[2] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 1, pp. 1–10, Jan. 1976.
[3] A. D. Wyner, "The rate-distortion function for source coding with side information at the decoder-ii: General sources," *Inf. Contr.*, vol. 38, no. 1, pp. 60–80, Jul. 1978.
[4] Y. Steinberg and N. Merhav, "On successive refinement for the Wyner–Ziv problem," *IEEE Trans. Inf. Theory*, vol. 50, no. 8, pp. 1636–1654, Aug. 2004.
[5] C. Heegard and T. Berger, "Rate distortion when side information may be absent," *IEEE Trans. Inf. Theory*, vol. IT-31, no. 6, pp. 727–734, Nov. 1985.
[6] A. H. Kaspi, "Rate-distortion function when side-information may be present at the decoder," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 2031–2034, Nov. 1994.
[7] R. Zamir, "The rate loss in the Wyner–Ziv problem," *IEEE Trans. Inf. Theory*, vol. 42, no. 6, pp. 2073–2084, Nov. 1996.
[8] L. Lastras and T. Berger, "All sources are nearly successively refinable," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 918–926, Mar. 2001.
[9] H. Feng and M. Effros, "Improved bounds for the rate loss of multiresolution source codes," *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 809–821, Apr. 2003.
[10] ——, "On the rate loss of multiple description source codes," *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 671–683, Feb. 2005.
[11] S. Shamai (Shitz), S. Verdú, and R. Zamir, "Systematic lossy source/channel coding," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 564–579, Mar. 1998.

On the Use of Training Sequences for Channel Estimation

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

***Abstract*—Suppose $\mathcal{Q}$ is a family of discrete memoryless channels. An unknown member of $\mathcal{Q}$ will be available, with perfect, causal output feedback for communication. We study a scenario where communication is carried by first testing the channel by means of a training sequence, then coding according to the channel estimate. We provide an upper bound on the maximum achievable error exponent of any such coding scheme. If we consider the Binary Symmetric and the Z families of channels this bound is much lower than Burnashev's exponent. For example, in the case of Binary Symmetric Channels this bound has a slope that vanishes at capacity. This is to be compared with our previous result that demonstrates the existence of coding schemes that achieve Burnashev's exponent (that has a nonzero slope at capacity) even though the channel is revealed neither to the transmitter nor to the receiver. Hence, the present result suggests that, in terms of error exponent, a good universal feedback scheme entangles channel estimation with information delivery, rather than separating them.**

***Index Terms*—Channel estimation, error exponent, feedback communication, training sequence, universal channel coding.**

Manuscript received March 24, 2005; revised August 17, 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, CH-1015 Lausanne, Switzerland.

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, School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland (e-mail: emre.telatar@epfl.ch).

Communicated by K. Kobayashi, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2005.864468

I. Introduction and Preliminaries

When considering information transmission over a channel that is partially known to either the transmitter or the receiver or both, it is common to employ a *training sequence*. This sequence is sent prior to the data to be conveyed and its purpose is to help the decoder (for channels without feedback) or both the encoder and the decoder (for channels with feedback) to adjust its/their parameters for the upcoming communication. For example, in slow fading channels without feedback, a training sequence can be sent at the beginning of each coherence interval, so that the receiver can estimate the channel characteristics (see, e.g., [3], [7], [10]).

Here we study feedback communication over a time invariant discrete memoryless channel (DMC) with perfect feedback, i.e., noiseless and instantaneous (causal) feedback. We assume that the transmitter and the receiver are not aware of the transition probability matrix Q of the channel, however, both know that Q belongs to some subset $\mathcal{Q}$ of DMCs.

In principle, the sending of a training sequence before the information need not affect the rates achievable by the communication system: the test sequence length can be made negligible compared to the length of the subsequent information sequence. However, and this is the main concern of this paper, the separation of the channel estimation from the information coding may result in a penalty in terms of error exponent.

In the case without feedback Feder and Lapidoth [5] show that, if a family of channels satisfy certain conditions, there exist universal decoders that are optimal in the sense that they perform (asymptotically) as well as the maximum-likelihood decoder tuned for the channel over which transmission is carried out. They also show that the combination of a training sequence and a maximum-likelihood decoder designed for the estimated channel is not optimal. The result presented in this paper, while concerning feedback channels, has the same flavor.

We end this section by reminding some definitions related to feedback communication and state an important result due to Burnashev that gives the maximum error exponent that can be achieved over a DMC with perfect feedback. In Section II we present our result and illustrate it with two examples, and in Section III we prove our result.

Definition 1 (Coding Scheme): Given a channel Q with input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$, and a message set $\mathcal{M}$ of size $M \geq 1$, an encoder (or codebook) is a sequence of functions

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

The symbol to be sent at time n is obtained by evaluating f_n for the message and the feedback sequence received until that time, i.e., $f_n(m, y^{n-1})$. A codeword for message m is the sequence of functions $\{f_n(m, \cdot)\}_{n \geq 1}$. A decoder (ϕ, T) is a sequence of functions

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

together with a stopping time T relative to the received symbols $Y_1, Y_2, \ldots$.[1] The decoded message is $\phi_T(y^T)$. A coding scheme is a tuple $c = (f, \phi, T)$.

In the sequel we will be concerned with sequences of coding schemes indexed by the message set size M. A sequence of coding schemes S is a sequence $\{c^M\}_{M \geq 1}$ where $c^M \triangleq (f^M, \phi^M, T^M)$.

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

Definition 2 (Rate): Given a channel Q, an integer $M \geq 1$, and a coding scheme $c = (f, \phi, T)$, the transmission rate is[2]

$$R(c, Q) \triangleq \frac{\ln M}{\mathbb{E}_Q T} \tag{3}$$

where $\mathbb{E}_Q T$ denotes the expected decision time, under the channel Q, over uniformly chosen messages, i.e.,

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

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

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

whenever the limit exists.

The error event is denoted by $\mathcal{E}$ and the average (over uniformly chosen messages) error probability given a coding scheme c and a channel Q is defined as

$$\mathbb{P}_Q(\mathcal{E}|c) \triangleq \tag{6}$$

$$\frac{1}{M} \sum_{m \in \mathcal{M}} \mathbb{P}_Q(\phi_T(Y^T) \neq m \,|\, \text{message } m \text{ is sent}). \tag{7}$$

Definition 3 (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

$$E(\mathtt{S}, Q) \triangleq \liminf_{M \to \infty} -\frac{1}{\mathbb{E}_Q T^M} \ln \mathbb{P}_Q(\mathcal{E} \,|\, c^M). \tag{8}$$

Theorem (Burnashev 1976): Let Q be a DMC with input and output alphabet $\mathcal{X}$ and $\mathcal{Y}$, and with capacity $C(Q)$. Let R be any constant in $[0, C(Q)]$. There exists $\mathtt{S} = \{c^M\}_{M \geq 1}$ such that $R(\mathtt{S}, Q) = R$ and

$$E(S, Q) = E_B(R, Q) \tag{9}$$

where

$$E_B(R, Q) \triangleq \max_{(x, x') \in \mathcal{X} \times \mathcal{X}} D(Q(\cdot \,|\, x) \,\|\, Q(\cdot \,|\, x')) \left(1 - \frac{R}{C(Q)}\right) \tag{10}$$

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'.[3] Further, for any $\mathtt{S}$ such that $R(\mathtt{S}, Q) = R$,

$$\limsup_{M \to \infty} -\frac{1}{\mathbb{E}T^M} \ln \mathbb{P}_Q(\mathcal{E} \,|\, c^M) \leq E_B(R, Q). \tag{11}$$

From now on $E_B(R, Q)$ will be referred as the Burnashev's exponent.

II. Statement of the Result

Let $\mathcal{Q}$ be a set of DMCs. We suppose that communication is carried out over a channel $Q \in \mathcal{Q}$ that is revealed neither to the transmitter nor to the receiver. The coding schemes we shall focus on are referred as "training based schemes" and admit two phases: a first phase of fixed length t, the "training period" (or "test period") during which the channel parameter is estimated and no information is conveyed, and, a second phase used to carry information. The training policy may depend on feedback.[4] We require training based schemes to satisfy the following asymptotic properties:

A sequence of training based schemes $\{c^M = (f^M, \phi^M, T^M)\}_{M \geq 1}$ is such that:

I) For each $M \geq 1$ there exists a "rate function" $n_t : \mathcal{Y}^t \to \mathbb{R}_+$ that associates to each output y^t obtained during the training period an approximate (average) length of the second phase in the sense that, for all $Q \in \mathcal{Q}$,

$$\mathbb{E}_Q(T^M | Y^t = y^t) = (t + n_t(y^t))(1 + o(1))$$

as $t \to \infty$.

II) There exists $\gamma \in (0, 1)$ such that, for any $a > 0$ and $Q \in \mathcal{Q}$[5]

$$\lim_{M \to \infty} \mathbb{P}_Q\left(\left|\frac{\ln M}{T^M} - \gamma C(Q)\right| > a\right) = 0.$$

III) There exists $b < \infty$ such that, for all $M \geq 1$

$$T^M \leq b \ln M.$$

A few comments are in order. Motivated by communications scenarios used in practice, we sought for a definition that captures the fact that training based schemes separate the channel estimation from the information transmission. To that aim we impose the condition I that requires to employ for the second phase a coding scheme whose rate essentially depends upon the output sequence y^t obtained during the training period. This rate is approximately equal to $\ln M / n_t(y^t)$. In particular one cannot use, as a second phase, a coding scheme that would fully adapt its rate on the run according to the channel under use, implicitly estimating the channel (see, e.g., [8], [9]). However note that, for a given y^t, two different channels may have a slight difference in the expected length of the second phase. Hence, variable length codes can be used for the second phase provided that, once the training period is over, the average decoding time is (approximately) set. Observe that the requirement I imposes a restriction neither on the channel estimation itself nor on the decision that results from it.

We introduce condition II in order to have some control on the rate through the "normalized rate" γ. This condition allows us to compute a bound on the maximum error exponent that can be achieved by any training based schemes operating at a given rate. Condition II may be satisfied, for example, if for the second phase one uses a fixed length block code together with the maximum-likelihood decoder both tuned for the empirical channel that results from the training period. Finally the restriction III is a mild technical requirement if $\inf_{Q \in \mathcal{Q}} C(Q) > 0$.

Our result stands in the following theorem.

Theorem: Let $\mathcal{Q}$ be a family of DMC's with same input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$, and such that $\inf_{Q \in \mathcal{Q}} C(Q) > 0$. Let $\{c^M\}_{M \geq 1}$ be a sequence of training-based schemes for $\mathcal{Q}$, and with parameter $\gamma \in (0, 1)$. For any $Q \in \mathcal{Q}$

$$\limsup_{M \to \infty} -\frac{1}{\mathbb{E}_Q T^M} \ln \mathbb{P}_Q(\mathcal{E}|c^M) \leq E_{\text{tbs}}(\gamma, Q) \tag{12}$$

where

$$E_{\text{tbs}}(\gamma, Q) \triangleq C(Q) \min_{V \in \mathcal{Q}} \frac{1}{C(V)} \times \max\left\{\max_{x \in \mathcal{X}} D(V(\cdot \,|\, x) \,\|\, Q(\cdot \,|\, x)), E_B(\gamma C(V), Q)\right\}. \tag{13}$$

A. Example: Binary Symmetric Channels

For $L \in [0, 1/2)$ let $\mathcal{Q} = \text{BSC}_L$ where BSC_L is the set of binary symmetric channels (BSCs) with crossover probability $\varepsilon \in [0, L]$. For simplicity let ε denote both the crossover probability and the BSC with this crossover probability. The function (13) reduces to[6]

$$E_{\text{tbs}}(\gamma, \varepsilon) = C(\varepsilon) \min_{\delta \in [0, L]} \frac{1}{C(\delta)} \max\{D(\delta \,\|\, \varepsilon), E_B(\gamma C(\delta), \varepsilon)\} \tag{14}$$

and in Section III we show that, for all $\varepsilon \in (0, L]$

$$\lim_{\gamma \uparrow 1} \frac{E_{\text{tbs}}(\gamma, \varepsilon)}{1 - \gamma} = 0. \tag{15}$$

[2] $\ln$ denotes the logarithm to the base e.

[3] We define $E_B(R, Q) = 0$ for $R \geq C(Q)$.

[4] In other words we do not assume the training sequence to be set prior to communication.

[5] Recall that $C(Q)$ denotes the capacity of the channel Q.

[6] $D(\delta \| \varepsilon)$ denotes $\delta \ln(\delta/\varepsilon) + (1 - \delta) \ln((1 - \delta)/(1 - \varepsilon))$.

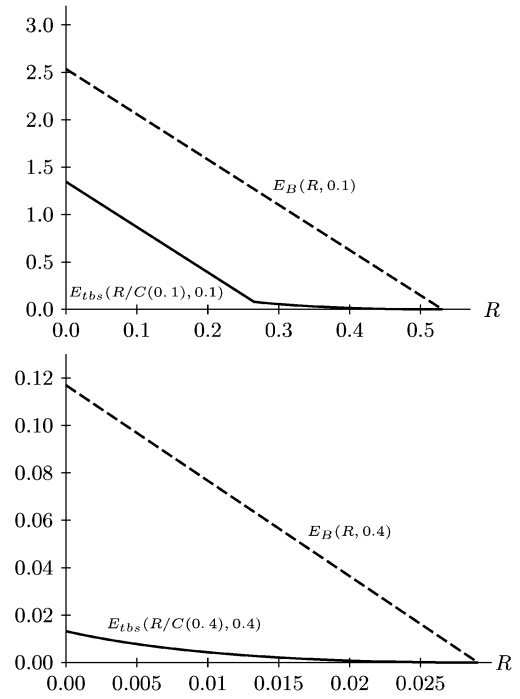

Fig. 1. Upper bound on the error exponent of training based schemes (lower curve) and Burnashev's error exponent (dashed line) for the BSCs with crossover probabilities 0.1 and 0.4.

In Fig. 1 we plot for two channels ($\varepsilon = 0.1$ and $\varepsilon = 0.4$) the function $R \mapsto E_{\rm tbs}(R/C(\varepsilon), \varepsilon)$ (lower curve) and Burnashev's exponent given by (10) (upper line).

In order to discuss the above result, let us first briefly refer to earlier results obtained in [9] for BSCs. Theorem 1 [9] claims that, given any $\gamma \in [0, 1)$ and the BSC_L family, there exist coding schemes that achieve Burnashev's exponent, simultaneously, on every channel $\varepsilon \in \mathrm{BSC}_L$, at a rate at least equal to $\gamma C(\varepsilon)$ and strictly less than $C(\varepsilon)$. Suppose now one is interested in having a low error probability instead of a high communication rate. Similarly, there exist coding schemes that universally achieve a rate that is guaranteed to be now at most γ times the capacity of the channel and with a corresponding error exponent that is also maximum.

In contrast with these results, training based schemes cannot achieve Burnashev's exponent for BSCs. While feedback does not increase capacity Burnashev's result tells us that feedback is of particular help at rates close to capacity: a little drop in the rate results in a linear augmentation of the error exponent. Training based schemes fail precisely in having this property: the slope of their error exponent equals to zero at capacity. Hence, at high rates, the situation becomes essentially the same as if no feedback were available and the channel were revealed to both the transmitter and the receiver (since in this case the maximum achievable error exponent is the sphere packing bound, for rates above the critical rate [6]).

Note however that, for BSCs, the comparison between training based schemes and the optimal coding schemes derived in [9] is not completely fair. For training based schemes we require an exact control on the rate through the parameter γ, whereas in [9] the parameter γ yields only an upper or a lower bound on the rate. Nevertheless the comparison is fair at small rates, in which case there is a significant difference between the error exponent of the optimal coding schemes and training based schemes (see, e.g., Fig. 1).

B. Example: Z Channels

For $L \in [0, 1)$ let $\mathcal{Q} = \mathrm{Z}_L$ where Z_L denotes the set of Z channels with crossover probabilities $\varepsilon \in [0, L]$. Pick a particular channel $Q \in \mathrm{Z}_L$ with nonzero crossover probability. One can find a $\gamma \in (0, 1)$ sufficiently close to 1 as well as a channel $W \in \mathrm{Z}_L$ such that $\gamma C(W) > C(Q)$. Therefore, we have

$$\begin{aligned} E_{\rm tbs}(\gamma, Q) &\triangleq C(Q) \min_{V \in \mathcal{Q}} \frac{1}{C(V)} \\ &\qquad \times \max\left\{ \max_{x \in \mathcal{X}} D(V(\cdot \mid x) \,\|\, Q(\cdot \mid x)), E_B(\gamma C(V), Q) \right\} \\ &\le \frac{C(Q)}{C(W)} \max\left\{ \max_{x \in \mathcal{X}} D(W(\cdot \mid x) \,\|\, Q(\cdot \mid x)), E_B(\gamma C(W), Q) \right\} \\ &= \frac{C(Q)}{C(W)} \max_{x \in \mathcal{X}} D(W(\cdot \mid x) \,\|\, Q(\cdot \mid x)) \\ &< \infty. \end{aligned} \tag{16}$$

The second equality holds since Burnashev's exponent equals to zero above capacity, and the last inequality holds since Q has a nonzero crossover probability. Hence, training based schemes for the Z_L family have a finite error exponent for any $Q \in \mathrm{Z}_L$ with nonzero crossover probability, and for γ sufficiently close to 1. This is in contrast with a result obtained in [9]. Theorem 2 [9] claims that, given the Z_L family and any $\gamma \in [0, 1)$, there exist coding schemes that, universally over Z_L, achieve a rate equal to $\gamma C(Q)$ and a corresponding error exponent equal to Burnashev's, in this case infinite.

Finally, one can easily show that, for the family of Erasure channels with erasure probability $\varepsilon \in [0, L]$ (with $L \in [0, 1)$), a similar result as for the Z family holds: training based schemes yield a finite error exponent on any channel with a nonzero erasure probability and γ sufficiently close to 1. This is in contrast with the "sent until a non-erasure occurs" strategy for a 1-bit message [6, p. 506, Problem 2.10]. This universal strategy is error-free (hence has an infinite error exponent) on any channel with erasure probability different from 1.

III. Analysis

Proof of the Theorem: We will first prove the theorem for the case where $\mathcal{Q} = \mathrm{BSC}_L$. The general case goes along the same main lines. For simplicity, as in the BSC example of Section II-A, let ε denote both the crossover probability and the BSC with this crossover probability.

We start with a short description of the main idea of the proof. We first show that the rate function of training based schemes has to "strongly" rely on the empirical channel $\hat{Q}_{y^t|x^t}$ that results from the training period. More precisely, the length of the second phase has to be approximately equal to $\frac{\ln M}{\gamma C(\hat{Q}_{y^t|x^t})} - t$. Due to this fact a large probability of error occurs because of atypical behavior of the channel during the training period.

Suppose the underlying channel is ε and let S denote the event that this channel behaves as a BSC with crossover probability δ. We lower bound the error probability of a training based scheme c as

$$\begin{aligned} \mathbb{P}_\varepsilon(\mathcal{E} \mid c) &\ge \mathbb{P}_\varepsilon(\mathcal{E} \cap S \mid c) \\ &= \mathbb{P}_\varepsilon(\mathcal{E} \mid S, c) \mathbb{P}_\varepsilon(S \mid c). \end{aligned} \tag{17}$$

By a principle of large deviations we have

$$\mathbb{P}_\varepsilon(S \mid c) \approx e^{-tD(\delta \,\|\, \varepsilon)}. \tag{18}$$

Now, conditioned on the event that the channel behaves like δ, the average length of the second phase is approximately equal to $\frac{\ln M}{\gamma C(\delta)} - t$. Since Burnashev's exponent yields a lower bound to the error probability we have

$$\mathbb{P}_\varepsilon(\mathcal{E} \mid S, c) \gtrapprox e^{-\left(\frac{\ln M}{\gamma C(\delta)} - t\right) E_B\left(\frac{\ln M}{\frac{\ln M}{\gamma C(\delta)} - t}, \varepsilon\right)}. \tag{19}$$

From the requirement II and III one deduces that $\mathbb{E}_\varepsilon T \approx \ln M / \gamma C(\varepsilon)$. Hence, using (17), (18), and (19) one gets the desired result by optimizing the fraction of the communication time dedicated to the training and noting that δ is arbitrary in $[0, L]$.

We now turn to the proof. Let $\mathtt{S} = \{c^M\}_{M\geq 1}$ be a sequence of training based schemes with parameter $\gamma \in (0,1)$. Let $t = t(\gamma, L, M)$ denote the length of the training period of c^M. Without loss of generality we make the following assumptions:

- the training sequence is the all-zero sequence;
- $t(\gamma, L, M)$ tends to infinity as M tends to infinity.

That the first assumption is without loss of generality is clear. For the second assumption, consider a training sequence of length t with a particular rate function. On the one hand, to any longer training sequence one can associate the same rate function that only depends on the results of the first t output symbols. On the other hand, by letting $t(\gamma, L, M)$ grow sub-logarithmically in M one can render the contribution of the testing part to the overall rate equal to zero in the limit $M \to \infty$. Therefore assuming the training sequence length to grow with M has asymptotically no effect on the rate and error probability, thus also no effect on the reliability function, which justifies the second assumption.

Assume that communication is carried out over some channel ε with $\varepsilon \in [0, L]$. Pick some $\delta \in [0, L]$ and let

$$S = S(a,\delta,t) \triangleq \left\{ y^t \in \{0,t\}^t : \mathbb{P}_\delta\left(T^M > \frac{\ln M}{\gamma C(\delta) - a} \middle| Y^t = y^t\right) \leq a \right\}. \quad (20)$$

For the moment the parameter a is chosen such that $0 < a \ll \gamma C(\delta)$. We have

$$\begin{aligned}\mathbb{P}_\varepsilon(\mathcal{E} \cap \{Y^t \in S\} \,|\, c^M) &= \mathbb{P}_\varepsilon(Y^t \in S \,|\, c^M)\mathbb{P}_\varepsilon(\mathcal{E} \,|\, Y^t \in S, c^M) \\ &= \mathbb{P}_\varepsilon(Y^t \in S)\mathbb{P}_\varepsilon(\mathcal{E} \,|\, Y^t \in S, c^M) \end{aligned} \quad (21)$$

where the last equality holds since, during the test period, the same symbol ("0") is sent irrespectively of the channel output. We will now derive lower bounds on

$$\mathbb{P}_\varepsilon(Y^t \in S) \quad (22)$$

and

$$\mathbb{P}_\varepsilon(\mathcal{E} \,|\, Y^t \in S, c^M) \quad (23)$$

and combine these bounds to prove the theorem for the BSC case.

We write $f(x) = o_{x,0}(1)$ if $|f(x)| \overset{x\to 0}{\to} 0$ and $f(x) = o_{x,\infty}(1)$ if $|f(x)| \overset{x\to\infty}{\to} 0$.[7]

From the requirement II, $\mathbb{P}_\delta(Y^t \in S) \geq 1 - a$ for M large enough (remember that t grows with M). Now an event of high probability under measure $\mathbb{P}_\delta$ cannot have too small a probability under $\mathbb{P}_\varepsilon$. More precisely, the data processing inequality for divergence[8] yields, for M large enough

$$\mathbb{P}_\varepsilon(Y^t \in S) \geq e^{-tD(\delta\,\|\,\varepsilon)(1+o_{a,0}(1))(1+o_{t,\infty}(1))}. \quad (26)$$

In order to compute a lower bound on $\mathbb{P}_\varepsilon(\mathcal{E}|Y^t \in S, c^M)$, let us first derive an upper bound on $\mathbb{E}_\varepsilon(T^M|Y^t \in S)$. From the definition of set S and the requirement III satisfied by training-based schemes we deduce that, for all $y^t \in S$

$$\mathbb{E}_\delta(T^M \,|\, Y^t = y^t) \leq \frac{\ln M}{\gamma C(\delta)}(1 + o_{a,0}(1)). \quad (27)$$

Combining (27) with the requirement I we get, for all $y^t \in S$

$$\mathbb{E}_\varepsilon(T^M \,|\, Y^t = y^t) \leq \frac{\ln M}{\gamma C(\delta)}(1 + o_{a,0}(1))(1 + o_{t,\infty}(1)) \quad (28)$$

hence,

$$\mathbb{E}_\varepsilon(T^M \,|\, Y^t \in S) \leq \frac{\ln M}{\gamma C(\delta)}(1 + o_{a,0}(1))(1 + o_{t,\infty}(1)). \quad (29)$$

Now, given a certain expected communication delay, Burnashev's exponent yields (asymptotically) a lower bound to the error probability. Therefore, from (29) we get

$$\mathbb{P}_\varepsilon(\mathcal{E} \,|\, Y^t \in S, c^M) \geq e^{-n\left(E_B\left(\frac{\ln M}{n},\varepsilon\right)+o_{M,\infty}(1)\right)} \quad (30)$$

where

$$n \triangleq \frac{\ln M}{\gamma C(\delta)}(1 + o_{a,0}(1))(1 + o_{t,\infty}(1)) - t.$$

Combining (21), (26), and (30) one obtains[9]

$$\begin{aligned} &-\ln \mathbb{P}_\varepsilon(\mathcal{E}|c^M) \\ &\leq tD(\delta\,\|\,\varepsilon) + \left(\frac{\ln M}{\gamma C(\delta)} - t\right) E_B\left(\frac{\ln M}{\frac{\ln M}{\gamma C(\delta)} - t}, \varepsilon\right) \\ &\quad + o_{a,0}(1)\mathcal{O}_{M,\infty}(\ln M). \end{aligned} \quad (31)$$

From the requirement II and III we have, for M large enough

$$\mathbb{E}_\varepsilon T^M = \frac{\ln M}{\gamma C(\varepsilon)}(1 + o_{a,0}(1)) \quad (32)$$

hence, from (31)

$$\begin{aligned} &-\frac{1}{\mathbb{E}_\varepsilon T^M}\ln \mathbb{P}_\varepsilon(\mathcal{E} \,|\, c^M) \\ &\leq \frac{C(\varepsilon)\left[(1-\alpha_M)D(\delta\,\|\,\varepsilon) + \alpha_M E_B\left(\frac{\gamma C(\delta)}{\alpha_M}, \varepsilon\right)\right]}{C(\delta)} \\ &\quad + o_{a,0}(1) + o_{M,\infty}(1) \end{aligned} \quad (33)$$

where

$$\alpha_M = \alpha_M(\gamma,\delta) \triangleq \frac{\frac{\ln M}{\gamma C(\delta)} - t}{\frac{\ln M}{\gamma C(\delta)}}.$$

Inequality (33) holds for any $a > 0$ and M large enough. Therefore by first taking the $\limsup_{M\to\infty}$ then $\lim_{a\downarrow 0}$ on both sides of (33) we get

$$\begin{aligned} &\limsup_{M\to\infty} -\frac{1}{\mathbb{E}_\varepsilon T^M}\ln \mathbb{P}_\varepsilon(\mathcal{E} \,|\, c^M) \\ &\leq \frac{C(\varepsilon)}{C(\delta)} \max_{\alpha\in[0,1]}\left[(1-\alpha)D(\delta\,\|\,\varepsilon) + \alpha E_B\left(\frac{\gamma C(\delta)}{\alpha}, \varepsilon\right)\right] \end{aligned} \quad (34)$$

where the right-hand side is now independent of $\{c^M\}_{M\geq 1}$. Since $\delta \in [0, L]$ is arbitrary, we may minimize the right-hand side of (34) and obtain

$$\begin{aligned} &\limsup_{M\to\infty} -\frac{1}{\mathbb{E}_\varepsilon T^M}\ln \mathbb{P}_\varepsilon(\mathcal{E} \,|\, c^M) \\ &\leq C(\varepsilon) \min_{\delta\in[0,L]} \frac{1}{C(\delta)} \\ &\quad \times \max_{\alpha\in[0,1]}\left[(1-\alpha)D(\delta\,\|\,\varepsilon) + \alpha E_B\left(\frac{\gamma C(\delta)}{\alpha}, \varepsilon\right)\right]. \end{aligned} \quad (35)$$

[7] In particular the $o(1)$ that appears in $\mathbb{E}_Q(T^M|Y^t = y^t) = (t + n_t(y^t))(1 + o(1))$ of the requirement I is equivalent in our notation to $o_{t,\infty}(1)$.

[8] Let $(\Omega, \mathcal{F})$ be a probability space, let P_1 and P_2 be two probability measures on $(\Omega, \mathcal{F})$ and let $B \in \mathcal{F}$. From the data processing inequality for divergence [2, p. 167], we have

$$D(P_1\,\|\,P_2) \geq D(P_1(B)\,\|\,P_2(B)) \quad (24)$$

where $D(P_1(B)\|P_2(B)) \triangleq P_1(B)\ln\frac{P_1(B)}{P_2(B)} + (1 - P_1(B))\ln\frac{(1-P_1(B))}{(1-P_2(B))}$. Expanding (24) we deduce that

$$P_2(B) \geq e^{-\frac{D(P_1\,\|\,P_2)+H(P_1(B))}{P_1(B)}} \quad (25)$$

where $H(P_1(B)) \triangleq P_1(B)\ln P_1(B) + (1-P_1(B))\ln(1-P_1(B))$. In order to derive (26), we set $\Omega = \{0,1\}^t$, $B = S$, $P_1 = \mathbb{P}_\delta$, and $P_2 = \mathbb{P}_\varepsilon$.

[9] Similarly to the notation introduced after (23), we write $f(x) = \mathcal{O}_{x,\infty}(g(x))$ if there exists $\mu > 0$ such that $|f(x)| \leq \mu g(x)$ for x large enough. In the sequel we shall also write $f(x) = \mathcal{O}_{x,0}(g(x))$ if there exists $\mu > 0$ such that $|f(x)| \leq \mu g(x)$ for x small enough.

Now observe that the term in squared brackets in (35) is convex in α, hence is maximized at either $\alpha = 0$ or $\alpha = 1$. Therefore we have, for all $\varepsilon \in [0, L]$

$$\begin{aligned}\limsup_{M\to\infty} -\frac{1}{\mathbb{E}_\varepsilon T^M}\ln \mathbb{P}_\varepsilon(\mathcal{E}\mid c^M) &\le C(\varepsilon)\min_{\delta\in[0,L]}\frac{1}{C(\delta)}\max_{\alpha\in\{0,1\}}\Big[(1-\alpha)D(\delta\,\|\,\varepsilon)\\ &\qquad +\alpha E_B\left(\frac{\gamma C(\delta)}{\alpha},\varepsilon\right)\Big]\\ &= C(\varepsilon)\min_{\delta\in[0,L]}\frac{1}{C(\delta)}\max\{D(\delta\,\|\,\varepsilon), E_B(\gamma C(\delta),\varepsilon)\}\end{aligned} \tag{36}$$

which concludes the proof of the theorem for the BSC case.

We now turn to the general case where $\mathcal{Q}$ is a family of DMC's with same input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$ and where $\inf_{Q\in\mathcal{Q}} C(Q) > 0$. Since this case is a straightforward extension of the BSC case we will only present the main steps.

We assume that communication is carried out over a channel $Q \in \mathcal{Q}$. Pick a channel V in $\mathcal{Q}$ and define

$$S = S(a,V,t) \triangleq \left\{y^t \in \{0,t\}^t : \mathbb{P}_V\left(T^M > \frac{\ln M}{\gamma C(V) - a}\,\middle|\, Y^t = y^t\right) \le a\right\}. \tag{37}$$

We have

$$\mathbb{P}_Q(\mathcal{E}\cap\{Y^t\in S\}\mid c^M) = \mathbb{P}_Q(Y^t\in S\mid c^M)\,\mathbb{P}_Q(\mathcal{E}\mid Y^t\in S, c^M) \tag{38}$$

where $\mathbb{P}_Q(Y^t \in S \mid c^M) \neq \mathbb{P}_Q(Y^t \in S)$ since the training sequence may depend now on feedback.[10]

Let $P^Q_{X^t,Y^t}$ be the distribution on $\mathcal{X}^t \times \mathcal{Y}^t$ induced by the training policy and the channel Q. One easily show that, without loss of generality, we have

$$\begin{aligned}P^Q_{X^t,Y^t}(x^t,y^t) &= \prod_{i=1}^t \mathbb{P}(X_i = x_i \mid Y^{i-1} = y^{i-1})\mathbb{P}(Y_i = y_i \mid X_i = x_i)\\ &= \prod_{i=1}^t \mathbb{P}(X_i = x_i \mid Y^{i-1} = y^{i-1})Q(y_i\mid x_i)\end{aligned} \tag{39}$$

with

$$\mathbb{P}(X_1 = x_1 | Y^0 = y^0) \triangleq \mathbb{P}(X_1 = x_1).$$

The training policy is completely specified by the family of probabilities $\{\mathbb{P}(X_i = x_i | Y^{i-1} = y^{i-1})\}$ with $x_i \in \mathcal{X}, y^{i-1} \in \mathcal{Y}^{i-1}$, and $i \in [1,t]$. Note that the Y_i's determine the X_i's during the training phase and that the event $Y^t \in S$ is the same as $(X^t, Y^t) \in \tilde{S}$ for some $\tilde{S} \in \mathcal{X}^t \times \mathcal{Y}^t$. Observe that $\tilde{S}$ has a high error probability undr $P^V_{X^t,Y^t}$, hence. similarly as for (26), the data processing inequality for divergence yields

$$\mathbb{P}_Q(Y^t\in S\mid c^M) \ge e^{-D(P^V_{X^t,Y^t}\,\|\,P^Q_{X^t,Y^t})(1+o_{a,0}(1))(1+o_{t,\infty}(1))}. \tag{40}$$

We now compute an upper bound on $D(P^V_{X^t,Y^t} \,\|\, P^Q_{X^t,Y^t})$. Using (39) we have

$$D\left(P^V_{X^t,Y^t}\,\|\,P^Q_{X^t,Y^t}\right) = \sum_{x^t\in\mathcal{X}^t} P^V_{X^t}(x^t) \tag{41}$$

$$\times \sum_{y^t\in\mathcal{Y}^t} P^V_{Y^t\mid X^t}(y^t\mid x^t)\log\left(\frac{\prod_{i=1}^t V(y_i\mid x_i)}{\prod_{i=1}^t Q(y_i\mid x_i)}\right). \tag{42}$$

[10]Notice the difference with the BSC case.

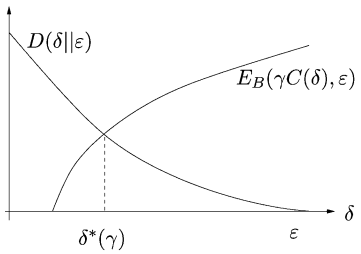

Fig. 2.

Then, since $P^V_{Y_i\mid X^t}(y_i\mid x^t) = V(y_i\mid x_i)$ for all $1 \le i \le t$, we deduce that

$$\begin{aligned}\sum_{y^t\in\mathcal{Y}^t} P^V_{Y^t\mid X^t}(y^t\mid x^t)\log\left(\frac{\prod_{i=1}^t V(y_i\mid x_i)}{\prod_{i=1}^t Q(y_i\mid x_i)}\right) &= \sum_{i=1}^t\sum_{y_i\in\mathcal{Y}} V(y_i\mid x_i)\log\frac{V(y_i\mid x_i)}{Q(y_i\mid x_i)}\\ &= \sum_{i=1}^t D(V(\cdot\mid x_i)\,\|\,Q(\cdot\mid x_i))\\ &\le t\max_{x\in\mathcal{X}} D(V(\cdot\mid x)\,\|\,Q(\cdot\mid x)).\end{aligned} \tag{43}$$

From (41) and (43) we get

$$D(P^V_{X^t,Y^t}\,\|\,P^Q_{X^t,Y^t}) \le t\max_x D(V(\cdot\mid x)\,\|\,Q(\cdot\mid x)) \tag{44}$$

and from (40) we conclude that, for any $V \in \mathcal{Q}$,

$$\mathbb{P}_Q(Y^t\in S\mid c^M) \ge e^{-t\max_{x\in\mathcal{X}} D(V(\cdot\mid x)\,\|\,Q(\cdot\mid x))(1+o_{a,0}(1))(1+o_{t,\infty}(1))}. \tag{45}$$

From the definition of S and the requirements I and III we get

$$\mathbb{E}_Q(T^M\mid Y^t\in S, c^M) \le \frac{\ln M}{\gamma C(V)}(1+o_{a,0}(1))(1+o_{t,\infty}(1)). \tag{46}$$

Finally, since $\mathbb{E}_Q T^M = \frac{\ln M}{\gamma C(Q)}(1+o_{a,0}(1))$ by the requirement II, a computation along the lines of (30)–(36) yields

$$\begin{aligned}\limsup_{M\to\infty} -\frac{1}{\mathbb{E}_Q T^M}\ln\mathbb{P}_Q(\mathcal{E}\mid c^M) &\le C(Q)\min_{V\in\mathcal{Q}}\frac{1}{C(V)}\\ &\quad\times\max\left\{\max_{x\in\mathcal{X}} D(V(\cdot\mid x)\,\|\,Q(\cdot\mid x)), E_B(\gamma C(V),Q)\right\}.\end{aligned} \tag{47}$$

□

To prove that $E_{\rm tbs}$ has a slope that vanishes at capacity in the case where $\mathcal{Q} = \text{BSC}_L$ we proceed as follows. First note that

$$E_{\rm tbs}(\gamma,\varepsilon) \le \min_{\delta\in[0,\varepsilon]}\max\{D(\delta\,\|\,\varepsilon), E_B(\gamma C(\delta),\varepsilon)\}. \tag{48}$$

Now pick some $\varepsilon \in (0, L]$ and some $\gamma \in (0,1)$. We refer the reader to Fig. 2 in which we draw $D(\delta\,\|\,\varepsilon)$ and $E_B(\gamma C(\delta),\varepsilon)$ as functions of δ. The value $\delta^*(\gamma)$ is defined as the δ such that

$$D(\delta\,\|\,\varepsilon) = E_B(\gamma C(\delta),\varepsilon). \tag{49}$$

Hence, $\delta^*(\gamma)$ satisfies

$$\min_{\delta\in[0,\varepsilon]}\max\{D(\delta\,\|\,\varepsilon), E_B(\gamma C(\delta),\varepsilon)\} = D(\delta^*(\gamma)\,\|\,\varepsilon). \tag{50}$$

Since $E_B(\gamma C(\delta),\varepsilon)$ is concave in the range of δ for which $E_B(\gamma C(\delta),\varepsilon)$ is positive, one deduces that

$$\begin{aligned}\varepsilon - \delta^*(\gamma) &\le \frac{E_B(\gamma C(\varepsilon),\varepsilon)}{\frac{dE_B(\gamma C(\delta),\varepsilon)}{d\delta}\Big|_{\delta=\varepsilon}}\\ &= (1-\gamma)\frac{C(\varepsilon)}{\gamma\ln\frac{1-\varepsilon}{\varepsilon}}.\end{aligned} \tag{51}$$

On the other hand, as $\gamma \uparrow 1$, the quantity $\delta^*(\gamma)$ tends to ε. Since $D(\delta^*(\gamma) \,\|\, \varepsilon) = \mathcal{O}_{|\varepsilon-\delta^*(\gamma)|,0}(|\varepsilon - \delta^*(\gamma)|^2)$, using (48), (50), and (51) gives

$$\begin{aligned} 0 &\le \lim_{\gamma\uparrow 1} \frac{E_{tbs}(\gamma,\varepsilon)}{1-\gamma} \\ &\le \lim_{\gamma\uparrow 1} \frac{\mathcal{O}_{|\varepsilon-\delta^*(\gamma)|,0}(|\varepsilon-\delta^*(\gamma)|^2)}{1-\gamma} \\ &\le \lim_{\gamma\uparrow 1} \frac{C(\varepsilon)^2}{\gamma^2\left(\ln\frac{1-\varepsilon}{\varepsilon}\right)^2}\mathcal{O}_{1-\gamma,0}(1-\gamma) \\ &= 0 \end{aligned} \tag{52}$$

yielding the desired result.

IV. Conclusion

We proposed a definition of a training based scheme for universal communication, and, given any class of channels, we derived an upper bound on the error exponent of any such scheme. We then compared this bound with the maximum error exponent that can universally be achieved over a certain class of channels, which is known for the Binary Symmetric, Z, and Binary Erasure families. In these cases our result shows that, in particular for high rate communication, good universal coding strategies do not separate channel estimation from information delivery.

Acknowledgment

The authors wish to thank M. V. Burnashev, J. L. Massey, and B. Rimoldi for valuable comments. They are grateful to an anonymous referee for an important observation made on the definition of a training-based scheme.

References

[1] 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.

[2] I. Csiszar and J. Körner, *Information Theory*. Budapest, Hungary: Akademiai Kiado, 1986.

[3] P. Dayal, M. Brehler, and M. K. Varanasi, "Leveraging coherent space-time codes for noncoherent channels via training," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 2058–2080, Sep. 2005.

[4] 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, 1962.

[5] M. Feder and A. Lapidoth, "Universal decoding for channels with memory," *IEEE Trans. Inf. Theory*, vol. 44, no. 5, pp. 1726–1745, Sep. 1998.

[6] R. G. Gallager, *Information Theory and Reliable Communications*. New York: Wiley, 1968.

[7] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?," *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 951–963, Apr. 2003.

[8] N. Shulman, "Communication over an Unknown Channel via Common Broadcasting," Ph.D. dissertation, Tel-Aviv Univ., Tel Aviv, Israel, 2003.

[9] A. Tchamkerten and İ. E. Telatar, "Variable length coding over an unknown channel," I*EEE Trans. Inf. Theory*, to be published.

[10] T. F. Wong and B. Park, "Training sequence optimization in MIMO systems with colored interference," *IEEE Trans. Commun.*, vol. 52, no. 11, pp. 1939–1947, Nov. 2004.

Optimized Diversity Combining With Imperfect Channel Estimation

Ranjan K. Mallik, *Senior Member, IEEE*

***Abstract*—In a communication system using receive diversity and linear combining in the presence of cochannel interference (CCI), optimum combining (OC) is known to give the best error performance since it maximizes the instantaneous signal-to-interference-plus-noise ratio of the combiner output, and consequently, in the presence of Gaussian interference plus noise, it minimizes the error rate. However, this is based on the assumption that a perfect estimate of the channel is available. Channel estimation methods in reality use some overhead. When the channel is time-invariant, the estimation error decreases with increase in the amount of overhead, like the number of pilot symbols. With the growing need for high data rate applications, the amount of overhead that can be allocated for the estimation of the channel needs to be reduced, and the channel estimation error cannot be ignored. In this situation, replacing the channel by its imperfect estimate in the OC weight vector no longer results in an optimum scheme. We have to find an optimum scheme based on the channel estimation method and the detection criterion, which results in what we call optimized diversity combining (ODC). Here we focus on ODC resulting from a pilot symbol based maximum likelihood (ML) channel estimation method applied to a correlated flat Rayleigh fading channel in the presence of CCI and additive noise. The channel is randomly time-invariant during the reception of pilot and data symbols. The decision rule, which is optimum in the ML sense, is derived using concepts of Gaussian and Wishart statistics. Numerical results show that ODC can perform significantly better than OC with imperfect channnel estimates by appropriate choice of system parameters.**

***Index Terms*—Characteristic function, imperfect channel estimation, maximum likelihood estimate, optimized diversity combining (ODC), probability density function (pdf), pseudo-Wishart distribution, symbol error probability, Wishart distribution.**

I. Introduction

In diversity reception systems, combining methods which require estimates of the channel (that is, the diversity branch gains), like maximal-ratio combining (MRC), outperform those which do not need channel estimates, like postdetection equal-gain combining. In MRC (which is a linear combining scheme), for example, if we consider complex baseband processing at the receiver, then the combiner weights are the conjugates of the complex diversity branch gains. The values of the branch gains are not known to the receiver apriori and need to be estimated. This calls for overheads in the data transmitted, like the insertion of pilot symbols (which are known to the receiver) to be used for channel estimation. When there is no pressing need to have high data rates, a good amount of overhead can be used, achieving almost perfect channel estimates. However, the growing need for high data rate applications limits the amount of overhead that can be allocated for channel estimation, resulting in imperfect estimates, the estimation errors of which cannot be neglected during system design or analysis. Pioneering work on the effect of estimation errors in combiner weights of an MRC system has been done by Bello and Nelin [1], Proakis [2], and Gans [3]. The general problem of the estimation errors having Gaussian distributions is analyzed in [3], whereas pilot symbol based estimation is investigated in [1] and [2]. A general form of the bit error rate in an MRC system with Gaussian distributed weighting errors is presented in [4]. For the case of rake

Manuscript received June 16, 2004; revised May 26, 2005.

The author is with the Department of Electrical Engineering, Indian Institute of Technology-Delhi, Hauz Khas, New Delhi 110016, India (e-mail: rkmallik@ee.iitd.ernet.in).

Communicated A. Kavčić, Associate Editor for Detection and Estimation.
Digital Object Identifier 10.1109/TIT.2005.864444