Binary additive channels with individual noise sequences and limited active feedback

Krishnan Eswaran
Anand D. Sarwate
Anant Sahai
Michael Gastpar

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2007-5
http://www.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-5.html

January 8, 2007

Acknowledgement

We thank Ofer Shayevitz and Meir Feder for providing a preprint of their paper after their presentation of it at the Kailath Colloquium. This work grew out of a presentation of that work for UC Berkeley's advanced information theory course EE290S. Special thanks go to the other students in that class for helpful discussions.

Binary additive channels with individual noise sequence and limited active feedback [1]

Krishnan Eswaran, Anand D. Sarwate, Anant Sahai, and Michael Gastpar
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
Berkeley, CA 94720, USA
Email: {keswaran, asarwate, sahai, gastpar}@eecs.berkeley.edu

Abstract

Recently, Shayevitz and Feder introduced an individual sequence formulation of channel coding and an elegant scheme that adapts Horstein's scheme to this setting. Shayevitz and Feder's scheme requires both full-rate passive channel output feedback as well as a lower-rate active feedback channel. We show how to eliminate the need for full-rate passive channel output feedback by using common randomness and limited active feedback in the style of Hybrid ARQ while still asymptotically achieving the empirical capacity.

I. Introduction

Consider a channel that takes binary inputs and produces binary outputs, where the output is produced by potentially flipping some bits of the channel input. If the positions at which bits are flipped do not depend on the channel input symbol, then one way to express the output $\mathbf{y} \in \{0,1\}^N$ is

$$\mathbf{y} = \mathbf{x} \oplus \mathbf{z}, \tag{1}$$

where $\mathbf{x} \in \{0,1\}^N$ is the channel input, $\mathbf{z} \in \{0,1\}^N$ is the noise sequence, and addition is carred out modulo-2. Let p_{emp} be the empirical fraction of 1's in $\mathbf{z}$.

A simple model for $\mathbf{z}$ is the binary symmetric channel (BSC), in which $\mathbf{z}$ iid with $\mathbb{P}(z_j = 1) = p_{\text{emp}}$. The capacity for such channels is $1 - h(p_{\text{emp}})$, which we will call the *empirical capacity* for sequences with noise frequency p_{emp}. Another model is the arbitrarily varying channel (AVC) [1] in which $p_{\text{emp}} \leq p$, where p is known in advance, and codes are designed to target the worst-case capacity $1-h(p)$. This model and these coding strategies do not exploit the case when $p_{\text{emp}} \ll p$, and so are suboptimal when the noise is not chosen adversarially.

Shayevitz and Feder recognized this problem and recently proposed a coding strategy that uses passive output feedback and limited active feedback to achieve the empirical capacity $1 - h(p_{\text{emp}})$ for all p_{emp} and $\mathbf{z}$ [2]. In their model, the encoder maintains an infinite queue of bits to transmit. Given a blocklength of N, the decoder choose to decode the first NR bits of the message, where R is not fixed in advance. Their coding scheme guarantees that the NR decoded bits are equal to the first NR bits of the message with high probability for all noise sequences $\mathbf{z}$ as the blocklength $N \to \infty$.

Feedback is a necessary component for any coding strategy attempting to achieve $1 - h(p_{\text{emp}})$ without prior knowledge of p_{emp}. Without feedback the encoder and decoder must choose a rate R in advance that will be larger than $1 - h(p_{\text{emp}})$ for some p_{emp} and by the strong converse the decoding error will tend to 1. Shayevitz and Feder use passive output feedback to allow the encoder to estimate the noise frequency and use a sequential coding strategy to control the decoder. In order to avoid requiring common randomness, training sequence positions are determined via an additional active feedback channel.

In this work we do away with the passive feedback link and consider only active feedback from the decoder to encoder along with common randomness. To do this, we adapt the feedback-reducing block/chunk strategies used earlier in the context of reliability functions [3], [4] and most specifically in [5]. They are in turn inspired by Hybrid ARQ [6].

[0]The work of A.D. Sarwate and M. Gastpar was supported in part by the National Science Foundation under award CCF-0347298.

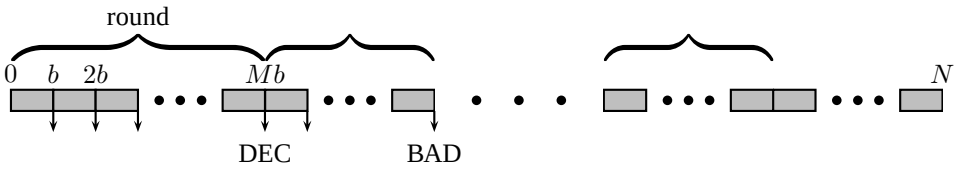

Fig. 1. After each chunk of length b feedback can be sent. Rounds end by decoding a message or declaring the noise to be bad.

Theorem 1: When used over a binary channel whose noise sequence has $p_{\text{emp}}N$ ones, with probability $1-\delta(N)$ the algorithm below achieves a rate $1-h_b(p_{\text{emp}})-\beta(N)$ for $p_{\text{emp}} \notin [1/2-\tau(N), 1/2+\tau(N)]$, where $\delta(N) \to 0$, $\beta(N) \to 0$, $\tau(N) \to 0$ as $N \to \infty$.

In Section II, the coding strategy is described in detail. The different error events are described in Section III, along with a sketch of the key proof ideas showing that these go to zero. Section IV concludes the paper with some discussion.

II. The coding strategy

We divide the blocklength N into *chunks* of length $b = b(N)$. The encoder attempts to send $k = k(N)$ bits over several chunks comprising a *round*. Prior to transmission, the decoder and encoder use common randomness to choose $n_1 = n_1(N)$ *training positions* T_n for chunk n. The encoder and decoder will also choose a random codebook for each round. In a round, the encoder divides the codebook into segments of length $b-n_1$ and transmits the i-th segment over the $b-n_1$ non-training positions in chunk i.

The decoder uses the training positions to estimate the empirical noise distribution in that chunk. After each chunk the decoder will either (a) decide to decode the k bits and tell the encoder to terminate the round, (b) decide that the empirical noise is too bad and tell the encoder to terminate the round and start over, or (c) decide that it cannot decode yet and tell the encoder to send another chunk.

A more formal description of the scheme follows, and an illustration is provided in Figure 1. For each round, the following steps are repeated for each chunk:

1) The encoder sends the b bits in chunk n. For $j \in T_n$, it sends $x_j = 0$ and for $j \notin T_n$ the bits $\{x_j : j \notin T_i\}$ are the next $b-n_1$ entries in the code for the current round.
2) Decoder estimates $p_{\text{emp}}^{(n)}$, the empirical frequency of 1's outside of T_n, and $\bar{p}_{\text{emp}}^{(n)}$, the average over chunks in the round:

$$\hat{p}_{\text{emp}}^{(n)} = \frac{1}{n_1}\sum_{j\in T_n} z_j \qquad\qquad \tilde{p}_{\text{emp}}^{(n)} = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_{\text{emp}}^{(i)} \ , \tag{2}$$

where $c_n = \{b(n-1)+1, \ldots, bn\}$ and

$$p_{\text{emp}}^{(n)} = \frac{1}{b}\sum_{j\in c_n} z_j \qquad\qquad \bar{p}_{\text{emp}}^{(n)} = \frac{1}{n}\sum_{i=1}^{n} p_{\text{emp}}^{(i)} \ . \tag{3}$$

3) The decoder makes a decision based on $\tilde{p}_{\text{emp}}^{(n)}$ and n:
   a) if

$$\tilde{p}_{\text{emp}}^{(n)} \in \left[\frac{1}{2}-\tau(N), \frac{1}{2}+\tau(N)\right] \tag{4}$$

   then decoder feeds back "BAD NOISE."
   b) if

$$\frac{k}{(b-n_1)\times n} < 1 - h\left(\tilde{p}_{\text{emp}}^{(n)}\right) - \epsilon_1(N) \tag{5}$$

   then decoder decodes and feeds back "DECODED."
   c) decoder feeds back "KEEP GOING" and goes to 1).

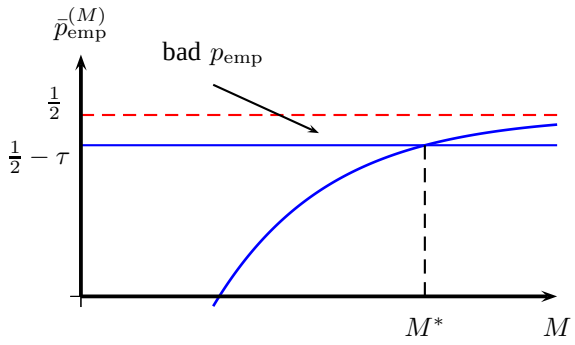

Fig. 2. Curve illustrating why M is finite.

III. Analysis

The analysis, carried out below, consists of two parts. In the first part we show that the training positions provide a good estimate of the empirical noise frequency (Lemma 2) and that the condition in (5) is sufficient to decode the k bits for a round with small probability of error (Lemma 3). The second part of the analysis shows that the loss in rate from our scheme is negligible as the blocklength increases. The rate loss within a round is small (Lemma 5). The overall rate loss across rounds is also small (Lemma 6). We show that all of the bounds can be satisfied by setting the parameters at the end of this section.

A. Error analysis

We will declare an error if one of following two events occurs:

1) (E_1) We declare an error if

$$\left|\tilde{p}_{\text{emp}}^{(M)} - \bar{p}_{\text{emp}}^{(M)}\right| > \epsilon_2(N). \tag{6}$$

2) (E_2) We will have an error if the decoder has a failure.

We must first find a bound on the length of a round in chunks. Let

$$M = \inf_{n>0} \left\{ \tilde{p}_{\text{emp}}^{(n)} \in \left[\frac{1}{2} - \tau, \frac{1}{2} + \tau\right] \right. \\ \left. \text{or } \frac{k}{(b-n_1)n} < 1 - h\left(\tilde{p}_{\text{emp}}^{(n)}\right) - \epsilon_1(k) \right\} \tag{7}$$

We now argue that M cannot be too large.

Lemma 1 (Bounds on M): We have $M \leq M^*$, where

$$M^* := \left\lceil \frac{k}{(b-n_1) \cdot \left(1 - h(\frac{1}{2} - \tau) - \epsilon_1\right)} \right\rceil . \tag{8}$$

If the decoder attempted to decode, then $M \geq M_*$, where

$$M_* = \left\lceil \frac{k}{b-n_1} \right\rceil \tag{9}$$

Proof: The argument is illustrated in Figure 2. We must simply find the point where the curve defined by (5) intersects the "BAD NOISE" threshold. This gives the bound in (8). The lower bound is trivial from the definition in (5). ■

Lemma 2 (Bound on E_1): We have

$$\mathbb{P}\left(E_1\right) \leq \frac{2N}{b} \exp\left(-2n_1 \cdot \epsilon_2^2\right) \tag{10}$$

Proof: We will first prove that our estimate of $\bar{p}_{\text{emp}}^{(n)}$ improves with n. We can view each $\hat{p}_{\text{emp}}^{(i)}$ as an independent random variable with mean $p_{\text{emp}}^{(i)}$. By Hoeffding's inequality [7]:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\hat{p}_{\text{emp}}^{(i)} - \frac{1}{n}\sum_{i=1}^{n}p_{\text{emp}}^{(i)}\right| > \epsilon_2(k)\right) \leq 2\exp(-2\cdot n\cdot\epsilon_2^2)\ . \tag{11}$$

Because we can terminate a round after any number of chunks, it is sufficient to bound the probability of E_1 only for a single chunk.

A result of Hoeffding [7, Theorem 4] states that the exponential inequalities for sampling with replacement hold for sampling without replacement as well, so $\hat{p}_{\text{emp}}^{(i)}$ converges to $p_{\text{emp}}^{(i)}$ and

$$\mathbb{P}\left(\left|\hat{p}_{\text{emp}}^{(i)} - p_{\text{emp}}^{(i)}\right| > \epsilon_2\right) \leq 2\exp\left(-2n_1\epsilon_2^2\right) \tag{12}$$

Taking a union bound over all N/b chunks, we obtain (10). ■

Lemma 3 (Bound on E_2): Assume $\tau > \epsilon_2$ and $\tilde{p}_{\text{emp}}^{(M)} \notin [1/2-\tau, 1/2+\tau]$, and let $\epsilon_3 = h(\frac{b}{b-n_1}\epsilon_2)$ Then

$$\mathbb{P}\left(E_2|E_1\right) \leq \exp\left(-k\left(\frac{(\epsilon_1-\epsilon_3-M^*/k)^2}{8/e^2+4(\ln 2)^2}\right)\right)\ . \tag{13}$$

Proof: Let us define the empirical noise in the non-training positions by

$$\bar{q}_{\text{emp}}^{(M)} = \frac{1}{M(b-n_1)}\sum_{n=1}^{M}\sum_{j\notin T_i} z_j \tag{14}$$

Then $M(b-n_1)\bar{q}_{\text{emp}}^{(M)} + Mn_1\tilde{p}_{\text{emp}}^{(M)} = \bar{p}_{\text{emp}}^{(M)}$. Then E_1^c implies

$$\left|M(b-n_1)\tilde{p}_{\text{emp}}^{(M)} - \sum_{n=1}^{M}\sum_{j\notin T_i} z_j\right| < Mb\epsilon_2 \tag{15}$$

So

$$\left|\tilde{p}_{\text{emp}}^{(M)} - \bar{q}_{\text{emp}}^{(M)}\right| < \frac{b}{b-n_1}\epsilon_2 \tag{16}$$

Thus under E_1^c we know that the empirical noise frequency in the non-training positions is also close to $\tilde{p}_{\text{emp}}^{(M)}$.

Since after every chunk the decoder must decide whether to decode, the actual rates that can be realized in a round fall in the discrete set $\{k/M(b-n_1) : M_* \leq M \leq M^*\}$. Under E_1^c we have from Lemma 7 that $|h(\tilde{p}_{\text{emp}}^{(M)}) - h(\bar{q}_{\text{emp}}^{(M)})| < \epsilon_3$. From the definition of the decoding rule in (5) we have

$$\begin{aligned}\frac{k}{M(b-n_1)} &< 1 - h(\tilde{p}_{\text{emp}}^{(M)}) - \epsilon_1 \\ &\leq 1 - h(\bar{q}_{\text{emp}}^{(M)}) - (\epsilon_1-\epsilon_3)\end{aligned} \tag{17}$$

Thus our code will be operating at a gap at least $(\epsilon_1-\epsilon_3)$ from the empirical capacity.

Since the capacity-achieving input distribution for any binary symmetric channel is Bernoulli$(1/2)$, the encoder and decoder will choose an iid random codebook of blocklength $M^*(b-n_1)$ with 2^{k+M^*} messages from that distribution using common randomness. Because M is not known in advance, we must guarantee that the codebook and its truncations to blocklengths $M(b-n_1)$ perform well, for $M_* \leq M \leq M^*$. From an exercise in Gallager [8, Exercise 5.23, p.539], we have that the error averaged over the ensemble and messages is upper bounded by

$$\exp\left(-M(b-n_1)\left(\frac{(C-\frac{k+M^*}{M(b-n_1)})^2}{8/e^2+4(\ln 2)^2}\right)\right)\ . \tag{18}$$

To turn this into a bound on maximum error, we remove the half of the messages with highest error probability. We must guarantee that the error probability is small for all $M \in \{M_*, \ldots, M^*\}$, so it is sufficient to perform M^*

such thinning operations to obtain a codebook with 2^k messages. The loosest bound in (18) is for $M(b-n_1)=k$ in the exponent and $C-k/M(b-n_1)=(\epsilon_1-\epsilon_3)$, so

$$\mathbb{P}\left(E_2|E_1\right) \leq \exp\left(-k\left(\frac{(\epsilon_1-\epsilon_3-M^*/k)^2}{8/e^2+4(\ln 2)^2}\right)\right) \ . \tag{19}$$

Under the assumptions $\tau>\epsilon_1$, $\tilde{p}_{\rm emp}^{(M)} \notin [1/2-\tau, 1/2+\tau]$, and E_1^c, we know that $\bar{p}_{\rm emp}^{(M)}$ and $\tilde{p}_{\rm emp}^{(M)}$ must lie on the same side of $1/2$. Therefore our decoding rule can be a simple minimum-distance (if $\tilde{p}_{\rm emp}^{(M)}<1/2$) or maximum -distance (if $\tilde{p}_{\rm emp}^{(M)}>1/2$) decoding rule, which is as good as maximum-likelihood for these channels. ■

Remark: In order for the error events E_1 and E_2 to have small probability, we must set the parameters b, k, n_1, τ, ϵ_1, and ϵ_2 so that the errors in (10) and (13) go to zero as $N\to\infty$.

B. Rate analysis

There two types of rate loss – loss within a round and loss for the final uncompleted round.

Lemma 4 (Rate loss for uncompleted round): The fraction of channel uses lost $\gamma(N)$ due to a nonterminating final round is upper bound by

$$\gamma(N) \leq \frac{M^* b}{N} \tag{20}$$

Proof: The proof follows immediately from Lemma 1 since no round can be larger than M^* chunks. ■

Lemma 5 (Rate loss within a round): Suppose we are under the event E_1^c. Then for a round with a empirical noise frequency $\bar{p}_{\rm emp}^{(M)}$ ones in the noise sequence, the rate is at least

$$\frac{k}{(b-n_1)M} \geq 1-h(\bar{p}_{\rm emp}^{(M)})-\beta_0(N), \tag{21}$$

where $\beta_0(N)=\max\{h(M_*^{-1})+\epsilon_3+\epsilon_1+\frac{M_*}{(M_*-1)^2}, h(1/2-\tau-\epsilon_2)\}$.

Proof: We experience rate loss in both rounds that terminate due to "BAD NOISE" and ones in which we decode. In the "BAD NOISE" rounds, we set the rate to 0. Under E_1^c, this can only happen when $\bar{p}_{\rm emp}^{(M)} \in [\frac{1}{2}-\tau-\epsilon_2, \frac{1}{2}+\tau+\epsilon_2]$, so we have the lower bound

$$\frac{k}{(b-n_1)M} \geq 1-h(\bar{p}_{\rm emp}^{(M)})-h(1/2-\tau-\epsilon_2). \tag{22}$$

In other rounds, our loss comes from ϵ_1 and deviations in $\tilde{p}_{\rm emp}^{(M)}$. While satisfying (5) gives us an upper bound on the rate, we want a lower bound of the form given in (21). To get this, we will use the fact that (5) is not satisfied in chunk $M-1$, so

$$\frac{k}{(b-n_1)(M-1)} \geq 1-h(\bar{p}_{\rm emp}^{(M-1)})-\epsilon_1-\epsilon_3. \tag{23}$$

To get the bound, we must bound the change in rate from $M-1$ to M and the change in $\bar{p}_{\rm emp}^{(M-1)}$ to $\bar{p}_{\rm emp}^{(M)}$ to get the bound. From Lemmas 9 and 1, we have

$$\frac{k}{(b-n_1)M} \geq \frac{k}{(b-n_1)(M-1)} - \frac{M_*}{(M_*-1)^2}. \tag{24}$$

Combining Lemmas 8 and 1 with equations (23) and (24), we get

$$\begin{aligned}\frac{k}{(b-n_1)M} \geq & 1-h(\bar{p}_{\rm emp}^{(M)})-h(M_*^{-1}) \\ & -\epsilon_3-\epsilon_1-\frac{M_*}{(M_*-1)^2}.\end{aligned} \tag{25}$$

Combining the lower bounds (22) and (25), we get (21). ■

An example of the bound above is given in Figure 3.

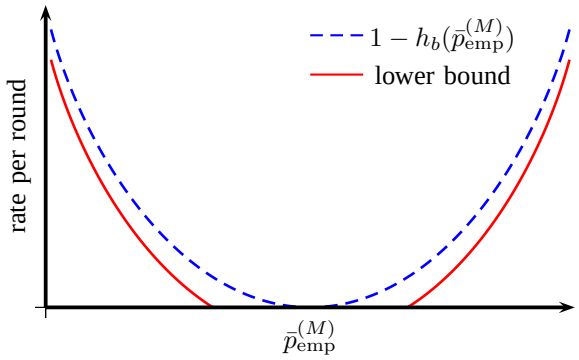

Fig. 3. Example of convex lower bound.

Lemma 6 (Overall rate loss): Under E_1^c, we have

$$\frac{k}{bM} \geq 1 - h(\bar{p}_{\rm emp}^{(M)}) - \beta(N), \tag{26}$$

where $\beta(N) = \max\{h(M_*^{-1}) + \epsilon_3 + \epsilon_1 + \frac{M_*}{(M_*-1)^2}, h(1/2 - \tau - \epsilon_2)\} + h(\frac{M^*b}{N}) + \frac{M^*b}{N}$.

Proof: One can express $p_{\rm emp}$ as a weighted sum of the $\bar{p}_{\rm emp}^{(M)}$ from each of the completed rounds and a penalty for the nonterminating round of no more than $\frac{M^*b}{N}$ by Lemma 4, where the weights correspond exactly to the fraction of chunks used for that round. Using the same weights on the empirical rates from each round gives us gives us the rate over the entire block N. Lemmas 4 and 5 along with the convexity of the $1 - h(p_{\rm emp})$ in $p_{\rm emp}$ gives

$$\frac{k}{bM} \geq 1 - h(\bar{p}_{\rm emp}^{(M)}) - \beta_0(N) - h(\frac{M^*b}{N}) - \frac{M^*b}{N} \ . \tag{27}$$

■

C. Setting the parameters

The algorithm as described has a large number of parameters which must satisfy various asymptotic conditions if we are to achieve $1 - h_b(p_{\rm emp})$. By way of an example, let us set

$k(N)$	$N^{1/2}$		$b(N)$	$N^{1/4}$
$n_1(N)$	$N^{1/8}$		ϵ_2	$\frac{1}{2}N^{-1/32}$
$\tau(N)$	$\max\{N^{-1/8}, \alpha(N)\}$		ϵ_1	$\frac{1}{2}N^{-1/8}$

Here, $\alpha(N)$ is the nonnegative solution to the equation $2\epsilon_4(N) = 1 - h_b(1/2 - \alpha(N))$. Since $h_b(\epsilon) = O(\epsilon \log \epsilon^{-1})$, these parameters give $\epsilon_3 = O(N^{-1/32} \log N)$ so $\epsilon_1 - \epsilon_3 > 0$. Therefore the errors in (10) and (13) both converge to 0. Furthermore, the rate loss $\beta(N) \to 0$ by Lemma 6. This is sufficient to complete the proof of Theorem 1.

IV. Discussion

In the model proposed by Shayevitz and Feder [2], the encoder has access to (passive) output feedback from the decoder that allows the encoder to provide control and estimation information in a set of training sequences that can be selected via common randomness. The flavor of the algorithm described here is different – in our scheme the decoder uses the feedback link to terminate rounds that are too noisy but otherwise attempts to correct the error in less noisy rounds. This can be viewed as a kind of hybrid ARQ [6] combining forward error correction and decision feedback on chunks.

A. Some comments on randomization

In our description of the algorithm, we assume that the encoder and decoder use $N/b \log 3$ bits of active feedback to sent the decisions after each chunk as well as common randomness to choose the training positions T_i and the random codebooks for each round. The decoder could use active feedback to inform the encoder of the n_1 training positions chunk i before chunk i is sent. Over the N/b chunks this would require an additional $N(n_1/b) \log N$ bits, which in our example of parameter setting is sublinear in N.

A more troublesome issue is the randomness used to generate the codebooks for each round. We can use a single codebook and choose a random permutation of the entries in each round, which would require $O(M^*(b - n_1) \log(M^*(b - n_1)))$ bits of common randomness per round, or $O(N \log N)$ bits overall. One way of reducing this randomness would be to prove the existence of nested *list-decodable codes* and use active feedback to provide a small secret key that can disambiguate the list. The decoder could feedback a request for retransmission of certain bits to disambiguate the list and the encoder could send the disambiguation information using a repetition code, as in the work of Shayevitz and Feder.

B. Future directions

As in the model and algorithm considered by Shayevitz and Feder, our algorithm does not rely heavily on the assumption of binary inputs and noise. The major feature needed is that the parameters of the channel can be estimated at the decoder using training sequences. Channels in which the capacity is achievable with a single input distribution fall into this category. Feedback may also be useful to change the codebook when there is not one universal capacity-achieving input distribution.

Rather than the very general individual sequence approach, we could consider a class of noise models whose p_{emp} varies in a piecewise-constant fashion. One could view this as a kind of "block fading" model for the binary additive channel. These models may also be related to the on-line estimation problems studied by Kozat and Singer [9]. For such models we could consider modifying out algorithm to adapt the value of k by trying to learn the coherence time of the channel. In the sense of competitive optimality, the competition class could be coding algorithms that know the coherence intervals exactly.

As in the work of Shayevitz and Feder [2], the algorithm described in this paper exploits the *local* variations in the empirical noise distribution $\bar{p}_{\mathrm{emp}}^{(M)}$ on the order of the round length. This in turn implies that our scheme should perform better than $1 - h_b(p_{\mathrm{emp}})$ for noise sequences which have local variation on that order. By decreasing k we can exploit finer variations at the expense of the error performance. Future work will include finer analysis of the scheme to make explicit these gains from local variability.

ACKNOWLEDGMENTS

We thank Ofer Shayevitz and Meir Feder for providing a preprint of their paper after their presentation of it at the Kailath Colloquium [10]. This work grew out of a presentation of that work for UC Berkeley's advanced information theory course EE290S. Special thanks go to the other students in that class for helpful discussions.

APPENDIX

Lemma 7: If

$$|p_1 - p_2| \leq \epsilon, \tag{28}$$

then

$$|h_b(p_1) - h_b(p_2)| \leq h_b(\epsilon). \tag{29}$$

Proof: $|h_b(p_1) - h_b(p_2)| = h_b(\epsilon)$ when $p_1 = 0$ and $p_2 = \epsilon$ as well as when $p_1 = 1 - \epsilon$ and $p_2 = 1$. For $p_1 \in [0, 1/2 - \epsilon/2]$, $|h_b(p_1) - h_b(p_1 + \epsilon)|$ is decreasing in p_1 and thus

$$|h_b(p_1) - h_b(p_1 + \epsilon)| \leq h_b(\epsilon). \tag{30}$$

For $p_1 \in [1/2 - \epsilon/2, 1 - \epsilon]$, $|h_b(p_1) - h_b(p_1 + \epsilon)|$ is increasing in p_1 and thus

$$|h_b(p_1) - h_b(p_1 + \epsilon)| \leq h_b(\epsilon). \tag{31}$$

Swapping p_1 for p_2 and vice versa along with identical arguments completes the proof. ■

Lemma 8:

$$h_b(\bar{p}_{\text{emp}}^{(n-1)}) \in [h_b(\bar{p}_{\text{emp}}^{(n)}) - h_b(1/n), h_b(\bar{p}_{\text{emp}}^{(n)}) + h_b(1/n)] \tag{32}$$

Proof: From the recursion

$$\bar{p}_{\text{emp}}^{(n)} = \frac{n-1}{n}\bar{p}_{\text{emp}}^{(n-1)} + \frac{1}{n}p_{\text{emp}}^{(n)}, \tag{33}$$

we know that

$$\bar{p}_{\text{emp}}^{(n)} \leq \bar{p}_{\text{emp}}^{(n-1)} + \frac{1}{n}, \tag{34}$$

$$\bar{p}_{\text{emp}}^{(n)} \geq \bar{p}_{\text{emp}}^{(n-1)} - \frac{1}{n}. \tag{35}$$

The result then follows immediately from Lemma 7. ■

Lemma 9: For $n \geq n_0$,

$$\frac{1}{n} \geq \frac{1}{n-1} - \frac{1}{(n_0-1)^2}. \tag{36}$$

Proof: This is easily verifiable. ■

REFERENCES

[1] I. Csiszár and P. Narayan, "Arbitrarily Varying Channels with Constrained Inputs and States," *IEEE Transactions on Information Theory*, vol. 34, no. 1, pp. 27–34, 1988.

[2] O. Shayevitz and M. Feder, "Achieving the empirical capacity using feedback part i: Memoryless additive models," submitted to IEEE Transactions on Information Theory. [Online]. Available: http://www.eng.tau.ac.il/˜ofersha/empirical_capacity_part1.pdf

[3] A. Sahai, "Why block length and delay are not the same thing," *IEEE Transactions on Information Theory*, Submitted 2006. [Online]. Available: http://www.eecs.berkeley.edu/˜sahai/Papers/FocusingBound.pdf

[4] A. Sahai and S. Draper, "Beating the Burnashev bound using noisy feedback," in *Proceedings of the Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Sept. 2006.

[5] A. Sahai, "Balancing forward and feedback error correction," in *Proc. of the Information Theory and Applications Workshop*, Jan. 2007.

[6] E. Soljanin, "Hybrid arq in wireless networks," in *DIMACS Workshop on Network Information Theory*, March 2003.

[7] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 1, pp. 13–30, March 1963.

[8] R. Gallager, *Information Theory and Reliable Communication*. New York: John Wiley and Sons, 1968.

[9] S. Kozat and A. Singer, "Universal Switching Linear Least Squares Prediction," in *Proc. of the 2006 Information Theory and its Applications Workshop*. La Jolla, CA: UCSD, February 2006.

[10] M. Feder, "Achieving the empirical capacity of individual noise channels using feedback," 2006 Kailath Symposium, July 2006.