Zero-rate feedback can achieve the empirical capacity

Krishnan Eswaran, *Student Member, IEEE*, Anand D. Sarwate, *Member, IEEE*, Anant Sahai, *Member, IEEE*, and Michael Gastpar, *Member, IEEE*

***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

Many contemporary communication systems can be modeled via a time-varying state. For example, in wireless communications, the channel variation may be caused by neighboring systems, mobility, or other factors that are difficult to model. In order to design robust communication strategies, engineers should adopt an appropriate model for the channel dynamics. One such model is the so-called arbitrarily varying channel (AVC), in which the state can depend on the communication strategy and is selected in the worst possible manner. One interpretation of this model is that there is a fixed rate (e.g. for voice) that one wants to support over the worst possible channel states. An alternative and perhaps more relevant approach (e.g. for data traffic) is an individual sequence model, where the state is fixed but unknown and not dependent on the communication strategy. Here, a natural requirement is for a strategy to perform well whenever the state sequence is favorable, while for less favorable state sequences, inferior performance is acceptable. Essentially, this model considers the case in which one wants to adapt the rate to one which the specific state sequence can support.

In order to achieve this variation in performance, the encoder must obtain some measure of the quality of the state sequence. This requires additional resources, and the most natural model is to introduce *feedback* from the receiver to the transmitter. A second resource is joint *randomization* between the encoder and the decoder, which can also be enabled via feedback. The encoder can use feedback to estimate the channel quality and hence communicate at rates commensurate with the channel quality. Two fundamental questions are the following: first, how good a performance (in terms of achievable rate) can one expect for favorable state sequences? Second, how much feedback is required to attain this performance? Many of the works in this area can be understood in terms of how they answer these two questions.

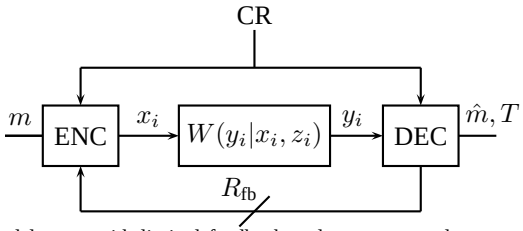

Fig. 1. Model setup with limited feedback and common randomness.

The main trade-off for the channel model at hand is the correct balance between the resources spent on communication versus those spent on channel estimation. One extreme is the case where the channel state sequence is fully revealed to the receiver, as shown in the work of Draper *et al.* [2]. Regarding the first question, for any fixed input distribution, their scheme can achieve rates arbitrarily close to the mutual information of the channel with the state known to both the transmitter and receiver. They also provide an interesting answer to the second question: a feedback link of vanishing rate is sufficient to attain this performance. To sum up, when channel estimation at the receiver is free, feedback of vanishing rate is enough.

Shayevitz and Feder [3] consider the more realistic case where the decoder has only the channel outputs. They develop a scheme in which the receiver keeps estimating the state sequence. The transmitter has full (causal) output feedback and can thus also track the state sequence. For the class of channels they consider, Shayevitz and Feder establish an achievable rate that they call 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. Therefore if feedback is free, then rates arbitrarily close to the "empirical capacity" are achievable.

This paper is a commentary on this development: we consider the same notion of "empirical capacity," but provide an answer to the second question. Specifically, for a fixed input distribution, we show that if common randomness is

Manuscript received October XX, 2007; revised XXXXXXXXXXXXXX. Part of this work was presented at the 2007 International Symposium on Information Theory in Nice, France [1].

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. The work of A. Sahai was also supported by the National Science Foundation under award CCF-0729122.

K. Eswaran, A. Sahai, and M. Gastpar are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley CA 94720-1770 USA. A. D. Sarwate was with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley. He is now with the Information Theory and Applications Center at the University of California, San Diego, La Jolla, CA 92093-0447 USA.

available, a feedback link of vanishing rate is sufficient to achieve the empirical mutual information, which in some settings, such as the class of channels considered by Shayevitz and Feder, coincides with the "empirical capacity". To do this, we adapt the feedback-reducing block/chunk strategies used earlier in the context of reliability functions [4], [5], and most specifically in [6]. They are in turn inspired by Hybrid ARQ [7]. Thus, the flavor of our algorithm is different from [3]. By doing away with the output feedback, we lose the simplicity of the scheme in [3], but we show that similar rates can still be obtained with almost negligible feedback.

The strategy developed in this paper fits in the category of rateless codes, which are a class of coding strategies that use limited feedback to adapt to unknown channel parameters. Most studies about feedback for rate and reliability have centered around full output feedback [4], [8]–[14]; however, recent work has started to improve our understanding of how limited feedback affects these performance measures. For instance, limited feedback can be used to improve reliability [6], [15]. Furthermore, in some multiuser Gaussian channels, noisy feedback increases the achievable rates [16]–[18] and the reliability [5], [19]. In a rateless code the decoder can use a low-rate feedback link to inform the encoder when it decodes. These codes were first studied in the context of the erasure channel [20], [21]. Later work focused on compound channels [22]–[24]. The work of Draper et al. [2] is to our knowledge the first step towards adapting rateless codes to time-varying states.

We are now in a position to compare the modeling assumptions in these previous works with the current investigation; the comparisons are summarized in Table I. The initial studies of rateless coding by Shulman [22] and Tchamkerten and Telatar [24] used feedback to tune the rate to the realized parameter governing the channel behavior. The study of time-varying states was first introduced by Draper et al. [2], but they assumed full state information at the decoder, which leads to higher rates. Most recently, Shayevitz and Feder [3] showed an explicit coding algorithm based on Horstein's method [8] that achieves the empirical capacity. Their scheme uses full feedback, but in turn works for a larger class of channel models. Moreover, it is a horizon-free scheme.

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 pilot sequence. The decoder uses the training positions to estimate the channel. As soon as the total *empirical mutual information* over the aggregate channel sufficiently exceeds k bits, 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

The following two simple examples will prove useful in explaining the meaning of the main result of this paper, and help motivate the present study. The first is the model considered in [3] – a binary modulo-additive channel with a noise sequence whose empirical frequency of 1's is unknown. In this example, the "empirical mutual information" under all state sequences is maximized by the uniform distribution, so our algorithm achieves the "empirical capacity". In the second example we consider the Z-channel for which the input distribution maximizing the empirical mutual information is not identical for all state sequences, so our scheme will not in general achieve rates as high as the empirical capacity.

A. Binary modulo-additive channels

The simplest example of a channel with an individual noise sequence is the binary modulo-additive channel. This channel takes binary inputs and produces binary outputs, where the output is produced by 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

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

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. The noise $\mathbf{z}$ is arbitrary but fixed, and we let $p \in [0,1]$ be the empirical fraction of 1's in $\mathbf{z}$, which is also arbitrary but fixed.

Because the state sequence $\mathbf{z}$ is arbitrary and unknown, it is not clear how to find the highest possible rate of reliable communication. For any *fixed* $\mathbf{z}$, we could say naïvely that the capacity is one bit, because the channel is deterministic. However, $\mathbf{z}$ is unknown and may, in fact, have been generated i.i.d. according to a Bernoulli distribution with parameter p, in which case the capacity should be no larger than $1 - h(p)$, namely, the capacity of a binary symmetric channel (BSC) with crossover p. The algorithm in this paper guarantees a rate close to $1 - h(p)$ for any state sequence $\mathbf{z}$ with an empirical fraction of 1's equal to p. This rate can be thought of as the empirical mutual information of the channel with a uniform input distribution. Since the uniform input distribution achieves the capacity for all BSCs, this rate can also be called the *empirical capacity*, as in the work of Shayevitz and Feder [3].

B. Z-channels with unknown crossover

Whereas the example above can be thought of as an XOR operation with the channel state, in our second example, we consider a binary channel in which the output is the logical OR of the input and state. For input x and noise z, the output y is given by the following:

$$y = \begin{cases} x & z = 0 \\ 0 & z = 1 \ . \end{cases}$$

	channel model	feedback	state information	common randomness
Shulman [22]	compound	full	none	none
Tchamkerten and Telatar [24]	compound	full	none	none
Draper, Frey, and Kschischang [2]	AVC	0-rate	at decoder	none
Shayevitz and Feder [3]	individual sequence	full	none	yes, some
This paper	individual sequence	0-rate	none	yes, lots

TABLE I
RELATED RESULTS AND ASSUMPTIONS ON CHANNEL MODEL, FEEDBACK, STATE INFORMATION AND COMMON RANDOMNESS

Again, the noise sequence $\mathbf{z}$ is arbitrary but fixed. Let q denote the empirical fraction of 1's in $\mathbf{z}$.

The algorithm in this paper achieves rates close to the mutual information induced by a fixed input distribution P to a Z-channel with crossover probability q. The channel is the average $W_{\mathbf{z}}$ of $W(y|x,z_i)$ over $\mathbf{z}$. Unlike the binary modulo-additive example, this channel has a capacity-achieving input distribution that depends on q. The algorithm proposed in this paper chooses a fixed input distribution P and achieves the mutual information $I(P,W_{\mathbf{z}})$ of a Z-channel with that input distribution. This leaves open the question of how to choose P. One method is to choose the P that minimizes the gap between $\max_Q I(Q,W_{\mathbf{z}}) - I(P,W_{\mathbf{z}})$ over all $\mathbf{z}$. However, in many cases the uniform distribution is not a bad choice, as shown by Shulman and Feder [25]. In our results we leave the choice of P open for the designer.

III. THE CHANNEL MODEL AND CODING STRATEGY

A. Notation

Script letters will generally be used to denote sets and alphabets and boldface to denote vectors. For a vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, we write $\mathbf{x}_i^j$ for the tuple $(x_i, x_{i+1}, \ldots, x_j)$ and $\mathbf{x}^j$ for the tuple $(x_1, x_2, \ldots, x_j)$. The notation $[J]$ will be used as shorthand for the set $\{1, 2, \ldots, J\}$. The probability distribution $T_{\mathbf{z}}$ is the type of a sequence $\mathbf{z}$. For a distribution Q, the set $T_N(Q)$ is the set of all length N sequences of type Q.

B. Channel model and coding

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 channel model we consider consists of a family of channels $\mathcal{W} = \{W(y|x,z) : z \in \mathcal{Z}\}$ indexed by a state variable in a finite set $\mathcal{Z}$. For any state sequence $\mathbf{z} = (z_1, z_2, \ldots, z_N)$, and output y_i, we assume

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

That is, the channel output depends only on the current input and state.

We consider coding for this channel using the setup shown in Figure 1. We think of the rate-limited feedback link as a noiseless channel that can be used every $n_{\rm fb}$ uses of the forward channel to send $B_{\rm fb}$ bits. The rate of the feedback is thus $R_{\rm fb} = B_{\rm fb}/n_{\rm fb}$. To avoid integer effects, we will consider only integer values for $n_{\rm fb}$ and $B_{\rm fb}$. We assume that the encoder and decoder have access to a common random variable G distributed uniformly over the unit interval $[0,1]$. This random variable can be used to generate common randomness that is shared between the encoder and decoder.

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 decoding truncation $T \in \mathbb{N}$ and a message $m \in \{0,1\}^T$. The truncation T is itself a random variable that will depend on the state sequence $\mathbf{z}$, the common randomness G, and the randomness in the channel.

An $(N, n_{\rm fb}, B_{\rm fb})$ *coding strategy* for blocklength N consists of a sequence of (possibly random) encoding functions for $i = 1, 2, \ldots, N$,

$$\eta_i : \{0,1\}^{NC_{\max}} \times \{0,1\}^{\lfloor (i-1)/n_{\rm fb} \rfloor B_{\rm fb}} \times [0,1] \to \mathcal{X} \ ,$$

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

$$\phi_i : \mathcal{Y}^i \times [0,1] \to \{0,1\}^{B_{\rm fb}} \ ,$$

and a decoding function

$$\psi : \mathcal{Y}^N \times [0,1] \to \{0,1,\ldots,NC_{\max}\} \times \{0,1\}^{NC_{\max}} \ .$$

We say a message $\mathbf{m} \in \{0,1\}^{NC_{\max}}$ is *encoded* into a codeword $\mathbf{x} \in \mathcal{X}^N$ if for $i \in [N]$,

$$x_i = \eta_i(\mathbf{m}, \phi_1(y^{n_{\rm fb}}, G), \ldots, \phi_{\lfloor \frac{i-1}{n_{\rm fb}} \rfloor}(y^{\lfloor \frac{i-1}{n_{\rm fb}} \rfloor \cdot n_{\rm fb}}, G), G) \ .$$

For an $(N, n_{\rm fb}, B_{\rm fb})$ coding strategy, let $\psi(\mathbf{y}, G) = (T, \hat{\mathbf{m}})$. The first output $T \in \{0,1,\ldots,NC_{\max}\}$ is the *decoding truncation* and $\hat{\mathbf{m}}^T$ is the *message estimate*. Both of these quantities are random variables.

For a state sequence $\mathbf{z}$, the *maximal error probability* of an $(N, n_{\rm fb}, B_{\rm fb})$ coding strategy, is defined as

$$\varepsilon_{\rm dec}(\mathbf{z}) = \max_{\mathbf{m} \in \{0,1\}^{NC_{\max}}} \mathbb{P}_{G,\mathcal{W}} \left(\mathbf{m}^T \neq \hat{\mathbf{m}}^T \ \middle| \ \mathbf{z}, \mathbf{m} \right) \ .$$

where the probability is taken over the common randomness G and randomness in the channel. For a state sequence $\mathbf{z}$, a *rate* R is said to be *achievable* with probability $1 - \varepsilon_{\rm ach}(\mathbf{z})$ if

$$\varepsilon_{\rm ach}(\mathbf{z}) = \max_{\mathbf{m} \in \{0,1\}^{NC_{\max}}} \mathbb{P}_{G,\mathcal{W}} \left(R \geq T/N, \ \mathbf{m}^T \neq \hat{\mathbf{m}}^T \ \middle| \ \mathbf{z}, \mathbf{m} \right) \ .$$

Note that we can upper bound $\varepsilon_{\rm ach}(\mathbf{z})$:

$$\varepsilon_{\rm ach}(\mathbf{z}) \leq \varepsilon_{\rm dec}(\mathbf{z}) + \max_{\mathbf{m} \in \{0,1\}^{NC_{\max}}} \mathbb{P}_{G,\mathcal{W}} \left(R \geq T/N \ \middle| \ \mathbf{z}, \mathbf{m} \right) \ .$$

Note that this channel model assumes a known finite horizon N, unlike the infinite horizon model of Shayevitz and Feder [3]. Furthermore, the basic model assumes an unbounded amount of common randomness in the form of the real number G. This point is discussed further in Section V.

C. Mutual information definitions

The results in this paper are stated in terms of mutual information quantities involving time-averaged channels dependent on the individual state sequence $\mathbf{z}$. For fixed $\mathbf{z}$ define the *state-averaged channel* to be

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

Note that if $\mathbf{z}$ and $\mathbf{z}'$ have the same type, then the state-averaged channels generated by them are the same. Define the empirical channel for a distribution Q on $\mathcal{Z}$:

$$W_Q(y|x) = \sum_{z\in\mathcal{Z}} W(y|x, z)Q(z) \ .$$

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

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

For an individual state sequence $\mathbf{z}$ the *empirical mutual information* is given by $I\left(P, W_{\mathbf{z}}\right)$.

D. Optimality versus empirical capacity

We are interested in analyzing strategies that can adapt their rates depending on the state sequence, and in our analysis, we want to consider the rates achieved by a strategy as a function of the state sequence. Unlike the compound channel setting (see e.g. [26] for definitions), which considers the worst-case behavior of a strategy over a class of channels, we instead want strategies that perform universally well over all sequences. However, this raises the problem of finding a notion of optimality that does not depend on the worst-case performance.

One possibility is to define an optimal strategy as one that, for every state sequence, achieves a rate at least as large as any other strategy for that sequence, and then define the capacity as the rates achieved by this strategy. However, this means comparing a strategy for all sequences against all strategies tailored to a fixed sequence. In the example in Section II-A, for each $\mathbf{z}$ there exists a decoding strategy which adds $\mathbf{z}$ to the output, undoing all of the bit flips. Each strategy achieves rate 1 for the specific choice of $\mathbf{z}$, but this is clearly an unreasonable target.

Instead, for each sequence we can consider a set of reference strategies and measure the "regret" of our strategy with respect to the reference strategies for each sequence. We take an approach inspired by source coding for individual sequences, in which we have a benchmark rate for each state sequence and then test whether a coding strategy attains the benchmark for each state sequence.

One such benchmark that we consider in this paper is the *empirical capacity* – for a fixed $\mathbf{z}$, the empirical capacity is defined as the supremum over all input distributions of the empirical mutual information:

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

First used by Shayevitz and Feder [3], empirical capacity is given its name not because it is purported to be optimal, but instead because of its resemblance to the capacity of point-to-point discrete memoryless channels.

There are two points that are worth mentioning before proceeding to describe the results in this paper. First, it is easy to see that the empirical capacity is a weaker target than the best possible strategy for a given sequence. It is possible that a strategy can achieve rates larger than the empirical capacity. In the binary modulo-additive example in Section II-A, if the sequence $\mathbf{z}$ were all 0 for the first half and all 1 for the second half, the empirical capacity is 0, whereas the coding strategy presented in this paper is expected to achieve rates close to 1.

Second, there may exist examples for which no strategy is guaranteed to achieve the empirical capacity. The coding strategy proposed in this paper uses a fixed input distribution P, and in general, the maximizing $P(x)$ may not be the same for all $\mathbf{z}$.[1] In these cases our strategy can achieve rates close to the empirical mutual information $I\left(P, W_{\mathbf{z}}\right)$ but not the empirical capacity $\bar{C}(\mathbf{z})$. It may be possible to adapt P over time, but at present we neither have a good strategy for achieving $\bar{C}(\mathbf{z})$ nor a counterexample showing that for some channels it is impossible to achieve $\bar{C}(\mathbf{z})$.

E. 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 $I\left(P, W_{\mathbf{z}}\right)$ for a large set of state sequences $\mathbf{z}$.

Theorem 1: Let $\{W(y|x, z) : z \in \mathcal{Z}\}$ be a given family of channels. Then given any $\rho > 0$, $\varepsilon > 0$, $\lambda^* > 0$, and channel input distribution P, there exists an N sufficiently large and an $(N, n_{\text{fb}}, B_{\text{fb}})$ coding strategy with feedback rate

$$R_{\text{fb}} = \frac{B_{\text{fb}}}{n_{\text{fb}}} < \lambda^* \ , \tag{3}$$

[1] A question then arises of how one chooses the input distribution P. One possibility could be to choose P to be uniform over the input alphabet. However, depending on the setting, other approaches might be preferable. Inspired by the theory of AVCs, one may choose the input distribution to be

$$P = \operatorname*{argmax}_{P'} \inf_{Q: I\left(P', W_Q\right) > \rho} I\left(P', W_Q\right) \ , \tag{2}$$

where ρ is a parameter governing the gap between the rates guaranteed by the algorithm and the empirical mutual information of the channel. This approach can run into problems in some situations in which for the P chosen, $I\left(P, W_Q\right) = 0$ for a large subset of state distributions Q, but there exists a distribution $\tilde{P}$ for which $I\left(\tilde{P}, W_Q\right) \geq \rho$ for all Q. On the other hand, if one were to remove the condition that $I\left(P', W_Q\right) > \rho$, for the example in Section II-A, $\inf_Q I\left(P', W_Q\right) = 0$ for all choices of P', and the choice of P' would be arbitrary. Because of such issues, we will leave the question of how to choose the input distribution P unanswered in this work. The problem of choosing P is similar to that studied by Shulman and Feder [25].

such that for all $\mathbf{z} \in T_Q(N)$, the rate

$$R \geq I\left(P, W_Q\right) - \rho \tag{4}$$

is achievable with probability $1 - \varepsilon$.

Binary modulo-additive channels, revisited: For the binary additive example in Section II-A, p denoted the fraction of ones in the noise sequence $\mathbf{z}$. Then, the empirical capacity is $1 - h(p)$, the capacity of the binary symmetric channel with crossover probability p. Theorem 1 implies the existence of strategies employing asymptotically zero-rate feedback such that for all $\rho, \varepsilon > 0$ and sufficiently large N,

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

is achievable with probability at least $1 - \varepsilon$.

Z-channels with unknown crossover, revisited: For the example in Section II-B with q equal to the fraction of 1's in the crossover sequence, the capacity-achieving input distribution is a function of q, so the theorem cannot guarantee a scheme achieving the empirical capacity. Despite this, it still provides achievable rates in this setting. If the channel input distribution has $P(X = 1) = p_x$ for this channel, then the empirical mutual information for this channel can be written as

$$I\left(P, W_q\right) = h(p_x) - (1 - p_x + p_x q) h\left(\frac{p_x q}{1 - p_x + p_x q}\right) \ ,$$

and is asymptotically achievable from Theorem 1. As discussed briefly at the end of Section III-D, the question of how to select p_x is outside the framework of this paper.

F. Proposed coding strategy: Randomized rateless code

The achievability result in Theorem 1 relies on the following coding strategy, which can be thought of as an iterated rateless code with randomized training (or, for short, randomized rateless code). The overall scheme is illustrated in Figure 2. The scheme divides time into chunks of $b(N)$ channel uses and in each round attempts to send $k(N)$ bits using a randomized rateless code. Each chunk contains a randomly interleaved training sequences, so the decoder can estimate the empirical channel. The decoder chooses to decode when the empirical rate falls below the estimated empirical mutual information calculated from the channel estimates. The round ends after the $k(N)$ bits are decoded, and the encoder starts a new round to send the next $k(N)$ bits. The length of each round is variable and depends on the empirical state sequence.

We now describe each component of the scheme in more detail.

1) Feedback: Divide the blocklength N into *chunks* of length $b = b(N)$ channel uses. Feedback occurs at the end of chunks, so $n_{\text{fb}} = b$ with three possible messages: "BAD NOISE," "DECODED," and "KEEP GOING," which correspond to the feedback messages 00, 01, and 10, respectively. Thus, $B_{\text{fb}} = 2$, so the feedback rate $R_{\text{fb}} = \lambda(N)$ is given by the expression

$$R_{\text{fb}} = \frac{B_{\text{fb}}}{b(N)} \ . \tag{5}$$

If the chunk size $b(N)$ goes to infinity as $N \to \infty$, the feedback rate $\lambda(N) \to 0$.

2) Rateless coding: A rateless code is a variable-length coding scheme to send a fixed number of bits. In the algorithm proposed here, the encoder attempts to send $k = k(N)$ bits over several chunks comprising a *round*. Rounds vary in length and terminate at the end of chunks in which the decoder feeds back either "BAD NOISE" or "DECODED." Let ℓ_r denote the time index at the end of round r:

$$\begin{aligned} \ell_r = \min\{ & j = i \cdot b(N) > \ell_{r-1} \\ & : \phi_i = \text{"BAD NOISE" or "DECODED"} \} \ , \end{aligned} \tag{6}$$

and set $\ell_0 = 0$.

An (M^*, c, k) *rateless code* is a sequence of maps $\{(\mu_i, \nu_i) : i = 1, 2, \ldots M^*\}$, where

$$\mu_i : \{0,1\}^k \to \mathcal{X}^c \tag{7}$$

$$\nu_i : \mathcal{Y}^{i \cdot c} \to \{0,1\}^k \ . \tag{8}$$

The encoding maps μ_i produce successive chunks of a codeword for a given message, and the decoding maps attempt to decode the message based on the channel outputs. An (M^*, c, k) *randomized* rateless code is a random variable that takes values in the set of (M^*, c, k) rateless codes. The *maximal error probability* $\hat{\varepsilon}(M, \mathbf{z}) = \hat{\varepsilon}(M, \mathbf{z}, \mathcal{D})$ for a randomized rateless code $\mathcal{D}$ decoded at time Mc with state sequence $\mathbf{z} \in \mathcal{Z}^{Mc}$ is

$$\hat{\varepsilon}(M, \mathbf{z}, \mathcal{D}) \tag{9}$$

$$= \max_{m \in \{0,1\}^k} \mathbb{E}\left[W^{Mc}\left(\left\{ \nu_M(\mathbf{y}_1^{Mc}) \neq m \right\} \ \middle| \ \mu_i(m), \mathbf{z} \right) \right]$$

$$= \max_{m \in \{0,1\}^k} \varepsilon_m(M, \mathbf{z}, \mathcal{D}) \ , \tag{10}$$

where the expectation is taken over the randomness in the code. We will suppress dependence on $\mathcal{D}$ when it is clear from context. The randomized rateless code used in this paper has codewords with constant composition $P(x)$ on $\mathcal{X}$ and uses a maximum mutual information (MMI) decoder.

3) Training : The coding strategy analyzed in this paper uses a randomized rateless code in conjunction with randomly located training symbols. The training allows the decoder to estimate the channel and choose an appropriate decoding time. For each chunk of b channel uses, the scheme uses $t = t(N)$ positions for training. Using the common randomness G, the encoder and decoder select t *training positions* $T_{r,n}$ for the n-th chunk[2] of round r. Formally, $T_{r,n}$ is uniformly distributed over subsets of $\{\ell_{r-1} + (n-1)b + 1, \ldots, \ell_{r-1} + nb\}$ of cardinality t. This set is further randomly partitioned into $|\mathcal{X}|$ subsets $T_{r,n}(x)$ for $x \in \mathcal{X}$.

4) Encoding: The encoder attempts to send a message $m \in \{0,1\}^{NC_{\max}}$ over several rounds. In each round it attempts to send a sub-message $m_r \in \{0,1\}^k$ consisting of k bits of m. The sub-message m_1 is the first k bits of m. If the round $r-1$ ended with "BAD NOISE" then $m_r = m_{r-1}$, and if round $r-1$ ended with "DECODED" then m_r is the next k bits of the message m.

The encoder and decoder share an $(M^*, b-t, k)$ randomized rateless code. Using the common randomness G, at the start of

[2]There is a slight abuse of notation with the type $T_N(Q)$, but the double subscript in $T_{r,n}$ should make the distinction unambiguous.

each round the encoder and decoder choose an $(M^*, b-t, k)$ rateless code $\{(\mu_j, \nu_j) : j = 1, 2, \dots M^*\}$ according to the distribution of this randomized code. Define the encoding map η_i in the n-th chunk of the r-th round:

$$\eta_i(m_r, G) = x \qquad i \in T_{r,n}(x) \tag{11}$$

$$\eta_i(m_r, G) = \mu_n(m_r) \qquad i \notin T_{r,n} \ . \tag{12}$$

That is, the n-th chunk transmitted by the scheme is created by taking the $b-t$ piece of the codeword $\mu_n(m_r)$ and inserting the t randomly chosen training positions, as illustrated in Figure 2. The dependence of η_i on the feedback is suppressed here because a round r is terminated as soon as the feedback message is no longer "KEEP GOING."

5) Decoding: The decoder uses the training symbols $\{y_i : i \in T_{r,n}(x)\}$ to estimate the channel transition probabilities $W_{\mathbf{z}}(y|x)$ and thereby obtain an estimate of the empirical mutual information $I(P, W_{\mathbf{z}})$ during the chunk and over the round so far. If the estimated mutual information is too low, then it feeds back "BAD NOISE." If the estimated mutual information is above the empirical rate $k/(n(b-t)) + \epsilon_1$ then it decodes the code using the MMI decoder ν_n of the rateless code and feeds back "DECODED." Otherwise, it feeds back "KEEP GOING." The parameter ϵ_1 ensures that with high probability the empirical rate is below the true empirical mutual information of the channel.

6) Algorithm : The parameters of the algorithm are the chunk size $b(N)$, training size $t(N)$, number of bits per round k, and decoding thresholds ϵ_1 and τ.

Given an $(M^*, b-t, k)$ randomized rateless code and message bits m_r, the encoder and decoder first use common randomness to choose a realization of the randomized rateless code. The following steps are then repeated for each chunk in round r:

1) Using common randomness, the encoder and decoder choose $t = t(N)$ positions $T_{r,n}$ and a random partition of $T_{r,n}$ into $|\mathcal{X}|$ subsets $T_{r,n}(x)$ of size $t/|\mathcal{X}|$ for training in chunk n.
2) The encoder transmits the n-th chunk using the encoding map as defined in Equations (11)–(12). In particular, the symbol x is sent during the training positions $T_{r,n}(x)$.
3) The decoder estimates the empirical channel in chunk n and the empirical channel over the round so far:

$$\hat{w}_r^{(n)}(y|x) = \frac{|\mathcal{X}|}{t} \cdot |\{j \in T_{r,n}(x) : y_j = y\}| \tag{13}$$

$$\hat{W}_r^{(n)}(y|x) = \frac{1}{n} \sum_{i=1}^{n} \hat{w}_r^{(i)}(y|x) \ . \tag{14}$$

4) The decoder makes a decision based on $\hat{W}_r^{(n)}$ and n.
   a) If

$$I\left(P, \hat{W}_r^{(n)}\right) - \epsilon_1 < \tau \ , \tag{15}$$

   where $\tau > 0$ is a parameter of the algorithm, 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

$$I\left(P, \hat{W}_r^{(n)}\right) - \epsilon_1 > \frac{k}{(b-t) \times n} \ , \tag{16}$$

   where t is defined in Section III-F3, 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).

Thus, where $\phi_{r,n}$ denotes feedback in chunk n of round r, we have that

$$\phi_{r,n}(\mathbf{y}, G) = \begin{cases} \text{"BAD NOISE"} , & I\left(P, \hat{W}_r^{(n)}\right) - \epsilon_1 < \tau, \text{ and} \\ & I\left(P, \hat{W}_r^{(n)}\right) - \epsilon_1 \leq \frac{k}{(b-t)\times n} \\ \text{"DECODED"} , & I\left(P, \hat{W}_r^{(n)}\right) - \epsilon_1 > \frac{k}{(b-t)\times n} \\ \text{"KEEP GOING"} , & \text{otherwise} \end{cases} \tag{17}$$

This 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 rate is smaller than the empirical mutual information in order guarantee a small error probability.

We note again that the channel model and problem formulation involve a fixed overall blocklength N and other parameters of the coding strategy are defined in terms of this parameter. However, in practice it may be more desirable to fix a number of bits $k(N)$ to send per round and then define the coding parameters in terms of k. We have chosen the former method because it is convenient for our mathematical analysis, but we believe that in principle the problem could be formulated in an "infinite-horizon" manner as well. This may require developing appropriate tree-structured anytime codes [27].

IV. Analysis

Showing that the strategy proposed in the previous section satisfies the conditions of Theorem 1 requires some more notation. For each round r, let the random variable $M(r)$ be the number of chunks in that round:

$$M(r) = \inf_{n>0} \left\{ I\left(P, \hat{W}_r^{(n)}\right) - \epsilon_1 < \tau \right. \\ \left. \text{or} \ \ \frac{k}{(b-t)n} < I\left(P, \hat{W}_r^{(n)}\right) - \epsilon_1 \right\} . \tag{18}$$

Let $U_{n,r}$ denote the time indices in the n-th chunk of round r that are not in the training set $T_{n,r}$.

The scheme depends on a number of parameters – the overall blocklength N, the number of bits per round $k(N)$, the chunk size $b(N)$, the number of training positions per chunk $t(N)$, the *rate gap* $\epsilon_1(N)$, the error bound ε, and the

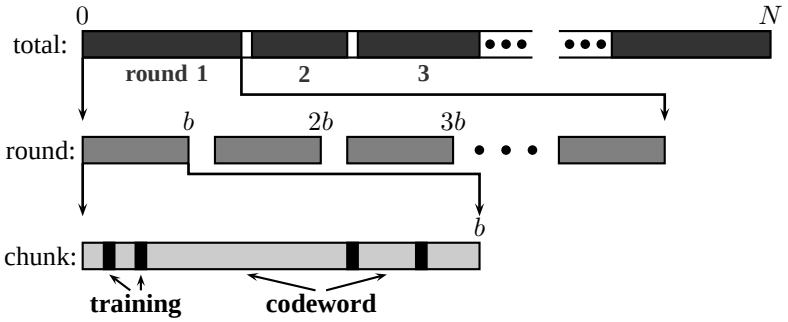

Fig. 2. After each chunk of length b feedback can be sent. Rounds end by decoding a message or declaring the noise to be bad.

feedback rate $\lambda(N)$. In order to make the proof of the result clear, assume that there exist real constants $g_1, g_2, g_3 \in (0, \frac{1}{2})$ with $g_1 > g_2 > g_3$ and set

$$k(N) = \Theta(N^{2g_1}),\ b(N) = \Theta(N^{g_2}),\ t(N) = \Theta(N^{g_3})\ . \tag{19}$$

In particular, this means that the ratios $k(N)/N \to 0$, $(b(N))^2/k(N) \to 0$, and $t(N)/b(N) \to 0$.

A. Error events

The scheme requires that the channel estimates $\hat{W}_r^{(M(r))}$ in (14) be "good" in two senses. First, $\hat{W}_r^{(M(r))}$ should be close to the average channel seen by the codeword in the non-training positions $\{U_{n,r}\}$ (defined after (18) above), and it should also be close to the channel averaged over the entire round. The former guarantees that the estimates provided by training are close enough to guarantee that the rateless code is decodable, and the latter guarantees the gap between the rates achieved by the scheme and the empirical mutual information is small. A *channel estimation error* $E_1(r)$ occurs for round r if

$$\left| I\left(P, \hat{W}_r^{(M(r))}\right) - I\left(P, \frac{1}{M(r)(b-t)} \sum_{n=1}^{M(r)} \sum_{i \in U_{n,r}} W(y|x, z_i)\right) \right| > \frac{\epsilon_1}{2} \tag{20}$$

or

$$\left| I\left(P, \hat{W}_r^{(M(r))}\right) - I\left(P, \frac{1}{M(r)b} \sum_{n=1}^{M(r)} \sum_{i \in U_{n,r} \cup T_{n,r}} W(y|x, z_i)\right) \right| > \frac{\epsilon_1}{2}\ . \tag{21}$$

A *decoding error* $E_2(r)$ happens in round r if the rateless code selected by the encoder and decoder experiences an error.

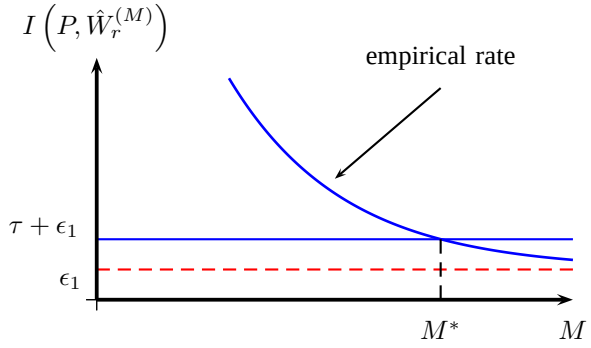

Fig. 3. Curve of the empirical rate illustrating the bounds on M. The upper bound M^* is given by (22).

B. Preliminaries: Bounding the length of a round

Before proceeding to bound the probabilities of the error events, we will provide bounds on the length of a round. Our reasons for establishing these are two-fold. First, if a round fails to terminate or does not result in successful decoding, the round length should be sufficiently small so that its impact on the overall rate should be small. Second, when taking union bounds over chunks in a round, the round length should be small enough to guarantee the corresponding error probabilities are small. Moreover, it helps set the maximum length for the randomized rateless code. Lemma 1 provides bounds on $M(r)$, the number of chunks in round r, which can be expressed equivalently as $(\ell_r - \ell_{r-1})/b(N)$, where ℓ_r is defined in (6). For simplicity, we will use M to denote $M(r)$ when the round r is clear from context.

Lemma 1 (Bounds on M): Fix $\epsilon_1 > 0$ and $\tau > 0$. Then for the scheme described in Section III-F6, the stopping time M for any round satisfies $M \le M^*$, where

$$M^* := \left\lceil \frac{k(N)}{(b(N) - t(N)) \cdot \tau} \right\rceil\ . \tag{22}$$

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

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

Proof: The argument is illustrated in Figure 3. The empirical rate given by (16) is shown in the curve. The empirical rate $\frac{k}{(b-t)\times M}$ decreases monotonically with M. In

order for the algorithm to continue at time M, from (17) we must have $\frac{k}{(b-t)\times M} \geq I\left(P, \hat{W}_r^{(n)}\right) - \epsilon_1 \geq \tau$. Rearranging shows that M must be less than M^* in (22). The lower bound is trivial from the definition in (16) and the cardinality bound on mutual information. ∎

C. Channel estimation for a single round

In this section, we provide an upper bound on the error event $E_1(r)$. The argument relies on the following observation: if sufficiently many samples are collected to estimate the channel, these estimates converge to the overall average channel. Lemmas 2 and 3 make this precise. That is, with a modest number of randomly chosen training symbols, the decoder can estimate the empirical mutual information of the channel such that the probability of the channel estimation error event $E_1(r)$ is small.

Lemma 2 (Simple channel estimation): Recall the chunk training estimates defined in (13), and let parameters satisfy the conditions in (19). Then for any $\epsilon_4 > 0$ there exists an N sufficiently large and constant a_1 such that for the j-th chunk the training estimates satisfy:

$$\mathbb{P}\left(\exists\ x, y \text{ s.t. } \left|\hat{w}_r^{(j)}(y|x) - W_{\mathbf{z}(U_{r,j}\cup T_{r,j})}(y|x)\right| \geq \epsilon_4\right) \leq \exp\left(-a_1\epsilon_4^2 t\right)$$

$$\mathbb{P}\left(\exists\ x, y \text{ s.t. } \left|\hat{w}_r^{(j)}(y|x) - W_{\mathbf{z}(U_{r,j})}(y|x)\right| \geq \epsilon_4\right) \leq \exp\left(-a_1\epsilon_4^2 t\right)\ ,$$

where t is the size of the training set $T_{r,j}$.

Proof: Proving the claim requires two applications of Hoeffding's inequality [28] to the training data. The first uses the sampling with replacement version of the inequality to show that the training estimates are close to the state-averaged channel at those training positions. The second uses the sampling without replacement version to show that the state-averaged channel in the training positions is close to the state-averaged channel over the entire chunk. An application of the triangle inequality and our parameter assumptions in (19) complete the argument.

We now make this precise. First consider the random variables $\{\mathbf{1}(y_i = y) : i \in T_{r,j}(x)\}$ for each x and y. Their expectations over the channel are $\{W(y|x, z_i) : i \in T_{r,j}(x)\}$. Applying Hoeffding's inequality to these variables shows that their empirical mean, which is $\hat{w}_r^{(j)}(y|x)$, is close to $W_{\mathbf{z}(T_{r,j}(x))}$, the average channel during the training:

$$\mathbb{P}\left(\left|\hat{w}_r^{(j)}(y|x) - W_{\mathbf{z}(T_{r,j})}(y|x)\right| \geq \epsilon_5\right) \leq 2\exp\left(-2\frac{t}{|\mathcal{X}|}\epsilon_5^2\right)\ . \tag{23}$$

Now, recall that the training positions $T_{r,n}$, defined in Section III-F3, are sampled uniformly without replacement from the whole chunk, so the average channel $W_{\mathbf{z}(T_{r,j}(x))}(y|x)$ is itself a random variable formed by averaging the random variable $\{W(y|x, z_i) : i \in T_{r,j}(x)\}$. The mean of each of these variables is $W_{\mathbf{z}(U_{r,j}\cup T_{r,j})}$, the state averaged channel over the whole chunk. For sampling without replacement, another result of Hoeffding [28, Theorem 4] states that the same exponential inequalities for sampling with replacement hold, so the channel during the training is a good approximation to the entire channel during the chunk:

$$\mathbb{P}\left(\left|W_{\mathbf{z}(T_{r,j}(x))} - W_{\mathbf{z}(U_{r,j}\cup T_{r,j})}\right| \geq \epsilon_5\right) \leq 2\exp\left(-2\frac{t}{|\mathcal{X}|}\epsilon_5^2\right)\ . \tag{24}$$

By applying the triangle inequality to equations (23) and (24), we have the following:

$$\mathbb{P}\left(\left|\hat{w}_r^{(j)}(y|x) - W_{\mathbf{z}(U_{r,j}\cup T_{r,j})}\right| \geq 2\epsilon_5\right) \leq 4\exp\left(-2\frac{t}{|\mathcal{X}|}\epsilon_5^2\right)\ . \tag{25}$$

Finally, observe the following:

$$\begin{aligned}
&\left|W_{\mathbf{z}(U_{r,j}\cup T_{r,j})} - W_{\mathbf{z}(U_{r,j})}\right| \\
&= \left|\frac{1}{b}\sum_{i\in U_{r,j}\cup T_{r,j}} W(y|x, z_i) - \frac{1}{b-t}\sum_{i\in U_{r,j}} W(y|x, z_i)\right| \\
&= \left|\frac{1}{b}\sum_{i\in T_{r,j}} W(y|x, z_i) - \frac{t}{b(b-t)}\sum_{i\in U_{r,j}} W(y|x, z_i)\right| \\
&\leq 2\frac{t}{b}\ .
\end{aligned} \tag{26}$$

The assumptions in (19) imply that (26) can be made small for sufficiently large N. Thus for N sufficiently large, another application of the triangle inequality to (25) and (26) gives the following:

$$\mathbb{P}\left(\left|\hat{w}_r^{(j)}(y|x) - W_{\mathbf{z}(U_{r,j})}\right| \geq 3\epsilon_5\right) \leq 4\exp\left(-2\frac{t}{|\mathcal{X}|}\epsilon_5^2\right)\ . \tag{27}$$

Choosing $\epsilon_4 = 3\epsilon_5$ and a union bound over all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ we get

$$\begin{aligned}
&\mathbb{P}\left(\exists\ x, y \text{ s.t. } \left|\hat{w}_r^{(j)}(y|x) - W_{\mathbf{z}(U_{r,j})}\right| \geq \epsilon_4\right) \\
&\qquad \leq \exp\left(-(2/9)\frac{t}{|\mathcal{X}|}\epsilon_4^2 + \log|\mathcal{X}||\mathcal{Y}| + \log 4\right)\ . \\
&\qquad \leq \exp\left(-a_1\epsilon_4^2 t\right)\ ,
\end{aligned}$$

where the last inequality follows from taking N sufficiently large and the fact that $t(N)$ increases with N. ∎

Lemma 3 (Channel estimation): Recall the error event $E_1(r)$ defined in Section IV-A, and let the parameters satisfy the conditions in (19). Then for any $\epsilon_1 > 0$ there exists N sufficiently large and an $a_2 > 0$ such that for any round r and

any state sequence $\mathbf{z} \in \mathcal{Z}^{M(r)b}$,

$$\mathbb{P}\left(\left| I\left(P, \hat{W}_r^{(M(r))}\right) - I\left(P, \frac{1}{M(r)(b-t)} \sum_{n=1}^{M(r)} \sum_{i \in U_{n,r}} W(y|x, z_i)\right)\right| > \frac{\epsilon_1}{2}\right) \leq \exp\left(-a_2 t\right) \quad (28)$$

$$\mathbb{P}\left(\left| I\left(P, \hat{W}_r^{(M(r))}\right) - I\left(P, \frac{1}{M(r)b} \sum_{n=1}^{M(r)} \sum_{i \in U_{n,r} \cup T_{n,r}} W(y|x, z_i)\right)\right| > \frac{\epsilon_1}{2}\right) \leq \exp\left(-a_2 t\right) \ . \quad (29)$$

Therefore $\mathbb{P}(E_1(r)) \leq 2\exp(-a_2 t)$.

Proof: For all (x, y), Lemma 2 guarantees that for any $\epsilon_4 > 0$ the channel estimated during the training of any chunk is within ϵ_4 of the average channel during the whole chunk and during the codeword positions with probability $\exp\left(-a_1 \epsilon_4^2 t\right)$. For a round of length $M(r)$, a union bound over chunks shows that

$$\mathbb{P}\left(\exists\ x, y \text{ s.t. } \left|\frac{1}{M(r)} \sum_{j=1}^{M(r)} \hat{w}_r^{(j)}(y|x) - \frac{1}{M(r)} \sum_{j=1}^{M(r)} W_{\mathbf{z}(U_{r,j} \cup T_{r,j})}(y|x)\right| \geq \epsilon_4\right) \leq M(r) \exp\left(-a_1 \epsilon_4^2 t\right) \quad (30)$$

$$\mathbb{P}\left(\exists\ x, y \text{ s.t. } \left|\frac{1}{M(r)} \sum_{j=1}^{M(r)} \hat{w}_r^{(j)}(y|x) - \frac{1}{M(r)} \sum_{j=1}^{M(r)} W_{\mathbf{z}(U_{r,j})}(y|x)\right| \geq \epsilon_4\right) \leq M(r) \exp\left(-a_1 \epsilon_4^2 t\right) \ . \quad (31)$$

Since $M(r)$ is at most M^*, for N sufficiently large the effect of the union bound is negligible.

The remainder of the proof is to show that if the channel estimated from the training is close with high probablity to both the average channel during the codeword positions and the average channel during the whole round, then the empirical mutual informations must be close as well. Lemma 7 in the Appendix shows exactly this. For any $\epsilon_1 > 0$ there exists a $\epsilon_4 > 0$ and N sufficiently large such if the events in (30) and (31) fail to hold then the events in (29) and (28) also fail to hold. This completes the proof. ∎

Remark: Under the parameter assumptions in equation (19), the number of bits of common randomness needed in Lemmas 2 and 3 to specify the training positions is sublinear in the blocklength N. Note that a similar conclusion was reached by Shayevitz and Feder for their scheme, which also uses training positions to the estimate the channel [3]. This point is discussed in more detail in Section V.

D. Rateless coding

The last ingredient in our strategy is the rateless code used during each round. The key property we need is that if the empirical rate drops below the empirical mutual information of the channel, then the code can be decoded with small probability of error.

Lemma 4 (Rateless codes): For any $\delta' > 0$ and distribution P, there exists an integer c sufficiently large, $\epsilon_8 > 0$ and an (M^*, c, k) randomized rateless code defined in Section III-F such that if at decoding time M the state sequence $\mathbf{z}_1^{Mc}$ satisfies

$$\frac{k}{Mc} \leq I\left(P, W_{\mathbf{z}_1^{Mc}}\right) - \delta' \ ,$$

then its maximal error $\hat{\varepsilon}(M, \mathbf{z})$, defined in (9), satisfies

$$\hat{\varepsilon}(M, \mathbf{z}) < \exp(-Mc\epsilon_8) \ .$$

Proof: Fix δ' and a distribution P. We can approximate P arbitrarily closely with a type of a sufficiently large denominator, so without loss of generality, we assume P is a type and choose c to be large enough so that the denominator of type P divides c. Let $\mathcal{C}_M(J)$ be a randomized rateless code. Specifically, $\mathcal{C}_M(J)$ is a random variable distributed on the set of rateless codes of blocklength Mc whose J codewords are drawn independently and uniformly from the composition-P set $T_{Mc}(P)$ and with a maximum mutual information (MMI) decoder. The remainder of the proof can be sketched as follows: we verify that the codebook $\mathcal{C}_M(J)$ has satisfactory error performance under the assumptions of this Lemma. Then, we construct a codebook $\mathcal{D}_M(K)$ by keeping only those codewords in $\mathcal{C}_M(J)$ whose composition is P in each chunk of c symbols. We then show that the distribution of $\mathcal{D}_M(K)$ is the same as that of a codebook $\mathcal{E}_{M^*}(K)$ truncated to blocklength Mc.

Codebook properties. Before proceeding to construct $\mathcal{D}_M(K)$, we first examine properties of the constant-composition codebook $\mathcal{C}_M(J)$ of composition P. Recall the definition of maximal error for randomized rateless codes in (9) and (10). A result of Hughes and Thomas [29, Theorem 1] shows that for sufficiently large Mc, there exists a function E_r such that for all $J > 0$, $\delta > 0$, and distribution Q on $\mathcal{Z}$,

$$\max_{\mathbf{z} \in T_{Mc}(Q)} \max_{j \in [J]} \varepsilon_j(M, \mathbf{z}, \mathcal{C}_M(J)) \leq \exp\left(-Mc\left[E_r((Mc)^{-1}\log J + \delta, W, P, Q) - \delta\right]\right) \quad (32)$$

$$E_r((Mc)^{-1}\log J + \delta, W, P, Q) \geq \max\left\{0,\ I\left(P, W_Q\right) - \delta - \frac{1}{Mc}\log J\right\} \ . \quad (33)$$

Fix $\epsilon_7 = \frac{\delta'}{4}$ and let $\mathcal{Q}(M)$ be the set of all Q such that

$$0 < \frac{\delta'}{4} \leq I\left(P, W_Q\right) - 2\delta - \frac{1}{Mc}\log J \ . \quad (34)$$

If $Q \in \mathcal{Q}(M)$, then we can rewrite the bound in (32) as follows:

$$\max_{\mathbf{z}\in T_{Mc}(Q):Q\in\mathcal{Q}(M)} \max_{j\in[J]} \varepsilon_j(M,\mathbf{z},\mathcal{C}_M(J)) \le \exp\left(-Mc\epsilon_7\right) \ .$$

In particular, this gives the following bound on the expectation over $\mathcal{C}_M(J)$ of the *average error*:

$$\max_{\mathbf{z}\in T_{Mc}(Q):Q\in\mathcal{Q}(M)} \mathbb{E}_{\mathcal{C}_M(J)}\left[\frac{1}{J}\sum_{j=1}^{J}\varepsilon_j(M,\mathbf{z},\mathcal{C}_M(J))\right] \le \exp\left(-Mc\epsilon_7\right) \ .$$

Use Markov's inequality to bound the probability that the average error exceeds a given value α_1:

$$\max_{\mathbf{z}\in T_{Mc}(Q):Q\in\mathcal{Q}(M)} \mathbb{P}_{\mathcal{C}_M(J)}\left(\frac{1}{J}\sum_{j=1}^{J}\varepsilon_j(M,\mathbf{z},\mathcal{C}_M(J)) \ge \alpha_1(c,M)\right) \le \frac{\exp\left(-Mc\epsilon_7\right)}{\alpha_1(c,M)} \ .$$

This establishes that for any $\delta > 0$ the codebook has average error no more than $\alpha_1(M)$ with high probability.

Expurgation. We define a thinning operation on the codebook $\mathcal{C}_M(J)$ to form the codebook $\mathcal{D}_M(K)$ as follows: remove all codewords in $\mathcal{C}_M(J)$ which are not in the piecewise constant-composition set $\{T_c(P)\}^M$. That is, we keep only those codewords which have type P in each chunk. If there are fewer than K remaining codewords after this expurgation, declare an encoding error – if there are more than K then keep the first K codewords. The decoding rule is the same MMI rule as before.

The probability of this encoding error can be bounded using Lemma 8, which states that the probability that a codeword drawn uniformly from $T_{Mc}(P)$ is also in the set $\{T_c(P)\}^M$ is at least $\beta_0(c,M) = \exp(-\eta M\log c)$ for c sufficiently large. Therefore the expected number of codewords in $\mathcal{C}_M(J)$ that survive the thinning is at least $J\exp(-\eta M\log c)$. Since the codewords are i.i.d., the probability that the number of codewords surviving the thinning is at least βJ can be bounded:

$$\mathbb{P}\left(\left|\mathcal{C}_M(J)\cap\{T_c(P)\}^M\right| \le \beta J\right) \le J\cdot\exp\left(-J\cdot D\left(\beta\middle\|\beta_0(c,M)\right)\right) \ .$$

By choosing $K = \beta_0(c,M)^2 J$, which corresponds to $\beta = \beta_0(c,M)^2$, the probability of encoder error can be made arbitrarily small. The rate of codebook $\mathcal{D}_M(K)$ is

$$\frac{1}{Mc}\log K = \frac{1}{Mc}\log J - \frac{2\eta\log c}{c} \ .$$

Setting $k = \log K$, note from (34), for sufficiently large c the error can be made small as long as

$$\frac{k}{Mc} \le \min_{Q\in\mathcal{Q}(M)} I\left(P,W_Q\right) - 3\delta - \frac{\delta'}{4} \ . \tag{35}$$

Setting $\delta = \delta'/4$ in the original construction of $\mathcal{C}_M(J)$, for sufficiently large c, equation (35) guarantees a bound on the error. In particular, since the codewords of $\mathcal{D}_M(K)$ are a subset of the codewords of $\mathcal{C}_M(K)$, the average error can increase at most by a factor of J/K:

$$\max_{\mathbf{z}\in T_{Mc}(Q):Q\in\mathcal{Q}(M)} \mathbb{P}_{\mathcal{C}_M(J)}\left(\frac{1}{K}\sum_{j=1}^{K}\varepsilon_j(M,\mathbf{z},\mathcal{D}_M(K)) \ge \frac{\alpha_1(c,M)}{\beta_0(c,M)^2}\right) \le \frac{\exp\left(-Mc\epsilon_7\right)}{\alpha_1(c,M)} \ . \tag{36}$$

This shows that for any $\delta' > 0$ the average error can be bounded.

Nesting. Consider the codebook $\mathcal{E}_M(K)$ formed by drawing K codewords independently uniformly distributed on $\{T_c(P)\}^M$ together with the MMI decoding rule. It is clear that $\mathcal{D}_M(K)$ has the same distribution as $\mathcal{E}_M(K)$, so the bound (36) holds for $\mathcal{E}_M(K)$ as well:

$$\max_{\mathbf{z}\in T_{Mc}(Q):Q\in\mathcal{Q}(M)} \mathbb{P}_{\mathcal{E}_M(K)}\left(\frac{1}{K}\sum_{j=1}^{K}\varepsilon_j(M,\mathbf{z},\mathcal{E}_M(K)) \ge \frac{\alpha_1(c,M)}{\beta_0(c,M)^2}\right) \le \frac{\exp\left(-Mc\epsilon_7\right)}{\alpha_1(c,M)} \ . \tag{37}$$

Note that $\mathcal{E}_M(K)$ has the same distribution as the codebook $\mathcal{E}_{M^*}(K)$ truncated to blocklength Mc. The set of $\mathbf{z}\in\mathcal{Z}^{M^*c}$ for which the bounds (37) hold is

$$\mathcal{Z}(K) = \Big\{\mathbf{z}\in\mathcal{Z}^{M^*c} : (z_1,\ldots,z_{Mc})\in T_{Mc}(Q), \ Q\in\mathcal{Q}(M),\ M\in\{M_*,\ldots,M^*\}\Big\} \ .$$

For any $\mathbf{z}$ in this set and decoding time M such that $(z_1,\ldots,z_{Mc})\in T_{Mc}(Q)$ for some $Q\in\mathcal{Q}(M)$, the probability that the random codebook $\mathcal{E}_{M^*}(K)$ truncated to blocklength M has average error probability exceeding $\frac{\alpha_1(c,M)}{\beta_0(c,M)^2}$ can be made arbitrarily small.

Back to maximal error. The equation (37) says that the average error under the randomized code $\mathcal{E}_M(K)$ can be made arbitrarily small. Standard results on AVCs [26, Exercise 2.6.5] show that by permuting the message index the same bound holds for the maximal error. Thus with probability $1-\exp(-Mc\epsilon_7)/\alpha_1(c,M)$ the randomly selected codebook has maximal error smaller than $\frac{\alpha_1(c,M)}{\beta_0(c,M)^2}$. The probability of encoding error is vanishingly small with respect to these quantities, so the total probability of error can be upper bounded:

$$\hat{\varepsilon}(M,\mathbf{z}) < \max\left(\frac{\exp\left(-Mc\epsilon_7\right)}{\alpha_1(c,M)}, \frac{\alpha_1(c,M)}{\beta_0(c,M)^2}\right) < \max\left(\frac{\exp\left(-Mc\epsilon_7\right)}{\alpha_1(c,M)}, \frac{\alpha_1(c,M)}{\exp(-2\eta M\log c)}\right) \ .$$

Selecting $\alpha_1(c,M) = \exp(-Mc\epsilon_7/2)$ yields the following bound for sufficiently large c:

$$\hat{\varepsilon}(M,\mathbf{z}) < \exp(-Mc\epsilon_7/3) \ .$$

Setting $\epsilon_8 = \epsilon_7/3$ yields the result. ■

Remark: As stated, the codebook constructed in Lemma 4 requires a very large amount of common randomness shared between the encoder and decoder. This issue is discussed in more detail in Section V.

E. Proof of Theorem 1

We now combine the results in the previous sections to prove Theorem 1. Namely, in Section IV-A, we defined error events $E_1(r)$ and $E_2(r)$. We then provided bounds on $E_1(r)$ in Lemma 3 and proved the existence of a randomized rateless code with a small maximal error probability in Lemma 4. As will be seen in the proof, Lemmas 3 and 4 provide a bound on $E_2(r)$. By combining this bound with the bound on $E_1(r)$ and parameter assumptions in (19), the result follows straightforwardly.

Proof: The proof is divided into three parts. We first establish in equation (38) that for sufficiently large N, the feedback rate can be made arbitrarily small. In the second part, we bound the error probability in (44). In the third part, we give a lower bound on the rate under the assumption the error event does not occur, which leads to equation (49). These parts establish all necessary components in the statement of the result.

We use the coding strategy proposed in Section III-F. Note that under the parameter assumptions in (19), for all $\lambda^* > 0$, there exists sufficiently large N such that the feedback rate (5) satisfies the following bound:

$$R_{\text{fb}} < \lambda^* \ . \tag{38}$$

Fix a sequence $\mathbf{z}$. The scheme induces a random partition of $\mathbf{z}$ into rounds $r = 1, 2, \ldots$ at times $\{\ell_r\}$. Let $\mathbf{z}(r) = \mathbf{z}_{\ell_{r-1}+1}^{\ell_r}$ be the state sequence during the r-th round. The type of $\mathbf{z}$ can be written as:

$$T_{\mathbf{z}} = \sum_r \frac{\ell_r - \ell_{r-1}}{N} T_{\mathbf{z}(r)} \ ,$$

where ℓ_r is the length of a round, as defined in equation (6). Lemma 3 shows that for any $\epsilon_1 > 0$ there exists an N sufficiently large such that the channel estimation error probability $\mathbb{P}(E_1(r))$ is exponentially small. Taking a union bound over all rounds, the probability of estimation error is

$$\mathbb{P}\left(\bigcup_r E_1(r)\right) \leq 2\frac{N}{b}\exp\left(-a_2 t\right) \ . \tag{39}$$

By the parameter assumptions in (19), N/b and t grow polynomially in N, so for large N the exponential term dominates and the probability of an estimation error in any round goes to 0. Given any $\varepsilon > 0$, for sufficiently large N, equation (39) gives the following bound:

$$\mathbb{P}\left(\bigcup_r E_1(r)\right) \leq \frac{\varepsilon}{2} \ . \tag{40}$$

Suppose round r was terminated due to "BAD NOISE." In this case, from (15) we have the following:

$$I\left(P, \hat{W}_r^{(M(r))}\right) - \epsilon_1 < \tau \ .$$

By Lemma 3, $I\left(P, \hat{W}_r^{(M(r))}\right)$ is close to $I\left(P, W_{\mathbf{z}(r)}\right)$. That is, there exists an N sufficiently large such that with probability $1 - \exp\left(-a_2 t\right)$, we have that $I\left(P, W_{\mathbf{z}(r)}\right) < \tau + 3\epsilon_1/2$. For any $\rho > 0$, we can choose a large N and small τ such that the following holds for all "BAD NOISE" rounds:

$$I\left(P, W_{\mathbf{z}(r)}\right) < \rho/2 \ . \tag{41}$$

Therefore, for rounds which are terminated due to bad noise, the state sequence $\mathbf{z}(r)$ has a type $T_{\mathbf{z}(r)}$ such that $I\left(P, W_{\mathbf{z}(r)}\right)$ is small.

Now suppose the decoder attempted to decode at the end of round r. Then (16) implies that the estimated empirical mutual information from the training satisfies a different inequality:

$$I\left(P, \hat{W}_r^{(M(r))}\right) - \epsilon_1 > \frac{k}{(b-t)\cdot M(r)} \ .$$

If the event $E_1(r)$ does not happen, then $I\left(P, \hat{W}_r^{(M(r))}\right)$ is within $\epsilon_1/2$ of the empirical mutual information during the non-training positions:

$$\frac{k}{(b-t)\cdot M(r)} < I\left(P, \frac{1}{r(b-t)} \sum_{n=1}^{M(r)} \sum_{i \in U_{n,r}} W(y|x, z_i)\right) - \frac{\epsilon_1}{2} \ . \tag{42}$$

Thus, conditioned on $E_1^c(r)$ and under our assumption (19), (42) and Lemma 4 imply that for $\delta' = \epsilon_1/2$ there exists a sufficiently large N, exponent $\epsilon_8 > 0$, and an $(M^*, b-t, k)$ randomized rateless code with error $\hat{\varepsilon}(M, \mathbf{z}) < \exp(-M(b-t)\epsilon_8)$ for every round r in which decoding occurs. A union bound then implies the decoding error probability over all rounds in which decoding occurs can be bounded:

$$\mathbb{P}\left(\bigcup_r E_2(r) \ \middle| \ \bigcap_r E_1^c(r)\right) \leq \frac{N}{b}\exp(-(b-t)\epsilon_8) \ . \tag{43}$$

By (19), this can be made arbitrarily small for sufficiently large N, and therefore for any $\varepsilon > 0$, (40) and (43) imply there exists an N sufficiently large such that the estimation error and decoding error can be made smaller than ε:

$$\mathbb{P}\left(\bigcup_{r,i=1,2} E_i(r)\right) \leq \varepsilon \ . \tag{44}$$

The remaining thing is to calculate the rate, given that none of the error events occur. If the decoder attempted to decode after $M(r)$ chunks, then after $M(r)-1$ chunks the threshold condition in (16) was not satisfied:

$$\frac{k}{(b-t)\cdot(M(r)-1)} \geq I\left(P, \hat{W}_r^{(M(r)-1)}\right) - \epsilon_1 \ ,$$

Our assumption in equation (19) that $(b(N))^2/k(N) \to 0$ and our lower bound on the length of a round in Lemma 1 is $\Theta(k(N)/b(N))$ channel uses imply that for sufficiently large N, the amount that the estimated mutual information can change over the course of a the final chunk in a round ($b(N)$

channel uses) can be made arbitrarily small. More formally, for any $\epsilon_6 > 0$, for sufficiently large N,

$$\left| I\left(P, \hat{W}_r^{(M(r)-1)}\right) - I\left(P, \hat{W}_r^{(M(r))}\right) \right| < \epsilon_6 \ .$$

Thus

$$\begin{aligned} \frac{k}{(b-t)\cdot M(r)} &= \left(1 - \frac{1}{M(r)}\right) \frac{k}{(b-t)\cdot (M(r)-1)} \\ &\geq \left(1 - \frac{1}{M(r)}\right) \left(I\left(P, \hat{W}_r^{(M(r)-1)}\right) - \epsilon_1\right) \\ &\geq \left(1 - \frac{1}{M(r)}\right) \left(I\left(P, \hat{W}_r^{(M(r))}\right) - \epsilon_6 - \epsilon_1\right) \ . \end{aligned}$$

Finally, the overall empirical rate for the round is slightly lower because of overhead from training:

$$\begin{aligned} \frac{k}{bM(r)} \geq \left(1 - \frac{1}{M(r)}\right) \left(1 - \frac{t}{b}\right) \\ \left(I\left(P, \hat{W}_r^{(M(r))}\right) - \epsilon_6 - \epsilon_1\right) \end{aligned}$$

Under the assumptions in (19) and conditioned on (21) not occurring, for any $\rho > 0$ there exists an N sufficiently large such that

$$\frac{k}{bM(r)} \geq I\left(P, \hat{W}_r^{(M(r))}\right) - \rho/2 \ . \tag{45}$$

The final source of rate loss is the last round r^*, which may not conclude within the overall blocklength, since $\ell_{r^*} = N$. The maximum length of this round is M^*b, and

$$\frac{\ell_{r^*} - \ell_{r^*-1}}{N} I\left(P, W_{\mathbf{z}(r^*)}\right) \leq \frac{M^*b}{N} \max\{|\mathcal{X}|, |\mathcal{Y}|\} \ . \tag{46}$$

By (19), for sufficiently large N, (46) can be made to satisfy the following condition:

$$\frac{\ell_{r^*} - \ell_{r^*-1}}{N} I\left(P, W_{\mathbf{z}(r^*)}\right) \leq \rho/2 \ . \tag{47}$$

To summarize, for sufficiently large N and each round r in which the decoder feeds back "BAD NOISE" or "DECODED", the rate at which the scheme decodes can be lower bounded by

$$R(r) \geq I\left(P, W_{\mathbf{z}(r)}\right) - \rho/2 \ , \tag{48}$$

which follows from (41) and (45). Finally, we use (47), (48), and the convexity of mutual information to provide a lower bound on the overall rate of the scheme:

$$\begin{aligned} R &\geq \sum_{r=1}^{r^*-1} \frac{\ell_r - \ell_{r-1}}{N} \left(I\left(P, W_{\mathbf{z}(r)}\right) - \rho/2\right) \\ &\geq I\left(P, \sum_r \frac{\ell_r - \ell_{r-1}}{N} W_{\mathbf{z}(r)}\right) - \rho \\ &= I\left(P, W_{\mathbf{z}}\right) - \rho \ . \end{aligned} \tag{49}$$

As mentioned above, the result now follows immediately from (38), (44), and (49). ■

V. Discussion

The central question we tried to address in this paper was how much feedback is needed to achieve the channel mutual information in the individual sequence setting of [3]. Limited feedback in two-way and relaying systems have been studied before [30]–[32] and are used in many modern-day communication protocols for control information. Research interest on limited feedback for multiuser and multiantenna models has grown tremendously (see [33] and references therein). Quantifying the role and possible benefits of limited feedback is an important step in understanding how to structure adaptive communication systems.

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 each element in 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 the empirical capacity, which is defined as 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 is a commentary on an earlier investigation by Shayevitz and Feder [3] 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. Furthermore, their scheme does not require a fixed blocklength in advance and hence has an infinite horizon. By contrast, our strategy can be viewed as a kind of incremental redundancy hybrid ARQ [7], in which the decoder uses the feedback link to terminate rounds that are too noisy while less noisy rounds are individually decoded. In order to set the parameters for our scheme we must fix a total blocklength in advance, although it may be possible to redefine the scheme to operate without a horizon, as in [3].

An interesting point is that our basic algorithm uses standard "tricks" for communication systems, such as channel estimation via pilot signals, ARQ with rateless codes, and randomization. By adapting or reusing technologies that have already been developed, these gains can be realized more easily. Several open questions and extensions of the algorithm presented here would be of interest, two of which are the following:

1) *The necessary amount of common randomness.* Common randomness serves at least three roles in coding arguments. Firstly, standard probabilistic method arguments to show the existence of good codes can be thought of as a use of common randomness. Secondly, common randomness can be used as a modeling tool to temper the inherently adversarial assumption that the state sequence is arbitrary while still preserving the notion that the channel is unknown. In our work,

common randomness enforces the requirement that the state selector act independently of the coding scheme. Finally, common randomness is an operational resource that is used as a secret key to combat malicious jammers or prevent two nearby systems from using the same codebook (e.g. spreading sequences in CDMA). Of these three roles it is important to quantify the *amount* of this third type of common randomness. In our scheme it is used by the encoder and decoder to choose *(i),* the channel training positions, and *(ii),* the codebook used in each round.

For *(i),* the training positions, under our parameter assumptions in (19), $\log N$ bits are required to indicate the position of each of the $t = \Theta(N^{g_3})$ training positions for each chunk of length $b = \Theta(N^{g_2})$, where $\frac{1}{2} > g_2 > g_3 > 0$. Since there are N/b chunks, this requires at total of

$$N \cdot \frac{t(N)}{b(N)} \cdot \log N = \Theta\left(N^{1-(g_2-g_3)} \cdot \log N\right) \text{ bits },$$

which, under our parameter assumptions is sublinear in N. For *(ii),* the selection of a codebook for each round can require as much as $M^* \cdot C_{\max}$ bits of common randomness per codeword for a total of $M^* \cdot C_{\max} \cdot 2^{M^* \cdot C_{\max}}$ bits of common randomness, where $C_{\max} = \log \min\{|\mathcal{X}|, |\mathcal{Y}|\}$. The total number of rounds can be as large as $\frac{N}{M_*}$, where M^* and M_* are defined in Lemma 1. Thus, codebook selection requires

$$M^* \cdot C_{\max} \cdot 2^{M^* \cdot C_{\max}} \cdot \frac{N}{M_*} = \frac{(C_{\max})^2}{\tau} \cdot N \cdot 2^{M^* \cdot C_{\max}} \text{ bits },$$

where τ, defined in (15), is a parameter of the algorithm that does not depend on N. Thus, the total common randomness required is superlinear in N.

Reducing this operational common randomness is outside the scope of the current work. However, if common randomness were not available between the encoder and decoder, it could be provided by the feedback link, but then the strategy considered in this paper would require a prohibitively large feedback rate that would increase with the blocklength N. To show instead that the feedback rate could be made asymptotically negligible in such a setting, one would need to prove the existence of a strategy for which the total bits of common randomness required would be sublinear in the blocklength N.

A potential technique that might be useful could be to adapt tools from the theory of arbitrarily varying channels [34] to find nested code constructions that use a limited amount of common randomness [35]. Such an argument would require showing that a randomized code with support on $T = (M^* b)^2$ codes can be made from i.i.d. sampling of the randomized code of Lemma 4. This new randomized code could then be used to establish a sublinear number of bits. Specifically, in each round, this new randomized code could be used by selecting one of the T codes for use. This would require $\log T = O(\log N)$ bits per round for a total cost of at most $O((N/M_*) \log N)$, which would be sublinear in N.

Another potential method, more in the interactive coding spirit of feedback systems, could be to show the existence of deterministic list-decodable codes with small list sizes. If the list is of size L, the decoder could find L bits in the message, which could be used to disambiguate the list [6]. By using $L \log k$ bits in the feedback, the decoder could request those L bits from the encoder. By sacrificing just $O(L)$ more forward channel uses, the encoder could send the L bits with negligible impact to the rate. If the empirical mutual information in the next round were above τ, this would be sufficient for success.

2) *Adaptation of the channel input, and thus, codebook distribution.* An apparent limitation of the algorithm presented here is that the channel input distribution is selected once and kept fixed throughout, irrespective of the behavior of the state sequence. Adaptation of the channel input distribution may lead to higher *or* lower rates. One interesting question would be whether universal prediction techniques [36] can be used in conjunction with channel coding to adapt the channel input. Another set of interesting questions emerges if we consider performance on a sequence that comes from a certain class of sequences. For example, if one were to consider an alternate notion of empirical capacity in which the empirical sequences were estimated as finite-order Markov models, adapting the channel input distribution may give quantifiable benefits.

The individual sequence model considered in this paper is by no means the only way of modeling channel uncertainty. One model which does away with modeling the channel state was recently proposed by Lomnitz and Feder [37]. An alternative model within the state sequence framework is 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 [38] 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 model uncertainty regarding the feedback link. 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 [39]. This work grew out of a presentation

of that work for UC Berkeley's Fall 2006 advanced information theory course EE290S. Special thanks go to the other students in that class for helpful discussions. We also thank our Associate Editor Ioannis Kontoyiannis and the reviewers for their insightful comments and Hari Palaiyanur and Jiening Zhan for their comments on the manuscript.

APPENDIX

We provide here the proofs of the lemmas used in the analysis of our algorithm[3].

A. Bounds on entropy and mutual information

We need a short technical lemma about concave functions.

Lemma 5: Let f be a concave increasing function on $[a, b]$. Then if $a \le x \le x + \epsilon \le b$, we have

$$f(x+\epsilon) - f(x) \le f(a+\epsilon) - f(a) \ . \tag{50}$$

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

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

Therefore

$$f(x) + f(\epsilon) \ge f(x+\epsilon) \ ,$$

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

$$|P(s) - Q(s)| \le \epsilon \qquad \forall s \in \mathcal{S} \ ,$$

then

$$\begin{aligned} |H(P) - H(Q)| \le (|\mathcal{S}| - 1)\cdot h_b(\epsilon) \\ + (|\mathcal{S}| - 1)\log(|\mathcal{S}| - 1)\cdot\epsilon \ , \end{aligned}$$

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

Proof: Let $\mathcal{S} = \{s_1, s_2, \ldots\}$. 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 5 to each interval, we obtain the bound:

$$|h_b(x+\epsilon) - h_b(x)| \le h_b(\epsilon) \ .$$

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

Now suppose that the lemma holds for $|\mathcal{S}| \le m - 1$, 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 $P' = \lambda^{-1}(P(s_1), \ldots, P(s_{m-1}))$ and $Q' = \lambda^{-1}(Q(s_1), \ldots, Q(s_{m-1}))$, so that

$$\begin{aligned} P &= (\lambda P', (1-\lambda)) \\ Q &= (\mu Q', (1-\mu)) \ . \end{aligned}$$

Therefore,

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

Now we 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 $H(Q')$ yields the result:

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

∎

Lemma 7: Let $W(y|x)$ and $V(y|x)$ be two channels with finite input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$. If

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

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

$$\begin{aligned} |I(P, W) - I(P, V)| \le 2(|\mathcal{Y}| - 1)\cdot h_b(\epsilon) \\ + 2(|\mathcal{Y}| - 1)\log(|\mathcal{Y}| - 1)\cdot\epsilon \ , \end{aligned}$$

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

Proof: We simply apply Lemma 6 twice. Let Q_W and Q_V be the marginal distributions on $\mathcal{Y}$ under channels W and V respectively. Then

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

Now we can break apart the mutual information and use Lemma 6 on each term:

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

∎

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

B. Properties of concatenated fixed composition sets

Let $\mathbf{T}_n(P) = \{\mathbf{x} \in \mathcal{X}^n : T_{\mathbf{x}} = P\}$ be the set of 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 8: 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 sufficiently large n, for all $M > 0$:

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

Proof: We begin with the following [26, p. 39] :

$$\begin{aligned} kH(P) - \frac{|\mathcal{X}|-1}{2}\log(2\pi k) - \nu_1(P) & \\ \leq \log|\mathbf{T}_k(P)| & \\ \leq kH(P) - \frac{|\mathcal{X}|-1}{2}\log(2\pi k) - \nu_2(P) \ , & \end{aligned}$$

for $0 < \nu_1(P) < \infty$ and $0 < \nu_2(P) < \infty$ since $p_x \geq p_0$ for all x. From this we can take the ratio:

$$\begin{aligned} \log\frac{|\mathbf{T}_n(P)|^M}{|\mathbf{T}_{Mn}(P)|} \geq -M\frac{|\mathcal{X}|-1}{2}\log(2\pi n) - M\nu_1(P) & \\ + \frac{|\mathcal{X}|-1}{2}\log(2\pi M n) + \nu_2(P) \ . & \end{aligned}$$

For fixed P and sufficiently large n, this lower bound is $\Omega(M\log n)$, which establishes the result. ∎

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] S. Draper, B. Frey, and F. Kschischang, "Rateless coding for non-ergodic channels with decoder channel state information," submitted to IEEE Transactions of Information Theory.

[3] O. Shayevitz and M. Feder, "Achieving the empirical capacity using feedback: Memoryless additive models," *IEEE Transactions on Information Theory*, vol. 55, no. 3, pp. 1269–1295, March 2009.

[4] A. Sahai, "Why block-length and delay behave differently if feedback is present?" *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 1860–1886, May 2008.

[5] A. Sahai and S. Draper, "Beating the Burnashev bound using noisy feedback," in *Proceedings of the 44th Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Sep. 2006.

[6] A. Sahai, "Balancing forward and feedback error correction for erasure channels with unreliable feedback," submitted to IEEE Transactions of Information Theory.

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

[8] M. Horstein, "Sequential transmission using noiseless feedback," *IEEE Transactions on Information Theory*, vol. 9, no. 3, pp. 136–143, July 1963.

[9] 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.

[10] M. Burnashev, "Data transmission over a discrete channel with feedback, random transmission time," *Problems of Information Transmission*, vol. 12, no. 4, October–December 1976.

[11] T. Cover and S. Pombra, "Gaussian feedback capacity," *IEEE Transactions on Information Theory*, vol. 35, pp. 37–43, 1989.

[12] 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.

[13] J. Ooi, *Coding for channels with feedback*. Boston, MA: Kluwer Academic Publishers, 1998.

[14] Y.-H. Kim, "Feedback capacity of stationary Gaussian channels," 2006, submitted to IEEE Transactions of Information Theory. [Online]. Available: http://arxiv.org/abs/cs.IT/0602091

[15] S. Draper and A. Sahai, "Variable-length channel coding with noisy feedback," *European Transactions on Telecommunications*, vol. 19, no. 4, pp. 355–370, April 2008.

[16] M. Gastpar and G. Kramer, "On noisy feedback for interference channels," in *Proceedings of the 2006 Asilomar Conference on Signals, Systems, and Computers*, 2006.

[17] M. Wigger, "Noisy feedback is strictly better than no feedback on the Gaussian MAC," 2006 Kailath Symposium, July 2006.

[18] A. Lapidoth and M. A. Wigger, "On the Gaussian MAC with imperfect feedback," *IEEE Transactions on Information Theory*, To appear.

[19] 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.

[20] M. Luby, "LT codes," in *Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science*, 2002.

[21] A. Shokrollahi, "Fountain codes," in *Proceedings of the 41st Allerton Conference on Communication, Control, and Computing*, October 2003, pp. 1290–1297.

[22] N. Shulman, "Communication over an unknown channel via common broadcasting," Ph.D. dissertation, Tel Aviv University, 2003.

[23] 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.

[24] 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.

[25] N. Shulman and M. Feder, "The uniform distribution as a uniform prior," *IEEE Transactions on Information Theory*, vol. 50, no. 6, pp. 1356–1362, June 2004.

[26] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Budapest: Akadémi Kiadó, 1982.

[27] A. Sahai and S. Mitter, "The necessity and sufficiency of anytime capacity for control over a noisy communication link: Part I," *IEEE Transactions on Information Theory*, vol. 52, no. 8, pp. 3369–3395, August 2006.

[28] 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.

[29] B. Hughes and T. Thomas, "On error exponents for arbitrarily varying channels," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 87–98, 1996.

[30] L. Schwartz, "Feedback for error control and two-way communication," *IEEE Transactions on Communications Systems*, vol. 11, no. 1, pp. 49–56, March 1963.

[31] J. Hayes, "Adaptive feedback communications," *IEEE Transactions on Communications Technology*, vol. 16, no. 1, pp. 29–34, February 1968.

[32] N. Ahmed, M. Khojastepour, A. Sabharwal, and B. Aazhang, "Outage minimization with limited feedback for the fading relay channel," *IEEE Transactions on Communications System*, vol. 54, no. 4, pp. 659–669, April 2006.

[33] D. Love, R. Heath, Jr., V. Lau, D. Gesbert, B. Rao, and M. Andrews, "An overview of limited feedback in wireless communication systems," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 8, pp. 1341–1365, October 2008.

[34] R. Ahlswede, "Elimination of correlation in random codes for arbitrarily varying channels," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 44, no. 2, pp. 159–175, 1978.

[35] A. Sarwate and M. Gastpar, "Rateless codes for AVC models," November 2007, submitted to IEEE Transactions of Information Theory.

[36] N. Merhav and N. Feder, "Universal prediction," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2124–2147, October 1998.

[37] Y. Lomnitz and M. Feder, "Feedback communication over individual channels," in *Proceedings of the 2009 International Symposium on Information Theory*, Seoul, South Korea, 2009.

[38] 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.

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