

# Unequal Error Protection: Some fundamental limits and optimal strategies

Shashi Borade Baris Nakiboglu Lizhong Zheng  
EECS, Massachusetts Institute of Technology  
Email: { spb, nakib, lizhong } @mit.edu

**Abstract**—<sup>1</sup> Various scenarios are considered where some information is more important than other and needs better protection. A general theoretical framework for unequal error protection is developed in terms of exponential error bounds. It provides some fundamental limits and optimal strategies for such problems. A new class of problems called *message-wise* unequal error protection is also analyzed.

Even for data-rates approaching the channel capacity, we show how a crucial part of information can be protected with exponential reliability. Channels without feedback are analyzed first, which is useful later in analyzing channels with feedback. A new channel parameter, called its *pseudo-capacity*, is fundamentally important in such problems.

## I. INTRODUCTION

Classical theoretical framework for communication assumes that all information is equally important. In this framework, the communication system aims to provide a uniform error protection to all messages: any particular message being mistaken as any other is viewed to be equally costly. With such uniformity assumptions, the reliability of a communication scheme is measured by either the average or the worst case probability of error, over all possible messages to be transmitted. In information theory literature, a communication scheme is said to be *reliable* if this error probability can be made small. Communication schemes designed with this framework turn out to be optimal in sending any source over any channel, provided that long enough codes can be employed. This homogeneous view of information motivates the universal interface of “bits” between any source and any channel [1], and is often viewed as Shannon’s most significant contribution.

In many communication scenarios, such as wireless networks, interactive systems, and control applications, where sufficient error protection becomes a luxury, providing such a uniform protection for all the information may be either a wasteful or an infeasible approach. Instead, it is more efficient here to protect a (crucial) part of information better than the rest. For example

- In a wireless network, control signals including channel state, power control, and scheduling information are often more important than the payload data, and should be protected more carefully. Thus even though the final objective is delivering the payload data, the physical layer should provide a better protection to such protocol

information. Similarly for the Internet, packet headers are more important for delivering the packet and need better protection to ensure that the actual data gets through.

- Another example is when a multiple resolution source code is transmitted over a wireless channel. The coarse resolution needs a better protection than the fine resolution because then user can at least have some crude reconstruction after bad channel realizations.

These examples demonstrate the heterogeneous nature of information in contrast with the classical homogeneous view. For these situations, unequal error protection (UEP) is a natural generalization to the conventional content-blind information processing.

The simplest method of unequal error protection is to allocate different channels for different types of data. For example, wireless systems allocate a separate “control channel”, often with short codes and low spectral efficiency, to transmit control signals with high reliability. The well known Gray code, assigning similar bit strings to close by constellation points, can be viewed as UEP: even if there is some error in identifying the transmitted symbol, there is a good chance that some of the bits are correctly received. More systematic designs for UEP can be found in [12] and references therein. For erasure channels, this problem is known as “priority encoded transmission” (PET) [11]. For wireless channels, [13] analyzes this problem in terms of diversity-multiplexing tradeoffs. Most of these approaches focus on designing good codes for specific channel models. The optimality of these designs was established in only limited cases. This paper aims to provide a general information theoretic framework for understanding fundamental limits in UEP.

A general formulation of the unequal error protection problem requires some different definitions of decoding error than the commonly used ones. Consider a channel encoder which takes the input of  $k$  information bits,  $\mathbf{b} = [b_1, b_2, \dots, b_k]$ , which is equivalent to a random variable  $M$  taking values from the set  $\{1, 2, 3, \dots, 2^k\}$ . Each message in this set corresponds to a particular value of the bit-sequence  $\mathbf{b}$ . This set of possible values of  $M$  are referred to as “messages”. After a message is encoded and transmitted over the channel, a decoding error is defined as the event that the receiver decodes to a different message than the transmitted one. In most information theory texts, when a decoding error occurs, the entire bit sequence  $\mathbf{b}$  is rejected. That is, errors in decoding the message and in decoding the information bits are treated similarly.

<sup>1</sup>This research is supported by DARPA ITMANET project and an AFOSR grant FA9550-06-0156. Initial part of this paper was submitted to IEEE International Symposium on Information Theory, 2008.

In the existing formulations of unequal error protection codes, the information bits are divided into subsets, and the decoding errors in different subsets of bits are viewed as different kinds of errors. For example, one might want to provide a better protection to one subset of bits by ensuring that errors in these bits are less probable than the other bits. We call such problems “bit-wise UEP”. Previous examples of packet headers, multiple resolution codes, etc. belong to this category of UEP.

However, in some situations, instead of *bits* one might want to provide a better protection to a subset of *messages*. For example, one might consider embedding a special message in a normal  $k$ -bit code, i.e., transmitting one of  $2^k + 1$  messages, where the extra message has a special meaning and requires a smaller error probability. Note that the error event for the special message is not associated to error in any particular bit. Instead, it corresponds to a particular bit-sequence (i.e., message) being decoded as some other bit-sequence. Borrowing from hypothesis testing, we can define two kinds of errors corresponding a special message.

- We say that *missed-detection* of a message occurred when that message was transmitted but the receiver missed it by decoding to some other message. For example, consider a special message indicating some system emergency, which is too costly to be missed. Such special messages demand a small missed detection probability. Note that missed detection probability of a message is the same as the conditional error probability after its transmission.
- The special messages could instead demand small *false-alarm* probability, the event when the receiver erroneously chooses that message although some other message was sent. For example consider the reboot message for a remote-controlled system such as a robot or satellite. Its false-alarm could cause unnecessary shutdowns and other system troubles.

We call such problems as “message-wise UEP”. In conventional data communication, there is no need to distinguish between bit-errors and message-errors, as all information is “created equal”, its meaning (and importance) is separated from the engineering problem of communication [1]. In the UEP problems however, bits and messages are different as some are labeled as “special” or “high priority”. Now it becomes necessary to differentiate the two notions of special information: special bits and special messages.

The main contribution of this paper is a set of results, identifying the performance limits and optimal coding strategies, for several new formulations of UEP. We will focus on a few simplified notions of UEP, most with immediate practical applications, and try to illustrate the main insights for them. One can imagine using these UEP strategies for embedding protocol information within the actual data. By eliminating a separate control channel, this can enhance the overall bandwidth and/or energy efficiency.

For conceptual clarity, this article focuses on situations where the data-rate is a crucial resource and we cannot afford

to keep it away from capacity<sup>2</sup>. In these cases, no positive error exponent in the conventional sense can be achieved. That is, if we are aiming for a uniform protection for the entire information, one cannot achieve an error probability that decreases exponentially with the code length. We ask the question then “can we make the error probability for a particular bit, or for a particular message, to decay exponentially fast with block length?”

The question of fundamental limits of UEP was clearly of interest in previous works on code designs for UEP. To the best of our knowledge, however, there was no general characterization of these limits in terms of error exponents; partially due to the difficulty in proving converses. In this paper and [9], we develop such converses as well as optimal strategies. More importantly, the formulation of message-wise UEP is new. Conceptually, when we break away from the conventional framework by providing better protection to selected parts, these parts of information need not be only separate bits. Thus message-wise UEP can be viewed as one further step in differentiating the priorities of various parts of information.

In the following, Section II discusses bit-wise UEP and message-wise UEP for the no-feedback case. Theorem 1 shows that for data-rates approaching capacity, even a single bit cannot achieve any positive error exponent. Thus in bit-wise UEP, the data-rate must back-off from capacity for achieving any error exponent even for a single bit. On the contrary, in message-wise UEP, positive error exponents can be achieved even at capacity. If only one message in a capacity achieving code was special and demanded an error exponent, Theorem 2 shows its optimal value is equal to a new fundamental channel parameter called pseudo-capacity. We then consider situations where an exponentially large subset of messages is special and demands a positive error exponent. Theorem 3 shows a surprising result that these special messages can achieve the same exponent as if all the other (not special) messages were absent. In other words, a capacity achieving code and an error exponent-optimal code below capacity can coexist without hurting each other. These results shed some new light on the structure of capacity achieving codes.

Insights from the no-feedback case become useful in Section III for the case with full feedback, which shows that full feedback creates some fundamental connections between bit-wise UEP and message-wise UEP. Now even for bit-wise UEP, positive error exponent can be achieved at capacity. In fact, Theorem 5 shows that a single special bit can achieve the same exponent as a single special message, which equals the pseudo-capacity. As the number of special bits increases, the achievable exponent for them decays linearly with their rate as shown in Theorem 6. Then Theorem 7 generalizes this result to the case when there are multiple levels of speciality—most special, second most special and so on. It uses a strategy similar to onion-peeling and achieves error exponents which are successively refinable over multiple layers.

For a single special message however, Theorem 11 shows

<sup>2</sup>In another write-up [9], we have analyzed similar problems in a more general framework to allow data-rates below capacity.

that feedback does not improve the achievable exponent. The case of exponentially many messages is resolved in Theorem 9. Of course, many special messages cannot achieve a better exponent compared to a single special message. We will see that the special messages can achieve the same error exponent with feedback as if all other messages were absent.

Section IV-A then addresses message-wise UEP situations where special messages demand small probability of false-alarms instead of missed-detections. It considers the case of no-feedback as well as full feedback. This analysis was postponed in earlier sections to avoid confusion with the missed-detection results. Lastly, some future directions are discussed briefly in Section V.

## II. ERROR EXPONENTS AT CAPACITY: NO-FEEDBACK CASE

Consider a discrete memoryless channel  $W$  from input  $\mathbf{x}$  to output  $\mathbf{y}$  and let  $\mathcal{X}, \mathcal{Y}$  denote their alphabets, respectively. The output distribution conditioned on a particular input  $x \in \mathcal{X}$  is denoted by  $W_{\mathbf{y}|\mathbf{x}=x}(\cdot)$  and the channel capacity is denoted by  $C$ . We assume that all entries of the channel transition matrix are non-zero, *i.e.*, every output is reachable from every input. Let us first review the classical definition of an error exponent when all information is treated equally [2],[3],[4],[5],[6].

*Definition 1:* A  $(n, R, \epsilon_n)$  code denotes a length  $n$  code of rate  $R$ , which has  $e^{nR}$  messages and the average error probability of the overall code equals  $\epsilon_n$ .

$$\Pr(\hat{M} \neq M) = \frac{1}{e^{nR}} \sum_i \sum_{j \neq i} \Pr(\hat{M} = j | M = i) = \epsilon_n$$

where  $M, \hat{M} \in \{1, 2, \dots, e^{nR}\}$  denote the randomly chosen transmitted message and the decoded message, respectively.

At a given rate  $R$ , a sequence of  $(n, R, \epsilon_n)$  codes with increasing blocklength  $n$  is said to achieve an error exponent if its error probability  $\epsilon_n$  can decay exponentially with  $n$ .

*Definition 2:* The classical error exponent  $E(R)$  at rate  $R$  is the maximum value of  $E$  such that a sequence of  $(n, R, \epsilon_n)$  codes exists for which  $\epsilon_n$  satisfies  $\epsilon_n \doteq e^{-nE}$ . We use  $\doteq$  as a shorthand notation for

$$E = \lim_{n \rightarrow \infty} \frac{-\log \epsilon_n}{n} \quad (1)$$

*Reliable communication at capacity* means that for arbitrarily small gap to capacity  $C - R \triangleq \xi > 0$ , a sequence of  $(n, R, \epsilon_n)$  codes exists for which  $\epsilon_n$  vanishes for large  $n$ . However,  $\epsilon_n$  cannot decay exponentially in this case. That is, no positive exponent is achievable for this error probability as the gap  $\xi$  to capacity vanishes [3].

$$\xi \rightarrow 0 \quad \Rightarrow \quad E(R) = E(C - \xi) \rightarrow 0$$

### A. Special bit

We first address the situation where one particular (say the first) information bit out of the total  $nR/\log 2$  information bits is a special bit—it needs a superior error protection compared

to other bits. If this first bit is denoted as  $b_0$  and its decoded value is denoted by  $\hat{b}_0$ , we require that

$$\Pr(\hat{b}_0 \neq b_0) \ll \Pr(\hat{M} \neq M) = \epsilon_n$$

More precisely, we require the error probability for  $b_0$  to decay exponentially while ensuring reliable communication at capacity for the remaining bits. Let us define its exponent.

*Definition 3:* Let  $E_b(\xi)$  be the largest value such that a sequence of  $(n, C - \xi, \epsilon_n)$  codes exists for which  $\epsilon_n$  vanishes for large  $n$  and  $\Pr(\hat{b}_0 \neq b_0) \doteq e^{-nE_b(\xi)}$ . We define  $E_b$  as the infimum of  $E_b(\xi)$  over all positive  $\xi$ .

$$E_b = \inf_{\xi > 0} E_b(\xi)$$

This is equivalent to this simpler version used in the remaining paper.

*Definition 4:*  $E_b$  is the largest number such that a sequence of  $(n, C - \xi, \epsilon_n)$  codes exists

- for arbitrarily small capacity gap  $\xi > 0$ ,
- for which  $\epsilon_n$  vanishes with  $n$ ,
- and  $\Pr(\hat{b}_0 \neq b_0) \doteq e^{-nE_b}$ .

As noted earlier, the overall information cannot achieve any positive error exponent  $E(R)$  near capacity. However, it was not clear whether a single special bit can steal an error exponent  $E_b$  near capacity.

*Theorem 1:*  $E_b = 0$

**Geometric Interpretation:** Let the dotted regions in Fig. 1 denote the decoding regions of the  $e^{n(C-\xi)}$  messages. These decoding regions are large enough to ensure that the overall error probability  $\epsilon_n$  is vanishingly small. The decoding regions on the left of the thick line correspond to  $\hat{b}_0 = 1$  and those on the right half correspond to the same when  $\hat{b}_0 = 0$ . Each of these halves has  $e^{n(C-\xi)}/2$  decoding regions.

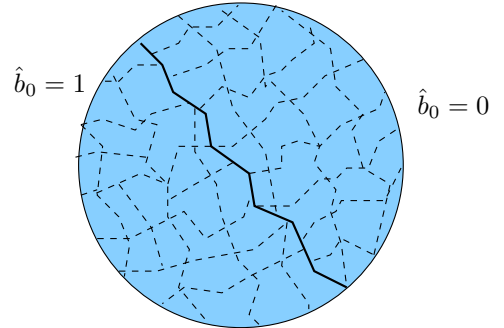


Fig. 1. Splitting the output space into 2 clusters.

A positive error exponent was achievable for the special bit if the codewords in the two halves were sufficiently separated from each other. This is possible if most of the codewords are outside a thick patch around the solid-line equator. Then the probability of landing in a half when a message from the other half was transmitted would have been exponentially small. However, above theorem implies that if we have to fill essentially  $e^{nC}$  codeword in this output space, then they cannot be separated into two distant clusters. Leaving a thick patch empty around the equator does not leave enough space for filling so many codewords.

## B. Special message

Now we focus on situations where one particular message (say  $M = 1$ ) out of the total  $e^{nR}$  messages is a special message—it needs a superior error protection. The missed detection (*i.e.*, conditional error) probability  $\Pr(\hat{M} \neq 1 | M = 1)$  for this ‘emergency’ message needs to be minimized.

*Definition 5:*  $E_{\text{md}}$  is the largest number such that a sequence of  $(n, C - \xi, \epsilon_n)$  codes exists

- for arbitrarily small  $\xi > 0$ ,
- for which  $\epsilon_n$  vanishes,
- and  $\Pr(\hat{M} \neq 1 | M = 1) \doteq e^{-nE_{\text{md}}}$ .

*Theorem 2:*  $E_{\text{md}} = \max_{x \in \mathcal{X}} D(P_{\mathbf{y}}^* \| W_{\mathbf{y}|x=x}) \triangleq \tilde{C}$ , where  $P_{\mathbf{y}}^*$  denotes the capacity achieving output distribution and where  $D(\cdot \| \cdot)$  denotes the Kullback-Liebler (KL) divergence between two distributions.

Compare this with the corresponding result for classical communication near capacity. If all the messages demand equally small missed detection probability, then no positive error exponent is achievable for them near capacity. This follows from the previous discussion of the classical error exponent  $E(R)$ . The above theorem shows the improvement in this exponent if we only demand it for a single message instead of all.

*Definition 6:* Parameter  $\tilde{C}$  of a channel is defined as its *pseudo-capacity*.

$$\tilde{C} = \max_{x \in \mathcal{X}} D(P_{\mathbf{y}}^* \| W_{\mathbf{y}|x=x}) \quad (2)$$

Define  $\mathcal{X}_p$  as the set of inputs achieving the maximum above.

Notice the relation between  $\tilde{C}$  and  $C$ : the arguments to KL divergence are flipped. It is because Karush-Kuhn-Tucker conditions for achieving capacity imply the following expression for  $C$  [4].

$$C = \max_{x \in \mathcal{X}} D(W_{\mathbf{y}|x=x} \| P_{\mathbf{y}}^*)$$

This is the reason for naming  $\tilde{C}$  as the pseudo-capacity. If capacity  $C$  represents the best possible data-rate over a channel, then pseudo-capacity  $\tilde{C}$  represents the best possible protection of a message for data-rates near capacity.

It is worth mentioning here the ‘‘very noisy’’ channel in [2]. In this formulation [10], the KL divergence is symmetric, which implies  $D(P_{\mathbf{y}}^* \| W_{\mathbf{y}|x=x}) \approx D(W_{\mathbf{y}|x=x} \| P_{\mathbf{y}}^*)$ . Hence the pseudo-capacity and capacity become essentially equal.

**Optimal strategy:** The special codeword is a repetition sequence of an input  $x_p \in \mathcal{X}_p$ . Its decoding region  $S$  contains every output sequence with empirical distribution (output type) different than the capacity achieving  $P_{\mathbf{y}}^*$ . For the ordinary codewords, use a capacity achieving code and apply ML decoding over them for output sequences outside  $S$ .

For a symmetric channel like BSC, all inputs can be used as  $x_p$ . Since the  $P_{\mathbf{y}}^*$  is the uniform distribution (denoted by  $U_{\mathbf{y}}$ ) for these channels,  $\tilde{C} = D(U_{\mathbf{y}} \| W_{\mathbf{y}|x=x})$  for any input  $x$ . This is the sphere-packing exponent  $E_{\text{sp}}(0)$  of this channel at rate 0.

**Geometric Interpretation:** Missed detection exponent for the special message corresponds to having a large decoding region  $S$  for the special message. This ensures that when the special message is transmitted, probability of landing outside  $S$  is exponentially small. In essence,  $E_{\text{md}}$  indicates how large this  $S$  could be made, while still filling essentially  $e^{nC}$  small decoding regions in the remaining space. The red region in Fig. 2 denotes such a large region. Note that the actual decoding region  $S$  is much larger than this illustration, because it consists of all output types except  $P_{\mathbf{y}}^*$  whereas the ordinary decoding regions only contain the output type  $P_{\mathbf{y}}^*$ .

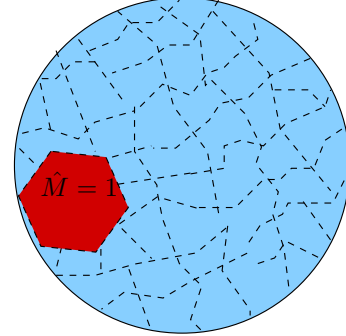


Fig. 2. Avoiding missed-detection

Utility of this result is two folds: first, the optimality of such a simple scheme was not obvious before; second, protecting a single special message can be a key building block for many other problems when feedback is available.

## C. Many special messages

Now consider that instead of a single special message, exponentially many of the  $e^{n(C-\xi)}$  total messages are special. Let these special messages be the first  $e^{nr}$  messages. Define  $E_{\text{MD}}(r)$  as the best missed detection exponent for these special messages.

*Definition 7:* For a fixed  $r < C$ , define  $E_{\text{MD}}(r)$  as the largest number such that a sequence of  $(n, C - \xi, \epsilon_n)$  codes exists

- for arbitrarily small  $\xi > 0$ ,
- for which  $\epsilon_n$  vanishes, and
- $\forall i \in \{1, 2, \dots, 2^{nr}\}$ ,  $\Pr(\hat{M} \neq i | M = i) \doteq e^{-nE_{\text{MD}}(r)}$ .

If there were only  $e^{nr}$  messages in the code (instead of  $e^{n(C-\xi)}$ ), their best missed detection exponent equals  $E(r)$ . This is the classical exponent defined in Eq. (1) earlier.

*Theorem 3:*  $E_{\text{MD}}(r) = E(r) \quad \forall r \in [0, C)$ .

Thus whatever  $E(r)$  is achievable for only  $e^{nr}$  messages, is also achievable when there are  $e^{n(C-\xi)} - e^{nr} \approx e^{nC}$  additional ordinary messages requiring reliable communication.

**Optimal strategy:** Start with an optimal code-book for  $e^{nr}$  messages which achieves error exponent  $E(r)$ . These codewords are used for the special messages. Now the ordinary codewords are added using random coding. The ordinary codewords which land close to a special codeword may be discarded without essentially any effect on the rate of

communication. At the decoder, a two-stage decoding rule is employed. The first stage decides that some special codeword was sent if at least one of the special codewords is within a threshold distance from the received sequence. Otherwise, the first stage decides that an ordinary codeword was sent. Depending on the first stage decision, the second stage ignores all codewords of one kind and applies ML decoding to the rest.

The overall missed detection exponent  $E_{\text{MD}}(r)$  is bottlenecked by the second stage errors. It is because the first stage error exponent equals the sphere-packing exponent  $E_{\text{sp}}(r)$ , which is never smaller than the second stage error exponent  $E(r)$ .

**Geometric Interpretation:** Thus we can communicate reliably at a rate near capacity and still protect the special messages as if we are only communicating the special messages. This means that we can start with a code of (say)  $2e^{nr}$  messages, where the decoding regions are large enough (see Fig. 3) to provide a missed detection exponent of  $E(r)$ . We empty out half of these  $2e^{nr}$  decoding regions and add  $e^{n(C-\xi)}$  ordinary codewords in this empty space.

Fig. 3 also shows how large decoding regions of the special codewords are interspersed within small decoding regions of ordinary codewords. This is analogous to filling sand particles in a box of large rocks. This theorem is like saying that the number of sand particles remains unaffected (exponentially) in spite of the large rocks in the box: channel capacity is achieved in spite of  $e^{nr}$  large decoding regions in the output space.

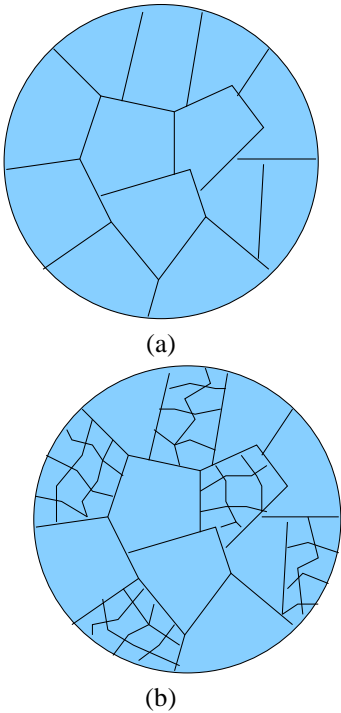


Fig. 3. (a) Original code (b) Modified code

#### D. Allowing erasures

In some situations, A decoder may be allowed declare an erasure when it is not sure about the transmitted message.

These erasure events are not counted as errors and are usually followed by a retransmission using a decision feedback protocol like Hybrid-ARQ.

*Definition 8:* A  $(n, R, \epsilon_n, e_n)$  erasure code denotes a length  $n$  code of rate  $R$ , which has  $e^{nR}$  messages, erasure probability  $\epsilon_n$ , and error probability  $\epsilon_n$ ,

$$e_n \triangleq \frac{1}{e^{nR}} \sum_{i=1}^{e^{nR}} \Pr(\text{erasure} | M = i)$$

and  $\epsilon_n = \Pr(\hat{M} \neq M)$  where,  $M, \hat{M} \in \{1, 2, \dots, e^{nR}\}$ .

Erasures are not counted as errors and an error event is defined as decoding to a wrong message without declaring an erasure. This event is sometimes called as an undetected error, but we will simply call it an error.

If the erasure probability is small, then average number of retransmissions needed is also small. Thus the effective rate of the decision feedback protocol remains essentially unchanged in spite of retransmissions when  $e_n$  tends to 0.

We could redefine all previous exponents by allowing erasures with small probabilities. This will add another condition in those definitions:  $e_n$  vanishes for large  $n$ . For example, the missed-detection exponent for  $e^{nr}$  special messages is defined as follows for erasure decoding.

*Definition 9:* For a given  $r < C$ , define  $E_{\text{MD}}^c(r)$  as the largest number such that a sequence of  $(n, C - \xi, \epsilon_n, e_n)$  erasure codes exists

- for arbitrarily small  $\xi > 0$ ,
- for which  $\epsilon_n$  and  $e_n$  vanish with  $n$ , and
- $\forall i \in \{1, 2, \dots, 2^{nr}\}$ ,  $\Pr(\hat{M} \neq i | M = i) \doteq e^{-nE_{\text{MD}}^c(r)}$ .

In all the problems discussed so far except one, this provision of erasures with vanishing probability does not improve the achievable exponents. This implies that decision feedback protocols such as Hybrid-ARQ cannot improve  $E_b$  and  $E_{\text{md}}$ . The only exception is the problem of  $e^{nr}$  special messages demanding a missed-detection exponent. Next theorem shows that compared to  $E_{\text{MD}}(r)$  in the no-erasure case, allowing erasures increases the missed-detection exponent at rates below critical rate.

*Theorem 4:*

$$E_{\text{MD}}^c(r) \geq E_{\text{sp}}(r) \quad \forall r \in [0, C).$$

For  $r = 0$ , where the number of special messages grows sub-exponentially, this lower bound is tight.

For  $r > 0$ , it is not clear (though unlikely) whether an exponent larger than  $E_{\text{sp}}(r)$  could be achieved. In contrast to the case without erasures, the obvious upper bound by ignoring the ordinary messages is not tight here.

**Optimal strategy:** It is similar to the no-erasure case. We first start with an erasure code in [7] for  $e^{nr}$  messages. Then add randomly generated ordinary codewords to it. Again a two-stage decoding is performed where the first stage decides between ordinary and special codewords using a threshold distance. If this first stage chooses special codewords, the second stage applies the decoding rule in [7] for choosing

a particular special codeword. Otherwise, the second stage chooses the ML ordinary codeword.

The overall missed detection exponent  $E_{\text{MD}}^e(r)$  is bottlenecked by the first stage errors. It is because the first-stage error exponent  $E_{\text{sp}}(r)$  is smaller than the second stage error exponent  $E_{\text{sp}}(r) + C - r$ . This is in contrast with the case without erasures.

### III. EFFECTS OF FULL FEEDBACK

Now we revisit the previous problems assuming perfect feedback at the transmitter: before transmitting each input, it knows all the past outputs. Feedback allows us to use variable time decoding schemes. Similar to Burnashev [15], we focus on block encoding schemes where transmission of a new message begins only after decoding of the old message is finished. Since the decoding time  $n$  could be a random variable now, let  $\bar{n}$  denote its average.

*Definition 10:* A  $(\bar{n}, R, \epsilon_{\bar{n}})$  feedback code denotes an encoding strategy which has  $e^{\bar{n}R}$  messages and error probability  $\epsilon_{\bar{n}}$ , where  $\bar{n}$  equals the average decoding delay assuming uniformly distributed messages.

$$\bar{n} \triangleq \frac{1}{e^{\bar{n}R}} \sum_{i=1}^{e^{\bar{n}R}} \mathbb{E}[n|M=i]$$

where  $\mathbb{E}[n|M=i]$  is average decoding time for message  $i$ .

#### A. Special bit

First consider the situation where the first bit  $b_0$  out of the  $\bar{n}R/\log 2$  bits is special. The error exponent for the special bit at capacity is defined as follows.

*Definition 11:*  $E_b^f$  is the largest number such that a sequence of  $(\bar{n}, R, \epsilon_{\bar{n}})$  feedback codes exists

- for arbitrarily small  $\xi > 0$ ,
- for which  $\epsilon_n$  and  $e_n$  vanish with  $n$ , and
- $\forall i \in \{1, 2, \dots, 2^{nr}\}$ ,  $\Pr(\hat{M} \neq 1|M=1) \doteq e^{-nE_{\text{MD}}^e(r)}$ .

for arbitrary  $\xi > 0$ , for which  $\epsilon_n$  vanishes, and  $\Pr(\hat{b}_0 = b_0) \doteq e^{-\bar{n}E_b^f}$ .

*Theorem 5:*  $E_b^f = \tilde{C}$ .

Recall that without feedback, the single bit could not achieve a positive error exponent near capacity. The following strategy shows how feedback connects message-wise UEP with bit-wise UEP: strategy for protecting a special message becomes useful for protecting special bits. This special message indicates incorrect decisions at the receiver.

**Optimal strategy:** We achieve this exponent using the missed detection exponent of  $\tilde{C}$  for a special message. This special message aims to notify the receiver when  $\hat{b}_0$  is incorrect. More specifically, first transmit  $b_0$  using a short repetition code of length  $\sqrt{\bar{n}}$ . If  $\hat{b}_0$  is correct after this repetition code, transmit the remaining bits with a capacity achieving code of length  $\bar{n} - \sqrt{\bar{n}}$ . If  $\hat{b}_0$  is incorrect after the repetition code, transmit a ‘buzzer’ codeword of length  $\bar{n} - \sqrt{\bar{n}}$ . For this buzzer, we use the same codeword that achieved the missed detection exponent  $E_{\text{md}} = \tilde{C}$ . It was repetition of  $x_p \in \mathcal{X}_p$ .

An erasure is declared (only) if the decoder detects the buzzer in the last  $\bar{n} - \sqrt{\bar{n}}$  symbols. Then the encoder retransmits by repeating the same strategy afresh. The erasure probability is vanishingly small, which ensures the effective rate of communication approaches capacity in spite of such retransmissions. Decoding error for  $b_0$  happens when the buzzer is not detected, which happens with the missed-detection exponent  $\tilde{C}$ .

#### B. Many special bits

Now consider the situation where many initial bits from the total  $\bar{n}R/\log 2$  bits are special. Let this initial string of special bits be denoted by  $\mathbf{b}$  and  $r$  be the rate of this string. For  $r > 0$ , the length of this string grows linearly in  $\bar{n}$  as  $\bar{n}r/\log 2$ . For  $r = 0$ , it grows sub-linearly in  $\bar{n}$  although it could still grow to infinity with  $\bar{n}$ , e.g., as  $\sqrt{\bar{n}}$ .

*Definition 12:* For a given  $r \in [0, C)$ , define  $E_{\text{bits}}^f(r)$  as the largest number such that a sequence of  $(\bar{n}, R, \epsilon_{\bar{n}})$  feedback codes exists

- for arbitrarily small  $\xi > 0$ ,
- for which  $\epsilon_{\bar{n}}$  vanishes for large  $\bar{n}$
- and  $\Pr(\hat{\mathbf{b}} \neq \mathbf{b}) \doteq \exp(-\bar{n}E_{\text{bits}}^f(r))$ .

The following theorem shows how this exponent decays linearly with rate  $r$  of the special bits.

*Theorem 6:* The following linear tradeoff can be achieved between the rate  $r$  of special bits and  $E_{\text{bits}}^f(r)$ .

$$E_{\text{bits}}^f(r) = \left(1 - \frac{r}{C}\right) \tilde{C}$$

Notice that for  $r = 0$ , the same exponent  $\tilde{C}$  as the single bit case in previous subsection could be achieved, although here the number of bits could be growing to infinity with  $\bar{n}$ . This linear tradeoff between rate and reliability reminds us of Burnashev’s result [15].

**Optimal strategy:** This is similar to the strategy for a single special bit. We first transmit  $\mathbf{b}$  using a capacity achieving code of length  $\approx \frac{\bar{n}r}{C}$ . If  $\mathbf{b}$  is decoded correctly, transmit the remaining bits with a capacity achieving code of length  $\approx \bar{n}(1 - \frac{r}{C})$ . Otherwise, transmit a ‘buzzer’ codeword of the same length. If the decoder detects this buzzer then an erasure is declared and the same strategy is repeated afresh. Error happens only if the buzzer is not detected, which happens with exponent  $E_{\text{MD}} = \tilde{C}(1 - \frac{r}{C})$ . The factor of  $(1 - \frac{r}{C})$  arises because the buzzer is only sent in that fraction of  $\bar{n}$ .

#### Multiple levels of speciality

We can generalize this result to the case of multiple levels of speciality. The most special layer of bits has rate  $r_1$ , the second most special layer of bits has rate  $r_2$  so on. The sum rate of all these layers equals  $C - \xi$ .

On similar lines of  $E_{\text{bits}}^f(\cdot)$  defined earlier, let  $E_{\text{bits},1}^f$  denote the error exponent of the most special layer (which we call as layer 1) for arbitrarily small  $\xi > 0$ ,  $E_{\text{bits},2}^f$  denote the same for the second most special layer (which we call layer 2) and so on. The least important layer cannot achieve any error exponent as  $\xi$  becomes small.

*Theorem 7:* The best error exponent for each layer  $i$  (except the least important) equals

$$E_{\text{bits},i}^f = \left(1 - \frac{\sum_{k=1}^i r_k}{C}\right) \tilde{C}$$

**Optimal strategy:** We first transmit the most important layer using a capacity achieving code of length  $\approx \bar{n}r_1/C$ . If it is decoded correctly, then transmit the next layer with a capacity achieving code of length  $\approx \bar{n}r_2/C$ . Otherwise, start the ‘buzzer’ sequence of input  $x_p$ . Repeat the same strategy for future layers too: press buzzer if wrong, next layer if right.

Once the block of  $\bar{n}$  symbols is received at the decoder, after decoding a layer (except the least important), it decides whether a buzzer was transmitted in remaining symbols. The next layer is decoded if no buzzer is detected and an erasure is declared otherwise. Thus layer after layer is decoded till a buzzer has been detected. Intuitively, this is similar to peeling an onion layer by layer and checking if it was rotten after peeling each layer.

Thus for each layer  $i$ , we can achieve the same exponent as if there were only two kinds of bits (as in Theorem 6): bits in layer  $i$  and more important layers  $k < i$  are special and bits in less important layers than layer  $i$  are ordinary. Hence this could be considered as a successively refinable version of Theorem 6.

The most important layer can achieve an exponent close to  $\tilde{C}$  if its rate is zero. As we move to across layers with decreasing importance, the achievable error exponent gradually decays in a linear manner.

### C. A special message

Now consider one particular message ( $M = 1$ ) which requires small missed-detection probability. Similar to the no-feedback case, define  $E_{\text{md}}^f$  as its missed-detection exponent near capacity.

*Definition 13:*  $E_{\text{md}}^f$  is the largest number such that a sequence of  $(\bar{n}, C - \xi, \epsilon_{\bar{n}})$  feedback-codes exists

- for arbitrarily  $\xi > 0$ ,
- for which  $\epsilon_{\bar{n}}$  vanishes, and
- $\Pr(\hat{M} \neq 1 | M = 1) \doteq \exp(-\bar{n}E_{\text{md}}^f)$ .

*Theorem 8:* Feedback does not improve the missed detection exponent for a single special message:  $E_{\text{md}}^f = E_{\text{md}} = \tilde{C}$ .

If pseudo-capacity was defined as the best protection of a special message (for data-rates near capacity), then this result could be thought of as an analog the “feedback does not increase capacity” for pseudo-capacity. Also note that with feedback,  $E_{\text{md}}^f$  for the special message and  $E_b^f$  for the special bit became equal.

### D. Many special messages

Now let us reconsider the problem where the first  $e^{\bar{n}r}$  messages are special. Henceforth, we will require that average decoding delay  $E[n|M=i]$  is equal across all messages—special and ordinary—and hence equals  $\bar{n}$ . This uniformity

constraint reflects a system requirement for ensuring a robust delay performance, which is invariant of the transmitted message<sup>3</sup>. Let us define the missed-detection exponent  $E_{\text{MD}}^f(r)$  under this uniform delay constraint.

*Definition 14:* For a given  $r < C$ , define  $E_{\text{MD}}^f(r)$  as the largest number such that a sequence of  $(\bar{n}, C - \xi, \epsilon_{\bar{n}})$  feedback codes exists

- for arbitrarily small capacity gap  $\xi > 0$ ,
- for which  $\epsilon_{\bar{n}}$  vanishes for large  $\bar{n}$ ,
- $E[n|M=j] = \bar{n} \quad \forall j$  (uniform delay constraint),
- $\forall i \in \{1, 2, \dots, e^{\bar{n}r}\}, \Pr(\hat{M} \neq i | M = i) \doteq e^{-\bar{n}E_{\text{MD}}^f(r)}$ .

*Theorem 9:* Let  $D_{\text{max}} \equiv \max_{x_1, x_2} D(W_{\mathbf{y}|\mathbf{x}=x_1} \| W_{\mathbf{y}|\mathbf{x}=x_2})$ ,

$$E_{\text{MD}}^f(r) = \min \left\{ \tilde{C}, (1 - r/C)D_{\text{max}} \right\}, \quad \forall r < C.$$

Thus  $E_{\text{MD}}^f(r)$  is the minimum of  $\tilde{C}$  and the Burnashev exponent at rate  $r$ . For  $r$  at which  $\tilde{C} \leq (1 - r/C)D_{\text{max}}$ , all  $e^{\bar{n}r}$  special messages achieve the best missed detection exponent  $\tilde{C}$  for a single special message. For larger  $r$  where  $\tilde{C} > (1 - r/C)D_{\text{max}}$ , the special messages achieve the Burnashev exponent as if the ordinary messages were absent.

The optimal strategy is based on transmitting a special bit first. It again shows how feedback connects bit-wise UEP with message-wise UEP: now however the strategy for protecting a special bit is used for protecting special messages. This is the exact opposite of the earlier strategy for achieving  $E_b^f$ .

**Optimal strategy:** We combine the strategy for achieving  $\tilde{C}$  for a special bit and the Yamamoto-Itoh strategy for achieving Burnashev exponent [16]. In the first phase, a special bit  $b_0$  is sent with a repetition code of  $\sqrt{\bar{n}}$  symbols. This is an indicator bit for special messages: it is 1 when a special message is to be sent and 0 otherwise. If it is decoded incorrectly as  $\hat{b}_0 = 0$ , then a missed-detection buzzer is sent for the remaining  $\bar{n} - \sqrt{\bar{n}}$  symbols. If it is decoded correctly as  $\hat{b}_0 = 0$ , then the ordinary codeword is sent using a capacity achieving code.

If it is decoded correctly as  $\hat{b}_0 = 1$ , then the particular special message is sent using the Yamamoto-Itoh scheme: transmit it at capacity using  $\approx \frac{\bar{n}r}{C}$  symbols and confirm the decoded  $\hat{M}$  in the last  $\approx \bar{n}(1 - \frac{r}{C})$  symbols. Declare an erasure when no confirmation.

## IV. AVOIDING FALSE ALARMS

### A. No-feedback case

We first consider the no-feedback case where false-alarm of a special message is a critical event, e.g., the “reboot” instruction. Now the false alarm probability  $\Pr(\hat{M} = 1 | M \neq 1)$  for this message should be minimized. By Baye’s rule and assuming uniformly chosen messages,

$$\begin{aligned} \Pr(\hat{M} = 1 | M \neq 1) &= \frac{\Pr(\hat{M} = 1, M \neq 1)}{\Pr(M \neq 1)} \\ &= \frac{\sum_{j \neq 1} \Pr(\hat{M} = 1 | M = j)}{(e^{nR} - 1)/e^{nR}} \end{aligned}$$

<sup>3</sup>Optimal exponents in all previous problems (remain unchanged irrespective of this uniformity constraint.

In classical error exponent analysis [2], the error probability for a given message usually means its missed detection probability. However, examples such as the ‘‘reboot’’ message necessitate this notion of false alarm probability.

*Definition 15:*  $E_{\text{fa}}$  is the largest number such that a sequence of  $(n, C - \xi, \epsilon_n)$  codes exists

- for arbitrarily small  $\xi > 0$ ,
- for which  $\epsilon_n$  vanishes for large  $n$
- and  $\Pr(\hat{M} = 1 | M \neq 1) \doteq e^{-nE_{\text{fa}}}$ .

*Theorem 10:* Let  $P_{\mathbf{x}}^*$  denote the capacity achieving input distribution,

$$E_{\text{fa}} = \max_{x_f \in \mathcal{X}} D(W_{\mathbf{y}|\mathbf{x}=x_f} \| W_{\mathbf{y}|\mathbf{x}} | P_{\mathbf{x}}^*) \quad (3)$$

$$\triangleq \max_{x_f \in \mathcal{X}} \left( \sum_x P_{\mathbf{x}}^*(x) D(W_{\mathbf{y}|\mathbf{x}=x_f} \| W_{\mathbf{y}|\mathbf{x}=x}) \right) \quad (4)$$

where  $D(\cdot \| \cdot | P)$  denotes conditional expectation of the KL divergence given  $\mathbf{x}$ , which is distributed as  $P$ . Define  $\mathcal{X}_f$  as the set of inputs achieving the maximum above.

This false-alarm exponent is larger than channel capacity  $C$  due to convexity of KL divergence.

$$E_{\text{fa}} \geq \max_{x_f} D(W_{\mathbf{y}|\mathbf{x}=x_f} \| \sum_x P_{\mathbf{x}}^*(x) W_{\mathbf{y}|\mathbf{x}=x}) \quad (5)$$

$$= \max_{x_f} D(W_{\mathbf{y}|\mathbf{x}=x_f} \| P_{\mathbf{y}}^*) = C \quad (6)$$

where  $P_{\mathbf{y}}^*$  denotes the capacity achieving output distribution. As discussed before, the last equality follows from KKT condition for achieving capacity [4].

Now we can compare this result for a special message with the similar result for classical situation where all messages are treated equally. It turns out that if every message in a capacity-achieving code demands equally good false-alarm exponent, then this uniform exponent cannot be larger than  $C$ . Reducing the demand of false-alarm exponent to only one message instead of all thus enhances it to  $E_{\text{fa}}$ .

**Optimal strategy:** Codeword for the special message  $M = 1$  is a repetition sequence of an input  $x_f \in \mathcal{X}_f$ . Its decoding region  $S$  is the typical ‘noise ball’ [8] around it. This noise ball consists of (only) the output sequences of type  $W_{\mathbf{y}|\mathbf{x}=x_f}$ . For the ordinary messages, we again use a capacity achieving code-book. The decoder chooses the ML ordinary codeword for output sequences outside  $S$ .

Note the difference between this strategy for achieving  $E_{\text{fa}}$  and the optimal strategy for achieving  $E_{\text{md}}$ . For achieving  $E_{\text{md}}$ , output sequences of any type other than  $P_{\mathbf{y}}^*$  were assigned to  $S$ .

For a symmetric channel like BSC, where divergence  $D(W_{\mathbf{y}|\mathbf{x}=x_1} \| W_{\mathbf{y}|\mathbf{x}=x_2})$  between any two distinct inputs is the same (say  $D$ ), every input can be used as  $x_f$ . In this case, any length  $n$  input sequence could be used as the special codeword. Since  $P_{\mathbf{x}}^*$  for symmetric channels is the uniform distribution,  $E_{\text{fa}} = \left(1 - \frac{1}{|\mathcal{X}|}\right) D$ .

**Remark:** This result seems to be directly connected with the problem of identification via channels [14]. We can prove the achievability part of their capacity theorem using an extension

of the achievability part of  $E_{\text{fa}}$ . Perhaps a new converse of their result is also possible using such results.

**Geometric Interpretation:** A false alarm exponent for the special message corresponds to having the smallest possible decoding region  $S$  for the special message. This ensures that when some ordinary message is transmitted, probability of landing in  $S$  is exponentially small. We cannot make it too small though, because when the special message is transmitted, the probability of landing outside it should be small. Hence it should at least contain the typical noise ball around the special codeword. The blue region in Fig. 4 denotes such a region.

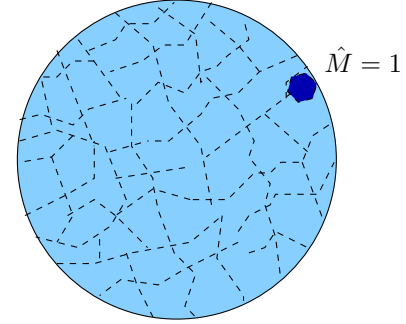


Fig. 4. Avoiding false-alarm

## B. Full feedback

Recall that feedback did not improve the missed-detection exponent for a special message. On the contrary, we will see that the false-alarm exponent for a special message can be improved with full feedback. We again restrict to uniform delay codes where average delay  $E[n | M = j]$  is equal to  $\bar{n}$  for every message.

*Definition 16:*  $E_{\text{fa}}^f$  is the largest number such that a sequence of  $(\bar{n}, C - \xi, \epsilon_{\bar{n}})$  feedback codes exists

- for arbitrarily small  $\xi > 0$ ,
- for which  $\epsilon_{\bar{n}}$  vanishes for large  $\bar{n}$ ,
- $E[n | M = j] = \bar{n} \quad \forall j$  (uniform delay constraint),
- and  $\Pr(\hat{M} = 1 | M \neq 1) \doteq \exp(-\bar{n}E_{\text{fa}}^f)$ .

*Theorem 11:*  $E_{\text{fa}}^f = D_{\text{max}}$ .

Thus feedback improves the false alarm exponent from  $E_{\text{fa}}$  to  $D_{\text{max}}$ , because  $D_{\text{max}}$  is always larger than  $E_{\text{fa}}$ .

**Optimal strategy:** We use the same strategy in subsection III-D that achieved  $E_{\text{MD}}^f(r)$ . First phase of  $\sqrt{\bar{n}}$  symbols sends a special bit  $b_0$ , which is an indicator bit of the special message. Second phase of length  $\bar{n} - \sqrt{\bar{n}}$  uses the Yamamoto-Itoh strategy for confirming or declining when  $\hat{b}_0 = 1$ . This strategy simultaneously achieves the optimal missed-detection exponent  $\tilde{C}$  and the optimal false-alarm exponent  $D_{\text{max}}$  for this special message.

## V. FUTURE DIRECTIONS

This framework provides a large variety of fundamental problems to be studied. For example, many fundamental

limits of bit-wise and message-wise UEP need to be understood for data-rates below capacity. In addition to theoretical understanding, constructing efficient coding mechanisms for achieving these tradeoffs is also crucial. One aspect of this task is designing LDPC-like and algebraic codes, which provide extra protection to the high priority information with small computational complexity. Another aspect is addressing the effects of some practical alternatives to classical decoding, *e.g.*, list decoding.

Information networks, such as, two-way channels, broadcast and relay channels provide another rich dimension for future research. Information theoretic understanding of such networks also provides a set of optimization problems to be studied: the achievable resources of reliability and rate need to be efficiently divided between multiple information layer ranging from the least special to the most special.

#### ACKNOWLEDGMENT

Shashi Borade is indebted to Bob Gallager for his insights and encouragement for this work in general. In particular, Theorem 3 was mainly inspired from his remarks. Helpful discussions with David Forney and Emre Telatar are also gratefully acknowledged.

#### REFERENCES

- [1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379-423 and 623-656, July and October, 1948.
- [2] R. Gallager, *Info. Theory and Reliable Communication*, Wiley, 1968.
- [3] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower Bounds to Error Probability for Coding on Discrete Memoryless Channels", *Information and Control*, pp. 65-103, December 1966.
- [4] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1981.
- [5] D. Forney, "On exponential error bounds for random codes on the BSC," unpublished manuscript.
- [6] A. Montanari and D. Forney, "On exponential error bounds for random codes on the DMC," unpublished manuscript.
- [7] D. Forney, "Exponential error bounds for erasure, list, and decision feedback schemes," *IEEE Trans. Info. Theory*, vol. 14, no. 2, pp. 206-220, Mar. 1968.
- [8] T. Cover, *Elements of Information Theory*, Wiley-Interscience, 1991.
- [9] S. Borade, B. Nakibolgu, L. Zheng "Fundamental limits of UEP: data-rates below capacity," in preparation.
- [10] S. Borade and L. Zheng, "Euclidean information theory," Allerton Conference, Monticello, Sept. 2007.
- [11] A. Albanese, J. Blomer, J. Edmonds, M. Luby, and M. Sudan, Priority encoding transmission, *IEEE Trans. Inform. Theory*, vol. 42, pp. 1737-1744, Nov. 1996.
- [12] A. R. Calderbank, N. Seshadri Multilevel codes for unequal error protection, *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1234-1248, July 1993.
- [13] S. Diggavi, D. Tse, "On successive refinement of diversity," Allerton Conference, Illinois, September 2004.
- [14] R. Ahlswede, G. Dueck Identification via channels, *IEEE Trans. Inform. Theory*, vol. 35, no. 1, pp. 15-29, Jan. 1989.
- [15] M. Burnashev, "Data transmission over a discrete channel with feedback, random trans. time", *Problems Info. Trans.*, vol. 12, pp. 10-30, 1976.
- [16] H. Yamamoto and K. Itoh, Asymptotic performance of a modified Schalkwijk-Barron scheme for channels with noiseless feedback, *IEEE Trans. Info. Theory*, vol. 25, pp. 729733, November 1979.