

Limited feedback achieves the empirical capacity

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

The utility of limited feedback for coding over an individual sequence of DMCs is investigated. This study complements recent results showing how limited or noisy feedback can boost the reliability of communication. A strategy with fixed input distribution P is given that asymptotically achieves rates arbitrarily close to the mutual information induced by P and the state-averaged channel. When the capacity achieving input distribution is the same over all channel states, this achieves rates at least as large as the capacity of the state averaged channel, sometimes called the empirical capacity.

I. INTRODUCTION

Feedback plays a significant role in the design of most practical communication systems and has also inspired many results in the information theory literature. Information-theoretic results on feedback can be grouped into two categories: the effect of feedback on the capacity or reliability of communication and the use of feedback to overcome uncertainty about the channel model. In most practical communication systems, physical constraints limit the amount of feedback one can send, and therefore the *amount of*

Manuscript received October XX, 2007; revised XXXXXXXXXXXXXXXX.

Part of this work was presented at the 2007 International Symposium on Information Theory in Nice, France [1]

The authors are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley CA 94720-1770 USA.

The work of A.D. Sarwate and M. Gastpar was supported in part by the National Science Foundation under award CCF-0347298. The work of K. Eswaran, A. Sahai, and M. Gastpar was supported in part by the National Science Foundation under award CNS-0326503.

feedback required is a key concern. Thus, an understanding of how limited feedback affects the above settings provides insights into the design of practical systems.

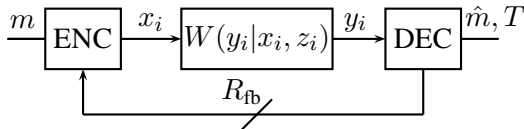


Fig. 1. Model setup with limited feedback.

Most studies about feedback for rate and reliability have centered around full output feedback [2]–[9]; however, recent work has started to improve our understanding of how limited feedback affects these performance measures. Noisy feedback increases reliability [10], [11] and the rate of some multiuser Gaussian channels [12], [13]. Furthermore, limited feedback can be used to improve reliability [14]. In contrast, channel uncertainty problems have focused almost exclusively on the case of full output feedback. Tchamkerten and Teletar explored one such model in which they show how to recover both the rate and reliability for an unknown discrete memoryless channel with feedback [15]. Shayevitz and Feder consider a more ambitious model, in which bits are sent across a modulo additive channel with a noise sequence that is fixed in advance but otherwise arbitrary. They present a strategy that uses full output feedback and limited common randomness to recover the *empirical capacity*, which they define as the capacity of an i.i.d. channel with transition probabilities corresponding to the empirical statistics of the noise sequence [16].

One attempt to understand the impact of limited feedback in overcoming channel uncertainty at the encoder is through rateless codes, which are a class of coding strategies that use limited feedback to adapt to unknown channel parameters. In a rateless code the decoder can use a low-rate feedback link to inform the encoder of when it has decoded. These codes were first studied in the context of the erasure channel [17], [18]. Later work focused on compound channels [15], [19], [20]. Draper et al. [21] investigated a model based on AVCs that is similar in spirit to the model considered by Shayevitz and Feder. However, the Draper et al. model assumes that full channel state information is available at the decoder, which does not capture the full extent of channel uncertainty studied by Shayevitz and Feder. Table I shows the relationship of the present work to the work of others.

In this paper, we show how we to achieve the empirical capacity for a model similar to Shayevitz and Feder’s but using a limited feedback strategy. To do this, we adapt the feedback-reducing block/chunk strategies used earlier in the context of reliability functions [9], [10], and most specifically in [14]. They

	channel model	feedback	state information
Shulman [19]	compound	full	none
Tchamkerten and Teletar [15]	compound	full	none
Draper, Frey, and Kschischang [21]	AVC	0-rate	at decoder
Shayevitz and Feder [16]	individual sequence	full	none
This paper	individual sequence	0-rate	none

TABLE I

RELATED RESULTS IN TERMS OF CHANNEL MODEL, FEEDBACK ASSUMPTIONS AND STATE INFORMATION ASSUMPTIONS.

are in turn inspired by Hybrid ARQ [22]. The flavor of our algorithm 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. By doing away with the output feedback, we lose some of the simplicity of the scheme in [16], but we show that a similar performance can still be obtained with almost negligible feedback.

In our scheme, the encoder attempts to send k bits over the channel during a variable length *round*. The encoder sends *chunks* of the codeword to the decoder, after which the decoder feeds back a decision as to whether it can decode. The encoder and decoder use common randomness to choose a set of randomly chosen *training* positions during which the encoder sends a fixed message. The decoder uses the training positions to estimate the channel. If the number of bits than can be transmitted over a channel with the estimated *empirical mutual information* exceeds k , then the decoder attempts to decode. Through this combination of training-based channel estimation and robust decoding we can exploit the limited feedback to achieve rates asymptotically equal to those with advance knowledge of the average channel.

In the next section, we motivate the study of this problem with some concrete examples. In Section III, we define the channel model, state our main result, and describe the coding strategy. Section IV contains the analysis of our strategy with most of the technical details reserved for the Appendix.

II. MOTIVATING EXAMPLES

Before introducing notation and stating our main result, we provide some examples to motivate the present study. We will revisit these examples in Section III after stating our main result.

A. Binary modulo-additive channels

The simplest example of a channel with individual noise sequences is the binary modulo-additive channel. This channel takes binary inputs and produces binary outputs, where the output is produced by potentially flipping some bits of the channel input. These flips do not depend on the channel input symbols. The output $\mathbf{y} \in \{0, 1\}^N$ can be written as

$$\mathbf{y} = \mathbf{x} \oplus \mathbf{z}, \quad (1)$$

where $\mathbf{x} \in \{0, 1\}^N$ is the channel input, $\mathbf{z} \in \{0, 1\}^N$ is the noise sequence, and addition is carried out modulo-2. \mathbf{z} is arbitrary but fixed, and we let p be the empirical fraction of 1's in \mathbf{z} , which is arbitrary but fixed over the $[0, 1]$ interval.

For this setup, we would like to compare the rates achievable with limited feedback to a binary symmetric channel with crossover probability p among the class of binary symmetric channels. Note that for this class of channels, the capacity achieving input distribution remains the same regardless of the value of the underlying parameter.

B. Spectrum sharing channel

Consider the following model of a wireless channel with additive interference:

$$y_i = x_i + \tilde{Z}_i + W_i. \quad (2)$$

We assume binary modulation, with input $x_i \in \{-\sqrt{P}, \sqrt{P}\}$ and iid noise $W_i \sim \mathcal{N}(0, 1)$. The interfering signal $\tilde{Z}_i \sim \mathcal{N}(0, \sigma_i^2)$ corresponds to interference from multiple systems. Since these systems may use the channel intermittently and the interference they generate can fluctuate over time, the noise variance σ_i^2 can be modeled as an arbitrary but fixed individual sequence.

For the spectrum sharing model just described, we would like to compare rates achievable with limited feedback to a corresponding channel among the class of binary input additive white Gaussian noise channels with noise variance $\sigma^2 + 1$, where σ^2 varies over the class. Note that the capacity achieving input distribution remains fixed over the class.

C. Z-channels with unknown crossover

Consider a channel for which $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0, 1\}$.

$$y = \begin{cases} x & z = 0 \\ 1 & z = 1. \end{cases} \quad (3)$$

Again, the state sequence \mathbf{z} is arbitrary but fixed, and we let q denote the empirical fraction of 1's in \mathbf{z} .

Again for reference, we want to compare the rates achievable with limited feedback against the corresponding the Z-channel with crossover probability q . Unlike the previous examples, this channel has a capacity achieving input distribution that depends on q . The optimal choice of input distribution will depend on other knowledge or design goals. For example, we may have bounds on q *a priori* and may wish to minimize the maximum rate loss.

III. THE CHANNEL MODEL AND CODING STRATEGY

A. Channel models

The problem we consider in this paper is that of communicating over a channel with an individual state sequence. Let the finite sets \mathcal{X} and \mathcal{Y} denote the channel input and output alphabets, respectively. The set $\mathcal{W} = \{W(y|x, z) : z \in \mathcal{Z}\}$ is a set of channels indexed by a state variable in a (not necessarily finite) set \mathcal{Z} . We model our channel as having an individual state sequence $\mathbf{z} = (z_1, z_2, \dots, z_N)$, so we can write the overall channel as

$$W(\mathbf{y}|\mathbf{x}, \mathbf{z}) = \prod_{i=1}^N W(y_i|x_i, z_i) .$$

Indeed, \mathcal{Z} can be large enough to accommodate all possible DMCs with input alphabet \mathcal{X} and output alphabet \mathcal{Y} .

Because the maximum capacity of this set of channels is $C_{\max} = \log \min\{|\mathcal{X}|, |\mathcal{Y}|\}$, we define the set of possible messages to be the set of all binary sequences $\{0, 1\}^{NC_{\max}}$. This message set is naturally nested – the truncated set $\{0, 1\}^T$ is a set of prefixes for $\{0, 1\}^{NC_{\max}}$. At the time of decoding, the decoder will decide on a truncation $T \in \mathbb{N}$ and a message $m \in \{0, 1\}^T$. We think of the rate-limited feedback link as a noiseless channel that can be used every n_{fb} uses of the forward channel to send B_{fb} bits. The rate of the feedback is $R_{\text{fb}} = B_{\text{fb}}/n_{\text{fb}}$. To avoid integer effects, we will consider only integer values for n_{fb} and B_{fb} . This will not affect our results.

A *coding strategy* for blocklength N consists of a sequence of (possibly random) encoding functions for $i = 1, 2, \dots, N$,

$$\eta_i : \{0, 1\}^{NC_{\max}} \times \{0, 1\}^{\lfloor i/n_{\text{fb}} \rfloor B_{\text{fb}}} \rightarrow \mathcal{X} , \quad (4)$$

a sequence of (possibly random) feedback functions for $i = n_{\text{fb}}, 2n_{\text{fb}}, \dots$:

$$\phi_i : \mathcal{Y}^i \rightarrow \{0, 1\}^{B_{\text{fb}}} , \quad (5)$$

and a decoding function

$$\psi : \mathcal{Y}^N \rightarrow \{0, 1, \dots, NC_{\max}\} \times \{0, 1\}^{NC_{\max}} , \quad (6)$$

We define the decoding threshold T and message estimate $\hat{\mathbf{m}}$ (which are random variables) by $\psi(Y^n) = (T, \hat{\mathbf{m}})$. We define the *maximal error probability* to be

$$\max_{\mathbf{m} \in \{0, 1\}^{NC_{\max}}} \mathbb{P}((m_1, m_2, \dots, m_T) \neq (\hat{m}_1, \hat{m}_2, \dots, \hat{m}_T) \mid \mathbf{m} \text{ was encoded}) , \quad (7)$$

where the probability is taken over the common randomness and randomness in the channel.

The *empirical rate* is T/N .

For an individual state sequence \mathbf{z} we can define the *state-averaged channel* to be

$$W_{\mathbf{z}}(y|x) = \frac{1}{N} \sum_{i=1}^N W(y|x, z_i) . \quad (8)$$

For a fixed input distribution $P(x)$ on \mathcal{X} and averaged channel $W_{\mathbf{z}}(y|x)$, the *mutual information* is given by the usual definition:

$$I(P, W) = \sum_{x, y} W(y|x) P(x) \log \frac{W(y|x) P(x)}{P(x) \sum_{x'} W(y|x') P(x')} .$$

For an individual state sequence we can define the *empirical mutual information* by $I(P, W_{\mathbf{z}})$. In our algorithm we will estimate $W_{\mathbf{z}}(y|x)$ to get an estimate of $I(P, W_{\mathbf{z}})$ so the decoder can choose an appropriate time to decode.

For a fixed \mathbf{z} , the *empirical capacity* is the supremum over all input distributions of the empirical mutual information:

$$\bar{C}(\mathbf{z}) = \sup_{P(x)} I(P, W_{\mathbf{z}}) .$$

In general, the maximizing $P(x)$ may not be the same for all \mathbf{z} , and in these cases our strategy can achieve rates close to $I(P, W_{\mathbf{z}})$ but not $\bar{C}(\mathbf{z})$. Some illuminating examples are given in the next section to make the distinction more clear.

Our coding strategy exploits local variation in the channel behavior. Let \mathbf{z} be an individual state sequence and let $\mathbf{z}_m = (z_{j_m+1}, \dots, z_{j_{m+1}})$, where $0 = j_1 < j_2 < \dots < j_M = N$. Averaging over the state as in (8), we can form the average channels $W_{\mathbf{z}_m}(y|x)$ for $m = 1, 2, \dots, M$. Because the mutual information is convex in the channel matrix, we have the following inequality:

$$I(P, W_{\mathbf{z}}) \leq \sum_{m=1}^{M-1} \frac{j_{m+1} - j_m}{N} I(P, W_{\mathbf{z}_m}) .$$

That is, time-sharing the empirical mutual information over over sub-blocks yields a higher rate than the empirical mutual information over all sub-blocks. Therefore if we can achieve rates close to $I(P, W_{\mathbf{z}_j})$ in each sub-block from j_{m-1} to j_m , the overall rate may even exceed $I(P, W_{\mathbf{z}})$

B. Main Result

The main result in this paper is that the algorithm given in the next section achieves rates that asymptotically approach the mutual information $I(P, W_{\mathbf{z}})$ for a large set of state sequences \mathbf{z} .

Theorem 1: When used over a family of channels $\{W(y|x, z) : z \in \mathcal{Z}\}$ with finite input and output alphabets, there is a coding strategy that with probability $1 - \varepsilon(N)$ achieves the rate

$$R \geq I(P, W_{\mathbf{z}}) - \rho(N) , \quad (9)$$

with feedback rate

$$R_{\text{fb}} = \lambda(N) . \quad (10)$$

Furthermore, as $N \rightarrow \infty$ we have $\rho(N) \rightarrow 0$, $\lambda(N) \rightarrow 0$, and $\varepsilon(N) \rightarrow 0$.

Let us return to our three motivating examples. For the binary modulo-additive channel, Theorem 1 implies the following result. For this case, the empirical capacity is $1 - h(p)$, the capacity of the binary symmetric channel with crossover probability p .

Corollary 1: For the binary modulo-additive channel with an individual noise sequence, there is a coding strategy that with probability $1 - \varepsilon(N)$ achieves the rate

$$R \geq 1 - h(p) - \rho(N) , \quad (11)$$

with feedback rate

$$R_{\text{fb}} = \lambda(N) . \quad (12)$$

and $h(\cdot)$ is the binary entropy function. Furthermore, as $N \rightarrow \infty$ we have $\rho(N) \rightarrow 0$, $\lambda(N) \rightarrow 0$, and $\varepsilon(N) \rightarrow 0$.

For the spectrum sharing channel, our results do not apply directly because the channel output alphabet is continuous for that setting. However, in a real system, the channel output may be quantized (perhaps to high accuracy), from which we can derive a corresponding channel with discrete inputs and outputs. In this case, we can apply the result in Theorem 1 to design a coding strategy that achieves the empirical capacity of the corresponding discrete output channel.

Finally, consider the channel that can force 0's to crossover to 1, with an arbitrary but fixed crossover sequence, where we let q be the arbitrary but fixed fraction of 1's in the crossover sequence. Note that if the sequence is chosen iid with crossover probability q , then this corresponds to a Z-channel with crossover probability q . For this channel, the capacity achieving input distribution is a function of q , so our scheme cannot achieve the empirical capacity. Despite this, our results still allow us to state the rates we can attain in this setting. If the channel input distribution $P(X = 1) = p$ for this channel, then the empirical mutual information for this channel can be written as

$$I(P, W_q) = h(p) - (1 - p + pq)h\left(\frac{pq}{1 - p + pq}\right), \quad (13)$$

and is achievable from Theorem 1.

C. Proposed coding strategy

We divide the blocklength N into *chunks* of length $b = b(N)$. Feedback occurs at the end of chunks, so $n_{\text{fb}} = b$ with three possible messages: “BAD NOISE”, “DECODED”, and “KEEP GOING”.

The encoder attempts to send $k = k(N)$ bits over several chunks comprising a *round*. Let $V_n = (n-1)b+1, (n-1)b+2, \dots, nb$ be the time indices in the n -th chunk within a round. For each chunk n , the decoder and encoder choose $t = t(N)$ *training positions* T_n (via common randomness) during which a known sequence is transmitted to enable the decoder to estimate the empirical channel. The remaining time indices $U_n = V_n \setminus T_n$ are used to transmit the codeword. Let $\mathcal{V}_n = V_1, \dots, V_n$, $\mathcal{T}_n = T_1, \dots, T_n$, and $\mathcal{U}_n = U_1, \dots, U_n$ be the time indices up to the n -th chunk for the round, training, and codeword positions, respectively.

We fix an input distribution $P(x)$ on \mathcal{X} . 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 $b - t$ and transmits the n -th segment over the $b - t$ non-training positions in U_n .

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

A formal description of the coding strategy follows, and an illustration is provided in Figure 2. At the beginning of round r , the encoder and decoder use common randomness to choose a random codebook of type P to be used in that round. Let $\mathbf{x}(r)$ denote the codeword to be sent in round r . Let $\mathbf{z}_n = \mathbf{z}_{n(r)}$ denote the state sequence during the n -th chunk of round r . We will suppress the dependence on r for

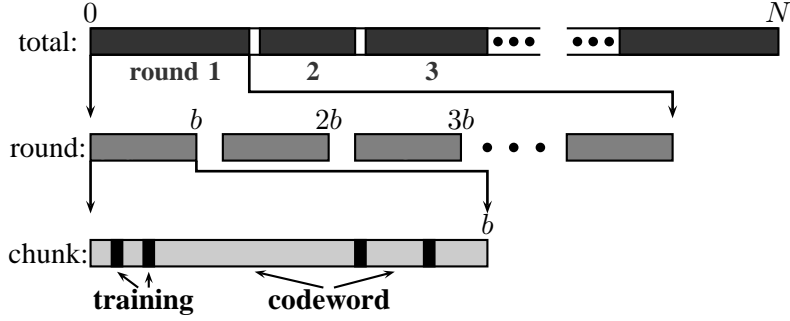


Fig. 2. After each chunk of length b feedback can be sent. Rounds end by decoding a message or declaring the noise to be bad.

simplicity. For indices $\mathcal{S} = s_1, s_2, \dots, s_{|\mathcal{S}|}$, we will let $\mathbf{z}(\mathcal{S}) = z_{s_1}, z_{s_2}, \dots, z_{s_{|\mathcal{S}|}}$, so $\mathbf{z}(T_n)$ is the state vector during the training in chunk n , $\mathbf{z}(\mathcal{T}_n)$ is the state during the training positions, $\mathbf{z}(\mathcal{U}_n)$ is the state during the non-training positions, and $\mathbf{z}(\mathcal{V}_n)$ is the state vector of the current round up to the n -th chunk.

For each round, the following steps are repeated for each chunk:

- 1) The encoder and decoder choose t positions T_n to use for the training in chunk n using common randomness. T_n is further partitioned into $|\mathcal{X}|$ subsequences $T_n(x)$ of size $t/|\mathcal{X}|$ positions.
- 2) The encoder transmits the chunk. At times $j \in T_n(x)$ the encoder sends x . In the $b - t$ remaining positions the encoder sends $(x_{(n-1)(b-t)+1}, x_{(n-1)(b-t)+1}, \dots, x_{n(b-t)})$, which are the next $b - t$ entries in the codeword corresponding to the k bits to be sent in the current round.
- 3) The decoder estimates the empirical channel $W_{\mathbf{z}(\mathcal{U}_n)}(y|x)$ in chunk n and the empirical channel over the round so far:

$$\hat{w}^{(n)}(y|x) = \frac{|\mathcal{X}|}{t} \cdot |\{j \in T_n(x) : y_j = y\}|$$

$$\hat{W}^{(n)}(y|x) = \frac{1}{n} \sum_{i=1}^n \hat{w}^{(i)}(y|x) .$$

- 4) The decoder makes a decision based on $\hat{W}^{(n)}$ and n :
 - a) if

$$I(P, \hat{W}^{(n)}) < \tau(N) , \quad (14)$$

then the decoder feeds back “BAD NOISE” and the round is terminated without decoding the k bits. In the next round, the encoder will attempt to resend the k bits from this round.

b) if

$$\frac{k}{(b-t) \times n} < I(P, W_{\mathbf{z}(\mathcal{I}_n)}) - \epsilon_1(N) , \quad (15)$$

then the decoder decodes, feeds back "DECODED," and the encoder starts a new round.

c) otherwise the decoder feeds back "KEEP GOING" and goes to 2).

The coding strategy uses $\log 3$ bits of feedback per chunk for the decision messages ("BAD NOISE," "DECODED," and "KEEP GOING") and common randomness to choose a new codebook for each round as well as training for each chunk. Observe that letting b get large with N causes the feedback rate to go to zero.

Our strategy has two main ingredients. First, the encoder uses random training sequences to let the decoder accurately estimate the empirical average channel. Given this accurate estimate, the decoder can track the empirical mutual information of the channel over the round. Second, the decoder only needs to know that the empirical mutual information is large enough in order guarantee a small error probability. To accomplish this we use a fixed-composition codebook and a maximum mutual information decoding rule.

IV. 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 average channel (Lemma 2) and that the condition in (15) 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 must first find a bound on the length of a round in chunks. Let M be the termination time for the round :

$$M = \inf_{n>0} \left\{ I(P, \hat{W}^{(n)}) < \tau \text{ or } \frac{k}{(b-t)n} < I(P, \hat{W}^{(n)}) - \epsilon_1 \right\} . \quad (16)$$

We now argue that M cannot be too large.

Lemma 1 (Bounds on M): If $\epsilon_1 < \tau$, we have $M \leq M^*$, where

$$M^* := \left\lceil \frac{k}{(b-t) \cdot (\tau - \epsilon_1)} \right\rceil . \quad (17)$$

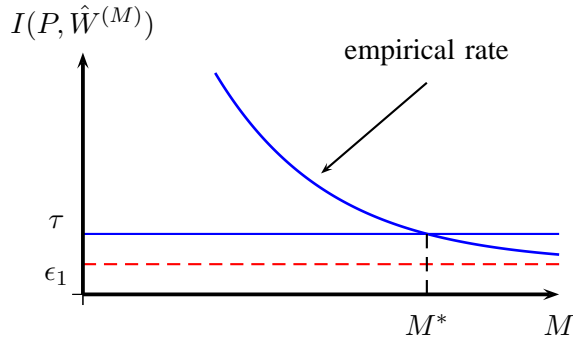


Fig. 3. Curve illustrating why M is finite.

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

$$M_* = \frac{k}{(b-t) \cdot C_{\max}} . \quad (18)$$

Proof: The argument is illustrated in Figure 3. We must simply find the point where the curve defined by (15) intersects the “BAD NOISE” threshold. This gives the bound in (17). The lower bound is trivial from the definition in (15) and the cardinality bound on mutual information. ■

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

- 1) (E_1) We declare an error if

$$\max_{x,y} \left| \hat{W}^{(M)}(y|x) - W_{\mathbf{z}(\mathcal{U}_M)}(y|x) \right| > \epsilon_2(N) . \quad (19)$$

This happens when the estimated channel is not sufficiently close to the average channel in the non-training positions \mathcal{U}_M .

- 2) (E_2) We will have an error if the decoder does not output the correct transmitted message. We call this a decoder *failure*.

We will need the following lemma, which states that with high probability the channel $\hat{W}^{(M)}$ estimated from the training is close to the average channel $W_{\mathbf{z}(\mathcal{V}_M)}(y|x)$ over the entire round and the average channel $W_{\mathbf{x}(\mathcal{U}_M)}(y|x)$ over which the codeword is transmitted. The proof is provided in the Appendix.

Lemma 2 (Channel estimation via training): Let $\epsilon_2 > t/b$. Then there exist constants α_1 and β_1 such that

$$\mathbb{P} \left(\max_{x,y} \left| \hat{W}^{(n)}(y|x) - W_{\mathbf{z}(\mathcal{V}_n)}(y|x) \right| > \epsilon_2 \right) \leq \alpha_1 n \exp(-\beta_1 t \epsilon_2^2) \quad (20)$$

$$\mathbb{P} \left(\max_{x,y} \left| \hat{W}^{(n)}(y|x) - W_{\mathbf{x}(\mathcal{U}_n)}(y|x) \right| > \epsilon_2 \right) \leq \alpha_1 n \exp(-\beta_1 t (\epsilon_2 - 2t/b)^2) . \quad (21)$$

Our next error lemma bounds the probability that the decoder fails conditioned on the channel estimates being accurate. We relegate the proof to the Appendix.

Lemma 3 (Bound on E_2): There exist constants $\alpha_3, \alpha_4, \alpha_5$, and $\alpha_6 > 0$, such that if

$$\epsilon_3(N) = \alpha_3 h_b(\epsilon_2) + \alpha_4 \epsilon_2, \quad (22)$$

then for $\epsilon_1 = \epsilon_1(N)$ chosen such that

$$\zeta(N) = \epsilon_1 - \epsilon_3 - \frac{\alpha_5 M^* \log(b-t) + \alpha_6 \log M^*}{M_*(b-t)} > 0, \quad (23)$$

where $0 < \eta(P) < \infty$, $\alpha = \alpha(|\mathcal{X}|, |\mathcal{Y}|) < \infty$, with probability

$$1 - M^* \exp\left(-\frac{1}{8} 2^{M^*+1} ((b-t)+1)^{-\eta M^*} \exp(M^*(b-t)R)\right), \quad (24)$$

a randomly chosen codebook will satisfy

$$\mathbb{P}(E_2|E_1^c, M) \leq 2^M 3((b-t)+1)^{\eta M} (M(b-t)+1)^\alpha \exp\left(-M(b-t) \frac{\zeta(N)^2}{8/e^2 + 4(\ln|\mathcal{Y}|)^2}\right). \quad (25)$$

Note that in order to guarantee our construction will work we must choose our parameters such that the probability in (24) is positive. We denote the overall error by $\varepsilon = \varepsilon(N) \leq \mathbb{P}(E_1) + \mathbb{P}(E_2|E_1^c)$. The bounds in (20) and (25) provide an upper bound on ε in terms of the other parameters. Note that we must also choose the parameters such that (70) will guarantee the existence of a codebook with our desired properties.

B. Rate analysis

In this section we analyze the gap between the rates achieved by our algorithm and the empirical mutual information $I(P, W_{\mathbf{z}})$. Within each round, the decoder is somewhat conservative, gathering enough packets so that the achieved rate is less than ϵ_1 of the the measured mutual information and less than $\epsilon_1 - \epsilon_3$ of empirical mutual information during the the non-training positions. In this section we will calculate the gap between the rates achieved by our algorithm and the overall empirical mutual information. This gap is due to two factors : the loss within each round, and the 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

$$\gamma(N) \leq \frac{M^* b}{N}. \quad (26)$$

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

The next lemma bounds the rate loss within a round, both for rounds terminated due to “BAD NOISE” and rounds in which the decoder decodes. The proof is given in the appendix.

Lemma 5 (Rate loss within a round): Suppose we are under the event $(E_1 \cup E_2)^c$. Let R be the rate achieved within a round, and define

$$\rho_0 = \max \left\{ \tau + \epsilon_3, \epsilon_1 + \epsilon_3 + |\mathcal{Y}|h_b(2M_*^{-1}) + 2|\mathcal{Y}| \log |\mathcal{Y}|M_*^{-1} + \frac{k}{b-t} \cdot \frac{1}{(M_* - 1)^2} \right\}, \quad (27)$$

where

$$\epsilon_3(N) = \alpha_3 h_b(\epsilon_2) + \alpha_4 \epsilon_2 \quad (28)$$

for constants $0 < \alpha_3, \alpha_4 < \infty$. Then the difference between the achieved rate in any round and the empirical mutual information is bounded:

$$I(P, W_{\mathbf{z}(\mathcal{V}_M)}) - R \leq \rho_0. \quad (29)$$

Lemma 6 (Overall rate loss): Suppose we are under $(E_1 \cup E_2)^c$, and let the total number of successfully decoded bits be RN . Let

$$\rho = \frac{M^*b}{N} \log |\mathcal{Y}| + \rho_0. \quad (30)$$

Then

$$I(P, W_{\mathbf{z}}) - R \leq \rho.$$

Proof: Let \mathbf{z}_j denote the state vector in round j , and suppose there were J rounds, with the J -th round possibly failing to terminate. Let R_j denote the rate achieved in round j and $M_j b$ the length of round j . Then using the convexity of the mutual information, Lemmas 4 and 5, we obtain:

$$I(P, W_{\mathbf{z}}) \leq \sum_{j=1}^J \frac{M_j b}{N} I(P, W_{\mathbf{z}_j}) \quad (31)$$

$$\leq \frac{M_j b}{N} I(P, W_{\mathbf{z}_j}) + \sum_{j=1}^{J-1} \frac{M_j b}{N} (R_j + \rho_0) \quad (32)$$

$$\leq \frac{M^* b}{N} \log |\mathcal{Y}| + R + \rho_0. \quad (33)$$

The lemma now follows from this inequality. ■

C. Setting the parameters

The coding strategy has a large number of parameters that must satisfy various asymptotic conditions if we are to achieve the empirical mutual information of the channel. By way of an example, let us set

transmitted bits per round	$k(N)$	$\Theta(N^{1/2})$
chunk size	$b(N)$	$\Theta(N^{1/4})$
training size	$t(N)$	$\Theta(N^{1/8})$
channel estimate gap	$\epsilon_2(N)$	$\Theta(N^{-1/32})$
bad round threshold	$\tau(N)$	$\Theta(N^{-1/8})$
decoding gap threshold	$\epsilon_1(N)$	$\Theta(N^{-1/16})$

With this setting, we have

minimum number of chunks	M_*	$\Theta(N^{1/4})$
maximum number of chunks	M^*	$\Theta(N^{3/8})$
estimation error probability	$\mathbb{P}(E_1)$	$O(\exp(-N^{1/16}))$
decoding failure probability	$\mathbb{P}(E_2 E_1^c)$	$O(\exp(-N^{7/16}))$
total error	$\epsilon(N)$	$O(\exp(-N^{1/16}))$
rate loss per round	ρ_0	$O(N^{-1/32} \log N)$
total rate loss	$\rho(N)$	$O(N^{-1/32} \log N)$

To complete the proof of Theorem 1 we must verify that we satisfy the conditions guaranteeing the existence of our codebook. The probability of a good codebook existing given in (24) is positive. The term $\exp(M^*(b-t)R)$ dominates the exponent, so the probability of the codebook existing goes to 1. The other condition to check is (23), which guarantees the positivity of the error exponent in the probability of error. This too, is satisfied by the settings above. Finally, we note that our rate loss $\rho(N)$ is defined to be at least as large as $\tau(N)$. For state sequences whose induced mutual information is smaller than $\tau(N)$, condition (9) is vacuous – the bound is negative, so the achieved rate is 0.

There is a range of parameter settings for which all of the relevant quantities can be driven to 0. The above is just one example, but others can be better depending on the code designer's priorities.

V. DISCUSSION

In this paper we described a coding strategy under a general channel uncertainty model that uses limited feedback to achieve rates arbitrarily close to an i.i.d. discrete memoryless channel with the same first-order statistics. Feedback allows the system to adapt the coding rate based on the channel conditions. When the class of channels over which we are uncertain has the same capacity achieving input distribution, the coding strategy achieves rates at least as large as empirical capacity, the capacity of an i.i.d. discrete memoryless channel with the same first-order statistics. Since the rates that we can guarantee for our

scheme are close to the average channel in a round, our total rate over many rounds may in fact exceed the empirical capacity. This is due to the convexity of mutual information in the channel.

The work extends an earlier investigation by Shayevitz and Feder [16] that considered the case in which the encoder has access to full output feedback from the decoder and allows the encoder to provide control and estimation information in a set of training sequences that can be selected via common randomness. By contrast, our strategy can be viewed as a kind of incremental redundancy hybrid ARQ [22], in which the decoder uses the feedback link to terminate rounds that are too noisy but otherwise attempts to correct the error in less noisy rounds.

Thus far, we have not accounted for the amount of common randomness required by our coding strategy. If the common randomness is sent via the feedback link, it is important for it to be sublinear in N to preserve asymptotically 0-rate feedback. We use common randomness to choose the locations for the training sequences as well as the codebook for each round. The decoder could use active feedback to inform the encoder of the t training positions chunk i before chunk i is sent. Over the N/b chunks this would require an additional $N \frac{t(N)}{b(N)} \log N$ bits, which in our example of parameter setting is sublinear in N .

We can also use a sublinear number of bits to choose a codebook to use in each round. One approach is to use tools from the theory of arbitrarily varying channels to find nested code constructions that use a limited amount of common randomness. This approach is taken in [23]. Another method, more in the interactive coding spirit of feedback systems, is to show the existence of deterministic list decodable codes with small list sizes. If the list is of size L , the decoder can find L bits in the message which can disambiguate the list. By using $L \log k$ bits in the feedback, the decoder can request those L bits from the encoder. By sacrificing just $\log N$ more forward channel uses, the encoder can send the L bits with negligible impact to the rate. Furthermore, success is guaranteed as long as the empirical mutual information in the next round is above τ . List coding constructions are also investigated in [23] and may be adapted for this purpose.

While it is true that the approach here can immediately be extended to exploit memory in the channel by looking at mini-segments of channel uses together and letting the mini-segments also grow slowly with N , such a result is not satisfying. In the memoryless context, being forced to declare an input distribution in advance seems reasonable. But the same assumption feels overly constraining when there is memory in the channel. The key underlying question seems to be how to adapt the input distribution appropriately and thus implicitly, what a reasonable class of competitors is.

The individual sequence model considered in this paper is by no means the only way of modeling

channel uncertainty. For the forward channel, an alternative is to consider a class of noise models that varies in a piecewise-constant fashion. This model is related to the on-line estimation problems studied by Kozat and Singer [24] and may be useful to understand block fading. For such models we could consider modifying our strategy 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 strategies that know the coherence intervals exactly. Variations on the model of the feedback link may also lead to interesting new results. Alternative channel models in which the feedback is noisy or allowed to have time-varying rate may present new issues to consider, particularly for the case in which there is uncertainty in the feedback link as well. One interesting model may be a two-way channel with individual noise, in which a feedback message may interact with a forward transmission. For future communications systems that must share common resources, such investigations may shed new light on strategies in these settings.

ACKNOWLEDGMENTS

We thank Ofer Shayevitz and Meir Feder for providing a preprint of their paper after their presentation of it at the Kailath Colloquium [25]. 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

We provide here the proofs of the lemmas used in the analysis of our algorithm¹.

A. Bounds on entropy and mutual information

We need a short technical lemma about concave functions.

Lemma 7: Let f be a concave increasing function on $[a, b]$. Then if $a \leq x \leq x + \epsilon \leq b$, we have

$$f(x + \epsilon) - f(x) \leq f(a + \epsilon) . \quad (34)$$

¹We were unable to find a standard reference for the entropy bounds below, which is why we provide the derivation.

Proof: Without loss of generality we can take $a = 0$, $b = 1$, and $f(a) = 0$. Now consider

$$\begin{aligned} f(x) &= f\left(\frac{x}{x+\epsilon} \cdot (x+\epsilon) + \frac{\epsilon}{x+\epsilon} \cdot 0\right) \geq \frac{x}{x+\epsilon}f(x+\epsilon) + \frac{\epsilon}{x+\epsilon}f(0) \\ &= \frac{x}{x+\epsilon}f(x+\epsilon) \\ f(\epsilon) &= f\left(\frac{x}{x+\epsilon} \cdot 0 + \frac{\epsilon}{x+\epsilon} \cdot (x+\epsilon)\right) \geq \frac{x}{x+\epsilon}f(0) + \frac{\epsilon}{x+\epsilon}f(x+\epsilon) \\ &= \frac{\epsilon}{x+\epsilon}f(x+\epsilon) . \end{aligned}$$

Therefore

$$f(x) + f(\epsilon) \geq f(x+\epsilon) , \quad (35)$$

as desired. ■

Using the preceding lemma, we can show that a bound on the total variational distance between two distributions gives a bound on the entropy between those two distributions.

Lemma 8: Let P and Q be two distributions on a finite set \mathcal{S} with $|\mathcal{S}| \geq 2$. If

$$|P(s) - Q(s)| \leq \epsilon \quad \forall s \in \mathcal{S} , \quad (36)$$

then

$$|H(P) - H(Q)| \leq (|\mathcal{S}| - 1) \cdot h_b(\epsilon) + (|\mathcal{S}| - 1) \log(|\mathcal{S}| - 1) \cdot \epsilon , \quad (37)$$

where $h_b(\cdot)$ is the binary entropy function.

Proof: Let $\mathcal{S} = \{s_1, s_2, \dots\}$. We proceed by induction on $|\mathcal{S}|$. Suppose $|\mathcal{S}| = 2$, and let $p = P(s_1)$ and $q = Q(s_1)$. The entropy function $h_b(x)$ is concave, increasing on $[0, 1/2]$ and decreasing on $[1/2, 1]$. Applying Lemma 7 to each interval, we obtain the bound:

$$|h_b(x+\epsilon) - h_b(x)| \leq h_b(\epsilon) . \quad (38)$$

Since $H(P) = h_b(p)$ and $H(Q) = h_b(q)$, this proves our result.

Now suppose that the lemma holds for $|\mathcal{S}| \leq m - 1$, and consider the case $|\mathcal{S}| = m$. Without loss of generality, let $P(s_m) > 0$ and $Q(s_m) > 0$. Let $\lambda = (1 - P(s_m))$ and $\mu = (1 - Q(s_m))$ and note that $|\lambda - \mu| < \epsilon$ by assumption. Define the $(m - 1)$ dimensional distributions $P' = \lambda^{-1}(P(s_1), \dots, P(s_{m-1}))$ and $Q' = \lambda^{-1}(Q(s_1), \dots, Q(s_{m-1}))$, so that

$$P = (\lambda P', (1 - \lambda))$$

$$Q = (\mu Q', (1 - \mu)) .$$

Therefore,

$$\begin{aligned} H(P) &= h_b(\lambda) + \lambda H(P') \\ H(Q) &= h_b(\mu) + \mu H(Q') . \end{aligned}$$

Now we can expand the difference of the entropies, using the fact that $\lambda < 1$, the induction hypothesis on $|H(P') - H(Q')|$ and $|h_b(\lambda) - h_b(\mu)|$, and the cardinality bound on the entropy $H(Q')$ to obtain

$$\begin{aligned} |H(P) - H(Q)| &= |\lambda H(P') - \mu H(Q') + h_b(\lambda) - h_b(\mu)| \\ &\leq \lambda |H(P') - H(Q')| + |\lambda - \mu| H(Q') + |h_b(\lambda) - h_b(\mu)| \\ &\leq (m-2) \cdot h_b(\epsilon) + (m-2) \log(m-2) \cdot \epsilon + \log(m-1) \cdot \epsilon + h_b(\epsilon) \\ &\leq (m-1) \cdot h_b(\epsilon) + (m-1) \log(m-1) \cdot \epsilon . \end{aligned}$$

■

Lemma 9: Let $W(y|x)$ and $V(y|x)$ be two channels with finite input and output alphabets \mathcal{X} and \mathcal{Y} . If

$$|W(y|x) - V(y|x)| \leq \epsilon \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y} , \quad (39)$$

then for any input distribution P on \mathcal{X} we have

$$|I(P, W) - I(P, V)| \leq 2(|\mathcal{Y}| - 1) \cdot h_b(\epsilon) + 2(|\mathcal{Y}| - 1) \log(|\mathcal{Y}| - 1) \cdot \epsilon , \quad (40)$$

where $h_b(\cdot)$ is the binary entropy function.

Proof: We simply apply Lemma 8 twice. Let Q_W and Q_V be the marginal distributions on \mathcal{Y} under channels W and V respectively. Then

$$|Q_W(y) - Q_V(y)| \leq \sum_x P(x) |W(y|x) - V(y|x)| \leq \epsilon .$$

Now we can break apart the mutual information and use Lemma 8 on each term:

$$\begin{aligned} |I(P, W) - I(P, V)| &\leq |H(Q_W) - H(Q_V)| + \sum_x P(x) |H(W(Y|X=x)) - H(V(Y|X=x))| \\ &\leq 2(|\mathcal{Y}| - 1) \cdot h_b(\epsilon) + 2(|\mathcal{Y}| - 1) \log(|\mathcal{Y}| - 1) \cdot \epsilon . \end{aligned}$$

■

B. Properties of concatenated fixed composition sets

Let $\tau(\mathbf{x})$ be the type of \mathbf{x} . Let $\mathbf{T}_n(P) = \{\mathbf{x} \in \mathcal{X}^n : \tau(\mathbf{x}) = P\}$ be the set of all length- n vectors of type P . For a vector \mathbf{x} , let \mathbf{x}_1^m be the first m elements of \mathbf{x} .

Lemma 10: For all finite sets \mathcal{X} , and all types P with $p_0 = \min_{x \in \mathcal{X}} P(x) > 0$, there exists $\eta = \eta(P) < \infty$ such that for all integers $M, n > 0$,

$$\frac{|\mathbf{T}_n(P)|^M}{|\mathbf{T}_{Mn}(P)|} \geq \exp(-\eta M \log(n+1)). \quad (41)$$

Proof: We begin by expanding the ratio:

$$\frac{|\mathbf{T}_n(P)|^M}{|\mathbf{T}_{Mn}(P)|} = \frac{\binom{n}{p_1 n, p_2 n, \dots, p_{|\mathcal{X}|} n}^M}{\binom{Mn}{p_1 Mn, p_2 Mn, \dots, p_{|\mathcal{X}|} Mn}}.$$

We can bound the multinomial coefficient using Stirling's approximation [26, pp. 50–53]:

$$\begin{aligned} \binom{n}{p_1 n, p_2 n, \dots, p_{|\mathcal{X}|} n} &= \frac{n!}{(p_1 n)! \cdot (p_2 n)! \cdot \dots \cdot (p_{|\mathcal{X}|} n)!} \\ &\geq (\sqrt{2\pi})^{-|\mathcal{X}|+1} \cdot \frac{n^n \sqrt{n}}{\prod_{x=1}^{|\mathcal{X}|} (p_x n)^{p_x n} \sqrt{p_x n}} \cdot \exp\left(\frac{1}{12n+1} - \sum_{x=1}^{|\mathcal{X}|} \frac{1}{12p_x n}\right) \\ \binom{Mn}{p_1 Mn, p_2 Mn, \dots, p_{|\mathcal{X}|} Mn} &= \frac{(Mn)!}{(p_1 Mn)! \cdot (p_2 Mn)! \cdot \dots \cdot (p_{|\mathcal{X}|} Mn)!} \\ &\leq (\sqrt{2\pi})^{-|\mathcal{X}|+1} \cdot \frac{(Mn)^{Mn} \sqrt{Mn}}{\prod_{x=1}^{|\mathcal{X}|} (p_x Mn)^{p_x Mn} \sqrt{p_x Mn}} \\ &\quad \cdot \exp\left(-\frac{1}{12Mn+1} + \sum_{x=1}^{|\mathcal{X}|} \frac{1}{12p_x Mn}\right). \end{aligned}$$

Now we can cancel some terms to get a further lower bound for some $0 < \nu(P) < \infty$:

$$\begin{aligned} \frac{|\mathbf{T}_n(P)|^M}{|\mathbf{T}_{Mn}(P)|} &\geq (\sqrt{2\pi})^{-(M-1)(|\mathcal{X}|-1)} \cdot \frac{n^{Mn}}{(Mn)^{Mn}} \cdot \left(\prod_{x=1}^{|\mathcal{X}|} \frac{(p_x Mn)^{p_x Mn}}{(p_x n)^{p_x Mn}}\right) \cdot \left(\frac{(Mn)^{(|\mathcal{X}|-1)}}{n^{M(|\mathcal{X}|-1)}} \cdot \prod_{x=1}^{|\mathcal{X}|} p_x^{-(M-1)}\right)^{1/2} \\ &\quad \cdot \exp\left(\frac{M}{12n+1} - \sum_{x=1}^{|\mathcal{X}|} \frac{M}{12p_x n} + \frac{1}{12Mn+1} - \sum_{x=1}^{|\mathcal{X}|} \frac{1}{12p_x Mn}\right) \\ &\geq \exp(-M|\mathcal{X}| \log \sqrt{2\pi}) \cdot \exp\left(\frac{1}{2}(|\mathcal{X}|-1)(\log Mn - M \log n)\right) \cdot \exp(-\nu(P)M/n) \\ &\geq \exp\left(-M\left(|\mathcal{X}| \log \sqrt{2\pi} + \frac{1}{2}(|\mathcal{X}|-1) \log n - \frac{(|\mathcal{X}|-1) \log Mn}{2M} + \frac{\nu(P)}{n}\right)\right) \\ &\geq \exp(-\eta M \log(n+1)), \end{aligned}$$

where $\eta = \eta(P) < \infty$. ■

C. Codebook construction

We restate useful results that we will be useful for our proof. We first present a result in Csiszár and Körner about MMI decoding. To make it easier to apply to our proofs and bring out the parameter dependencies that play a significant role in this work, our statements vary slightly from the original version.

Lemma 11: (Csiszár and Körner [27, Lemma 5.1, Theorem 5.2, pp. 162–166]) For every DMC W and all $R > 0$, random uniform selection with replacement of

$$J = (n + 1)^{-\alpha} \exp(nR) \quad (42)$$

codewords $\{\mathbf{u}(i) : i = 1, \dots, J\}$ from the set $\mathbf{T}_n(P)$ of fixed composition $P = (p_1, p_2, \dots, p_{|\mathcal{X}|})$ has an expected error probability under MMI decoding

$$\varepsilon_r \leq (n + 1)^\alpha \exp(-nE_r(R, W)) , \quad (43)$$

where $E_r(R, W)$ is Gallager's random coding exponent [27, Problem 5.23, p. 192] and $\alpha = 5|\mathcal{X}| + 6|\mathcal{X}| \cdot |\mathcal{Y}|$. A fortiori, there exists a codebook with the same property.

Proof: This is a restatement of Theorem 5.2 in Csiszár and Körner which holds for any n instead of n sufficiently large. Their result states that for a fixed $\delta > 0$ and n sufficiently large, selecting $\exp(n(R - \delta))$ codewords will result in an error probability

$$\varepsilon \leq \exp(-n(E_r(R, W) - \delta)) . \quad (44)$$

To see why this is equivalent to our statement, first observe that the rate loss appears in the first two lines of their proof as a consequence of the Packing lemma [27, Lemma 5.1, p. 162-164]. A closer inspection of the packing lemma reveals the requirement that

$$(n + 1)^{2|\mathcal{X}| + 3|\mathcal{X}| \cdot |\mathcal{Y}|} \exp\left(-n\frac{\delta}{2}\right) \leq \frac{1}{2} . \quad (45)$$

Rather than fixing δ and taking large n , we can fix n and find a bound on δ . By rearranging terms in (45) we see that taking

$$\delta \geq (5|\mathcal{X}| + 6|\mathcal{X}| \cdot |\mathcal{Y}|)n^{-1} \log(n + 1) \quad (46)$$

is sufficient for the bound. The exact error probability derived in the proof of Theorem 5.2 is

$$(n + 1)^{2|\mathcal{X}| \cdot |\mathcal{Y}|} \exp(-nE_r(R, W)) . \quad (47)$$

Clearly, choosing $\alpha = 5|\mathcal{X}| + 6|\mathcal{X}| \cdot |\mathcal{Y}|$ is sufficient to prove the Lemma. ■

We want a bound on the error probability that depends only on the rate gap between our codebook and capacity.

Lemma 12: (Gallager [28, p. 539, Problem 5.23]) Define $C = I(P, W)$. For $R \leq C$,

$$E_r(R, W) \geq \frac{(C - R)^2}{8/e^2 + 4(\ln |\mathcal{Y}|)^2}, \quad (48)$$

where $E_r(R, W)$ is Gallager's random coding exponent.

Proof: The result is presented as an exercise in Gallager [28, p. 539, Problem 5.23] with a sketch of the proof that leaves out the tedious calculations required to show the result. While there is an error in that proof sketch, we note that a stronger result than the one presented here was recently shown by correcting the arguments outlined in that exercise [29]. ■

Lemma 13: For $R > 0$, with probability at least

$$1 - M \exp\left(-\frac{1}{8}2^{M+1}(n+1)^{-\eta M} \exp(MnR)\right) \quad (49)$$

there exists a set of K' codewords $\{\mathbf{x}(l) : l = 1, 2, \dots, K'\}$ of blocklength Mn and fixed composition $P = (p_1, p_2, \dots, p_{|\mathcal{X}|})$, where

$$K' \geq \exp(MnR), \quad (50)$$

for $\alpha = \alpha(|\mathcal{X}|, |\mathcal{Y}|) < \infty$ and $\eta = \eta(|\mathcal{X}|) < \infty$ that satisfies the following properties:

- 1) For each i we have $\mathbf{x}(i) \in \{\mathbf{T}_n(P)\}^M$.
- 2) The collection $\{\mathbf{x}_1^{\ell n}(i) : i = 1, 2, \dots, K'\}$ is a codebook with K' codewords and whose maximum probability of error under maximum mutual information decoding smaller than

$$\varepsilon_\ell \leq 2^\ell 3(n+1)^{\eta \ell} (\ell n + 1)^\alpha \exp\left(-\ell n E_r\left(\frac{M}{\ell} \left(R + \frac{\log 2}{n} + \eta \frac{\log(n+1)}{n} + \alpha \frac{\log(\ell n + 1)}{Mn}\right), W\right)\right), \quad (51)$$

where $E_r(R, W)$ is Gallager's random coding exponent [28].

Proof: Our construction will be as follows:

- 1) Draw J random vectors $\mathbf{U}(1), \mathbf{U}(2), \dots, \mathbf{U}(J)$ from $\mathbf{T}_{Mn}(P)$.

Let $J = 2^{M+1} \frac{|\mathbf{T}_{Mn}(P)|}{|\mathbf{T}_n(P)|^M} \exp(MnR)$ and define $\tilde{R}_{Mn} = (Mn)^{-1} \log J = R + \frac{\log 2}{n} + (Mn)^{-1} \log \frac{|\mathbf{T}_{Mn}(P)|}{|\mathbf{T}_n(P)|^M}$.

By Lemma 11, we know that the expected error probability over messages $\{\mathbf{U}(j) : j = 1, 2, \dots, J\}$

after M for the codebook $\{\mathbf{x}_1^{Mn}(i) : i = 1, 2, \dots, J\}$ is at most

$$\tilde{\varepsilon}_M \leq (Mn + 1)^\alpha \exp(-Mn E_r(\tilde{R}_{Mn}, W)), \quad (52)$$

where $\alpha = \alpha(|\mathcal{X}|, |\mathcal{Y}|) < \infty$ and $\bar{R}_{Mn} = \tilde{R}_{Mn} + \frac{\alpha \log(Mn+1)}{Mn}$.

2) Thin the set $\{\mathbf{U}(j) : j = 1, 2, \dots, J\}$ by removing all codewords not in $\mathbf{T}_n(P) \times \dots \times \mathbf{T}_n(P)$ to form a codebook $\{\mathbf{V}(k) : k = 1, 2, \dots, K\}$.

Let $f_m : \mathbf{T}_{mn}(P) \cup \{\perp\} \rightarrow \mathbf{T}_{(m-1)n}(P) \cup \{\perp\}$ be the thinning map defined by

$$f_m(\mathbf{u}) = \begin{cases} \mathbf{u}_1^{(m-1)n} & \mathbf{u}_1^{(m-1)n} \in \mathbf{T}_{(m-1)n}(P) \\ \perp & \text{otherwise} \end{cases}$$

We think of \perp as a null symbol that represents deleting a codeword from the codebook. That is, $f_m(\cdot)$ removes codewords whose type is not P in the first $(m-1)n$ positions. The thinned codebook is thus

$$\mathcal{C} = \{\mathbf{U}(k) : f_m(\mathbf{U}(k)) \neq \perp \text{ for } m = 2, 3, \dots, M\} . \quad (53)$$

Let us write $\mathcal{C} = \{\mathbf{V}(k) : k = 1, 2, \dots, K\}$. Furthermore, we can define a sequence of nested thinned codebooks. Let

$$\mathcal{C}_\ell = \{\mathbf{X}(k) : f_m(\mathbf{X}(k)) \neq \perp \text{ for } m = \ell + 1, \ell + 2, \dots, M\} . \quad (54)$$

Note that $\mathcal{C}_\ell \subseteq \mathcal{C}_{\ell+1}$ and $\mathcal{C}_1 = \mathcal{C}$. Furthermore, if $\mathbf{X}(k) \in \mathcal{C}_\ell$, then $f_{\ell+1}(\mathbf{X}(k))$ is uniformly distributed over $T_{\ell n}(P)$. Since there are at most J codewords in codebook \mathcal{C}_ℓ and these are uniform over $T_{\ell n}(P)$, by Lemma 11, the average error probability for \mathcal{C}_ℓ after ℓn channel uses is bounded by

$$\tilde{\varepsilon}_\ell \leq (\ell n + 1)^\alpha \exp\left(-\ell n E_r\left(\frac{M}{\ell} \bar{R}_{\ell n}, W\right)\right) , \quad (55)$$

where $\bar{R}_{\ell n} = \tilde{R}_{Mn} + \frac{\alpha \log(\ell n + 1)}{Mn}$. To understand the behavior on the codebook \mathcal{C} , we first need to provide a guarantee on the number of codewords after thinning. The probability that a randomly chosen codeword survives thinning is $\gamma = \frac{|\mathbf{T}_n(P)|^M}{|\mathbf{T}_{Mn}(P)|}$. We can use this to conclude

$$\mathbb{P}\left(\frac{K}{J} < \gamma/2\right) = \mathbb{P}\left(\frac{K}{J} - \gamma < -\gamma/2\right) \leq \exp(-J\gamma^2/8) , \quad (56)$$

where (56) follows from Hoeffding's inequality [30]. Let K_i be the number of codewords in \mathcal{C}_ℓ with boundary conditions $K = K_1$ and $J = K_M$. We want upper bounds on K_i . The probability of a randomly chosen codeword will survive the thinning process for codebook \mathcal{C}_ℓ is $\gamma_i = \frac{|\mathbf{T}_{\ell n}(P)| \cdot |\mathbf{T}_n(P)|^{M-\ell}}{|\mathbf{T}_{Mn}(P)|}$.

$$\mathbb{P}\left(\frac{K}{J} > 3\gamma_i/2\right) = \mathbb{P}\left(\frac{K}{J} - \gamma_i > \gamma_i/2\right) \leq \exp(-J\gamma_i^2/8) , \quad (57)$$

where (57) follows from Hoeffding's inequality [30]. Thus, by a union bound and the observation that $\gamma < \gamma_i$ for all i , with probability at least

$$1 - M \exp(-J\gamma^2/8) , \quad (58)$$

the thinned codebook \mathcal{C} satisfies

$$K \geq \frac{\gamma}{2} J = \frac{|\mathbf{T}_n(P)|^M}{|\mathbf{T}_{Mn}(P)|} \cdot \frac{J}{2} , \quad (59)$$

and thinned codebooks \mathcal{C}_ℓ satisfy

$$K_i \leq \frac{3\gamma_i}{2} J = \frac{|\mathbf{T}_{\ell n}(P)| \cdot |\mathbf{T}_n(P)|^{M-\ell}}{|\mathbf{T}_{Mn}(P)|} \cdot \frac{J}{2} . \quad (60)$$

By Lemma 10, (58) implies (49). Thus, with probability at least (49), the average error probability for codebook \mathcal{C} after ℓn channel uses is bounded by

$$\begin{aligned} \bar{\varepsilon}_\ell &\leq \frac{K_\ell}{K} \tilde{\varepsilon}_\ell \\ &\leq 3 \frac{|\mathbf{T}_{\ell n}(P)|}{|\mathbf{T}_n(P)|^\ell} \tilde{\varepsilon}_\ell \end{aligned} \quad (61)$$

$$\leq 3(n+1)^{\eta_\ell} \tilde{\varepsilon}_\ell , \quad (62)$$

where (62) follows from Lemma 10.

- 3) Expurgate the codebook $\{\mathbf{V}(j) : j = 1, 2, \dots, K_2\}$ to guarantee small decoding error to form our final codebook $\{\mathbf{X}(l) : l = 1, 2, \dots, L\}$.

We have far made guarantees about the average error probability $\bar{\varepsilon}_\ell$ after ℓn channel uses. If we remove half those codewords that result in the worst error probability for the first ℓn channel uses, Markov's inequality implies that the maximal error probability after n channel uses is at most

$$\varepsilon_1 \leq 2\bar{\varepsilon}_1 . \quad (63)$$

For $m = 2, \dots, M$, the average error probability after expurgation is at most $2\bar{\varepsilon}_m$. Similarly, for $\ell = 2, \dots, M$, we continue expurgating those codewords that yield the worst error probability for ℓn channel uses. This yields a maximum error probability

$$\varepsilon_\ell \leq 2^\ell \bar{\varepsilon}_\ell , \quad (64)$$

and the average error probability for $m = \ell, \ell + 1, \dots, M$, the average error probability after the previous ℓ expurgations is at most $2^\ell \bar{\varepsilon}_m$. Applying Lemma 10 to (55) and combining it with (62)

and (64) implies (51). Indeed, expurgation further reduces the number of codewords, and from our earlier bound, with probability at least (49), the number of codewords is at least

$$K' \geq 2^{-M} K \geq \exp(MnR), \quad (65)$$

thereby completing the proof. ■

In order to actually prove the existence of a good codebook, we must set M such that the probability bound in (49) is actually positive. This is shown in Section IV-C.

D. Proof of Lemma 2

Proof: (Proof of Lemma 2) We will first prove that for fixed (x, y) , $\hat{W}^{(n)}(y|x)$ is close to $W_{\mathbf{z}(\mathcal{T}_n)}(y|x)$, and that $W_{\mathbf{z}(\mathcal{T}_n)}(y|x)$ is close to $W_{\mathbf{z}(\mathcal{U}_n)}(y|x)$ and $W_{\mathbf{z}(\mathcal{V}_n)}(y|x)$.

Fix $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ and note that the random variable $\hat{w}^{(i)}(y|x)$ has expectation $\mathbb{E}[\hat{w}^{(i)}(y|x)] = W_{\mathbf{z}(T_i(x))}(y|x)$. We can now apply Hoeffding's inequality [30] to the sum:

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n \hat{w}^{(i)}(y|x) - \frac{1}{n} \sum_{i=1}^n W_{\mathbf{z}(T_i(x))}(y|x) \right| > \frac{\epsilon_2}{2} \right) \leq 2 \exp \left(-\frac{1}{2} \cdot n \cdot \epsilon_2^2 \right).$$

Rewriting this we have

$$\mathbb{P} \left(\left| \hat{W}^{(n)}(y|x) - W_{\mathbf{z}(\mathcal{T}_n)}(y|x) \right| > \frac{\epsilon_2}{2} \right) \leq 2 \exp \left(-\frac{1}{2} \cdot n \cdot \epsilon_2^2 \right).$$

Now let us consider the random variable $W_{\mathbf{z}(T_i(x))}(y|x)$. We can view this as drawing $t/|\mathcal{X}|$ samples without replacement from the set $\{W(y|x, z_j) : j \in V_i\}$. Another result of Hoeffding [30, Theorem 4] states that the exponential inequalities for sampling with replacement hold for sampling without replacement as well, so the channel during the training is a good approximation to the entire channel during the chunk:

$$\mathbb{P} \left(\left| W_{\mathbf{z}(T_i(x))}(y|x) - W_{\mathbf{z}(V_i(x))}(y|x) \right| > \frac{\epsilon_2}{2} \right) \leq 2 \exp \left(-\frac{1}{2} \cdot \frac{t}{|\mathcal{X}|} \cdot \epsilon_2^2 \right). \quad (66)$$

Therefore we can bound over the n chunks to get

$$\mathbb{P} \left(\left| W_{\mathbf{z}(\mathcal{T}_n(x))}(y|x) - W_{\mathbf{z}(\mathcal{V}_n(x))}(y|x) \right| > \frac{\epsilon_2}{2} \right) \leq 2n \exp \left(-\frac{1}{2} \cdot \frac{t}{|\mathcal{X}|} \cdot \epsilon_2^2 \right). \quad (67)$$

Now we have:

$$W_{\mathbf{z}(\mathcal{V}_n(x))}(y|x) = \frac{t}{b} W_{\mathbf{z}(\mathcal{T}_n(x))}(y|x) + \frac{b-t}{b} W_{\mathbf{z}(\mathcal{U}_n(x))}(y|x).$$

Therefore

$$\mathbb{P} \left(|W_{\mathbf{z}(\mathcal{T}_n(x))}(y|x) - W_{\mathbf{z}(\mathcal{U}_n(x))}(y|x)| > \frac{\epsilon_2}{2} \right) \leq 2n \exp \left(-\frac{1}{2} \cdot \frac{t}{|\mathcal{X}|} \cdot (\epsilon_2 - 2t/b)^2 \right). \quad (68)$$

Now, taking a union bound over \mathcal{X} and \mathcal{Y} in (67) and (68) yields (21) and (20) with $\alpha_1 = 2|\mathcal{X}||\mathcal{Y}|$ and $\beta_1 = 1/(2|\mathcal{X}|)$. \blacksquare

E. Proof of Lemma 3

Proof: Given that our estimated channel $\hat{W}^{(M)}(y|x)$ is close to the true channel $W_{\mathbf{z}(\mathcal{U}_M)}(y|x)$, we must ensure that the slack ϵ_1 is sufficient to guarantee successful decoding with high probability. Under E_1^c we know that $|\hat{W}^{(M)}(y|x) - W_{\mathbf{z}(\mathcal{U}_M)}(y|x)| < \epsilon_2$. Using Lemma 9, we can pick α_3 and α_4 so that

$$|I(P, \hat{W}^{(M)}) - I(P, W_{\mathbf{z}(\mathcal{U}_M)})| \leq \epsilon_3.$$

From the definition of the decoding rule in (15) we have

$$\begin{aligned} \frac{k}{M(b-t)} &< I(P, \hat{W}^{(M)}) - \epsilon_1 \\ &\leq I(P, W_{\mathbf{z}(\mathcal{U}_M)}) - (\epsilon_1 - \epsilon_3). \end{aligned} \quad (69)$$

Therefore the rate of our codebook at the decoding time M is within $(\epsilon_1 - \epsilon_3)$ of the empirical capacity.

We now turn to the codebook construction. 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 $\mathcal{M} = \{k/M(b-t) : M_* \leq M \leq M^*\}$. Using common randomness, the encoder and decoder will choose a constant composition random codebook with composition P of blocklength $M^*(b-t)$ whose truncation to lengths $m \in \mathcal{M}$ is also a constant composition codebook of blocklength m . Lemma 13 shows that such codebooks exist with probability

$$1 - M^* \exp \left(-\frac{1}{8} 2^{M^*+1} ((b-t)+1)^{-\eta M^*} \exp(M^*(b-t)R) \right), \quad (70)$$

where $\eta(|\mathcal{X}|) < \infty$. Furthermore, these codebooks have probability of error when decoded at blocklength M that is upper bounded by

$$\mathbb{P}(E_2|E_1^c, M) \leq 2^M 3((b-t)+1)^{\eta M} (M(b-t)+1)^\alpha \exp \left(-M(b-t)E_r \left(\frac{M^*}{M} (R+\chi), W_{\mathbf{z}(\mathcal{U}_M)} \right) \right), \quad (71)$$

where $\alpha(|\mathcal{X}|, |\mathcal{Y}|) < \infty$ and

$$\chi = \frac{\log 2}{(b-t)} + \eta \frac{\log((b-t)+1)}{(b-t)} + \alpha \frac{\log(M(b-t)+1)}{M^*(b-t)}. \quad (72)$$

Here we choose

$$R = \frac{k}{M^*(b-t)} . \quad (73)$$

Using Lemma 12, we can further lower bound the error exponent $E_r(\cdot, \cdot)$:

$$E_r \left(\frac{k}{M(b-t)} + \frac{M^*}{M} \chi, W_{\mathbf{z}(\mathcal{U}_M)} \right) \geq \frac{\left(C - \frac{k}{M(b-t)} - \frac{M^*}{M} \chi \right)^2}{8/e^2 + 4(\ln |\mathcal{Y}|)^2} . \quad (74)$$

for $\frac{k}{M(b-t)} \leq C - \frac{M^*}{M} \chi$, where $C = I(P, W_{\mathbf{z}(\mathcal{U}_M)})$. Then using (69)

$$E_r \left(\frac{k}{M(b-t)} + \frac{M^*}{M} \chi, W_{\mathbf{z}(\mathcal{U}_M)} \right) \geq \frac{\left(C - I(P, W_{\mathbf{z}(\mathcal{U}_M)}) + (\epsilon_1 - \epsilon_3) - \frac{M^*}{M} \chi \right)^2}{8/e^2 + 4(\ln |\mathcal{Y}|)^2} \quad (75)$$

$$= \frac{\left(\epsilon_1 - \epsilon_3 - \frac{M^*}{M} \chi \right)^2}{8/e^2 + 4(\ln |\mathcal{Y}|)^2} . \quad (76)$$

For the exponent to be positive at all valid decoding times, we must have

$$\epsilon_1 - \epsilon_3 - \frac{M^*}{M_*} \left(\frac{\log 2}{(b-t)} + \eta \frac{\log((b-t)+1)}{(b-t)} + \alpha \frac{\log(M^*(b-t)+1)}{M^*(b-t)} \right) > 0 . \quad (77)$$

We can simplify this for two finite positive constants η' and α' :

$$\epsilon_1 - \epsilon_3 - \frac{\eta' M^* \log(b-t) + \alpha' \log M^*}{M_*(b-t)} > 0 . \quad (78)$$

Setting $\alpha_5 = \eta'$ and $\alpha_6 = \alpha'$ completes the proof. ■

F. The rate loss within a round

Proof: (Proof of Lemma 5) Lemma 2 states that with high probability the channel $\hat{W}^{(n)}$ is within ϵ_2 of both $W_{\mathbf{z}(\mathcal{U}_n)}$ and $W_{\mathbf{z}(\mathcal{V}_n)}$. Therefore Lemma 9 shows that

$$\left| I(P, \hat{W}^{(n)}) - I(P, W_{\mathbf{z}(\mathcal{V}_n)}) \right| < \epsilon_3 . \quad (79)$$

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 have $I(P, \hat{W}^{(n)}) < \tau$ and the achieved rate is obviously 0, so the loss is

$$I(P, W_{\mathbf{z}(\mathcal{V}_n)}) - R < \tau + \epsilon_3 . \quad (80)$$

Now let us turn to the rounds in which the decoder decodes after receiving M chunks. Since the decoding criterion (15) was not met after $M-1$ chunks but was met after M chunks, we have

$$\frac{k}{(b-t)(M-1)} \geq I(P, \hat{W}^{(M-1)}) - \epsilon_1 . \quad (81)$$

We will further bound both sides of this inequality. First, we have

$$\frac{k}{(b-t)(M-1)} \leq \frac{k}{(b-t)M} + \frac{k}{b-t} \cdot \frac{1}{(M_*-1)^2}. \quad (82)$$

Turning to $I(P, \hat{W}^{(M-1)})$, note that

$$\begin{aligned} \left| \hat{W}^{(M)}(y|x) - \hat{W}^{(M-1)}(y|x) \right| &\leq \frac{2}{M} \\ &\leq \frac{2}{M_*}. \end{aligned}$$

Therefore from Lemma 9 we have

$$\left| I(P, \hat{W}^{(M-1)}) - I(P, \hat{W}^{(M)}) \right| \leq |\mathcal{Y}|h_b(2M_*^{-1}) + 2|\mathcal{Y}| \log |\mathcal{Y}|M_*^{-1}. \quad (83)$$

Plugging (82) and (83) into (81), we get

$$\frac{k}{(b-t)M} + \frac{k}{b-t} \cdot \frac{1}{(M_*-1)^2} \geq I(P, \hat{W}^{(M)}) - |\mathcal{Y}|h_b(2M_*^{-1}) + 2|\mathcal{Y}| \log |\mathcal{Y}|M_*^{-1} - \epsilon_1.$$

Rearranging this expression and using (79), we get

$$I(P, W_{\mathbf{z}(\mathcal{V}_n)}) - R \leq \epsilon_1 + \epsilon_3 + |\mathcal{Y}|h_b(2M_*^{-1}) + 2|\mathcal{Y}| \log |\mathcal{Y}|M_*^{-1} + \frac{k}{b-t} \cdot \frac{1}{(M_*-1)^2}. \quad (84)$$

The lemma now follows from the two bounds on the rate loss. ■

REFERENCES

- [1] K. Eswaran, A. Sarwate, A. Sahai, and M. Gastpar, "Binary additive channels with individual noise sequences and limited active feedback," in *Proceedings of the 2007 IEEE International Symposium on Information Theory*, Nice, France, 2007.
- [2] M. Horstein, "Sequential transmission using noiseless feedback," *IEEE Transactions on Information Theory*, vol. 9, no. 3, pp. 136–143, July 1963.
- [3] J. P. M. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback I: No bandwidth constraint," *IEEE Transactions on Information Theory*, vol. 12, pp. 172–182, 1966.
- [4] M. Burnashev, "Data transmission over a discrete channel with feedback, random transmission time," *Problems of Information Transmission*, vol. 12, no. 4, October–December 1976.
- [5] T. Cover and S. Pombra, "Gaussian feedback capacity," *IEEE Transactions on Information Theory*, vol. 35, pp. 37–43, 1989.
- [6] J. Ooi and G. Wornell, "Fast iterative coding techniques for feedback channels," *IEEE Transactions on Information Theory*, vol. 44, no. 7, pp. 2960–2976, November 1998.
- [7] J. Ooi, *Coding for channels with feedback*. Boston, MA: Kluwer Academic Publishers, 1998.
- [8] Y.-H. Kim, "Feedback capacity of stationary Gaussian channels," 2006. [Online]. Available: arXiv:cs/0602091v1
- [9] A. Sahai, "Why block-length and delay behave differently if feedback is present," *IEEE Transactions on Information Theory*, Submitted 2006. [Online]. Available: <http://www.eecs.berkeley.edu/~sahai/Papers/FocusingBound.pdf>
- [10] 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.

- [11] Y.-H. Kim, A. Lapidoth, and T. Weissman, "On reliability of Gaussian channels with noisy feedback," in *Proceedings of the 44th Allerton Conference on Communication, Control, and Computing*, September 2006.
- [12] M. Gastpar and G. Kramer, "On noisy feedback for interference channels," in *Proceedings of the 2006 Asilomar Conference on Signals, Systems, and Computers*, 2006.
- [13] M. Wigger, "Noisy feedback is strictly better than no feedback on the Gaussian MAC," 2006 Kailath Symposium, July 2006.
- [14] A. Sahai, "Balancing forward and feedback error correction for erasure channels with unreliable feedback," submitted to *IEEE Transactions of Information Theory*.
- [15] A. Tchamkerten and I. E. Telatar, "Variable length coding over an unknown channel," *IEEE Transactions on Information Theory*, vol. 52, no. 5, pp. 2126–2145, May 2006.
- [16] 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
- [17] M. Luby, "LT codes," in *Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science*, 2002.
- [18] A. Shokrollahi, "Fountain codes," in *Proceedings of the 41st Allerton Conference on Communication, Control, and Computing*, October 2003, pp. 1290–1297.
- [19] N. Shulman, "Communication over an unknown channel via common broadcasting," Ph.D. dissertation, Tel Aviv University, 2003.
- [20] S. Draper, B. Frey, and F. Kschischang, "Efficient variable length channel coding for unknown DMCs," in *Proceedings of the 2004 International Symposium on Information Theory*, Chicago, USA, 2004.
- [21] ———, "Rateless coding for non-ergodic channels with decoder channel state information," submitted to *IEEE Transactions of Information Theory*.
- [22] E. Soljanin, "Hybrid ARQ in wireless networks," in *DIMACS Workshop on Network Information Theory*, March 2003.
- [23] A. Sarwate and M. Gastpar, "Rateless coding with partial side information at the decoder," November 2007, submitted to *IEEE Transactions of Information Theory*.
- [24] 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.
- [25] M. Feder, "Achieving the empirical capacity of individual noise channels using feedback," 2006 Kailath Symposium, July 2006.
- [26] W. Feller, *An Introduction to Probability Theory and Its Applications*. New York: John Wiley and Sons, Inc., 1968.
- [27] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Budapest: Akadémiai Kiadó, 1982.
- [28] R. Gallager, *Information Theory and Reliable Communication*. New York: John Wiley and Sons, 1968.
- [29] C. Chang, unpublished manuscript, 2006.
- [30] 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.