Forty-Seventh Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 30 - October 2, 2009

A training based scheme for communicating over unknown channels with feedback

Aditya Mahajan
Dept. of Electrical Engineering,
Yale University,
New Haven, CT -06520, USA
Email: aditya.mahajan@yale.edu

Sekhar Tatikonda
Dept. of Electrical Engineering,
Yale University,
New Haven, CT -06520, USA
Email: sekhar.tatikonda@yale.edu

***Abstract*—We consider a communication system with noiseless feedback where the channel is not known to the encoder or the decoder. The channel belongs to a known family of channels and remains constant over time. Using the noiseless feedback, the encoder can learn the channel over time and communicate at a rate equal to the capacity of the actual realization of the channel. Thus, not knowing the channel does not affect capacity. However, analyzing the error exponent (for variable length coding) is more challenging. Tchamkerten and Telatar (2006) showed that for certain families of channels, not knowing the channel does not change the error exponent; for other families, not knowing the channel results in a strict decrease in the error exponent. In general, the error exponent is not known. It is also known that simple training based schemes have poor error exponent behavior. In this paper, we show that a smart training based scheme can achieve an error exponent which is a multiplicative factor less than the error exponent for known channel. This shows that contrary to popular belief, smart training based schemes preserve the main advantage of feedback—an error exponent with non-zero slope at rates close to capacity.**

I. Introduction

We consider a communication system with noiseless feedback where the channel is not known to the encoder or the decoder. The channel $W_\circ$ belongs to a family of DMCs (discrete memoryless channels) defined over common input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$. We use $W_\circ$ to denote both the channel as well as its transition probability. Nature chooses a channel out of this family before start of communication, her choice is not revealed to the encoder and the decoder, and the choice of channel does not change with time. We are interested in characterizing the error exponent for this setup.

When $W_\circ$ is known to the encoder and the decoder, the above setup is identical to the classical channel coding problem [1]. The capacity of channel $W_\circ$ is given by

$$C(W_\circ) = \max_{P(x)} I(X;Y)$$

where

$$I(X;Y) = \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} P(x)W_\circ(y|x)\log\left(\frac{W_\circ(y|x)}{\sum\limits_{x'\in\mathcal{X}} P(x')W_\circ(y|x')}\right)$$

is the mutual information between the channel input and the channel output. Since the channel is memoryless, feedback does not increase capacity. Nevertheless, feedback allows the encoder to adapt to the channel variations and consequently boosts the error exponent. In particular, if variable length coding [3] is allowed, the error exponent at rate $R < C(W_\circ)$ is

$$E_B(R, W_\circ) = \left(\max_{(x,x')\in\mathcal{X}\times\mathcal{X}} D(W_\circ(\cdot|x)\|W_\circ(\cdot|x'))\right)\left(1 - \frac{R}{C(W_\circ)}\right)$$

where

$$D(W_\circ(\cdot|x)\|W_\circ(\cdot|x')) = \sum_{y\in\mathcal{Y}} W_\circ(y|x)\log\left(\frac{W_\circ(y|x)}{W_\circ(y|x')}\right)$$

is the Kullback-Liebler divergence between the output distributions induced by input letters x and x'. We call $E_B(R, W_\circ)$ the *Burnashev's exponent* for channel $W_\circ$. Unlike the sphere packing and random coding exponents, the Burnashev's exponent has a non-zero slope at rates close to capacity. This slope captures the main advantage of noiseless feedback—reducing the transmission rate by a small *fraction* of the capacity, linearly increases the error exponent and thus, exponentially decreases the probability of error.

We are interested in the scenario when the channel $W_\circ$ is not known. In such a scenario, simple training based schemes can guarantee[1] reliable transmission at any rate below $C(W_\circ)$ (even though $C(W_\circ)$ is not known before the start of transmission). Thus, not knowing the channel does not affect channel capacity. However, error exponents behave differently.

For some families of channels, appropriately chosen adaptive communication schemes [6] can have error exponent equal to the Burnashev's exponent of $W_\circ$. However, for other families [6], no communication scheme can have an error exponent equal to the Burnashev's exponent of $W_\circ$. Thus, for some families of channels, not knowing the channel does not affect the error exponent; for others, it does. In the latter case, error exponent is not completely characterized.

In view of this negative result, we relax our objective. As explained before, the usefulness of feedback in a communication systems can be characterized by the slope of the

[1] The result follows from large deviation bounds for channel estimation and uniform continuity of mutual information in the input distribution and the channel transition matrix.

978-1-4244-5871-4/09/$26.00 ©2009 IEEE

error exponent at capacity. So, instead of trying to completely characterize the error exponent for communicating over an unknown channel, we simply ask if the error exponent has a non-zero slope near capacity. In Section II, we show an example of a family of channels for which no scheme can achieve Burnashev's exponent, yet a simple training based scheme an error exponent up to a multiplicative factor of Burnashev's exponent.

This example illustrates two points. First, it shows (perhaps not surprisingly) that even when the channel is unknown, noiseless feedback boosts the error exponent. Second, it shows that for communicating over unknown channels, training based schemes have merit.

The second point puts the result of [5] and [7] in perspective. In both [5] and [7], a simple family $\mathcal{W}_p = \{\mathrm{BSC}(p), \mathrm{BSC}(1-p)\}$, with $0 \leq p < \frac{1}{2}$ was used to prove negative results about training based schemes. ($\mathrm{BSC}(p)$ denotes a binary symmetric channel with crossover probability p). With fixed length communication, training based schemes do no not achieve universal error exponent [5]. With variable length communication with noiseless feedback, training based scheme *that have a single training phase* will have error exponent with zero slope near capacity. These results, especially the result of [7], suggest that when we are concerned about error exponents, we should not use training based schemes. In Section II, we show that this need not be the case. We propose a training based scheme *with multiple training phases* for the family $\mathcal{W}_p$; the error exponent of this scheme is of the form

$$E_S(R, \mathcal{W}_p) = D_{\mathcal{W}_p}\left(1 - \frac{R}{C(W_\circ)}\right)$$

where $D_{\mathcal{W}_p}$ is a constant that depends on the family $\mathcal{W}_p$ and is strictly less than the corresponding constant in Burnashev's exponent. This exponent has a non-zero slope at capacity! We present the intuition behind our proposed scheme in Section IV.

II. A TRAINING BASED SCHEME AND THE MAIN RESULT

In this section we consider the family of channels $\mathcal{W}_p = \{\mathrm{BSC}(p), \mathrm{BSC}(1-p)\}$, where $0 \leq p < \frac{1}{2}$ is known. This family is used in [5] and [7] to prove negative results about training based schemes. We use this family to show that training based schemes with *multiple training phases* can achieve error exponents that have the same form as Burnashev's exponent.

Both $\mathrm{BSC}(p)$ and $\mathrm{BSC}(1-p)$ have capacity

$$C_p = 1 - h(p)$$

where $h(p) = -p \log p - (1-p)\log(1-p)$ is the binary entropy function. Since both the channels have the same capacity, we do not need to modulate the transmission rate according to nature's choice of the channel. Hence, we assume that the transmission rate has a predetermined value R. This assumption simplifies the description of the coding scheme. The coding scheme operates in multiple epochs of duration t. We choose four fractions β_1, β_2, β_3, and β_4 such that $0 \leq \beta_i < 1$, $i = 1, 2, 3, 4$, and $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$. These fractions are chosen before the start of communication; their actual values will be described later.

For every sufficiently large epoch duration t, the encoder and the decoder choose a codebook and "hypothesis testing regions" for channels $\mathrm{BSC}(p)$ and $\mathrm{BSC}(1-p)$ such that

1) The codebook is of length $\beta_2 t$ and rate R/β_2. When $R/\beta_2 < C_p$, the error exponent $E_r(R/\beta_2)$ of the codebook is positive.
2) The "hypothesis testing regions" optimally distinguish between the transmission of $\beta_4 t$ zeros from $\beta_4 t$ ones according to the Chernoff-Stein Theorem [2].

These codebooks and hypothesis testing regions are for the case then the channel is known. Thus, the codebook for $\mathrm{BSC}(p)$ assumes that the channel is $\mathrm{BSC}(p)$ and the codebook for $\mathrm{BSC}(1-p)$ assumes that the channel is $\mathrm{BSC}(1-p)$. Similar interpretation holds for the hypothesis testing regions.

Consider a variable length communication scheme that consists of multiple epochs of length t. Each epoch consists for four phases:

1) *A training phase of length $\beta_1 t$*: During this phase the encoder sends a training sequence of β_1 zeros. At the end of this phase, the encoder and the decoder choose a channel estimate θ_1. If the number of ones in the channel output is less than $\beta_1 t/2$, θ_1 equals $\mathrm{BSC}(p)$; otherwise θ_1 equals $\mathrm{BSC}(1-p)$.
2) *A communication phase of length $\beta_2 t$*: During this phase the encoder transmits of of $M^{(t)} = \lfloor 2^{tR} \rfloor$ messages using the codebook corresponding to θ_1. The decoder decodes according to the same codebook. Because of output feedback, the encoder knows the decoded message.
3) *A re-training phase of length $\beta_3 t$*: During this phase, the encoder sends $\beta_3 t$ zeros. At the end of this phase, the encoder and the decoder choose a channel estimate θ_2. If the number of ones in the channel output is less than $\beta_3 t/2$, θ_2 equals $\mathrm{BSC}(p)$; otherwise θ_2 equals $\mathrm{BSC}(1-p)$. The channel estimate θ_2 depends on the channel behavior in only the third phase and not the first phase.
4) *A control phase of length $\beta_4 t$*: If the decoding in the second phase was correct, the encoder sends an ACK consisting of $\beta_4 t$ zeros; otherwise it sends a NACK consisting of $\beta_4 t$ ones. The decoder uses the hypothesis testing regions of channel θ_2 to decode the ACK/NACK. If the decoded symbol is ACK, communication stops; otherwise, the encoder and the decoder discard the results of the current epoch and repeat the same process in the next epoch.

This communication scheme is similar to the scheme proposed by Yamamoto-Itoh [4]. That scheme achieves the Burnashev's exponent for variable length coding over a known channel. As in that scheme, a decoding error occurs if the message transmitted during the communication phase is decoded incorrectly, and the NACK sent by the encoder during the control phase is decoded as an ACK. All other erroneous situations are corrected by retransmission and increase only the duration of communication.

For epoch duration t, the above scheme has a probability of error $P_e^{(t)}$ and a random communication length $T^{(t)}$. Then, the main result of this paper is the following.

Proposition 1: The average transmission rate of the above communication scheme is

$$\lim_{t\to\infty} \frac{\log M^{(t)}}{\mathbb{E}\{T^{(t)}\}} = R. \tag{1}$$

Furthermore, the error exponent of the above scheme is

$$E_s(R, \mathcal{W}_p) = -\lim_{t\to\infty} \frac{\log P_e^{(t)}}{\mathbb{E}\{T^{(t)}\}} \geq D_p\left(1 - \frac{R}{C_p}\right) \tag{2}$$

where

$$D_p = \frac{D(0.5\|p)D(p\|1-p)}{D(0.5\|p) + D(p\|1-p)} \tag{3}$$

and

$$D(x\|y) = x\log(x/y) + (1-x)\log((1-x)/(1-y))$$

is the binary Kullback-Leibler divergence between binary probability distributions $[x, 1-x]$ and $[y, 1-y]$.

In contrast, when the channel is known the error exponent is

$$D(p\|1-p)\left(1 - \frac{R}{C_p}\right).$$

Therefore, in this example, a training based scheme for unknown channel achieves a fraction

$$\frac{D(0.5\|p)}{D(0.5\|p) + D(p\|1-p)}$$

of the Burnashev's exponent of the actual channel.

III. Proof of Proposition II

For epoch k, $k = 1, 2, \ldots$, with epoch duration t define the following events:

- $E_1^{(t,k)}$: the channel estimation at the end of the first phase is incorrect.
- $D^{(t,k)}$: the decoding at the end of the second phase is incorrect.
- $E_2^{(t,k)}$: the channel estimation at the end of the third phase is incorrect.
- $C_A^{(t,k)}$: an ACK is decoded as a NACK in the forth phase.
- $C_N^{(t,k)}$: a NACK is decoded as an ACK in the forth phase.

The assumptions on the codebook and the hypothesis testing regions can be stated more formally as follows. The error exponent of the codebook used in phase two is

$$E_r\left(\frac{R}{\beta_2}\right) = -\lim_{t\to\infty} \frac{\log \Pr(D^{(t,k)}|\bar{E}_1^{(t,k)})}{\beta_2 t}. \tag{4}$$

When $R/\beta_2 < C_p$, the error exponent $E_r(R/\beta_2) > 0$ and the probability of error

$$\lim_{t\to\infty} \Pr(D^{(t,k)}|\bar{E}_1^{(t,k)}) = 0. \tag{5}$$

Similarly, for the hypothesis testing region, Chernoff-Stein Theorem [2] implies that the probability of type I and type II errors are

$$\begin{aligned} &\lim_{t\to\infty} \Pr(C_A^{(t,k)}|\bar{E}_2^{(t,k)}) = 0, \\ &\lim_{t\to\infty} \Pr(C_N^{(t,k)}|\bar{E}_2^{(t,k)}) = 0, \end{aligned} \tag{6}$$

and the exponent of type I error is

$$-\lim_{t\to\infty} \frac{\log \Pr(C_A^{(t,k)}|\bar{E}_2^{(t,k)})}{\beta_4 t} = D(p\|1-p). \tag{7}$$

To evaluate the performance of the above communication scheme, we first determine the asymptotic behavior of the number of retransmission epochs.

Lemma 1: For epochs of duration t, let $\zeta^{(t)}$ be the epoch when communication stops, i.e.,

$$\zeta^{(t)} = \inf\{k \in \mathbb{N} : D^{(t,k)} \cap C_N^{(t,k)} \cup \bar{D}^{(t,k)} \cap \bar{C}_A^{(t,k)}\}.$$

Then, $T^{(t)} = t\zeta^{(t)}$ and

$$\lim_{t\to\infty} \mathbb{E}\{\zeta^{(t)}\} = 1$$

Proof: $T^{(t)} = t\zeta^{(t)}$ follows from the definition of $\zeta^{(t)}$. For the second part, we claim that $\zeta^{(t)}$ is stochastically dominated by a geometrically distributed random variable, i.e., for any t

$$\Pr(\zeta^{(t)} \geq k) \leq (\xi^{(t)})^{k-1}$$

where

$$\xi^{(t)} = \Pr(D(t,1)|\bar{E}_1(t,1)) + \Pr(C_N(t,1) \cap \bar{E}_2(t,1)).$$

We prove this claim by induction. The claim is trivially true for $k = 1$. Suppose it is true for some k in $\mathbb{N}$; then

$$\begin{aligned}
&\Pr(\zeta^{(t)} \geq k+1) \\
&\quad = \Pr(\zeta^{(t)} \geq k) \cdot \Pr(\zeta^{(t)} \geq k+1 | \zeta^{(t)} \geq k) \\
&\quad = \Pr(\zeta^{(t)} \geq k) \cdot \Pr(D^{(t,k)} \cap \bar{C}_N^{(t,k)} \cup \bar{D}^{(t,k)} \cap C_A^{(t,k)}) \\
&\quad = \Pr(\zeta^{(t)} \geq k) \\
&\qquad \cdot [\Pr(D^{(t,k)}) \cdot \Pr(\bar{C}_N^{(t,k)}) + \Pr(\bar{D}^{(t,k)}) \cdot \Pr(C_A^{(t,k)})] \\
&\quad \leq \Pr(\zeta^{(t)} \geq k) \cdot [\Pr(D^{(t,k)}|\bar{E}_1^{(t,k)}) + \Pr(C_A^{(t,k)} \cap \bar{E}_2^{(t,k)})] \\
&\quad \leq (\xi^{(t)})^k
\end{aligned}$$

where the first inequality follows from law of total probability, and the second inequality follows from the induction hypothesis. This completes the induction argument. Now, we use this claim to bound $\mathbb{E}\{\zeta^{(t)}\}$.

$$\mathbb{E}\{\zeta^{(t)}\} = \sum_{k\geq 1} \Pr(\zeta^{(t)} \geq k) \leq \sum_{k \geq 1} (\xi^{(t)})^{k-1} = \frac{1}{1 - \xi^{(t)}}.$$

(5) and (6) imply that $\lim_{t\to\infty} \xi^{(t)} = 0$. Thus, we have

$$\lim_{t\to\infty} \mathbb{E}\{\zeta^{(t)}\} \leq 1.$$

By definition $\zeta^{(t)} \geq 1$, which completes the proof. ■

We use the result of Lemma 1 to prove Proposition 1.

Proof of Proposition 1: The transmission rate is given by

$$\lim_{t\to\infty}\frac{\log M^{(t)}}{\mathbb{E}\{T^{(t)}\}}=\lim_{t\to\infty}\frac{\log\lfloor 2^{tR}\rfloor}{t\mathbb{E}\{\zeta^{(t)}\}}=R\lim_{t\to\infty}\frac{1}{\mathbb{E}\{\zeta^{(t)}\}}=R$$

where the first equality follows from Lemma 1.

Now, consider the probability of error,

$$\begin{aligned}
P_e^{(t)}&=\Pr(D^{(t,\zeta^{(t)})}\cap C_N^{(t,\zeta^{(t)})})\\
&\le\sum_{k\ge1}\Pr(D^{(t,k)}\cap C_N^{(t,k)}|\zeta^{(t)}\ge k)\cdot\Pr(\zeta^{(t)}\ge k)\\
&\le\sum_{k\ge1}\Pr(D^{(t,k)})\cdot\Pr(C_N^{(t,k)})\cdot(\xi^{(t)})^{k-1}\\
&=\Pr(D^{(t,1)})\cdot\Pr(C_N^{(t,1)})\cdot\Big[\sum_{k\ge1}(\xi^{(t)})^{k-1}\Big]\\
&=\Pr(D(t,1))\cdot\Pr(C_N(t,1))\cdot\frac{1}{1-\xi^{(t)}}.
\end{aligned}\tag{8}$$

The first inequality follows the union bound of probability, the second inequality is due to the independence of decoding and hypothesis testing and the claim in Lemma 1, and the second last equality is due to symmetry across transmission epochs. The error exponent is given by

$$\begin{aligned}
E_S&=-\lim_{t\to\infty}\frac{\log P_e^{(t)}}{\mathbb{E}\{T^{(t)}\}}\\
&=-\lim_{t\to\infty}\frac{\log P_e^{(t)}}{t}\lim_{t\to\infty}\frac{1}{\mathbb{E}\{\zeta^{(t)}\}}\\
&=-\lim_{t\to\infty}\frac{\log P_e^{(t)}}{t}\\
&\ge-\lim_{t\to\infty}\frac{\log\Pr(D^{(t,1)})}{t}-\lim_{t\to\infty}\frac{\log\Pr(C_N^{(t,1)})}{t}\\
&\quad+\lim_{t\to\infty}\frac{1-\xi^{(t)}}{t}\\
&=-\lim_{t\to\infty}\frac{\log\Pr(D^{(t,1)})}{t}-\lim_{t\to\infty}\frac{\log\Pr(C_N^{(t,1)})}{t}
\end{aligned}\tag{9}$$

where the first two equalities follow from Lemma 1, the third inequality follows from (8) and last equality follows from $\lim_{t\to\infty}\xi^{(t)}=0$.

Next, consider the above two term one by one. First consider $\Pr(D^{(t,1)})$. By the law of total probability

$$\begin{aligned}
\Pr(D^{(t,1)})&=\Pr(D^{(t,1)}|E_1^{(t,1)})\cdot\Pr(E_1^{(t,1)})\\
&\qquad+\Pr(D^{(t,1)}|\bar{E}_1^{(t,1)})\cdot\Pr(\bar{E}_1^{(t,1)})\\
&\le\Pr(E_1^{(t,1)})+\Pr(D^{(t,1)}|\bar{E}_1^{(t,1)})\\
&\le\exp(-\beta_1 tD(0.5\|p))+\exp(-\beta_2 tE_R(R/\beta_2)).
\end{aligned}\tag{10}$$

The first term in last inequality follows from Chernoff-Hoeffding Theorem [8] and the second term follows from the definition of error exponents and is true for sufficiently large values of t. Thus, when $R/\beta_2<C_p$,

$$\begin{aligned}
&-\lim_{t\to\infty}\frac{\log\Pr(D^{(t,1)})}{t}\\
&\quad\ge\lim_{t\to\infty}\min\{\beta_1 D(0.5\|p),\beta_2 E_R(R/\beta_2)\}\ge0.
\end{aligned}\tag{11}$$

Now consider $\Pr(C_N^{(t,1)})$. By the law of total probability

$$\begin{aligned}
\Pr(C_N^{(t,1)})&=\Pr(C_N^{(t,1)}|E_2^{(t,1)})\cdot\Pr(E_2^{(t,1)})\\
&\qquad+\Pr(C_N^{(t,1)}|\bar{E}_2^{(t,1)})\cdot\Pr(\bar{E}_2^{(t,1)})\\
&\le\Pr(E_2^{(t,1)})+\Pr(C_N^{(t,1)}|\bar{E}_2^{(t,1)})\\
&\le\exp(-\beta_3 tD(0.5\|p))+\exp(-\beta_4 tD(p\|1-p))
\end{aligned}\tag{12}$$

where the first term in the last inequality follows from Chernoff-Hoeffding Theorem [8] and the second term follows from (7) and is true for sufficiently large values of t. Thus,

$$\begin{aligned}
&-\lim_{t\to\infty}\frac{\log\Pr(C_N^{(t,1)})}{t}\\
&\quad\ge\lim_{t\to\infty}\min\{\beta_3 D(0.5\|p),\beta_4 D(p\|1-p)\}.
\end{aligned}$$

The minimum value is achieved when

$$\frac{\beta_3}{\beta_4}=\frac{D(p\|1-p)}{D(0.5\|p)}$$

and in that case

$$\begin{aligned}
&-\lim_{t\to\infty}\frac{\log\Pr(C_N^{(t,1)})}{t}\\
&\quad\ge\frac{D(0.5\|p)D(p\|1-p)}{D(0.5\|p)+D(p\|1-p)}\lim_{t\to\infty}(\beta_3+\beta_4).
\end{aligned}\tag{13}$$

Finally, choose β_1, β_2, β_3, and β_4 such that $\beta_1=o(1)$ and $\beta_2=R/C_p-o(1)$. Thus, $\beta_3+\beta_4=1-\beta_1-\beta_2=1-R/C_p+o(1)$. (For example, the $o(1)$ terms can be $\log t$). Combining this with (9), (11) and (13), we get,

$$E_S\ge\left(1-\frac{R}{C_p}\right)\frac{D(0.5\|p)D(p\|1-p)}{D(0.5\|p)+D(p\|1-p)}.\tag{14}$$

■

IV. Discussion

At first glance, the communication scheme proposed in this paper looks counter intuitive. In the retraining phase, we ignore what we learnt in the first training phase. To understand why ignoring this available information makes sense, we need to reconsider why training based schemes work badly for fixed length communication over an unknown channel and why Yamamoto-Itoh's scheme is able to achieve the Burnashev's exponent for variable length communication over a known channel.

First, lets reconsider the arguments of [5] on why training based schemes perform badly for fixed length communication over an *unknown* channel. In that scenario, the probability of decoding error is similar to expression (10) for $\Pr(D^{(t,1)})$: a sum of two exponential terms. To ensure that the exponent of the channel estimation error (the first term) does not dominate, the training length should be long enough so that the two exponents match. However, training for a non-negligible fraction of the communication length results in a loss in transmission rate. On the other hand, if we do not want a loss in transmission rate, the training length should be negligible; this in turn implies that the exponent of the channel estimation dominates, and the overall error exponent drops drastically (although it still remains positive). Thus, training based schemes will either result in a loss in transmission rate, or a loss in the error exponent.

Next, lets reconsider why Yamamoto-Itoh's scheme [4] achieves the Burnashev's exponent for variable length coding over a *known* channel. As in Lemma 1, the number of retransmissions in Yamamoto-Itoh's scheme is stochastically dominated by a geometric random variable whose parameter goes to zero with epoch length. As a result, the retransmissions do not affect the transmission rate. Furthermore, the dominant error event is an intersection of two *independent* events: the decoding error in the communication phase, and a NACK being decoded as an ACK in the control/hypothesis testing phase. As long as the exponent of the decoding error in the communication phase is positive, the overall error exponent is dominated by the Chernoff-Stein exponent for hypothesis testing (the Kullback-Leibler divergence term in the Burnashev's exponent) times the fraction of time spent in the hypothesis testing phase. If this fraction is less than $(1 - R/C)$, then the exponent of the decoding error is positive. This is why Yamamoto-Itoh's scheme achieves the Burnashev's exponent.

A careful examination of these two results shows that if we use Yamamoto-Itoh like scheme for variable length coding over an *unknown* channel which uses a training based scheme in the communication and the hypothesis testing phases then a happy coincidence occurs. In the communication phase, we need to ensure only that the exponent of the decoding error is positive. So, we can use a training based scheme that spends a negligible fraction of time on training. In the hypothesis testing phase, we want to maximize the exponent of hypothesis testing. So, we can use a training based scheme that spends an appropriate amount of time on training to ensure a good exponent. Such a training based scheme will result in a loss in transmission rate. However, in the hypothesis testing phase we are sending one of only two messages; so, the transmission rate is zero anyways. Thus, training based schemes for fixed length coding over unknown channels fits nicely with the requirements of a Yamamoto-Itoh like scheme. The only caveat for the analysis to go through is that the decoding error should be independent of a NACK being decoded as an ACK. To ensure this independence, we do not use any information from the first training phase in the second one.

In the example of Section II, the family of channels was such that all channels had the same capacity. As a result, we could fix the transmission rate in advance, and thereby fix the ratio of $\beta_1+\beta_2$ and $\beta_3+\beta_4$ *and* the epoch length t in advance. When the transmission rate depends on the channel, both the ratio $(\beta_1 + \beta_2)/(\beta_3 + \beta_4)$ and the epoch length t cannot be chosen in advance. So, the communication scheme proposed in Section II will not work in that case. Nonetheless, the example of Section II illustrates that training based schemes for communicating over unknown channels warrant further investigation.

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 22, pp. 379–423, Jul. 1948.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, ser. Wiley series in Telecommunication. John Wiley and Sons, 1991.

[3] M. V. Burnashev, "Data transmission over a discrete channel with feedback. Random transmission time," *Problemy peredachi informatsii*, vol. 12, no. 4, pp. 10–30, 1976.

[4] H. Yamamoto and K. Itoh, "Asymptotic performance of a modified Schalkwijk-Barron scheme for channels with noiseless feedback," *IEEE Trans. Inf. Theory*, vol. 25, no. 6, pp. 729–733, Nov. 1979.

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

[6] A. Tchamkerten and I. E. Telatar, "Variable length coding over an unknown channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 2126–2145, May 2006.

[7] A. Tchamkerten and I. E. Telatar, "On the use of training sequences for channel estimation," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1171–1176, Mar. 2006.

[8] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, Mar. 1963.