simplicity that we want a coding strategy that achieves Burnashev's exponent universally over some family $\mathcal{Q} = \{Q_1, Q_2\}$ at a rate strictly below capacity. In other words, we seek for a coding strategy that is optimal from the error exponent point of view, but that does not necessarily satisfy

[15] The set ∇ has been defined in (136).

the constraint that controls the rate.[16] From the alternative proofs given for Theorems 1 and 2 (see discussion before the proof of Theorem 3), the first phase may be carried out by any universal capacity-achieving coding scheme, such as training-based schemes. However, a major problem arises in finding a sequence of two-message coding schemes such that, for $Q \in \{Q_1, Q_2\}$

I. $\mathbb{P}_Q(N|x(A)) \stackrel{M\to\infty}{\longrightarrow} 0$,
II. $\mathbb{P}_Q(A|x(N)) \leq e^{-\mathbb{E}_Q T_2 E_B(0,Q)(1+o(1))}$ as $M \to \infty$.

For BSCs and Z channels, there exist sequences of two-message coding schemes that satisfy I and II. In addition, these coding schemes have the property of having constant codewords $x(A) = 1, 1, \ldots$ and $x(N) = 0, 0, \ldots$. To gain more insight on the limitation of universal coding schemes, and in connection with hypothesis testing, it might be interesting to consider the following situation. Suppose that $\mathcal{Q} = \{Q_1, Q_2\}$ does not satisfy the conditions (30) in Theorem 3, and that

$$\arg\max_{(x,x')} D(Q_1(\cdot|x)\|Q_1(\cdot|x')) \neq \arg\max_{(x,x')} D(Q_2(\cdot|x_1)\|Q_2(\cdot|x_2')). \quad (150)$$

Is there a sequence of two-message coding schemes such that conditions I and II are satisfied? Informally, the question can be rephrased as: when are adaptive encoding procedures necessary in order to achieve Burnashev's error exponent universally at zero rate? The motivation for studying two-message coding schemes (see [19] for a related study) is that the maximum achievable error exponent that can be obtained with two-message coding schemes corresponds to the Burnashev exponent at zero rate. This is in contrast with the situation without feedback where, in general, there is a significant difference between the maximum error exponents at zero rate and the one obtained with only two messages [10].

V. Conclusion

The main concern of this paper has been to show how much feedback may help when communication is carried out over a stationary DMC that is unknown to both the transmitter and the receiver. We have demonstrated that there are channels for which the ignorance of both the transmitter and the receiver of the channel in use is not a fundamental impediment to reliable communication (Theorems 1 and 2). The communicating parties can employ a universal coding strategy to asymptotically perform as well as the best communication schemes tuned for the channel over which communication is carried out. In these cases we may notice that, in terms of error exponent, it is better to have feedback while ignoring the channel rather than to know the channel and not having feedback.

However, in general, for a given family of channels such optimal coding schemes do not exist: there are simple families, namely, families with only two channels, for which universally optimal coding schemes do not exist (Theorem 3).

In order to further understand in which situations the presence of feedback helps in a communication setting, it might be interesting to consider the following research directions. The optimal blind schemes presented in this paper use full feedback, as Burnashev's optimal schemes [2]. Since in practice the feedback link may have a limited capacity, we are lead to the following questions: what error exponent can be achieved if we restrict ourselves to a low-rate error-free feedback link or to decision feedback? Is it necessary to have full feedback to attain Burnashev's exponent? To the best of our knowledge, Burnashev's exponent has been obtained only by using two-phase coding schemes, the basic structure of which was first introduced by Schalkwijk and Barron [16] for Gaussian channels. In these schemes, full feedback is needed to inform the transmitter about which message has been declared "most probable" by the receiver at the end of the first phase (see Section III-B). Perhaps this amount of feedback may be reduced, for example, by using the fact that the sent and received symbols are correlated.

In the case where the channel is known, Forney [9] showed that error in an exponent larger than $C - R$ can be achieved with decision feedback. In the case where the channel is unknown, we showed that, in some cases, $C - R$ can be achieved with 1-bit feedback (see Proposition 2).[17] For a study related to the tradeoff between feedback rate and error exponent we refer to [6]. There the authors consider two-phase schemes where the feedback channel is used at a low rate, but in a "bursty" way.

An aspect that has not been addressed in this paper is complexity. In this framework, we would like to mention Ooi's Ph.D. dissertation [15] in which practical low-complexity feedback schemes have been derived for different categories of known/unknown channels, such as discrete channels with and without memory, and multiple-access channels. It might be interesting to further study low-complexity coding schemes in the framework of a more general question that seeks the tradeoff between performance and complexity.

Appendix A
Two-Phase Coding Scheme Decoding Time

In this section, we show that the average decoding time of a two-phase scheme is approximatively equal to the sum of the average decoding time of the first phase and the average decoding time of the second phase, which justifies (116) and (133).[18]

Let $T = T(\alpha, \beta, M)$ be the overall decoding time, T_1 be the decoding time of the first phase (for brevity we write T_1 for $T_1(\alpha, M) \wedge n_*(\alpha, P, M)$), and T_2 be the decoding time of the first phase. Let $T^1 \stackrel{\Delta}{=} T_1 + T_2$, and S denote the number of cycles performed by the two-phase scheme (i.e., the number of "Nacks" before the final "Ack," plus one). We have

$$\begin{aligned}\mathbb{E}T &= \sum_{s\geq 1} \mathbb{E}(T|S=s)\mathbb{P}(S=s) \\ &= \sum_{s\geq 1} [(s-1)\mathbb{E}(T^1|N) + \mathbb{E}(T^1|A)]\mathbb{P}(S=s) \\ &= \mathbb{E}(T^1|N)\mathbb{E}(S-1) + \mathbb{E}(T^1|A) \end{aligned} \quad (151)$$

where $\mathbb{E}(T^1|l)$ denotes the expected value of the first cycle given that, at the end of the second phase, the decoder declares message $l \in \{\text{Ack}, \text{Nack}\}$.

[16] Formally, using the operator Δ' defined in (22), we ask whether $\Delta'(\mathcal{Q}) = 0$.

[17] Note that the information carried by this single bit through the feedback link is more than 1 bit since it is sent at a random time.

[18] Note that the assertion becomes trivial if one considers fixed-length block coding schemes for the first and second phase.

We will show that

$$\mathbb{E}(T^1|N)\mathbb{E}(S-1)+\mathbb{E}(T^1|A)=\mathbb{E}(T^1|A)(1+o(1)) \quad (152)$$

and that $\mathbb{E}(T^1|A)$ is approximatively equal to the average length of one cycle, i.e.,

$$\mathbb{E}(T^1|A)=\mathbb{E}(T^1)(1+o(1)). \quad (153)$$

We first prove (153). From the identity

$$\mathbb{E}T^1=\mathbb{E}(T^1|A)\mathbb{P}(A)+\mathbb{E}(T^1|N)\mathbb{P}(N) \quad (154)$$

and since $\mathbb{P}(A)=1+o(1)$, it suffices to show that

$$\mathbb{E}(T^1|N)\mathbb{P}(N)=o(\mathbb{E}(T^1|A)). \quad (155)$$

We have

$$\begin{aligned}\mathbb{E}(T^1|A)=&\mathbb{E}(T_1|x(A))\mathbb{P}(x(A)|A)\\&+\mathbb{E}(T_2|x(A),A)\mathbb{P}(x(A)|A)\\&+\mathbb{E}(T_1|x(N))\mathbb{P}(x(N)|A)\\&+\mathbb{E}(T_2|x(N),A)\mathbb{P}(x(N)|A)\end{aligned} \quad (156)$$

where $\mathbb{P}(x(m)|m')$ is the probability that, at the beginning of the second phase, $x(m)$ ($m\in\{\text{Ack},\text{Nack}\}$) is sent conditioned on the event that, at the end of the second phase, the decoder declares message m' ($m'\in\{\text{Ack},\text{Nack}\}$). Similarly

$$\begin{aligned}\mathbb{E}(T^1|N)\mathbb{P}(N)=&\mathbb{E}(T_1|x(A))\mathbb{P}(x(A)|N)\mathbb{P}(N)\\&+\mathbb{E}(T_2|x(A),N)\mathbb{P}(x(A)|N)\mathbb{P}(N)\\&+\mathbb{E}(T_1|x(N))\mathbb{P}(x(N)|N)\mathbb{P}(N)\\&+\mathbb{E}(T_2|x(N),N)\mathbb{P}(x(N)|N)\mathbb{P}(N).\end{aligned} \quad (157)$$

We now show that the four terms on the right-hand side of (157) are negligible compared to $\mathbb{E}(T^1|A)$.

Since $\mathbb{P}(N)=o(1)$, we have

$$\mathbb{E}(T_1|x(A))\mathbb{P}(x(A)|N)\mathbb{P}(N)=o(1)\mathbb{E}(T_1|x(A)). \quad (158)$$

Then, by symmetry of the two-message coding scheme we have $\mathbb{E}(T_2|x(N),N)=\mathbb{E}(T_2|x(A),A)$, and therefore,

$$\mathbb{E}(T_2|x(N),N)\mathbb{P}(x(N)|N)\mathbb{P}(N)=\mathbb{E}(T_2|x(A),A)o(1). \quad (159)$$

Now for the term $\mathbb{E}(T_1|x(N))$, let us define $\tilde{T}_1=\tilde{T}_1(\alpha,M)$ as

$$\tilde{T}_1(\alpha,M)=\inf\left\{1\le n\le n_*(\alpha,P,M):\exists m\in\mathcal{M}\backslash\{1\}\ \text{such that}\ I(\hat{P}_{x^n(m),Y^n})\ge\frac{\alpha\ln M}{n}\right\}. \quad (160)$$

Since $\tilde{T}_1\ge T_1$, it follows that

$$\begin{aligned}\mathbb{E}(T_1|x(N))\le&\mathbb{E}(\tilde{T}_1|x(N))\\=&\mathbb{E}(\tilde{T}_1|\tilde{T}_1\le T_1)\\\le&\mathbb{E}(\tilde{T}_1).\end{aligned} \quad (161)$$

Using Lemma 1 and the union bound we have[19]

$$\begin{aligned}\mathbb{E}(\tilde{T}_1)=&\sum_{1\le n\le n_*}n\mathbb{P}(\tilde{T}_1=n)\\\le&\sum_{1\le n\le n_*}nM^{-(\alpha-1)}(n+1)^{|\mathcal{X}||\mathcal{Y}|}\\\le&M^{-(\alpha-1)}(n_*+1)^{|\mathcal{X}||\mathcal{Y}|+2}.\end{aligned} \quad (162)$$

[19]We assume that message 1 is sent.

Since $n_*(\alpha,P,M)=\Theta(\ln M)$ (see paragraph after (94)) we conclude that $\mathbb{E}(\tilde{T}_1)=o(1)$, and therefore, (161) gives

$$\mathbb{E}(T_1|x(N))=o(1). \quad (163)$$

Finally, we show that

$$\mathbb{E}(T_2|x(A),N)\mathbb{P}(x(A)|N)\mathbb{P}(N)=o(1). \quad (164)$$

Let $\mathbb{1}_N$ be equal to 1 if at the end of the second phase a "Nack" is declared, and be equal to zero otherwise. Since $\mathbb{P}(N|x(A))=\mathbb{P}(A|x(N))$, we have

$$\mathbb{E}(T_2|x(A),N)\mathbb{P}(x(A)|N)\mathbb{P}(N)=\mathbb{E}(T_2\mathbb{1}_N|x(A))\mathbb{P}(x(A)). \quad (165)$$

Hence, since $\mathbb{P}(x(A))=1+o(1)$, to prove (164) it suffices to show that the term $\mathbb{E}(T_2\mathbb{1}_N|x(A))=o(1)$. To that aim we refer to [11, Ch. 7], where one can find the results that we use here and that concern the expected stopping times for random walks. Let $g(r)\triangleq\ln\mathbb{E}(e^{rZ})$ where Z is the binary random variable taking value in $\{+1,-1\}$ and such that $\mathbb{E}Z=2\varepsilon-1$ ($0\le\varepsilon\le L<1/2$). Let r_0 be the strictly positive root of $g(r)$ and let $g'(r)$ denote the derivative of g at r. Let $\bar{r}$ be any value in $(0,r_0)$ such that $0<g'(\bar{r})<g'(r_0)$. From [11, p. 236][20] one deduces that

$$\mathbb{E}(T_2\mathbb{1}_N|x(A))\le\bar{n}e^{-\bar{r}\beta\ln M}+\sum_{n>\bar{n}}e^{ng(\bar{r})} \quad (166)$$

where $\bar{n}\triangleq\lceil\beta\ln M\rceil/g'(\bar{r})$. Since $g(\bar{r})<0$, the right-hand side of (166) vanishes as M tends to infinity, i.e.,

$$\mathbb{E}(T_2\mathbb{1}_N|x(A))=o(1) \quad (167)$$

and therefore (164) follows from (165).

Combining (157), (158), (159), (163), and (164) we have

$$\mathbb{E}(T^1|N)\mathbb{P}(N)=o(\mathbb{E}(T^1|A)) \quad (168)$$

and (154) gives

$$\mathbb{E}(T^1|A)=\mathbb{E}(T^1)(1+o(1)) \quad (169)$$

yielding (153).

Now notice that, in order to prove (168), the only property of $\mathbb{P}(N)$ we used is that it tends to zero as M goes to infinity. In other words, we made no assumption on the speed of decay of $\mathbb{P}(N)$. Hence, since $\mathbb{E}(S-1)=o(1)$,[21] using (168) we also deduce that $\mathbb{E}(T^1|N)\mathbb{E}(S-1)=o(\mathbb{E}(T^1|A))$. From (151)–(153) we have

$$\begin{aligned}\mathbb{E}(T)=&\mathbb{E}(T^1)(1+o(1))\\=&(\mathbb{E}T_1+\mathbb{E}T_2)(1+o(1)).\end{aligned} \quad (171)$$

From (171), and since $\mathbb{E}T_2=O(\mathbb{E}T_1)$, we conclude that

$$\mathbb{E}T-\mathbb{E}T_2=\mathbb{E}T_1(1+o(1)) \quad (172)$$

[20]In [11, p. 236] there is a typo in (28). The first term on the right-hand side of (28) should be $e^{-r\alpha+n\gamma(r)}$ instead of $e^{-r\alpha+\gamma(r)}$.

[21]Letting N_i denote the event that a "Nack" is declared at the end of the ith cycle, we have that $S\ge i$ if and only if there has been $i-1$ times a "Nack" that was declared at the end of the second phase. Hence,

$$\begin{aligned}\mathbb{E}S=&\sum_{i\ge1}\mathbb{P}(S\ge i)\\=&1+\sum_{i\ge2}\mathbb{P}(N_{i-1})\\=&\frac{1}{1-\mathbb{P}(N_1)}\end{aligned} \quad (170)$$

where the last equality is justified after (100).

where we wrote T_1 for $T_1 \wedge n_*(\alpha, P, M)$, which proves the desired result.

APPENDIX B
THE CLOSURE OF A SET OF ACHIEVABLE ERROR EXPONENTS

We prove that if (118) and (119) hold for any $\alpha > 1$, then (118) and (119) also hold for $\alpha = 1$.

Let $\{\alpha_n\}_{n\geq 1}$ be a nonincreasing sequence such that $\lim_{n\to\infty} \alpha_n = 1$. For convenience, let us define the six quantities:

$$\begin{aligned} R(m, \alpha_n, \beta) &\triangleq \frac{\ln M}{\mathbb{E}_Q T(\alpha_n, \beta, m)} \\ E(m, \alpha_n, \beta) &\triangleq -\frac{1}{\mathbb{E}_Q T(\alpha_n, \beta, m)} \ln \mathbb{P}_Q(\mathcal{E}|c^m) \\ R(\alpha_n, \beta) &\triangleq \lim_{m\to\infty} R(m, \alpha_n, \beta) \\ E(\alpha_n, \beta) &\triangleq \liminf_{m\to\infty} E(m, \alpha_n, \beta) \\ R(\beta, \varepsilon) &\triangleq \frac{1}{1 + \frac{\beta C(\varepsilon)}{1-2\varepsilon}} C(Q) \\ E(\beta, \varepsilon) &\triangleq \left(1 - \frac{1}{1 + \frac{\beta C(\varepsilon)}{1-2\varepsilon}}\right) D(\varepsilon\|1-\varepsilon) \,. \end{aligned} \quad (173)$$

Then we introduce the sequence $\{M(n)\}_{n\geq 0}$ where $M(0) = 1$ and, for every $n \geq 1$, the quantity $M(n)$ is obtained by recursion as

$$\begin{aligned} M(n) = \min \Big\{ & M > M(n-1) : \\ & E(\alpha_n, \beta) - \inf_{m\geq M} E(m, \alpha_n, \beta) \leq \frac{1}{n} \text{ and} \\ & |R(\alpha_n, \beta) - R(m, \alpha_n, \beta)| \leq \frac{1}{n}, \text{ for all } m \geq M \Big\}. \end{aligned} \quad (174)$$

From (174), we deduce that

$$E(M(n), \alpha_n, \beta) - E(\beta, \varepsilon) \geq -\frac{1}{n} + E(\alpha_n, \beta) - E(\beta, \varepsilon) \,. \quad (175)$$

Using (119), we have $\liminf_{n\to\infty} E(\alpha_n, \beta) \geq E(\beta, \varepsilon)$, therefore from (175) we get

$$\liminf_{n\to\infty} E(M(n), \alpha_n, \beta) \geq E(\beta, \varepsilon) \,. \quad (176)$$

Similarly, we obtain

$$\lim_{n\to\infty} R(M(n), \alpha_n, \beta) = R(\beta, \varepsilon) \,. \quad (177)$$

Since the sequence $\{M(n)\}_{n\geq 0}$ is nondecreasing, by defining

$$n(M) = \max\{n \geq 1 : M(n) \leq M\} \quad (178)$$

we conclude that

$$\begin{aligned} &\liminf_{M\to\infty} E(M, \alpha_{n(M)}, \beta) = E(\beta, \varepsilon) \\ \text{and} \quad &\lim_{n\to\infty} R(n, \alpha_{n(M)}, \beta) = R(\beta, \varepsilon). \end{aligned} \quad (179)$$

Setting

$$T(\beta, M) = T(\alpha_{n(M)}, \beta, M)$$

we infer that there exists a sequence of coding schemes $\{c^M = (f^M, \phi^M, T(\beta, M))\}_{M\geq 1}$ such that, for all $Q \in \mathrm{BSC}_L$

$$\begin{aligned} \liminf_{M\to\infty} -\frac{\ln M}{\mathbb{E}_Q T(\beta, M)} \ln \mathbb{P}_Q(\mathcal{E}|c^M) \\ \geq D(\varepsilon\|1-\varepsilon)\left(1 - \frac{R(\beta, \varepsilon)}{C(\varepsilon)}\right) \end{aligned} \quad (180)$$

where

$$R(\beta, \varepsilon) = \lim_{M\to\infty} \frac{\ln M}{\mathbb{E} T_Q(\beta, M)} = \frac{1}{1 + \frac{\beta C(\varepsilon)}{1-2\varepsilon}} C(\varepsilon) \,. \quad (181)$$

ACKNOWLEDGMENT

We are grateful to M. V. Burnashev, J. L. Massey, U. Niesen, and B. Rimoldi for their the substantial comments on the manuscript, and to E. Arikan for many stimulating discussions. We thank the reviewers as well as the Associate Editor for their numerous comments and suggestions for helping to improve the manuscript.

REFERENCES

[1] D. Blackwell, L. Breiman, and A. J. Thomasian, "The capacitites of certain channel classes under random coding," *Ann. Math. Statist.*, vol. 31, pp. 558–567, 1960.
[2] M. V. Burnashev, "Data transmission over a discrete channel with feedback: Random transmission time," *Probl. Inf. Transm.*, vol. 12, no. 4, pp. 250–265, 1976.
[3] I. Csiszàr, "On the capacity of noisy channels with arbitrary signal costs," in *Probl. Control and Inf. Theory*, vol. 2, 1973, pp. 283–304.
[4] I. Csiszàr and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Channels*. New York: Academic, 1981.
[5] R. L. Dobrushin, "Asymptotic bounds on the probability of error for the transmission of messages over a memoryless channel using feedback," *Probl. Kibern.*, vol. 8, pp. 161–168, 1963.
[6] S. C. Draper, K. Ramchadran, B. Rimoldi, A. Sahai, and D. N. C. Tse, "Attaining maximal reliability with minimal feedback via joint channel-code and hash-function desing," in *Proc. Allerton Conf. Communication, Control and Computing*, Monticello, IL, Sep. 2005.
[7] A. Ephremides and B. Hajek, "Information theory and communication networks: An unconsummed union," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2416–2434, Oct. 1998.
[8] M. Feder and A. Lapidoth, "Universal decoding for channels with memory," *IEEE Trans. Inf. Theory*, vol. 44, no. 5, pp. 1726–1745, Sep. 1998.
[9] G. D. Forney, Jr., "Exponential error bounds for erasure, list and decision feedback schemes," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 2, pp. 206–220, Mar. 1968.
[10] R. G. Gallager, *Information Theory and Reliable Communication*. Budapest: Wiley, 1968, ch. Hungary.
[11] ——, *Discrete Stochastic Processes*. Norwell, MA: Kluwer, 1995.
[12] V. D. Goppa, "Nonprobabilistic mutual information without memory," *Probl. Contr. Inf. Theory*, vol. 4, pp. 97–102, 1975.
[13] D. G. Luenberger, *Optimization by Vector Space Methods*. New York: Wiley, 1969.
[14] E. E. Majani and H. Rumsey, "Two results on binary-input discrete memoryless channels," in *Proc. IEEE Int. Symp. Information Theory*, Budapest, Hungary, Jun. 1991, p. 104.
[15] J. M.-S. Ooi, "A Framework for low-complexity communication over channels with feedback," Ph.D. dissertation, MIT, Cambridge, MA, 1997.
[16] J. P. M. Schalkwijk and M. E. Barron, "Sequential transmission under a peak power constraint," *IEEE Trans. Inf. Theory*, vol. IT-17, no. 3, pp. 278–282, May 1971.
[17] C. E. Shannon, "The zero-error capacity of a noisy channel," *IRE Trans. Inf. Theory*, vol. PGIT-2, no. 3, pp. 8–19, Sep. 1956.
[18] N. Shulman, "Communication over an unknown channel via common broadcasting," Ph.D. dissertation, Tel-Aviv Univ., Tel-Aviv, Israel, 2003.
[19] A. Tchamkerten and İ. E. Telatar, "On the universality of Burnashev's error exponent," *IEEE Trans. Inf. Theory*, vol. 51, no. 8, pp. 2940–2944, Aug. 2005.
[20] ——, "On the use of training sequences for channel estimation," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1171–1176, Mar. 2006.