

Universal communication part II: channels with memory

Yuval Lomnitz, Meir Feder
Tel Aviv University, Dept. of EE-Systems
Email: {yuvall,meir}@eng.tau.ac.il

Abstract—Consider communication over a channel whose probabilistic model is completely unknown vector-wise and is not assumed to be stationary. Communication over such channels is challenging because knowing the past does not indicate anything about the future. The existence of reliable feedback and common randomness is assumed. In a previous paper it was shown that the Shannon capacity cannot be attained, in general, if the channel is not known. An alternative notion of “capacity” was defined, as the maximum rate of reliable communication by any block-coding system used over consecutive blocks. This rate was shown to be achievable for the modulo-additive channel with an individual, unknown noise sequence, and not achievable for some channels with memory. In this paper this “capacity” is shown to be achievable for general channel models possibly including memory, as long as this memory fades with time. In other words, there exists a system with feedback and common randomness that, without knowledge of the channel, asymptotically performs as well as any block code, which may be designed knowing the channel. For non-fading memory channels a weaker type of “capacity” is shown to be achievable.

Index Terms—Unknown channels, Universal communication, Feedback communication, Arbitrarily varying channels, Channels with memory.

I. INTRODUCTION

Consider communication over a channel which has a general probabilistic structure. In other words, the infinite length output \mathbf{Y}_1^∞ depends on the infinite length input \mathbf{X}_1^∞ through an arbitrary vector-wise probability function $P_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}_1^n|\mathbf{X}_1^\infty)$, $n = 1, 2, \dots$, which is unknown to the transmitter and the receiver. Particular cases of such a channel include any unknown functional relation between the input and output sequences, as well as arbitrarily varying channels, compound channels [1] and channels with an individual state sequence [2][3][4][5]. In the current paper, an attempt is made to keep the model as general as possible, i.e. minimize any assumptions on $P_{\mathbf{Y}|\mathbf{X}}$, except for causality. Without feedback, communication over such a channel is limited, as the communication rate, and the codebook would have to be selected in advance. Therefore, the existence of a reliable feedback link is assumed.

Two traditional models, which relate to particular cases of the current problem, are the arbitrarily varying channel (AVC) model [1] and the compound finite state channel (compound-FSC) model [6]. In the AVC model, the channel is assumed to be controlled by a sequence of states which is arbitrary and unknown to the transmitter and the receiver. In the compound channel model, the channel is assumed to be arbitrarily selected from a family of possible channels. In both models,

the capacity is the maximum rate of reliable communication that can be guaranteed. Both models do not give a satisfying answer to the current problem: the fundamental reason is that these models focus on capacity, i.e. before knowing the channel, one is required to find a rate of reliable transmission which can be guaranteed a-priori. Clearly, if the channel is completely general, the compound/AVC capacity is zero, as it is possible, for example, that a channel with zero capacity will be selected. In both models mentioned, constraints on the family of channels, or on the possible state sequences need to be defined, and these constraints do not seem suitable for natural channels. In addition to this fundamental gap, the models considered under the AVC and compound-FSC frameworks are quite limited, in a way that does not seem to capture the possible complexity of an unknown natural channel. For example, most papers on AVC consider only memoryless channels, and the compound-FSC is stationary.

Using feedback, the communication rate can be adapted, so that one does not have to commit to a communication rate a-priori. Several works by us and other authors considered the gains from such adaptation [2][3][7][5]. The first question to ask is, how the target communication rate should be defined? The sought rate $R(P_{\mathbf{Y}|\mathbf{X}})$ can be a function of the channel, but should be universally attainable without prior knowledge of the channel, and should have an operational meaning. Put simply, one would like to have a “universal modem” which can be connected over any channel, and would attain rates, which, may not be optimal, but would at least be justifiable and will not make one regret for not modeling the channel and using a modem optimized for the channel.

In a previous paper [4], the problem of determining such a communication rate was addressed. In general, the Shannon capacity [8] of the channel, $C(P_{\mathbf{Y}|\mathbf{X}})$ is not attainable universally with feedback, when the channel is unknown. This is exemplified in [4] through the simple example of the modulo-additive channel with an unknown noise sequence, where the Shannon capacity of each channel individually is positive (the logarithm of the alphabet size), while the maximum reliable communication rate that can be guaranteed a-priori is zero. The problem of determining a universally-achievable rate is similar to the source coding problem of setting a compression rate for an individual sequence. As in the universal source coding problem, due to the richness of the model family, there is a large gap between the performance that can be attained universally and the performance that can be attained without constraints, when knowing the specific model (the

Shannon capacity) and this gap requires limiting the abilities of the reference system. Following the spirit of the “finite state compressibility” of Lempel and Ziv [9], we proposed to set as a target, the best rate that can be reliably attained by a system employing finite block encoding (successively) over the infinite channel. The supremum of these rates is termed the Iterative-Finite-Block (IFB) capacity and denoted $C_{\text{IFB}}(P_{Y|X})$. When the channel is stationary and ergodic, then the IFB capacity equals the Shannon capacity. This motivates considering the IFB capacity as a goal.

It is easy to see that the IFB capacity is not universally achievable for completely general models. The counter example in [4] is of a family consisting of only two binary channels, termed “password” channels, where the first input bit X_1 determines whether the channel becomes “good” or “bad” for eternity, and where the values of X_1 matching each state are opposite in the two channels. There is no way for the universal system to correctly guess X_1 with high probability. The conclusion is that the IFB capacity is not universally attainable for some channels with infinite memory. On the other hand, the IFB capacity was shown to be asymptotically attainable for the class of modulo-additive channels with an individual, unknown noise sequence. In this case, it was further shown, that the IFB capacity is related to the finite state compressibility of the noise sequence, and the scheme attaining it uses the Lempel-Ziv source encoder [9] to generate decoding metrics. The result in [4] relies crucially on two properties of the modulo additive channel:

- 1) The channel is memoryless with respect to the input x_i (i.e. current behavior is not affected by previous values of the input).
- 2) The capacity achieving input distribution is fixed (uniform i.i.d.) regardless of the noise sequence.

To avoid these assumptions it is required to address the memory of the channel and the setting of the communication prior. The second limitation, raises the question, how the input distribution should be adapted, if the channel changes arbitrarily over time? This question was the center of [5], where universal prediction methods were used to set the communication prior. The focus of that paper is on channels which are memoryless in the input, and therefore can be defined by an unknown sequence of memoryless channels $P_{Y|X}(\mathbf{Y}_1^n | \mathbf{X}_1^n) = \prod_{i=1}^n W_i(Y_i | X_i)$. It is shown there that the capacity of the time-averaged channel $\bar{W}(y|x) = \frac{1}{n} \sum_{i=1}^n W_i(y|x)$ can be universally attained using feedback and common randomness without knowing $\{W_i\}$, and that this value is the maximum rate that can be achieved universally and does not depend on the order of the channels in the sequence. The notion of universality used in [5] is different and weaker than the IFB universality, since the rate is only compared with other rates that could have been universally attained.

In the current paper, ideas from [4] and [5] are combined to generalize the previous results. It is shown that the IFB capacity is asymptotically universally attainable for any channel with a fading memory, i.e. where the effect of the channel history on the far future is vanishing. In this sense, the two assumptions used in the previous paper [4] are avoided as

much as possible, and the assumptions made on the channel are significantly minimized. The fading memory condition includes as particular cases memoryless arbitrarily varying channels as well as compound indecomposable finite state channels [10]. Here, an example is given of a class of finite state channels where the state is a non-homogenous Markov chain, which satisfy the fading memory condition.

Considering channels where memory of the past is not necessarily fading, it may still be possible to communicate universally over the channel, if it is not maliciously designed like the password channel described above. The advantage of the IFB reference class which enables it to win over any universal system is its ability to determine such a codebook that will not only enable reliable transmission, but will also keep the channel in a favorable state, whereas the universal system does not know the long term effects of certain input symbols or distributions. An alternative formulation is proposed, where the reference system is crippled, so that it cannot enjoy the ability to shape the past: the encoder and decoder operate over finite blocks, however the error probability is required to be small in the worst case channel state (history) prior to each block, and average over blocks. This models a situation where the reference encoder and decoder are “thrown” each time into a different location in time, where the past state might have been arbitrary. It is not required to have good performance in each of these events, but only on average. This alternative reference system is termed “arbitrary-finite-block” (AFB) and the same universal system is shown to asymptotically approach the respective AFB capacity, without requiring that the channel memory is fading. This reference class is less natural than the IFB, yet it enables releasing constraints on the channel.

Note that there are several alternative definitions of a limited reference class for the universal communication problem [4]. Most notably, Misra and Weissman [11] generalized the main results of [4] to finite-state communication systems with feedback. For the sake of simplicity the current paper focuses on the basic model of reference systems using block coding. Although the current result is purely theoretical, it supplies motivation for using competitive universality in communication.

II. PROBLEM SETTING, DEFINITIONS AND MAIN RESULT

The definitions in Sections II-A-II-D repeat and extend the respective definitions in the previous paper [4]. Section II-E formalizes the main result of the paper.

A. Notation

Vectors are denoted by boldface letters. Sub-vectors are defined by superscripts and subscripts: $\mathbf{x}_j^i \triangleq [x_j, x_{j+1}, \dots, x_i]$. \mathbf{x}_j^i equals the empty string if $i < j$. The subscript is sometimes removed when it equals 1, i.e. $\mathbf{x}^i \triangleq \mathbf{x}_1^i$. For a vector \mathbf{x} , $\mathbf{x}_i^{[k]} \triangleq \mathbf{x}_{(i-1)k+1}^{(i-1)k+k}$ denotes the i -th block of length k in the vector. For brevity, vectors with similar ranges are sometimes joined together, for example, the notation $(\mathbf{xy})_1^k$ is used instead of $\mathbf{x}_1^k \mathbf{y}_1^k$. Exponents and logs are base 2. Random variables are distinguished from their sample values by capital letters. \mathbb{Z}^+ denotes the set of non-negative integers.

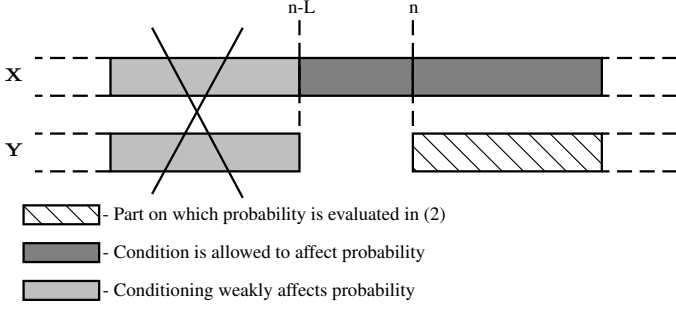


Fig. 1. An illustration of the fading memory condition (Definition 2).

$I(Q, W)$ denotes the mutual information obtained when using a prior Q over a channel W , i.e. it is the mutual information $I(Q, W) = I(X; Y)$ between two random variables with the joint probability $\Pr(X, Y) = Q(X) \cdot W(Y|X)$. $C(W)$ denotes the channel capacity $C(W) = \max_Q I(Q, W)$.

B. Channel model

Let \mathbf{x} and \mathbf{y} be infinite sequences denoting the input and the output respectively, where each letter is chosen from the alphabets \mathcal{X}, \mathcal{Y} respectively, $x_i \in \mathcal{X}, y_i \in \mathcal{Y}$. Throughout the current paper the input and output alphabets are assumed to be finite. A channel $P_{Y|X}$ is defined through the probabilistic relations $P_{Y|X}(\mathbf{y}^n | \mathbf{x}^\infty) = \Pr(\mathbf{Y}^n = \mathbf{y}^n | \mathbf{X}^\infty = \mathbf{x}^\infty)$ for $n = 1, 2, \dots$. A finite length output sequence is considered in order to make the probability well defined. Sometimes, this probability will be informally referred to as $\Pr(Y_1^n | X_1^\infty)$, and should be understood as the sequence of these distributions for $n = 1, 2, \dots$.

Definition 1. The channel defined by $\Pr(Y_1^n | X_1^\infty)$ is termed *causal* if for all n :

$$\Pr(\mathbf{Y}_1^n | \mathbf{X}_1^\infty) = \Pr(\mathbf{Y}_1^n | \mathbf{X}_1^n). \quad (1)$$

All the definitions below (including IFB/AFB capacity) pertain to causal channels. This characterization of a causal channel is similar to the definition used by Han and Verdú [8] (and references therein). This definition is also limited in assuming the channel starts from a known state (at time 0). However this does not limit the current setting, because an arbitrary initial state can be modeled by considering the family of channels with all possible initial states. Note that non causality that consists of bounded negative delays can always be compensated by applying a delay to the output.

Definition 2. The channel is termed a *fading memory channel* if for any $h > 0$ there exists L and a sequence of causal conditional vector distribution functions $\{P_n(\cdot | \cdot)\}$, such that for all n and $m \geq n$:

$$\|\Pr(\mathbf{Y}_n^m | \mathbf{X}_1^\infty, \mathbf{Y}_1^{n-L-1}) - P_n(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^\infty)\|_1 \leq h, \quad (2)$$

where the L_1 norm is calculated over \mathbf{Y}_n^m , and defined by $\|g(\mathbf{Y} | \cdot)\|_1 \triangleq \sum_{\mathbf{y}} |g(\mathbf{y} | \cdot)|$

The difference between the terms on the LHS of (2) is that P_n does not include $(\mathbf{X}\mathbf{Y})_1^{n-L-1}$ (see Fig.1), and thus the

fading memory condition asserts that the dependence of the conditional distribution of future outputs, on the channel state at the far past, decays. Notice that the conditional distribution $\Pr(\mathbf{Y}_n^m | \mathbf{X}_1^\infty, \mathbf{Y}_1^{n-L-1})$ is completely defined by the channel, since it is conditioned on the entire input \mathbf{X}_1^∞ . On the other hand, the conditional distribution $\Pr(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^\infty)$ may depend also on the input distribution (through the unspecified symbols \mathbf{X}_1^{n-L-1}). Therefore, the distribution P_n in Definition 2 is not identical to $\Pr(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^\infty)$. On the other hand, Proposition 1 shows that $\Pr(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^\infty)$ obtained with any input distribution yields a legitimate P_n .

The fading memory condition does not imply stationarity or ergodicity. The memoryless arbitrary varying channel model considered in [5] is fading memory, and so are the FSC [10, §4.6] or compound-FSC models [6], if the underlying FSC is indecomposable. An example of a non-homogeneous finite state channel with fading memory is presented in Section V.

C. IFB and AFB capacity

The following definitions lead to the definitions of IFB capacity and AFB capacity.

Definition 3 (Reference encoder and decoder). A finite length encoder E with block length k and a rate R is a mapping $E : \{1, \dots, M\} \rightarrow \mathcal{X}^k$ from a set of $M \geq \exp(kR)$ messages to a set of input sequences \mathcal{X}^k . A respective finite length decoder D is a mapping $D : \mathcal{Y}^k \rightarrow \{1, \dots, M\}$ from the set of output sequences to the set of messages.

Definition 4 (IFB error probability). The *average error probability in iterative mapping* of the k length encoder E and decoder D to b blocks over the channel $P_{Y|X}$ is defined as follows: b messages $\mathbf{m}_1, \dots, \mathbf{m}_b$ are chosen as i.i.d. uniformly distributed random variables $\mathbf{m}_i \sim U\{1, \dots, M\}$, $i = 1, \dots, b$. The channel input is set to $\mathbf{X}_i^{[k]} = E(\mathbf{m}_i)$, $i = 1, \dots, b$, and the decoded message is $\hat{\mathbf{m}}_i = D(\mathbf{Y}_i^{[k]})$ where \mathbf{Y} is the channel output. The iterative mapping is illustrated in Fig.2. The average error probability is $P_e = \frac{1}{b} \sum_{i=1}^b \Pr(\hat{\mathbf{m}}_i \neq \mathbf{m}_i)$.

Definition 5 (AFB error probability). The *average error probability in arbitrary mapping* of the k length encoder E and decoder D to b blocks over the channel $P_{Y|X}$ is defined as $P_e = \frac{1}{b} \sum_{i=1}^b P_e(i)$. $P_e(i)$ is the worst case per-block error probability, defined as:

$$P_e(i) = \max_{(\mathbf{X}\mathbf{Y})_1^{(i-1)k}} \left[\Pr \left\{ D(\mathbf{Y}_i^{[k]}) \neq \mathbf{m} \mid \mathbf{X}_i^{[k]} = E(\mathbf{m}), (\mathbf{X}\mathbf{Y})_1^{(i-1)k} \right\} \right], \quad (3)$$

where $\mathbf{m} \sim U\{1, \dots, M\}$.

Definition 6 (IFB/AFB achievability). A rate R is *iterated-finite-block (IFB) / arbitrary-finite-block (AFB) achievable* (resp.) over the channel $P_{Y|X}$, if for any $\epsilon > 0$ there exist $k, b^* > 0$ such that for any $b > b^*$ there exist an encoder E and a decoder D with block length k and rate R for which the average error probability in iterative/arbitrary mapping (resp.) of E, D to b blocks is at most ϵ .

This is equivalent to stating that the limsup of the average error probability with respect to b is at most ϵ .

Definition 7 (IFB/AFB capacity). The IFB/AFB capacity of the channel $P_{Y|X}$ is the supremum of the set of IFB/AFB achievable rates, and is denoted $C_{\text{IFB}}/C_{\text{AFB}}$ (resp.).

By definition, the AFB error probability is at least as large as the IFB error probability, and as a result, the AFB capacity is smaller than, or equal to the IFB capacity.

D. Competitive Universality

In the following, the properties of the adaptive system with feedback, and IFB/AFB-universality are defined. A randomized rate-adaptive transmitter and receiver for block length n with feedback are defined as follows (see also formal definitions in [12, §5]): the transmitter is presented with a message expressed by an infinite bit sequence, and following the reception of n symbols, the decoder announces the achieved rate R , and decodes the first $\lceil nR \rceil$ bits. An error means any of these bits differs from the bits of the original message sequence. Both encoder and decoder have access to a random variable S (the common randomness) distributed over a chosen alphabet, and a causal feedback link allows the transmitted symbols to depend on previously sent feedback from the receiver. The system is illustrated in Fig. 3.

The following definition states formally the notion of IFB/AFB-universality for rate adaptive systems:

Definition 8 (IFB/AFB universality). With respect to a set of channels $\{P_{Y|X}^{(\theta)}\}$, $\theta \in \Theta$ (not necessarily finite or countable), a rate-adaptive communication system (possibly using feedback and common randomness) is called IFB/AFB universal if for every channel in the family and any $\epsilon, \delta > 0$ there is n large enough such that when the system is operated over n channel uses, then with probability $1 - \epsilon$, the message is correctly decoded and the rate is at least $C_{\text{IFB}}(P_{Y|X}) - \delta$ or $C_{\text{AFB}}(P_{Y|X}) - \delta$ (resp.).

Notice that the definitions above (and specifically Definitions 6,8) do not require uniform convergence with respect to the channel, i.e. the number of channels uses n or blocks b for which the requirements hold may be a function of the channel.

E. The main result

Theorem 1. For any $\epsilon > 0$ there exists a sequence of adaptive rate systems over a block of size N with feedback and common randomness, for growing values of N , such that with a probability of at least $1 - \epsilon$ the message is received correctly with a rate of:

$$R_{\text{UNI}}[N] \geq \max[C_{\text{IFB}} - \delta_N^{\text{IFB}}, C_{\text{AFB}} - \delta_N^{\text{AFB}}], \quad (4)$$

where $\delta_N^{\text{AFB}} \xrightarrow[N \rightarrow \infty]{} 0$ for any causal channel, and $\delta_N^{\text{IFB}} \xrightarrow[N \rightarrow \infty]{} 0$ for any causal fading memory channel. Furthermore, this can be attained with any positive rate of the feedback link.

This implies that the system is IFB universal over the set of causal fading memory channels, and AFB universal over the set of causal channels, according to Definition 8. While the

system does not depend on the channel, the convergence rate of $\delta_N^{\text{IFB}}, \delta_N^{\text{AFB}}$ does.

III. COMMUNICATION SCHEME AND PROOF OUTLINE

A. The communication scheme

In [5], a communication scheme for adapting the prior over an arbitrarily varying channel which is memoryless in the input was described. Combining Theorem 3 and Lemma 9 of [5] yields:

Lemma 1. [Lemma 9 of [5]] For every $\tilde{\epsilon}, \tilde{\delta} > 0$ there exists n^* and a constant c_Δ , such that for any $n \geq n^*$ there is an adaptive rate system with feedback and common randomness, such that for any channel $\Pr(\mathbf{Y}_1^n | \mathbf{X}_1^n)$:

- 1) The probability of error is at most $\tilde{\epsilon}$
- 2) The rate satisfies $R \geq C(\bar{W}_{\text{SUBJ}}) - \tilde{\Delta}_C$ with probability at least $1 - \tilde{\delta}$

where

$$\bar{W}_{\text{SUBJ}} = \frac{1}{n} \sum_{i=1}^n \Pr(Y_i = y | X_i = x, \mathbf{X}^{i-1}, \mathbf{Y}^{i-1}), \quad (5)$$

and

$$\tilde{\Delta}_C = c_\Delta \cdot \left(\frac{\ln^2(n)}{n} \right)^{\frac{1}{4}}. \quad (6)$$

The universal communication scheme for attaining the claims of Theorem 1 is as follows. The infinite time is divided into epochs of increasing length, numbered $m = 1, 2, \dots$. In the first epoch, the scheme of Lemma 1 (described in [5]) is operated over N_1 symbols. In the second epoch, the channel inputs and outputs are joined into pairs, i.e. super-symbols of dimension 2, and the scheme is operated over N_2 such super-symbols. In epoch m , the scheme is operated over N_m super-symbols of dimension 2^{m-1} (Fig.4). Since all N_m are finite, the dimension of the super-symbols used grows indefinitely with time.

The parameters of the scheme are chosen as follows. Let $\epsilon > 0$ the chosen error probability. Choose any $\Delta_C > 0$, and let $\epsilon_m = \frac{1}{2}\epsilon \cdot 2^{-m}$. The length of the m -th epoch, N_m , is chosen such that:

- 1) It is equal to or larger than the value of n^* given by Lemma 1 for the parameters $\tilde{\epsilon} = \tilde{\delta} = \epsilon_m$.
- 2) The value of $\tilde{\Delta}_C$ given by Lemma 1 for $n = N_m$ is not larger than Δ_C (the chosen value).
- 3) If the end of the next epoch N_{m+1} would occur beyond symbol N , then the current epoch N_m is extended to reach symbol N .

The second requirement makes sure that there is no more than a constant loss from capacity per epoch, while the dimension of the super-symbol of each epoch is growing, and therefore the loss per symbol tends to zero. The values of $\tilde{\delta}$ and $\tilde{\epsilon}$ chosen per epoch, guarantee that the overall probability of error is not larger than $\sum_{m=1}^{\infty} \epsilon_m = \frac{1}{2}\epsilon$ and similarly the overall probability that at any epoch the rate falls below the rate declared in the lemma is at most $\frac{1}{2}\epsilon$. This way the overall probability of having an error or falling below the guaranteed

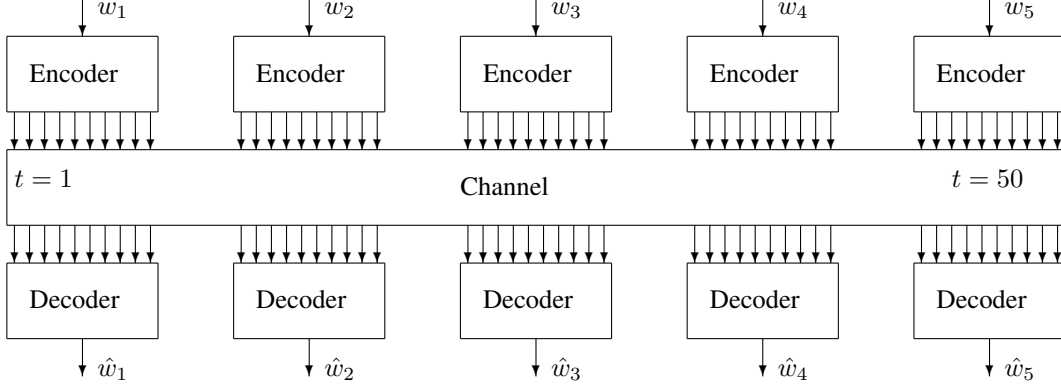


Fig. 2. An illustration of *iterative mapping* used for the definition of average error probability (see Definition 4). The same encoder and decoder are used over each of the $b = 5$ blocks of $k = 10$ channel uses, and the average error probability is computed.

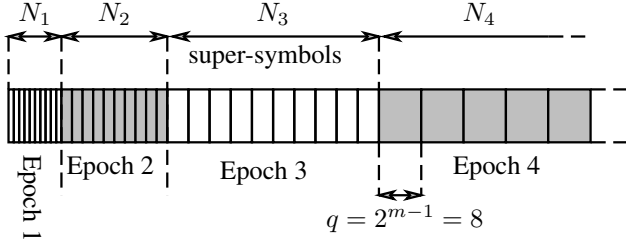


Fig. 4. Division into epochs and super-symbols in the universal scheme

rate is at most ϵ . Note that the epoch durations N_m are fixed and do not depend on the message or received signal.

The scheme does not need to know the IFB/AFB block length, rate and error probability, and the exact relation between L, h given by the fading memory condition (Definition 2). Its only parameters are the input and output alphabets, the number of symbols N , and the error probability ϵ .

The claim of Theorem 1, that any positive feedback rate is sufficient, simply follows from the fact [5] that this is true for the scheme of Lemma 1.

The scheme of Lemma 1 is a finite horizon scheme, i.e. n has to be set in advance, and there is no guarantee on the rate at the middle of an epoch. Due to this technical limitation, the universal scheme proposed here is also of a finite horizon, i.e. the symbol N in which the system's performance is to be measured is specified in advance. It is clear from the construction of the scheme that this limitation is technical and minor.

B. Proof outline

Following is the outline of the proof. The value $\Pr(Y_i = y | X_i = x, \mathbf{X}^{i-1}, \mathbf{Y}^{i-1})$ appearing in the definition of \bar{W}_{SUBJ} (5) is the probability of a certain output symbol to appear given a certain input symbol at time i , where the history of the channel $(\mathbf{X}\mathbf{Y})^{i-1}$ attains the specific value that occurred during the universal system's operation. $\Pr(Y_i = y | X_i = x, \mathbf{X}^{i-1}, \mathbf{Y}^{i-1})$ is a random variable and depends both on the channel and on

the universal communication system behavior. As a result, the rate \bar{W}_{SUBJ} guaranteed by Lemma 1 is also a random variable and depends on the joint input-output distribution induced by the universal communication scheme. This rate is termed “subjective” since it would be different had a different scheme operated on the same channel.

The baseline for comparison with the reference system is the “pessimistic average channel capacity”, (11) obtained by replacing the history $(\mathbf{X}\mathbf{Y})^{i-1}$ by an arbitrary state, and taking the worst-case state sequence (worst case history), i.e. the one that yields the minimum capacity. The rate attained by the universal system (for a particular state sequence) would be at least as large. For super-symbols, the averaged channel relates to the joint distribution over the super-symbol, where the state $(\mathbf{X}\mathbf{Y})^{(i-1)q}$ refers to the input and output sequences before the start of the super-symbol. The universal system is shown to asymptotically attain a rate which is at least the weighted average of the pessimistic average channel capacities measured over the epochs (Proposition 2).

Next, the reference system with block size k is compared to the universal system during epoch m , where the super-symbol length is $q = 2^{m-1}$. Consider a set of super-symbols in hops of k ($l \cdot k + j : l \in \mathbb{Z}^+, j = 1, \dots, k$). Since the number of symbols between the start of two successive super-symbols in each of these “alignment” sets divides by k , in each of these super-symbols, the reference system's blocks and the super-symbols align, i.e. the IFB/AFB blocks begin at the same location with respect to the beginning of the super-symbol (see Fig.5).

Therefore, there is an equivalence between the average error probability of the reference system over these super-symbols, and the error probability that would be attained for the “collapsed” channel, generated by randomly and uniformly drawing one of the super-symbols in the set and operating the reference system over this channel. Due to this equivalence, the reference system's rate, for a given average error probability, is limited by the capacity of the “collapsed” channel.

For the IFB case, this “collapsed” channel is induced not only by the channel law, but also by the behavior of the reference system in previous blocks. When replacing the

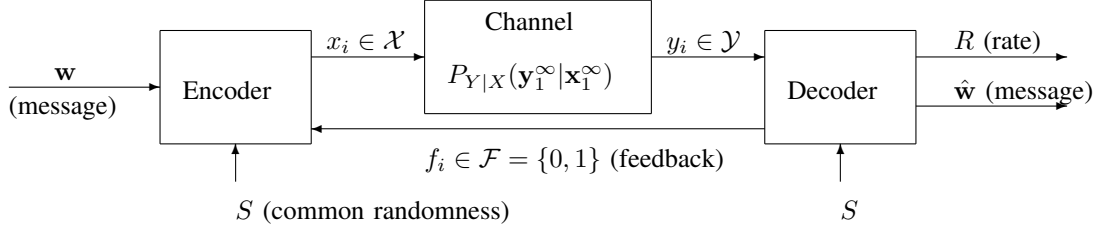


Fig. 3. Rate adaptive encoder-decoder pair with feedback, over an unknown channel

collapsed channel with a similar channel, where the history $(\mathbf{X}\mathbf{Y})^{(i-1)q}$ before each super-symbol is forced to a specific value, then due to the fading memory assumption, from some point in the block onward, the two channels become similar (in \mathcal{L}_1 sense). Due to this similarity, the increase in error probability, when exchanging the original “collapsed channel” with the new one, is small (Lemma 3). The new channel is not “subjective”, i.e. it is only a function of the channel $P_{Y|X}$ and not of the system operating over it. For the AFB case, this transition is not needed, as the desired relation stems immediately from the definition.

Using a variant of Fano’s inequality, the rate of the IFB/AFB system is related to the capacity of the pessimistic average channel measured over each of the k alignment sets of super-symbols (34). The pessimistic average channel over the epoch, is the average of the k average channels measured over the alignment sets. Averaging k channels may induce a loss of at most $\log k$ in capacity (Lemma 4). This results in a bound on the pessimistic average channel during each epoch, as a function of the IFB/AFB capacity, and the IFB/AFB error probability during the epoch. Note that at this stage, the error probability of the reference system cannot be dismissed as being small, it is guaranteed to be small only on average, over growing intervals in time. Taking the weighted average of the pessimistic capacities over the epochs enables relating the rate of the universal system to the rate and the average error probability of the reference system, where the latter tends to zero. All overheads, such as the ones related to alignment of the blocks to the super-symbols, the time it takes the channel memory to fade, the $\log k$ penalty for mixing k channels, vanish asymptotically as the super-symbol length increases indefinitely with time.

Although the result of this paper is simple, the proof is far from elegant, and the system, although simple, is not efficient in converging to its target. Let us hope that a more direct proof will be found in the future.

IV. PROOF OF THE MAIN RESULT

A. Additional notation for the proof

Additional notation required for the proof is defined below. The proof compares a situation where the reference (IFB) system operates on the channel to the universal system operating on the same channel. Although the channels are the same, the joint distribution of the input and the output is different due to the different encoders. The channel input and outputs when the universal system operates are denoted by \mathbf{X}, \mathbf{Y} , while $\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}$ denote the channel inputs and outputs when

the reference system operates. Since both systems operate on the same channel the conditional distribution is the same, i.e. $\Pr(\mathbf{Y}_1^n = \mathbf{y} | \mathbf{X}_1^\infty = \mathbf{x}) = \Pr(\tilde{\mathbf{Y}}_1^n = \mathbf{y} | \tilde{\mathbf{X}}_1^\infty = \mathbf{x})$

The following symbols have constant meaning throughout the proof. m denotes the epoch index, and q denotes the dimension of the super-symbol, which is a function of m ($q = 2^{m-1}$). k denotes the block length of the reference system. N denotes the overall number of symbols and M denotes the overall number of epochs.

B. Channel model preliminaries

The following simple conclusions follow from the definitions of the causal and the fading memory channel.

Regarding Definition 1 of a causal channel, note that the same holds for marginal distributions of \mathbf{Y}_1^n (e.g. the distribution of \mathbf{Y}_m^n) as is easily shown by summation over (1). Another consequence of Definition 1 is that for $n_1, n_2, n_3, n_4 \leq n$, the following conditional distribution can also be given as a function of a finite input:

$$\begin{aligned} \Pr(\mathbf{Y}_{n_1}^{n_2} | \mathbf{Y}_{n_3}^{n_4}, \mathbf{X}_1^\infty) &= \frac{\Pr(\mathbf{Y}_{n_1}^{n_2} \mathbf{Y}_{n_3}^{n_4} | \mathbf{X}_1^\infty)}{\Pr(\mathbf{Y}_{n_3}^{n_4} | \mathbf{X}_1^\infty)} \\ &= \frac{\Pr(\mathbf{Y}_{n_1}^{n_2} \mathbf{Y}_{n_3}^{n_4} | \mathbf{X}_1^n)}{\Pr(\mathbf{Y}_{n_3}^{n_4} | \mathbf{X}_1^n)} \\ &= \Pr(\mathbf{Y}_{n_1}^{n_2} | \mathbf{Y}_{n_3}^{n_4}, \mathbf{X}_1^n). \end{aligned} \quad (7)$$

Two simple consequences of Definition 2 (fading memory channel) are given below. The proof is simple and deferred to Appendix A.

Proposition 1. For a causal fading memory channel, the following holds

- 1) If (2) holds for a certain m , then it holds for any smaller m (as long as $m \geq n$). This implies that (2) only needs to be established for m “large enough”.
- 2) For any input distribution and for any $m > n$,

$$\|\Pr(\mathbf{Y}_n^m | \mathbf{X}_1^m, \mathbf{Y}_1^{n-L-1}) - \Pr(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^m)\|_1 \leq 2h. \quad (8)$$

In other words, the property applies when P_n is replaced with the true probability $\Pr(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^m)$, obtained with any input distribution.

C. A guarantee on the pessimistic rate

The rate W_{SUBJ} is subjective in the sense that it depends on the joint input-output distribution induced by the universal communication scheme, and would be different had a different scheme operated on the same channel.

In the following, a lower rate is defined, but such that is a function of the channel alone. Let $\overline{W}_{\text{SUBJ}}^{[q]}$ denote the subjective average channel over n super-symbols of dimension q :

$$\overline{W}_{\text{SUBJ}}^{[q]}(\mathbf{y}^q|\mathbf{x}^q) = \frac{1}{n} \sum_{i=1}^n \Pr\left(\mathbf{Y}_i^{[q]} = \mathbf{y} | \mathbf{X}_i^{[q]} = \mathbf{x}, (\mathbf{X}\mathbf{Y})^{(i-1)q}\right). \quad (9)$$

This channel is termed subjective since it depends on the specific input-output distribution induced by the universal scheme when operating on the channel (which is different, in general, from the joint distribution induced by a reference system). Furthermore, since $\Pr\left(\mathbf{Y}_i^{[q]} = \mathbf{y} | \mathbf{X}_i^{[q]} = \mathbf{x}, (\mathbf{X}\mathbf{Y})^{(i-1)q}\right)$ is a random variable depending on the history $(\mathbf{X}\mathbf{Y})^{(i-1)q}$, also $\overline{W}_{\text{SUBJ}}^{[q]}$ is a random variable, whose distribution depends on the joint distribution induced by the scheme. Also note that while the conditioning on $(\mathbf{X}\mathbf{Y})^{(i-1)q}$ represent what truly happened (as a random variable), this channel is not an empirical channel, since the probability $\Pr(\cdot)$ above represents what would have happened, hypothetically at the output, if one forced the input $\mathbf{X}_i^{[q]} = \mathbf{x}$.

Let $S_i^{[q]}$ denote the state before super-symbol i , $S_i^{[q]} = (\mathbf{X}\mathbf{Y})^{(i-1)q}$. As the channel is not a finite state channel, the alphabet size of $S_i^{[q]}$ increases with i . Consider the average channel when the history $\mathbf{X}^{(i-1)q}, \mathbf{Y}^{(i-1)q}$ obtains a specific value s_i :

$$\begin{aligned} \overline{W}^{[q]}(\mathbf{y}^q|\mathbf{x}^q; \{s_i\}_{i=1}^n) = \\ \frac{1}{n} \sum_{i=1}^n \Pr\left(\mathbf{Y}_i^{[q]} = \mathbf{y} | \mathbf{X}_i^{[q]} = \mathbf{x}, S_i^{[q]} = s_i\right). \end{aligned} \quad (10)$$

In other words, for fixed input and output, this is the average probability to see the specific output given the specific input when the channel had been in a specific state. This is no longer a random variable, but a function of $\{s_i\}$. The pessimistic average channel capacity is defined as the worst capacity of $\overline{W}^{[q]}(\cdot|\cdot; \{s_i\}_{i=1}^n)$ for any state sequence.

$$C_{\text{PMA}}^{[q]} = \inf_{\{s_i\}_{i=1}^n} C\left(\overline{W}^{[q]}(\mathbf{y}|\mathbf{x}; \{s_i\}_{i=1}^n)\right). \quad (11)$$

Note that in taking the minimum, (11) does not require that the state sequence satisfies the natural constraint given by the recursion $S_i = (S_{i-1}, X_i, Y_i)$, i.e. it is allowed to include so-called ‘‘contradictions’’. By definition, this rate lower bounds the capacity of the subjective averaged channel:

$$\begin{aligned} C\left(\overline{W}_{\text{SUBJ}}^{[q]}(\mathbf{y}^q|\mathbf{x}^q)\right) &= C\left(\overline{W}^{[q]}(\mathbf{y}|\mathbf{x}; \{s_i\}_{i=1}^n)\right) \Big|_{s_i=S_i} \\ &\geq \inf_{\{s_i\}_{i=1}^n} C\left(\overline{W}^{[q]}(\mathbf{y}|\mathbf{x}; \{s_i\}_{i=1}^n)\right) \\ &= C_{\text{PMA}}^{[q]}. \end{aligned} \quad (12)$$

Since in each epoch, the universal scheme asymptotically attains the capacity of $\overline{W}_{\text{SUBJ}}^{[q]}$ (measured over the epoch), it also attains $C_{\text{PMA}}^{[q]}$. The next proposition maintains that if it is guaranteed that the normalized pessimistic capacity, is asymptotically on average above some rate \overline{C} then the scheme will asymptotically approach the rate \overline{C} . The pessimistic capacity with super-symbol q measured over epoch m is denoted $C_{\text{PMA}}^{[q,m]}$.

Proposition 2. Assume that for each epoch m with super-symbol length $q = 2^{m-1}$, the pessimistic capacity satisfies:

$$\frac{1}{q} C_{\text{PMA}}^{[q,m]} \geq C_m - \delta_m, \quad (13)$$

where $\delta_m \xrightarrow{m \rightarrow \infty} 0$. Let $\overline{C} \triangleq \frac{1}{N} \sum_{m=1}^M 2^{m-1} N_m C_m$ denote the average of C_m weighted by the relative epoch durations. Then, for the universal scheme of Section III-A, over N symbols and M epochs, with probability at least $1 - \epsilon$, the message is correctly decoded and the rate satisfies:

$$R_{\text{UNI}}[N] \geq \overline{C} - \tilde{\delta}_N, \quad (14)$$

where $\tilde{\delta}_N \xrightarrow{N \rightarrow \infty} 0$.

Proof: By its construction and Lemma 1, in epoch m , with probability at least $1 - \epsilon_m$, the scheme attains the following rate, per super-symbol:

$$\begin{aligned} R_m &\geq C\left(\overline{W}_{\text{SUBJ}}^{[q,m]}(\mathbf{y}^q|\mathbf{x}^q)\right) - \Delta_C \stackrel{(12)}{\geq} C_{\text{PMA}}^{[q,m]} - \Delta_C \\ &\geq q(C_m - \delta_m). \end{aligned} \quad (15)$$

The number of bits sent during this epoch is at least $R_m \cdot N_m$. Let M denote the number of epochs until time N , where $N = \sum_{m=1}^M 2^{m-1} N_m$. With probability at least $1 - \epsilon$ (recall: $\epsilon = 2 \sum_{m=1}^M \epsilon_m$), there is no decoding error and the rate up to time N is at least:

$$\begin{aligned} R_{\text{UNI}}[N] &\geq \frac{\sum_{m=1}^M R_m \cdot N_m}{N} \\ &\stackrel{(15)}{\geq} \frac{1}{N} \sum_{m=1}^M (2^{m-1}(C_m - \delta_m) - \Delta_C) \cdot N_m \\ &= \overline{C} - \frac{1}{N} \sum_{m=1}^M (\delta_m + 2^{-m+1} \Delta_C) \cdot 2^{m-1} \cdot N_m \\ &= \overline{C} - \delta'_M. \end{aligned} \quad (16)$$

where $\delta'_M \xrightarrow{M \rightarrow \infty} 0$. The last step stems from the following simple lemma (see Appendix C):

Lemma 2. For a positive, monotonic non-decreasing sequence a_n and $0 \geq \delta_n \xrightarrow{n \rightarrow \infty} 0$, $\frac{\sum_{i=1}^n a_i \delta_i}{\sum_{i=1}^n a_i} \xrightarrow{n \rightarrow \infty} 0$. Furthermore, the convergence is uniform over the values of $\{a_n\}$.

Note that because the last epoch stretches to time N , the coefficients $a_m = 2^{m-1} N_m$ vary as N is increased. However, according to the lemma, it only matters that they remain monotonic and that the number of coefficients grows with N . \square

The next subsections relate $C_{\text{PMA}}^{[q,m]}$ to the rate obtained by the reference system for a certain error probability. The proof for the IFB and AFB cases is quite similar. For the purpose of clarity, the proof below focuses on the more complex IFB case, and at the end, the modifications required for the AFB case are discussed.

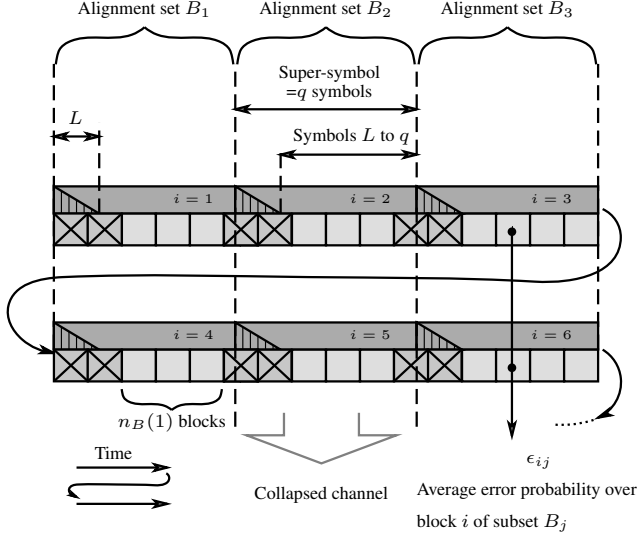


Fig. 5. The alignment of reference system blocks in the universal system's super-symbols. The large dark rectangles are the supersymbols of length q , with the triangles denoting the first L symbols. The light rectangles are the reference system blocks of length k where here $k = 3$. There are three alignment sets $B_j, j = 1, 2, 3$. In the example, $n_B(j) = 3, 3, 4$ for $j = 1, 2, 3$. The blocks that are not accounted for in $n_B(j)$ are marked with an 'x'. The error probability ϵ_{ij} refers to the same reference system block over different super-symbols in the alignment set. The collapsed channel is averaged across an alignment set.

D. IFB system performance during a single epoch

Consider the reference system composed of an encoder and a decoder operating over block size k , and the universal system in epoch m , with super-symbol length $q = 2^{m-1}$. For simplicity, as long as a single epoch is concerned, the symbols and super-symbols of the epoch are denoted by indices starting from 1 (i.e. $i = 1, 2, \dots, N_m \cdot q$ or $i = 1, 2, \dots, N_m$ respectively). In the following, the properties of the IFB system (such as rate and error probability) are linked to a channel averaged over super-symbols. First, let us consider the channel from the IFB system's point of view. $\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}$ denote the input and output vectors during the epoch, where the joint distribution depends on the joint behavior of the IFB encoder and the channel.

Consider the set of super-symbols with index $i \in B_j \triangleq \{i = l \cdot k + j : l \in \mathbb{Z}^+, i \leq N_m\}$ for $j = 1, \dots, k$, i.e. the set of super-symbols in hops of k (the reference block size) starting from the j -th super-symbol. B_j are not necessarily of the same size. In each of the super-symbols in a set B_j , the reference system's blocks begin at the same location with respect to the beginning of the super-symbol (see Fig.5). The sets B_j are termed "alignment sets".

To use the fading memory assumption, the reference system performance is considered only over symbols L through q out of the q symbols in the super-symbol. In each subset B_j , consider the blocks which completely overlap with symbols L through q . The number of such blocks per super-symbol in the set B_j is denoted $n_B(j)$. The number of symbols in epoch m which are not included in any of these blocks (for any B_j)

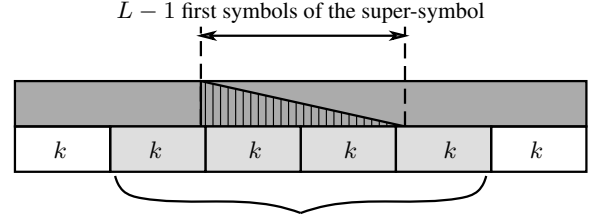


Fig. 6. At worst, $\frac{L-1}{k} + 2$ blocks may fully or partially overlap with the first $L - 1$ symbols of any super-symbol.

is denoted n_0 , where by the above definitions:

$$n_0 = N_m \cdot q - \sum_{j=1}^k |B_j| \cdot n_B(j) \cdot k, \quad (17)$$

i.e. n_0 equals the total number of symbols in the epoch, minus the number of symbols covered per super-symbol, summed over the subsets. n_0 can be bounded from above by considering that no more than $\frac{L-1}{k} + 2$ blocks may fully or partially overlap with the first $L - 1$ symbols of any super-symbol (see Fig.6), and therefore at most $L - 1 + 2k$ symbols per super-symbol are lost, hence

$$n_0 \leq (L - 1 + 2k) \cdot N_m. \quad (18)$$

This calculation accounts correctly for the special cases of the first and the last super-symbols in the epoch, as one may wrap the epoch around its tail, and imagine that the end of the epoch is cyclically connected to its beginning. It is convenient to normalize n_0 and the number of symbols in each set B_j by the total number of symbols in the epoch, and look at the relative sizes:

$$\lambda_j \triangleq \frac{|B_j| \cdot n_B(j) \cdot k}{N_m q}, \quad j = 1, \dots, k \quad (19)$$

$$\lambda_0 \triangleq \frac{n_0}{N_m q} \stackrel{(18)}{\leq} \frac{L - 1 + 2k}{q}, \quad (20)$$

where by (17):

$$\sum_{j=0}^k \lambda_j \stackrel{(17)}{=} 1. \quad (21)$$

Considering a specific alignment set B_j , because the reference system's operation is fixed during these blocks, its average error probability can be related to the mutual information of the averaged (collapsed) channel. Denote by $\tilde{\mathbf{X}}_i^{[q]}, \tilde{\mathbf{Y}}_i^{[q]}$ the channel input and output of the reference system during the i -th super-symbol, and by $(\tilde{\mathbf{Y}}_i^{[q]})_L^q$ the output during symbols L to q of the super-symbol. Let $(\tilde{\mathbf{X}}_{c,j}, \tilde{\mathbf{Y}}_{c,j})$ denote a random variable generated by a uniform selection over $i \in B_j$ of $(\tilde{\mathbf{X}}_i^{[q]}, (\tilde{\mathbf{Y}}_i^{[q]})_L^q)$, in other words,

$$(\tilde{\mathbf{X}}_{c,j}, \tilde{\mathbf{Y}}_{c,j}) = (\tilde{\mathbf{X}}_U^{[q]}, (\tilde{\mathbf{Y}}_U^{[q]})_L^q), \quad U \sim \mathbb{U}(B_j). \quad (22)$$

The joint distribution of $\tilde{\mathbf{X}}_{c,j}, \tilde{\mathbf{Y}}_{c,j}$ is:

$$\begin{aligned} \Pr(\tilde{\mathbf{X}}_{c,j} = \mathbf{x}, \tilde{\mathbf{Y}}_{c,j} = \mathbf{y}) \\ = \frac{1}{|B_j|} \sum_{i \in B_j} \Pr\left\{\tilde{\mathbf{X}}_i^{[q]} = \mathbf{x}, (\tilde{\mathbf{Y}}_i^{[q]})_L^q = \mathbf{y}\right\}. \end{aligned} \quad (23)$$

Because the reference system induces the same input distribution in all the super-symbols in B_j , i.e. $\Pr\{\tilde{\mathbf{X}}_i^{[q]} = \mathbf{x}\}$ is constant for all $i \in B_j$, the marginal distribution of $\tilde{\mathbf{X}}_{c,j}$ equals the per-block distribution (for all $i \in B_j$)

$$\Pr(\tilde{\mathbf{X}}_{c,j} = \mathbf{x}) = \frac{1}{|B_j|} \sum_{i \in B_j} \Pr\{\tilde{\mathbf{X}}_i^{[q]} = \mathbf{x}\} = \Pr\{\tilde{\mathbf{X}}_i^{[q]} = \mathbf{x}\}, \quad (24)$$

and hence the conditional distribution is:

$$\begin{aligned} \Pr(\tilde{\mathbf{Y}}_{c,j} = \mathbf{y} | \tilde{\mathbf{X}}_{c,j} = \mathbf{x}) &= \frac{\Pr(\tilde{\mathbf{X}}_{c,j} = \mathbf{x}, \tilde{\mathbf{Y}}_{c,j} = \mathbf{y})}{\Pr(\tilde{\mathbf{X}}_{c,j} = \mathbf{x})} \\ &= \frac{1}{|B_j|} \sum_{i \in B_j} \frac{\Pr\{\tilde{\mathbf{X}}_i^{[q]} = \mathbf{x}, (\tilde{\mathbf{Y}}_i^{[q]})_L^q = \mathbf{y}\}}{\Pr\{\tilde{\mathbf{X}}_i^{[q]} = \mathbf{x}\}} \\ &= \frac{1}{|B_j|} \sum_{i \in B_j} \Pr\{(\tilde{\mathbf{Y}}_i^{[q]})_L^q = \mathbf{y} | \tilde{\mathbf{X}}_i^{[q]} = \mathbf{x}\}. \end{aligned} \quad (25)$$

Denote the reference system rate by R_{IFB} . Consider employing the reference system over a super-symbol selected randomly and uniformly over B_j , where the message is encoded in the $n_B(j)$ blocks which are contained in symbols L through q of the super-symbol. Denote the average error probability which is attained for the i -th block (averaged over the channel and over all super-symbols in B_j) by ϵ_{ij} . The average error probability over the $n_B(j)$ blocks is denoted $\bar{\epsilon}_j \triangleq \frac{1}{n_B(j)} \sum_{i=1}^{n_B(j)} \epsilon_{ij}$. It is convenient to define the average error rate of the IFB system in the epoch m , $\bar{\epsilon}_{\text{IFB}}^{(m)}$ as the sum of error probabilities over all blocks that begin in the epoch, normalized by the approximate number of blocks in the epoch $\frac{N_m q}{k}$. This error probability can be bounded as:

$$\begin{aligned} \bar{\epsilon}_{\text{IFB}}^{(m)} &\geq \frac{\sum_{j=1}^k \sum_{i=1}^{n_B(j)} |B_j| \cdot \epsilon_{ij}}{\frac{N_m q}{k}} \\ &= \frac{k}{N_m q} \sum_{j=1}^k |B_j| n_B(j) \cdot \bar{\epsilon}_j \\ &= \sum_{j=1}^k \lambda_j \cdot \bar{\epsilon}_j, \end{aligned} \quad (26)$$

where the inequality is because the summation on the right side only accounts for the error probability over the blocks that are fully contained within symbols L to q of any super-symbol.

E. Operation of the IFB system over a modified channel

The random variables $(\tilde{\mathbf{X}}_{c,j}, \tilde{\mathbf{Y}}_{c,j})$ are induced by the channel and the behavior of the reference system. Not only is the distribution of $\tilde{\mathbf{X}}_{c,j}$ determined by the codebook distribution of the reference encoder, but the channel behavior determining $\tilde{\mathbf{Y}}_{c,j}$ is potentially affected by the input distribution induced by the reference encoder on all previous symbols. To account for the fading memory of the channel, and relate these variables to the ones seen by the universal system, let us consider alternative random variable, representing an alternative, specific, channel state at the beginning of the super-symbol. For a

given state sequence s_i , consider the random variable $\tilde{\mathbf{Y}}_{s,j}$ which depends on $\tilde{\mathbf{X}}_{c,j}$ through the following conditional distribution:

$$\begin{aligned} \Pr(\tilde{\mathbf{Y}}_{s,j} = \mathbf{y} | \tilde{\mathbf{X}}_{c,j} = \mathbf{x}) \\ = \frac{1}{|B_j|} \sum_{i \in B_j} \Pr\{(\tilde{\mathbf{Y}}_i^{[q]})_L^q = \mathbf{y} | \tilde{\mathbf{X}}_i^{[q]} = \mathbf{x}, (\tilde{\mathbf{X}}\tilde{\mathbf{Y}})^{(i-1)q} = s_i\} \end{aligned} \quad (27)$$

Because the current numbering refers only to epoch m , the notation $(\tilde{\mathbf{X}}\tilde{\mathbf{Y}})^{(i-1)q}$ formally refers to the input and output of the channel (with the reference system) during the epoch m . However the meaning of $(\tilde{\mathbf{X}}\tilde{\mathbf{Y}})^{(i-1)q}$ should be understood as the input and output of the channel from the beginning of time (potentially before epoch m). Also, considering that the reference encoder may not be able to emit all possible input sequences, the meaning of conditioning on $\tilde{\mathbf{X}}$ should be understood as if the encoder was disconnected and an input value was forced into the channel. Because the probability on the RHS of (27) is conditioned on the entire past of $\tilde{\mathbf{X}}$, and the channel is causal, this probability does not depend on the reference system, but only on the channel. Therefore, the same probability would be attained for the random variables \mathbf{X}, \mathbf{Y} representing the inputs and outputs of the channel when the universal system is applied:

$$\begin{aligned} \Pr(\tilde{\mathbf{Y}}_{s,j} = \mathbf{y} | \tilde{\mathbf{X}}_{c,j} = \mathbf{x}) \\ = \frac{1}{|B_j|} \sum_{i \in B_j} \Pr\{(\mathbf{Y}_i^{[q]})_L^q = \mathbf{y} | \mathbf{X}_i^{[q]} = \mathbf{x}, (\mathbf{X}\mathbf{Y})^{(i-1)q} = s_i\} \end{aligned} \quad (28)$$

Using the fading memory assumption, it is shown below, that the error probability obtained when applying the $n_B(j)$ blocks of the reference system to the channel defined by (27) is not significantly worse than its performance over the channel of (25).

Let

$$\begin{aligned} \Delta_P^{\mathbf{xy}}(i) &= \Pr\{(\tilde{\mathbf{Y}}_i^{[q]})_L^q = \mathbf{y} | \tilde{\mathbf{X}}_i^{[q]} = \mathbf{x}\} \\ &\quad - \Pr\{(\tilde{\mathbf{Y}}_i^{[q]})_L^q = \mathbf{y} | \tilde{\mathbf{X}}_i^{[q]} = \mathbf{x}, (\tilde{\mathbf{X}}\tilde{\mathbf{Y}})^{(i-1)q} = s_i\}. \end{aligned} \quad (29)$$

Because the channel is assumed to be causal and fading memory, using Proposition 1, for any $h > 0$ there exists L such that:

$$\|\Delta_P^{\mathbf{xy}}(i)\|_1 \leq 2h, \quad (30)$$

and therefore by the triangle inequality, the difference between the two channels is bounded by:

$$\begin{aligned} &\left\| \Pr(\tilde{\mathbf{Y}}_{s,j} = \mathbf{y} | \tilde{\mathbf{X}}_{c,j} = \mathbf{x}) - \Pr(\tilde{\mathbf{Y}}_{c,j} = \mathbf{y} | \tilde{\mathbf{X}}_{c,j} = \mathbf{x}) \right\|_1 \\ &\stackrel{(25),(27)}{=} \left\| \frac{1}{|B_j|} \sum_{i \in B_j} \Delta_P^{\mathbf{xy}}(i) \right\|_1 \\ &\leq \frac{1}{|B_j|} \sum_{i \in B_j} \|\Delta_P^{\mathbf{xy}}(i)\|_1 \leq 2h. \end{aligned} \quad (31)$$

From the \mathcal{L}_1 bound on the difference between the conditional probabilities, a bound on the increase in the error probability is easily derived as follows:

Lemma 3. *Let the error probability of a given encoder and decoder over the vector channel $W_i(\mathbf{y}|\mathbf{x})$ be ϵ_i for $i = 1, 2$. If for all \mathbf{X} , $\|W_1(\mathbf{Y}|\mathbf{X}) - W_2(\mathbf{Y}|\mathbf{X})\|_1 \leq h_W$ then $|\epsilon_1 - \epsilon_2| \leq h_W$*

Proof: Denote by E the event of error. Then,

$$\epsilon_i = \Pr(E|W_i) = \sum_{\mathbf{X}, \mathbf{Y}} \Pr(E|\mathbf{X}\mathbf{Y}) \cdot \Pr(\mathbf{X}) \cdot W_i(\mathbf{Y}|\mathbf{X}), \quad (32)$$

where the probability of error given $\mathbf{X}\mathbf{Y}$ does not depend on the channel (and for a deterministic encoder and decoder it is either 0 or 1 depending on whether \mathbf{Y} belongs to the decision region of \mathbf{X}). As a result,

$$\begin{aligned} |\epsilon_1 - \epsilon_2| &= \left| \sum_{\mathbf{X}} \Pr(\mathbf{X}) \sum_{\mathbf{Y}} \Pr(E|\mathbf{X}\mathbf{Y}) \cdot (W_1(\mathbf{Y}|\mathbf{X}) - W_2(\mathbf{Y}|\mathbf{X})) \right| \\ &\leq \sum_{\mathbf{X}} \Pr(\mathbf{X}) \sum_{\mathbf{Y}} \Pr(E|\mathbf{X}\mathbf{Y}) \cdot |W_1(\mathbf{Y}|\mathbf{X}) - W_2(\mathbf{Y}|\mathbf{X})| \\ &\leq \sum_{\mathbf{X}} \Pr(\mathbf{X}) \sum_{\mathbf{Y}} |W_1(\mathbf{Y}|\mathbf{X}) - W_2(\mathbf{Y}|\mathbf{X})| \\ &= \sum_{\mathbf{X}} \Pr(\mathbf{X}) \|W_1(\mathbf{Y}|\mathbf{X}) - W_2(\mathbf{Y}|\mathbf{X})\|_1 \\ &\leq \sum_{\mathbf{X}} \Pr(\mathbf{X}) \cdot h_W \\ &= h_W. \end{aligned} \quad (33)$$

□

Note that the lemma applies to any event whose probability is fixed as function of \mathbf{X}, \mathbf{Y} . In the following, it will be applied to the event of an error in each of the blocks separately (while the channel is the channel over the super-symbol defined in (27)).

F. The average capacity per alignment set

A lower bound on the capacity of the channel $\Pr\{\tilde{\mathbf{Y}}_{s,j}|\tilde{\mathbf{X}}_{c,j}\}$ is obtained by using the fact the IFB system delivers a certain rate with a small block error probability. Following is a variation of Fano's inequality, which takes into account that the errors are block errors rather than full message errors. Denote by E_i the indicator associated with the event of error in the i -th block out of $n_B(j)$ blocks, and $\epsilon'_{ij} = \mathbb{E}[E_i]$ the probability of error on this block (over all super-symbols in B_j), when the reference decoder is applied to the channel output $\tilde{\mathbf{Y}}_{s,j}$ (27). By applying Lemma 3 to the event of an error in the i -th block, $\epsilon'_{ij} \leq \epsilon_{ij} + 2h$ is obtained (where ϵ_{ij} is the error probability of the same block under the original channel (25)). The average error probability over the blocks is denoted $\bar{\epsilon}'_j \triangleq \frac{1}{n_B(j)} \sum_{i=1}^{n_B(j)} \epsilon'_{ij}$. Whenever $E_i = 0$, then given the channel output, $k \cdot R_{\text{IFB}}$ bits of the input become known, whereas when $E_i = 1$ these bits are unknown and have entropy at most $k \cdot R_{\text{IFB}}$. Denote by \mathbf{m} the transmitted

message (a sequence of $K_j = n_B(j) \cdot k \cdot R_{\text{IFB}}$ bits) and by $\hat{\mathbf{m}}$ the decoded message. The derivation below uses the fact conditioning reduces entropy, and the concavity of the binary entropy function $h_b(\cdot)$:

$$\begin{aligned} H(\mathbf{m}|\hat{\mathbf{m}}) &\leq H(\mathbf{m}, \{E_i\}|\hat{\mathbf{m}}) \\ &= H(\mathbf{m}|\{E_i\}, \hat{\mathbf{m}}) + H(\{E_i\}|\hat{\mathbf{m}}) \\ &\leq \sum_{\{e_i\}} H(\mathbf{m}|\forall i: E_i = e_i, \hat{\mathbf{m}}) \Pr(\{E_i = e_i\}) \\ &\quad + H(\{E_i\}) \\ &\leq \sum_{\{e_i\}} \sum_i e_i \cdot k \cdot R_{\text{IFB}} \Pr(\{E_i = e_i\}) + \sum_i H(E_i) \\ &= \sum_i E[E_i] \cdot k \cdot R_{\text{IFB}} + \sum_i h_b(\epsilon'_{ij}) \\ &= \sum_i \epsilon'_{ij} \cdot k \cdot R_{\text{IFB}} + \sum_i h_b(\epsilon'_{ij}) \\ &\leq n_B(j) \bar{\epsilon}'_j \cdot k \cdot R_{\text{IFB}} + n_B(j) h_b(\bar{\epsilon}'_j) \\ &= K_j \bar{\epsilon}'_j + n_B(j) h_b(\bar{\epsilon}'_j). \end{aligned} \quad (34)$$

Using the information processing inequality, the capacity of the channel is lower bounded as follows:

$$\begin{aligned} C\left(\Pr\left\{\tilde{\mathbf{Y}}_{s,j}|\tilde{\mathbf{X}}_{c,j}\right\}\right) &\geq I(\tilde{\mathbf{X}}_{c,j}; \tilde{\mathbf{Y}}_{s,j}) \\ &\geq I(\mathbf{m}; \hat{\mathbf{m}}) \\ &= H(\mathbf{m}) - H(\mathbf{m}|\hat{\mathbf{m}}) \\ &\geq K_j - K_j \bar{\epsilon}'_j - n_B(j) h_b(\bar{\epsilon}'_j). \end{aligned} \quad (35)$$

Define $h_b^\nearrow(p) \triangleq h_b(\min(p, \frac{1}{2})) \geq h_b(p)$ as the monotone continuation of $h_b(\cdot)$. $h_b^\nearrow(p)$ is non decreasing and concave. Then using $\bar{\epsilon}'_j \leq \bar{\epsilon}_j + 2h$:

$$\begin{aligned} C\left(\Pr\left\{\tilde{\mathbf{Y}}_{s,j}|\tilde{\mathbf{X}}_{c,j}\right\}\right) &\geq K_j - K_j \bar{\epsilon}'_j - n_B(j) h_b^\nearrow(\bar{\epsilon}'_j) \\ &\geq K_j(1 - \bar{\epsilon}_j - 2h) \\ &\quad - n_B(j) h_b^\nearrow(\bar{\epsilon}_j + 2h). \end{aligned} \quad (36)$$

G. A bound on the pessimistic averaged channel

To connect the capacity above to the pessimistic averaged channel, the bound on the capacity of each average channel over a set B_j needs to be linked with the capacity over the averaged channel over the sets. For this purpose the following simple lemma is used:

Lemma 4. *Let W_i be a set of channels, and $p_i \geq 0, \sum_i p_i = 1$ a probability distribution over the channels. Then*

$$\sum_i p_i C(W_i) - H(p) \leq C\left(\sum_i p_i W_i\right) \leq \sum_i p_i C(W_i). \quad (37)$$

The right inequality is based on convexity of the mutual information with respect to the channel and the left inequality is based on the fact the difference between knowing and not knowing the index i at the channel output is at most the entropy of this information. The simple proof is deferred to Appendix B.

The averaged channel over the epoch with a specific state sequence is:

$$\begin{aligned} \bar{W}^{[q]}(\mathbf{y}^q|\mathbf{x}^q;\mathbf{s}) &= \frac{1}{N_m} \sum_{i=1}^{N_m} \Pr \left\{ \mathbf{Y}_i^{[q]} = \mathbf{y} | \mathbf{X}_i^{[q]} = \mathbf{x}, (\mathbf{X}\mathbf{Y})^{(i-1)q} = s_i \right\} \end{aligned} \quad (38)$$

This channel's capacity is at least as large of the capacity of the next channel, where the first $L-1$ outputs are removed:

$$\begin{aligned} \bar{W}^{[q]\setminus L-1}(\mathbf{y}^{q-L+1}|\mathbf{x}^q;\mathbf{s}) &= \frac{1}{N_m} \sum_{i=1}^{N_m} \Pr \left\{ (\mathbf{Y}_i^{[q]})_L^q = \mathbf{y} | \mathbf{X}_i^{[q]} = \mathbf{x}, (\mathbf{X}\mathbf{Y})^{(i-1)q} = s_i \right\} \\ &= \frac{1}{N_m} \sum_{j=1}^k \sum_{i \in B_j} \Pr \left\{ (\mathbf{Y}_i^{[q]})_L^q = \mathbf{y} | \mathbf{X}_i^{[q]} = \mathbf{x}, (\mathbf{X}\mathbf{Y})^{(i-1)q} = s_i \right\} \\ &\stackrel{(28)}{=} \sum_{j=1}^k \frac{|B_j|}{N_m} \Pr \left\{ \tilde{\mathbf{Y}}_{s,j} = \mathbf{y} | \tilde{\mathbf{X}}_{c,j} = \mathbf{x} \right\}. \end{aligned} \quad (39)$$

The lemma implies that

$$\begin{aligned} C(\bar{W}^{[q]}(\cdot|\cdot;\mathbf{s})) &\geq C(\bar{W}^{[q]\setminus L-1}(\cdot|\cdot;\mathbf{s})) \\ &= C \left(\sum_{j=1}^k \frac{|B_j|}{N_m} \Pr \left\{ \tilde{\mathbf{Y}}_{s,j} = \mathbf{y} | \tilde{\mathbf{X}}_{c,j} = \mathbf{x} \right\} \right) \\ &\stackrel{\text{Lemma 4}}{\geq} \underbrace{\sum_{j=1}^k \frac{|B_j|}{N_m} C \left(\Pr \left\{ \tilde{\mathbf{Y}}_{s,j} = \mathbf{y} | \tilde{\mathbf{X}}_{c,j} = \mathbf{x} \right\} \right)}_{C_{avg}} \\ &\quad - H \left(\left\{ \frac{|B_j|}{N_m} \right\}_{j=1}^k \right). \end{aligned} \quad (40)$$

The last term is upper bounded by

$$H \left(\left\{ \frac{|B_j|}{N_m} \right\}_{j=1}^k \right) \leq \log k, \quad (41)$$

and the first term C_{avg} is bounded by (35) and substituting

$$K_j = n_B(j) \cdot k \cdot R_{\text{IFB}}:$$

$$\begin{aligned} C_{avg} &\stackrel{(35)}{\geq} \sum_{j=1}^k \frac{|B_j|}{N_m} \left(K_j(1 - \bar{\epsilon}_j - 2h) - n_B(j)h_b^{\nearrow}(\bar{\epsilon}_j + 2h) \right) \\ &= \sum_{j=1}^k \frac{|B_j|n_B(j)}{N_m} \left(kR_{\text{IFB}}(1 - \bar{\epsilon}_j - 2h) - h_b^{\nearrow}(\bar{\epsilon}_j + 2h) \right) \\ &\stackrel{(19)}{=} qR_{\text{IFB}} \cdot \sum_{j=1}^k \lambda_j(1 - \bar{\epsilon}_j - 2h) \\ &\quad - \frac{q(1 - \lambda_0)}{k} \sum_{j=1}^k \frac{\lambda_j}{1 - \lambda_0} h_b^{\nearrow}(\bar{\epsilon}_j + 2h) \\ &\stackrel{(21),(26)}{\geq} qR_{\text{IFB}} \cdot (1 - \lambda_0 - \bar{\epsilon}_{\text{IFB}}^{(m)} - 2h) \\ &\quad - \frac{q(1 - \lambda_0)}{k} h_b^{\nearrow} \left(\sum_{j=1}^k \frac{\lambda_j}{1 - \lambda_0} (\bar{\epsilon}_j + 2h) \right) \\ &\geq qR_{\text{IFB}} \cdot (1 - \lambda_0 - \bar{\epsilon}_{\text{IFB}}^{(m)} - 2h) \\ &\quad - \frac{q}{k} h_b^{\nearrow} \left(\frac{1}{(1 - \lambda_0)} \bar{\epsilon}_{\text{IFB}}^{(m)} + 2h \right). \end{aligned} \quad (42)$$

Combining (41),(42), dividing by q and taking infimum over \mathbf{s} yields:

$$\begin{aligned} \frac{1}{q} C_{\text{PMA}}^{[q,m]} &= \inf_{\mathbf{s}} C(\bar{W}^{[q]}(\cdot|\cdot;\mathbf{s})) \\ &\geq R_{\text{IFB}} - R_{\text{IFB}} \cdot (\lambda_0 + \bar{\epsilon}_{\text{IFB}}^{(m)} + 2h) \\ &\quad - \frac{1}{k} h_b^{\nearrow} \left(\frac{1}{(1 - \lambda_0)} \bar{\epsilon}_{\text{IFB}}^{(m)} + 2h \right) - \frac{\log k}{q}. \end{aligned} \quad (43)$$

For any L, k λ_0 can be made arbitrarily small by taking m large enough (equivalently q large enough). Taking m large enough such that $\frac{1}{(1 - \lambda_0)} \leq 2$. Thus, for m large enough:

$$\begin{aligned} \frac{1}{q} C_{\text{PMA}}^{[q,m]} &\geq R_{\text{IFB}} - R_{\text{IFB}} \cdot (\lambda_0 + \bar{\epsilon}_{\text{IFB}}^{(m)} + 2h) \\ &\quad - \frac{1}{k} h_b^{\nearrow} (2\bar{\epsilon}_{\text{IFB}}^{(m)} + 2h) - \frac{\log k}{q}. \end{aligned} \quad (44)$$

Alternatively, for all m :

$$\frac{1}{q} C_{\text{PMA}}^{[q,m]} \geq R_{\text{IFB}} - \Delta_1^{(k, R_{\text{IFB}})} (2\bar{\epsilon}_{\text{IFB}}^{(m)} + 2h) - \Delta_{2m}^{(k, R_{\text{IFB}})}, \quad (45)$$

where

$$\Delta_1^{(k, R_{\text{IFB}})}(t) = R_{\text{IFB}} \cdot t + \frac{1}{k} h_b^{\nearrow}(t), \quad (46)$$

and Δ_2 is defined as the remainder, i.e. R_{IFB} minus the RHS of (43) minus Δ_1 , and by (44), for large enough m ,

$$\Delta_{2m}^{(k, R_{\text{IFB}})}(L) \leq R_{\text{IFB}} \cdot \lambda_0 + \frac{\log k}{q} \leq R_{\text{IFB}} \cdot \frac{L - 1 + 2k}{q} + \frac{\log k}{q}. \quad (47)$$

$\Delta_1^{(k, R_{\text{IFB}})}(t)$ is concave in t , tends to zero with $t \rightarrow 0$ and decreases with k . $\Delta_{2m}^{(k, R_{\text{IFB}})}(L)$ tends to zero with m .

H. Conclusion of the proof for the IFB case

Now, multiple epochs are considered, and Proposition 2 is applied to bound the rate of the universal scheme. Suppose that over the N symbols (and M epochs) of the system's operation, the IFB system achieves rate R_{IFB} with an average error probability $\bar{\epsilon}_{\text{IFB}}$. The definition of $\bar{\epsilon}_{\text{IFB}}^{(m)}$ (above (26)) results in $\bar{\epsilon}_{\text{IFB}} = \frac{1}{N_{\text{blocks}}} \sum_{m=1}^M \frac{2^{m-1} N_m}{k} \bar{\epsilon}_{\text{IFB}}^{(m)}$ where N_{blocks} , the number of IFB blocks that begin in any symbol of the system's operation is upper bounded by $N_{\text{blocks}} \leq \frac{N}{k}$ and so

$$\bar{\epsilon}_{\text{IFB}} \geq \frac{1}{N} \sum_{m=1}^M 2^{m-1} N_m \bar{\epsilon}_{\text{IFB}}^{(m)}. \quad (48)$$

Choose $h = \bar{\epsilon}_{\text{IFB}}$ and determine the respective L to satisfy the fading memory property (note that this choice is for the purpose of analysis, and the scheme itself is not aware of these values). Applying Proposition 2 with $C_m = R_{\text{IFB}} - \Delta_1^{(k, R_{\text{IFB}})}(2\bar{\epsilon}_{\text{IFB}}^{(m)} + 2h)$ and $\delta_m = \Delta_{2m}^{(k, R_{\text{IFB}})}(L)$, there exists $\tilde{\delta}_N \xrightarrow[N \rightarrow \infty]{} 0$ such that

$$\begin{aligned} R_{\text{UNI}}[N] &\geq \bar{C} - \delta_N \\ &= R_{\text{IFB}} - \frac{1}{N} \sum_{m=1}^M 2^{m-1} N_m \Delta_1^{(k, R_{\text{IFB}})}(2\bar{\epsilon}_{\text{IFB}}^{(m)} + 2h) \\ &\quad - \tilde{\delta}_N \\ &\geq R_{\text{IFB}} - \Delta_1^{(k, R_{\text{IFB}})} \left(\frac{1}{N} \sum_{m=1}^M 2^{m-1} N_m (2\bar{\epsilon}_{\text{IFB}}^{(m)} + 2h) \right) \\ &\quad - \tilde{\delta}_N \\ &= R_{\text{IFB}} - \Delta_1^{(k, R_{\text{IFB}})}(2\bar{\epsilon}_{\text{IFB}} + 2h) - \tilde{\delta}_N \\ &= R_{\text{IFB}} - \Delta_1^{(k, R_{\text{IFB}})}(4\bar{\epsilon}_{\text{IFB}}) - \tilde{\delta}_N, \end{aligned} \quad (49)$$

where the inequality is due to the concavity of $\Delta_1^{(k, R_{\text{IFB}})}(t)$. For an arbitrarily small δ_C , choose $R_{\text{IFB}} = C_{\text{IFB}} - \delta_C$. By the definition of the IFB capacity, there is a k large enough and N large enough so that $\bar{\epsilon}_{\text{IFB}}$ can be made arbitrarily small. Therefore $\Delta_1^{(k, R_{\text{IFB}})}(4\bar{\epsilon}_{\text{IFB}})$ can be made arbitrarily small (note that it decreases with k) while $\tilde{\delta}_N \xrightarrow[N \rightarrow \infty]{} 0$. Therefore for large enough N , the RHS of (49) can be made arbitrarily close to C_{IFB} . This proves the IFB universality of the proposed universal system.

I. Modifications for the AFB case

The proof for the AFB case is similar and the required modifications are discussed below. The same definitions of Section IV-D are used for the alignment sets (up to Equation (21)), except L is set to $L = 1$. The definition of $(\tilde{\mathbf{X}}_{c,j}, \tilde{\mathbf{Y}}_{c,j})$ is not needed, as the performance of the AFB system is directly related to the constrained-state channel whose output is $\tilde{\mathbf{Y}}_{s,j}$. For the arbitrary sequence of states s_i , define the channel $\Pr(\tilde{\mathbf{Y}}_{s,j} = \mathbf{y} | \tilde{\mathbf{X}}_{c,j} = \mathbf{x})$ according to (28). This channel implies that the channel history is forced to s_i at the beginning of each super-symbol. In each alignment set B_j the $n_B(j)$ blocks of the AFB system are mapped to this averaged channel. Clearly, the error probability of the AFB

system when the state is forced to some value at the beginning of the super-symbol, is not worse than the error probability in arbitrary mapping defined in Definition 5, where the state is forced to its worst-case value just before the relevant block. Formally, let E_l denote an indicator of the event of error in the l -th block of the i -th supersymbol, a block which begins at symbol n_l of the supersymbol, then when mapping to the channel where only the initial state is forced, the error probability is:

$$\begin{aligned} \mathbb{E} [E_l | (\mathbf{X}\mathbf{Y})^{(i-1)q} = s_i] \\ = \mathbb{E} \left[\mathbb{E} \left[E_l | (\mathbf{X}\mathbf{Y})^{(i-1)q} = s_i, \right] | (\mathbf{X}\mathbf{Y})^{(i-1)q} = s_i \right] \end{aligned} \quad (50)$$

where the iterated expectation law is applied. The internal expectation is by definition upper bounded by $P_e(l)$, the error probability in arbitrary mapping (Definition 5) over the same block, and therefore the error probability with the current mapping is upper bounded by the error probability in arbitrary mapping.

Denote as before by ϵ_{ij} the average error probability over the i -th blocks in the B_j alignment set, when the AFB system is mapped to the channel $\Pr(\tilde{\mathbf{Y}}_{s,j} = \mathbf{y} | \tilde{\mathbf{X}}_{c,j} = \mathbf{x})$, and the average error probability over the $n_B(j)$ blocks by $\bar{\epsilon}_j \triangleq \frac{1}{n_B(j)} \sum_{i=1}^{n_B(j)} \epsilon_{ij}$, (26) now holds with respect to the average error in arbitrary mapping over the epoch, where now the inequality stems not only from the fact that not all errors are accounted for, but in addition because the error probabilities ϵ_{ij} , $\bar{\epsilon}_j$ are upper bounded by the respective errors obtained by arbitrary mapping.

The transition to a modified channel (Section IV-E) is not required in this case and the proof is continued with the value $h = 0$. The rest of the proof proceeds as before (Sections IV-F, IV-G, IV-H), where ϵ_{AFB} and R_{AFB} replace ϵ_{IFB} and R_{IFB} , except that in (49) h is chosen to be zero rather than equal $\bar{\epsilon}_{\text{IFB}}$. This concludes the proof of Theorem 1. \square

V. AN EXAMPLE OF A FADING MEMORY CHANNEL

In the definition of a fading memory channel (Definition 2), the overall probability of \mathbf{Y} over the infinite future (from n to ∞) is required to be close in \mathcal{L}_1 sense to a distribution that does not depend on the past. This raises the question how strict is the requirement and whether it is satisfied by broad family of channels. Below, an example is given of a family of finite state channels with non-homogeneous transition probabilities that, under the assumption that there is a non-zero probability to arrive from any state to any state, satisfies the fading memory requirement.

Consider a finite state channel where the state at each moment in time S_i belongs to the finite set \mathcal{S} . The probability of each output letter is given as a time-varying function of the input letter and the current state $\Pr(Y_i | X_i; S_i) = W_i(Y_i | X_i; S_i)$, and the state sequence is a non-homogeneous Markov chain which depends on the input via $\Pr(S_i | S_{i-1}, X_{i-1}) = T_i(S_i | S_{i-1}; X_{i-1})$. The joint probability is therefore

$$\Pr\{\mathbf{Y}_1^n \mathbf{S}_1^n | \mathbf{X}_1^n, S_0\} = \prod_{i=1}^n T_i(S_i | S_{i-1}; X_{i-1}) W_i(Y_i | X_i; S_i). \quad (51)$$

If the Markov chain determining the state transitions is such that, eventually, it is possible to move from any state to any state, then it is termed a indecomposable Markov chain. Similarly, for constant state transition and channel probabilities T_i, W_i , Gallager [10, 4.6] defined the resulting finite state channel as indecomposable if the memory of the initial state fades with time (Eq. (4.6.26) there). Here, for simplicity, a stricter condition is assumed: that it is possible to move from any state to any state within one step and with a certain, non-vanishing probability $\beta > 0$, i.e. that

$$\forall S_{i-1}; X_{i-1} : T_i(S_i|S_{i-1}; X_{i-1}) \geq \beta. \quad (52)$$

It appears that this condition can be relaxed and the results can be generalized to indecomposable Markov chains, under the assumption that there is some minimum probability to arrive from any state to any state with a finite number of steps, by simply treating a block of symbols as a new super-symbol. However for simplicity let us focus on this type of channels, which is also quite general.

Proposition 3. Any channel with the structure defined above is a causal fading memory channel. Specifically, the \mathcal{L}_1 distance in Definition 2 is $h \leq 2(1 - |\mathcal{S}| \cdot \beta)^{L+1}$, i.e. fades exponentially with L .

The rest of this section is devoted to the proof of this proposition. The transition probability may be written alternatively as follows:

$$T_i(S_i|S_{i-1}; X_{i-1}) = \lambda \cdot \frac{1}{|\mathcal{S}|} + (1 - \lambda)T_i^{(\text{rem})}(S_i|S_{i-1}; X_{i-1}), \quad (53)$$

where $\lambda = |\mathcal{S}| \cdot \beta$. Due to the condition (52), the remainder $T_i^{(\text{rem})}$ is non negative, and by summing both sides of (53) over S_i it is easily seen that $T_i^{(\text{rem})}$ is a legitimate probability distribution. This motivates the following formulation: consider a sequence of i.i.d. Bernully random variables $A_i \sim \text{Ber}(\lambda)$, which are drawn independently of \mathbf{X}_1^∞ and of previous S_i -s. The next state is determined as follows. If $A_i = 0$ then the next state is determined by $T_i^{(\text{rem})}(S_i|S_{i-1}; X_{i-1})$. Otherwise, it is selected uniformly with equal probabilities. This results in the same conditional probability $T_i(S_i|S_{i-1}; X_{i-1})$ due to (53). The fading memory property stems from the observation that whenever $A_i = 1$, the memory of the past disappears, and that over a long enough interval, the probability for such an event approaches one.

Due to the independence of A_i in the sequence \mathbf{X} and the previous states, it is also independent of the past of \mathbf{Y} . Hence

$$\begin{aligned} & \Pr(\mathbf{Y}_n^m | \mathbf{X}_1^\infty, \mathbf{Y}_1^{n-L-1}) \\ &= \sum_{\mathbf{A}_{n-L}^n} \Pr(\mathbf{Y}_n^m \mathbf{A}_{n-L}^n | \mathbf{X}_{n-L}^\infty (\mathbf{X}\mathbf{Y})_1^{n-L-1}) \\ &= \sum_{\mathbf{A}_{n-L}^n} \Pr(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^\infty (\mathbf{X}\mathbf{Y})_1^{n-L-1} \mathbf{A}_{n-L}^n) \cdot \Pr(\mathbf{A}_{n-L}^n). \end{aligned} \quad (54)$$

The distribution conditioned on \mathbf{A}_{n-L}^n is:

$$\begin{aligned} & \Pr(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^\infty (\mathbf{X}\mathbf{Y})_1^{n-L-1} \mathbf{A}_{n-L}^n) \\ &= \sum_{S_{n-L+1}, S_n} \Pr(\mathbf{Y}_n^m | S_{n-L-1} S_n \mathbf{X}_{n-L}^\infty (\mathbf{X}\mathbf{Y})_1^{n-L-1} \mathbf{A}_{n-L}^n) \\ & \quad \cdot \Pr(S_{n-L-1} | \mathbf{X}_{n-L}^\infty (\mathbf{X}\mathbf{Y})_1^{n-L-1} \mathbf{A}_{n-L}^n) \\ & \quad \cdot \Pr(S_n | \mathbf{X}_{n-L}^\infty (\mathbf{X}\mathbf{Y})_1^{n-L-1} \mathbf{A}_{n-L}^n S_{n-L-1}) \\ &= \sum_{S_{n-L+1}, S_n} \Pr(\mathbf{Y}_n^m | \mathbf{X}_n^\infty, S_n) \\ & \quad \cdot \Pr(S_{n-L-1} | (\mathbf{X}\mathbf{Y})_1^{n-L-1}) \\ & \quad \cdot \Pr(S_n | \mathbf{X}_{n-L}^n S_{n-L-1} \mathbf{A}_{n-L}^n). \end{aligned} \quad (55)$$

Focusing on the last term, it can be shown that it has a weak dependence on S_{n-L-1} . Given $\{A_i\}$, the sequence S_i remains a Markov chain, therefore for any $n - L \leq m < n$

$$\begin{aligned} & \Pr(S_n | S_{n-L-1} \mathbf{X}_{n-L}^n, \mathbf{A}_{n-L}^n) \\ &= \sum_{S_m} \Pr(S_m | S_{n-L-1} \mathbf{X}_{n-L}^n, \mathbf{A}_{n-L}^n) \\ & \quad \cdot \Pr(S_n | S_m \mathbf{X}_{n-L}^n, \mathbf{A}_{n-L}^n). \end{aligned} \quad (56)$$

If $A_m = 1$ then the first term is constant and independent of S_{n-L} and therefore $\Pr(S_n | S_{n-L-1} \mathbf{X}_{n-L}^n, \mathbf{A}_{n-L}^n)$ does not depend on S_{n-L-1} . The same is trivially true for $m = n$. The probability that none of \mathbf{A}_{n-L}^n would be 1 is $(1 - \lambda)^{L+1}$. Whenever any of \mathbf{A}_{n-L}^n is 1, because the last term in (55) is independent of S_{n-L-1} , the sum in (55) breaks into two independent sums and $\Pr(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^\infty (\mathbf{X}\mathbf{Y})_1^{n-L-1} \mathbf{A}_{n-L}^n)$ does not depend on $(\mathbf{X}\mathbf{Y})_1^{n-L-1}$. Therefore, considering the summation in (54), it can be written as:

$$\begin{aligned} & \Pr(\mathbf{Y}_n^m | \mathbf{X}_1^\infty, \mathbf{Y}_1^{n-L-1}) \\ &= (1 - (1 - \lambda)^{L+1}) \cdot P_1(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^\infty) \\ & \quad + (1 - \lambda)^{L+1} \cdot P_2(\mathbf{Y}_n^m | \mathbf{X}_1^\infty, \mathbf{Y}_1^{n-L-1}), \end{aligned} \quad (57)$$

where the probabilities P_1, P_2 are generated by splitting the sum (54) to the single component that depends on $\mathbf{X}_1^{n-L-1}, \mathbf{Y}_1^{n-L-1}$ and the other components that do not, and normalizing each part. From (57) the \mathcal{L}_1 distance can be bounded:

$$\begin{aligned} h &= \|\Pr(\mathbf{Y}_n^m | \mathbf{X}_1^\infty, \mathbf{Y}_1^{n-L-1}) - P_1(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^\infty)\|_1 \\ &= (1 - \lambda)^{L+1} \|P_1(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^\infty) - P_2(\mathbf{Y}_n^m | \mathbf{X}_1^\infty, \mathbf{Y}_1^{n-L-1})\|_1 \\ &\leq (1 - \lambda)^{L+1} \left(\|P_1(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^\infty)\|_1 \right. \\ & \quad \left. + \|P_2(\mathbf{Y}_n^m | \mathbf{X}_1^\infty, \mathbf{Y}_1^{n-L-1})\|_1 \right) \\ &= 2(1 - \lambda)^{L+1}. \end{aligned} \quad (58)$$

□

VI. DISCUSSION

A. Comparison with exiting results

Table I compares the current results with previous and new results of us and other authors.

Channel model	Achieved rate based on zero order statistics	Achieved rate based on competitive universality
Modulo additive with an individual noise sequence \mathbf{z}	$\log \mathcal{X} - \tilde{H}(\mathbf{z})$ (Shayevitz & Feder [2])	$R = (1 - \rho(\mathbf{z})) \log \mathcal{X} \geq C_{\text{IFB}}$ [4] $R = C_{\text{FS}}$ (Misra & Weissman [13])
Arbitrarily varying sequence of memoryless channels	$I(Q, \bar{W})$ (Eswaran <i>et al</i> [3], ignoring differences in formulation) $C(\bar{W})$ [5]	$\geq C_{\text{IFB}} = C_{\text{AFB}}$ (Current paper)
General vector channels	$C(\bar{W}_{\text{SUBJ}}) \geq C_{\text{PMA}}$, (5), (11) (Current paper)	$\geq C_{\text{IFB}}(\text{fading memory}), \geq C_{\text{AFB}}$ (Current paper)

TABLE I
SUMMARY OF NEW AND EXISTING RESULTS

B. Asymptotics

Although the current result is pleasing in terms of the asymptotical rate, it is theoretical in at least two senses related to asymptotical convergence rate. First, as the “finite state compressibility”, the definition of the IFB capacity relies on the order of limits – i.e. one first examines the performance of a finite-block code on the *infinite* channel and only then lets the block length go to infinity. The second sense is that the scheme proposed here only attempts to attain the asymptotical result, and does not endeavor to be efficient in terms of convergence rate. The best convergence rate, and more efficient schemes are left for further study.

There are several reasons for the scheme’s inefficiency. One is the use of a single super-symbol length. Due to alignment issues with the reference system’s blocks, the super-symbol length q is required to exceed the block length k significantly. It seems better to enhance the methods of [5] for learning communication priors over several possible k -s simultaneously. Another cause for inefficiency is the fact each epoch stands on its own and the information learned from the past is reset. Furthermore, in the asymptotical case one can always assume that q eventually becomes larger than L , the channel’s effective memory length. However, in a more efficient scheme it may be desired, instead of wasting L symbols of each super-symbol, to attempt learning and adapting to a conditional distribution which includes also the past (e.g. estimate the average over i of $\Pr(\mathbf{Y}_i^{[q]} | \mathbf{X}_i^{[q]}, \mathbf{X}_{(i-1)q-L}^{(i-1)q})$ and set the prior accordingly). The rate of convergence of the prior prediction scheme of [5] used as basis for the current universal scheme may be improved as well.

The channel assumed in this paper is very general, and the penalty for this generality is not captured in the asymptotical rates. However it surely induces a penalty in the rate of convergence. Probably, the ability to efficiently learn and utilize channel behavior would come from identifying similarities and repetitive behavior of channel occurrence, rather than slow increase of the super-symbol size as done here.

On the other hand, it seems inevitable that the overheads related to learning the decoding rule and the prior would grow at a rate which is at least linear in the super-alphabet size, i.e. exponential in the super-symbol length. Furthermore, it was already shown in [4] that even for the modulo-additive channel, to achieve a small redundancy, the transmission length of IFB-universal systems must grows exponentially with

the reference block size k . A rough analysis of the maximum convergence rate for general channels is given in Appendix D, and suggest that the transmission length must grow at least like $O(|\mathcal{X}|^k \cdot |\mathcal{Y}|^k)$ with k .

The difficulty of finding an input distribution to attain the IFB capacity may be exemplified by the following channel. Starting with an arbitrary IFB encoder of M codewords over block length k , and a decoder with an arbitrary decision region for each of the M messages, the channel is constructed to favor this IFB system. For each block of k symbols, if the input $\mathbf{X}_i^{[k]}$ is one of the codewords, then the output is randomly chosen inside the respective IFB decoder decision region, and otherwise, the output is random and independent of the input. In order to achieve the IFB capacity ($\frac{\log M}{k}$) over this channel, the universal system is required to “guess” most of the codewords in the reference encoder’s codebook. This channel is a fading memory channel but it is not causal, however this is easily fixed with a more elaborate structure presented in Appendix D.

The fact that the transmission length N required to obtain a small IFB redundancy, scales exponentially with k , combined with the fact that reasonable reference coding systems would have block sizes of at least 100-1000 symbols, raises the question: can such universal schemes ever become practical?

It is natural to compare the universal communication problem with the case of universal compression using the LZ algorithm, especially in view of the theoretical and practical success of this algorithm. The result [9] showing that LZ asymptotically beats every finite state machine, supplies motivation for the algorithm from an engineering perspective, since all digital computation machines are eventually finite state machines. However, as in the current case, this is only theoretical. Considering that a state machine with a state memory of k bits can simply memorize an individual sequence of 2^k bits, then the length of the sequence is required to be larger than this value in order to surpass the performance of a k -bit state machine.¹ In fact, Lempel and Ziv’s bound [9, Eq.(14)] would require the length of the sequence n to

¹To comply with the definitions of [9], the encoder may be designed knowing the individual sequence, but is required to encode any possible sequence. The encoder may keep a counter of the letter index and check for a deviation from this known sequence. If the input does not deviate from the known sequence, it is encoded to 1 bit, and if it does, the remainder of the sequence can be encoded in any uniquely decodable way (e.g. quoting the place of deviation and the remainder of the sequence).

scale faster than the squared number of states (2^{2k}) in order for the redundancy $\delta_s(n)$ [9, Eq.(10)] to vanish. In spite of this impractical asymptotical result, the LZ algorithm and newer algorithms that improve over it, work well. The reason is probably related to the fact the sequences encountered in practice are relatively simple and can be modeled by small state machines.

To summarize, in the case of LZ universal source coding there is a combination of an elegant scheme, a competitive universality result which is rather theoretical (if competent competitors are considered), and good performance for simple models and for practical scenarios. In the communication setting presented here, only the second property, i.e. a theoretical competitive universality result, was shown. Complementary results that present faster convergence rates under simpler models or reference systems are required, in order to show such schemes can have gains that are realizable in practice (such is the result of [5], for example).

A possible direction for improving asymptotical convergence rate is modifying the comparison class or the channel model. As an example, comparing the results of [4] and [5] regarding convergence rates, it is observed that the overheads related to learning the prior are larger than overheads of universal decoding, for the same block lengths. As the current bounds are not tight, this only a conjecture. In view of this, one may consider as reference, encoders and decoder which operate over a block of a certain size, however their codebook distributions are close to i.i.d. (e.g. in the sense of [14]) or have constrained structures, as practical codes do.

Another aspect related to convergence rate is the amount of time and data which are reasonable for training. One should take into account that the alternative process, of manually studying the channel model, coming up with simplified mathematical models, and designing systems optimized for these models, is also time consuming. Therefore, it is not unreasonable to allow a significant amount of time for training.

C. Time variations

One issue with the current definitions is that in competing against *static* coding systems the universal system does not take advantage of time variations in the channel, at least not explicitly. This is not only a matter of obtaining better rates: as an example, even a small frequency offset between the oscillators of the transmitter and the receiver may turn IFB capacity into zero, as a static decoder is not able to track and correct it. On the other hand, if the tracking mechanism is considered as external to the encoder/decoder, this raises the question how to perform these tasks over an unknown channel. This means that models have to be improved before these systems become practical.

This issue relates to the subject of convergence rate, because adaptation of the universal system over time is only possible if learning time is quick enough. It is possible to consider an extension of the current results by allowing adaptation (e.g. re-learning) of the model over time, where the simpler models have a faster refresh rate and the complex models have a slower one, thus balancing between overhead and the refresh rate.

D. Fading memory in the wide sense

In the definition of fading memory (Definition 2) there is a conditioning on $(\mathbf{XY})^{n-L+1}$ which is required to have a small effect. Similarly, the definition of AFB error probability (Definition 5) includes a conditioning on the past of both \mathbf{X}, \mathbf{Y} . It appears, at least intuitively, that the conditioning on \mathbf{Y} in both cases is redundant, and may be done without. After all, what the universal system does not know and the reference system does, is the effect of possible *inputs*. Therefore the definition of fading memory as

$$\Pr(\mathbf{Y}_n^\infty | \mathbf{X}_1^\infty) \stackrel{\mathcal{L}_1}{\approx} \Pr(\mathbf{Y}_n^\infty | \mathbf{X}_{n-L}^\infty), \quad (59)$$

instead of the current definition:

$$\Pr(\mathbf{Y}_n^\infty | \mathbf{X}_1^\infty, \mathbf{Y}_1^{n-L-1}) \stackrel{\mathcal{L}_1}{\approx} \Pr(\mathbf{Y}_n^\infty | \mathbf{X}_{n-L}^\infty), \quad (60)$$

seems more plausible. The first definition can be thought of as fading memory in the wide sense, or input only, while the current definition is narrower. To give an example, consider the channel where a coin is tossed at the beginning of time (irrespective of any input) and chooses between two channels memoryless in the input, which will last to eternity. This channel is fading memory according to (59) but not according to (60) and Definition 2. It is easy to see that although this channel is ruled out by the current fading-memory requirement, it does not pose a problem for competitive universality. Because the IFB system is required to deliver a given rate at a vanishing error probability, it will eventually tune to the worst channel. Therefore, the universal system should not have a problem to exceed the IFB system's performance. Note that in spite of the fact the channel is given as a single conditional probability, it is beneficial to treat it as an arbitrary choice between the two channels (seemingly a worst channel, as an arbitrary choice is worse than a probabilistic one), and see that the IFB system would attain either the IFB capacity of the good channel or the IFB capacity of the bad channel, according to whichever was drawn.

This conditioning on \mathbf{Y}_1^{n-L-1} appears also in the definition of the AFB capacity (through the definition of error probability in arbitrary mapping). It seems unfair that the AFB system is "punished" by considering the worst channel state, or history $(\mathbf{XY})^{n-L+1}$ (where \mathbf{Y}^{n-L+1} is controlled by the channel), and instead it would have been sufficient and more plausible to consider the worst case input \mathbf{X}^{n-L+1} .

Technically speaking, the conditioning on \mathbf{Y}_1^{n-L-1} stemmed from the analysis of the rate of the universal scheme in [5, Lemma 9], and is required in order to generate the martingale property which is used in the convergence analysis. Once the condition appears in \bar{W}_{SUBJ} it is required everywhere. It appears that removing this conditioning would require taking several steps back compared to the techniques developed here and in [5]. An example is that the "collapsed channel capacity" is no longer a useful bound: considering the example channel above, the collapsed channel is the average (across the "coin toss") of the per-block averaged channels, whereas in order to show universality one needs to bound the reference system by the capacity of the worst channel (over the "coin toss"). For example, if the time-averaged channels

over blocks of size k are $\overline{W}_{\text{good}}$ and $\overline{W}_{\text{bad}}$, and the coin is fair, then the collapsed channel capacity is $C(\frac{1}{2}\overline{W}_{\text{good}} + \frac{1}{2}\overline{W}_{\text{bad}})$, while the rate that can be guaranteed by the universal system is related to $C(\overline{W}_{\text{bad}})$. To solve this problem, the information density should be considered instead of the mutual information (its average), and the probability of the information density to fall below the rate of the IFB system should be used as a tighter bound for error probability [8, Thm.4,5]. This may require the universal system to base its decisions on the information density.

E. An alternative comparison class

The IFB/AFB comparison class is limited by having a relatively short block size, which implies the distance from capacity (e.g. for simple models such as DMC's) may be large. This is not utilized in the current bounds, as the IFB rate was only bounded by the collapsed channel capacity. However, the specific maximum IFB rate with a certain block size may be much smaller. The collapsed channel capacity bound would still hold, if the encoder and decoder were allowed to operate over multiple blocks, but treat each block in the same way.

One option to define an alternative class is to limit the encoder to be a random encoder over the entire transmission length n , with an i.i.d. prior of choice (alternatively, i.i.d. in blocks) and limit the decoder to use a memoryless decoding metric (or more generally, alpha decoding, i.e. type-based decoding, or more elaborate, e.g. finite state metrics). Another similar way is to let the encoder and decoder be general (over the entire n length transmission) but randomly permute the inputs and outputs of the channel. As before, the reference encoder and decoder are limited, but are designed based on full channel knowledge.

This comparison class is more contrived on one hand (includes many arbitrary details in its definition – the use of some randomization or permutation in the coding, constraint on the metric, etc), whereas the IFB class is more natural, but suffers from inefficiency. On the other hand it should be possible to compete with both classes simultaneously.

For the class of channels memoryless in the input (discussed in [5]), it should not be too hard to show that the rate that can be obtained by these reference classes cannot exceed the average channel capacity, which is obtained by the universal system of [5]. For the class of fading-memory channels, the system presented here can be applied to these classes as well: again using the claim that for each super-symbol, for most of the super-symbol duration, the channel (conditioned on the state at the beginning of the super-symbol) is similar to the channel seen by the reference system, and this way obtain a rate which is approximately the capacity of the averaged channel in blocks, as seen by the reference system, and this is more than the single-letter collapsed channel capacity which limits the rate of the reference system.

However note that also for these alternative classes, the infinite channel memory, or “password” issue is not resolved, and therefore universal communication is not possible over completely general channels where the memory is not restricted. This is shown by an example in [4].

VII. CONCLUSION

Communication over an unknown causal vector channel was considered, where the channel may include memory, and may change its behavior in an arbitrary way over time. It was demonstrated, that there exists a universal system with feedback, which without knowing the channel, asymptotically attains rates meeting or exceeding the rates of any finite block encoding system operating on the same channel, where the latter system may be designed with prior knowledge of the channel. The result holds for a finite block system mapped iteratively to sequential blocks, under a condition of fading-memory in the channel, and alternatively for any channel, but where the competing finite block system is required to start-off anywhere from an arbitrary channel state.

Compared to other models of unknown channels where there is an explicit model, here the assumptions on the channel are minimized. This general channel model includes as special cases many models previously considered.

This result marks the theoretical possibility of having a system which is not designed based on a channel model, made up by engineers, but rather learns the actual channel and automatically adapts to it. There are many theoretical and practical issues to resolve before such systems would be practical. However, similarly to the world of source coding, there is hope that universal systems would be implemented one day, and perhaps improve over systems optimized under specific channel model assumptions.

APPENDIX

A. Proof of Proposition 1

Property 1: let $M > m$ and assume (2) holds for M then:

$$\begin{aligned} & \sum_{\mathbf{Y}_n^m} \left| \Pr(\mathbf{Y}_n^m | \mathbf{X}_1^\infty, \mathbf{Y}_1^{n-L-1}) - P_n(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^\infty) \right| \\ &= \sum_{\mathbf{Y}_n^m} \left| \sum_{\mathbf{Y}_{m+1}^M} (\Pr(\mathbf{Y}_n^M | \mathbf{X}_1^\infty, \mathbf{Y}_1^{n-L-1}) - P_n(\mathbf{Y}_n^M | \mathbf{X}_{n-L}^\infty)) \right| \\ &\stackrel{(a)}{\leq} \sum_{\mathbf{Y}_n^m} \sum_{\mathbf{Y}_{m+1}^M} \left| \Pr(\mathbf{Y}_n^M | \mathbf{X}_1^\infty, \mathbf{Y}_1^{n-L-1}) - P_n(\mathbf{Y}_n^M | \mathbf{X}_{n-L}^\infty) \right| \\ &= \|\Pr(\mathbf{Y}_n^M | \mathbf{X}_1^\infty, \mathbf{Y}_1^{n-L-1}) - P_n(\mathbf{Y}_n^M | \mathbf{X}_{n-L}^\infty)\|_1 \leq h, \end{aligned} \quad (61)$$

where the triangle inequality (a) was used.

Property 2:

$$\begin{aligned} \Pr(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^m) &= \sum_{\mathbf{z}} \Pr(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^m, (\mathbf{X}\mathbf{Y})^{n-L-1} = \mathbf{z}) \\ &\quad \cdot \Pr((\mathbf{X}\mathbf{Y})^{n-L-1} = \mathbf{z} | \mathbf{X}_{n-L}^m). \end{aligned} \quad (62)$$

Defining for brevity $P_Z(\mathbf{z}) = \Pr((\mathbf{X}\mathbf{Y})^{n-L-1} = \mathbf{z} | \mathbf{X}_{n-L}^m)$, and using the triangle inequality $\|a(y) + b(y)\|_1 \leq \|a(y)\|_1 +$

$\|b(y)\|_1$ and causality, yields:

$$\begin{aligned}
& \|\Pr(\mathbf{Y}_n^m | \mathbf{X}_1^m, \mathbf{Y}_1^{n-L-1}) - \Pr(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^m)\|_1 \\
& \leq \|\Pr(\mathbf{Y}_n^m | \mathbf{X}_1^m, \mathbf{Y}_1^{n-L-1}) - P_n(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^m)\|_1 \\
& \quad + \|\Pr(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^m) - P_n(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^m)\|_1 \\
& \leq h + \left\| \sum_{\mathbf{z}} \left[\Pr(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^m (\mathbf{X}\mathbf{Y})^{n-L-1} = \mathbf{z}) \right. \right. \\
& \quad \left. \left. - P_n(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^m) \right] P_Z(\mathbf{z}) \right\|_1 \\
& \leq h + \sum_{\mathbf{z}} \left\| \Pr(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^m (\mathbf{X}\mathbf{Y})^{n-L-1} = \mathbf{z}) \right. \\
& \quad \left. - P_n(\mathbf{Y}_n^m | \mathbf{X}_{n-L}^m) \right\|_1 \cdot P_Z(\mathbf{z}) \\
& \leq h + \sum_{\mathbf{z}} h \cdot P_Z(\mathbf{z}) = 2h.
\end{aligned} \tag{63}$$

The last inequality stems from Definition 2, where, due to causality (Definition 1), conditioning on \mathbf{X}_{n-L}^∞ can be replaced by conditioning on \mathbf{X}_{n-L}^m .

B. Proof of Lemma 4

Let X be the channel input, Y the channel output, $J \sim p$ the channel index and Q an input distribution. The joint distribution is defined by $\Pr(XYJ) = p_J \cdot Q(X) \cdot W_J(Y|X)$. Then

$$I(X; Y|J) = \sum_i p_i I(X; Y|J=i) = \sum_i p_i I(Q, W_i), \tag{64}$$

and

$$I(X; Y) = I\left(Q, \sum_i p_i W_i\right). \tag{65}$$

On one hand, due to the convexity of the mutual information with respect to the channel

$$I\left(Q, \sum_i p_i W_i\right) \leq \sum_i p_i I(Q, W_i). \tag{66}$$

Maximizing with respect to Q yields the right inequality of (37). On the other hand,

$$\begin{aligned}
\sum_i p_i I(Q, W_i) &= I(X; Y|J) \\
&= H(X|J) - H(X|JY) \\
&\leq H(X) - (H(XJ|Y) - H(J|Y)) \\
&\leq H(X) - H(X|Y) + H(J) \\
&= I(X; Y) + H(J) \\
&= I\left(Q, \sum_i p_i W_i\right) + H(p).
\end{aligned} \tag{67}$$

Maximizing with respect to Q yields the left inequality of (37). \square

C. Proof of Lemma 2

Choose an ϵ and find N large enough so that for $n \geq N$ $\delta_n \leq \epsilon$, then for $n \geq N$:

$$\begin{aligned}
\frac{\sum_{i=1}^n a_i \delta_i}{\sum_{i=1}^n a_i} &\leq \frac{\sum_{i=1}^{N-1} a_i \delta_i}{\sum_{i=1}^n a_i} + \frac{\sum_{i=N}^n a_i \epsilon}{\sum_{i=1}^n a_i} \\
&\leq \frac{a_{N-1} \sum_{i=1}^{N-1} \delta_i}{a_{N-1} (n - N + 1)} + \frac{\sum_{i=N}^n a_i \epsilon}{\sum_{i=1}^n a_i} \\
&= \frac{\sum_{i=1}^{N-1} \delta_i}{(n - N + 1)} + \epsilon.
\end{aligned} \tag{68}$$

By taking n large enough, the first term can be made arbitrarily small, and therefore the RHS can be made arbitrarily small for n large enough. \square

D. A limit on the convergence rate

In [4], a rigorous analysis of the best possible convergence rate for the modulo-additive channel was performed. Here, considering the general vector channel, only rough estimates for the convergence rate are presented, without a rigorous proof. The main question is the value of $n^*(k, \delta)$, which is the minimum value of n required to obtain a redundancy δ with respect to an IFB system with block size k , and was shown in [4] to grow like $O(|\mathcal{X}|^k)$ for small δ in the case of the modulo additive channel. Below, a rough lower bound on n^* is shown for general causal fading-memory channels.

Consider a test channel defined as follows: let $\{\mathbf{x}^{(m)}\}_{m=1}^{|\mathcal{Y}|}$ be $|\mathcal{Y}|$ different arbitrary input strings of length k , and $F : \mathcal{Y}^{k-1} \rightarrow \mathcal{Y}$ be an arbitrary function from the set of $k-1$ length output strings to a single output letter. The channel operates independently over each block of k symbols. Let \mathbf{X}, \mathbf{Y} denote the input and output over these k symbols. For each block of k output symbols, the first $k-1$ output symbols Y_1, \dots, Y_{k-1} are drawn i.i.d. uniformly. The last output symbol is determined as follows: if $\mathbf{x}^{(m)}$ was the input (over the k input letters), for some m , then $Y_k = F(\mathbf{Y}^{k-1}) + m$, where the addition is modulo- $|\mathcal{Y}|$. Otherwise, Y_k is drawn randomly uniformly and independently of the previous outputs. An ensemble of such channels can be created by uniformly drawing $\{\mathbf{x}^{(m)}\}_{m=1}^{|\mathcal{Y}|}$ out of all possible sets of different words, and generating F as a random function, by drawing each of the $|\mathcal{Y}|^{k-1}$ values $F(\mathbf{y}^{k-1})$ i.i.d. and uniformly over \mathcal{Y} . The channel is causal and is fading memory (with memory of k symbols). The reference IFB system achieves a rate of $\log |\mathcal{Y}|$ bits per block $R_{\text{IFB}} = \frac{\log |\mathcal{Y}|}{k}$, without error, by encoding the message $m \in \mathcal{Y}$ into $\mathbf{x}^{(m)}$, and decoding using $\hat{m} = Y_k - F(\mathbf{Y}^{k-1})$. Note that this contrived construction is mainly aimed at achieving causality, and would be simplified if any block-wise channel law could be devised.

A universal system attempting to reach the rate of $\log |\mathcal{Y}|$ bit per block needs to be able to identify $\{\mathbf{x}^{(m)}\}_{m=1}^{|\mathcal{Y}|}$. Identification is meant in the sense, that eventually (by time n^*), most of the time, only $\{\mathbf{x}^{(m)}\}$ will be transmitted, so an agent viewing the transmitter's output will be able to infer $\{\mathbf{x}^{(m)}\}$. To see this, consider the Shannon capacity of the channel with $|\mathcal{Y}| + 1$ inputs $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(|\mathcal{Y}|)}, \text{"other input"}\}$ whose output is $Y_k - F(\mathbf{Y}^{k-1})$. It is easy to see that, for the purpose of

communication when the test channel is known, this channel is a sufficient description. The channel is noiseless for the first $|\mathcal{Y}|$ input letters, and completely noisy for the last input letter. Therefore, its capacity achieving prior places all the distribution on $\{\mathbf{x}^{(m)}\}$ and any significant deviation from this distribution will reduce the achieved rate. Now, the input words $\mathbf{x}^{(m)}$ are only special in the sense, that the last output letter is a function of the first $k-1$ outputs. To determine whether an arbitrary word \mathbf{x} is in this set, one has to observe multiple times the same sequence \mathbf{Y}^{k-1} , and see that they all yield the same Y_k . Thus, this identification takes $O(|\mathcal{Y}|^k)$ trials, in which \mathbf{x} is the input to the channel. This $O(\cdot)$ is in the sense that lower than $|\mathcal{Y}|^{k-1}$ are not sufficient for reliable decision, and some constant times $|\mathcal{Y}|^k$ is sufficient. The words $\mathbf{x}^{(m)}$ are randomly scattered in the set of $|\mathcal{X}|^k$ possible input sequences, and virtually, the detection of one sequence, does not give any significant information for the detection of others (it can only reduce the bound above by a small constant, by knowing which values of Y_k to expect). Hence, in order to identify $\{\mathbf{x}^{(m)}\}_{m=1}^{|\mathcal{Y}|}$, all $|\mathcal{X}|^k$ input sequences would have to be tested, i.e. appear at the encoder's input at least $O(|\mathcal{Y}|^k)$ times, which requires $n^* \geq O(|\mathcal{X}|^k \cdot |\mathcal{Y}|^k)$.

While this convergence rate is already slow, the actual convergence rate of the scheme presented here §III-A is far slower. This is not surprising, as the current scheme was not optimized for efficiency. As a result, unlike the modulo-additive case [4], we do not have an upper bound on n^* , with the same growth rate as the lower bound above. A rough analysis of the scheme's convergence rate is presented in [12, §D.5].

REFERENCES

- [1] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Trans. Information Theory*, vol. 44, no. 6, pp. 2148–2177, Oct. 1998.
- [2] O. Shayevitz and M. Feder, "Achieving the empirical capacity using feedback: Memoryless additive models," *IEEE Trans. Information Theory*, vol. 55, no. 3, pp. 1269–1295, Mar. 2009.
- [3] K. Eswaran, A. Sarwate, A. Sahai, and M. Gastpar, "Zero-rate feedback can achieve the empirical capacity," *IEEE Trans. Information Theory*, vol. 58, no. 1, Jan. 2010.
- [4] Y. Lomnitz and M. Feder. (2010, Dec.) Universal communication part I: modulo additive channels. arXiv:1012.2751v1 [cs.IT]. Submitted to IEEE-IT. [Online]. Available: <http://arxiv.org/abs/1012.2751>
- [5] —. (2011, Sep.) Universal communication over arbitrarily varying channels. arXiv:1102.0710 [cs.IT]. Accepted for publication in IEEE-IT. [Online]. Available: <http://arxiv.org/abs/1102.0710>
- [6] A. Lapidoth and I. Telatar, "The compound channel capacity of a class of finite-state channels," *IEEE Trans. Information Theory*, vol. 44, no. 3, pp. 973–983, May 1998.
- [7] Y. Lomnitz and M. Feder, "Communication over individual channels," *IEEE Trans. Information Theory*, vol. 57, no. 11, pp. 7333–7358, Nov. 2011.
- [8] S. Verdú and T. Han, "A general formula for channel capacity," *IEEE Trans. Information Theory*, vol. 40, no. 4, pp. 1147–1157, Jul. 1994.
- [9] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Information Theory*, vol. 24, no. 5, pp. 530–536, Sep. 1978.
- [10] R. Gallager, *Information Theory and Reliable Communication*. John Wiley & sons, 1968.
- [11] V. Misra and T. Weissman, "The porosity of additive noise sequences," in *IEEE Int. Symp. Information Theory (ISIT)*, 2012.
- [12] Y. Lomnitz, "Universal communication over unknown channels," Ph.D. dissertation, Tel Aviv University, Aug. 2012, available online http://www.eng.tau.ac.il/~yuval/publications/YuvalL_PhD_report.pdf.
- [13] V. Misra and T. Weissman. (2012) The porosity of additive noise sequences. [Online]. Available: <http://arxiv.org/abs/1205.6974>
- [14] S. Shamai and S. Verdú, "The empirical distribution of good codes," *IEEE Trans. Information Theory*, vol. 43, no. 3, pp. 836–846, May 1997.