Universal Communication—Part I: Modulo Additive Channels

Yuval Lomnitz, *Member, IEEE*, and Meir Feder, *Fellow, IEEE*

Abstract—Which communication rates can be attained over a channel whose output is an unknown (possibly stochastic) function of the input that may vary arbitrarily in time with no *a priori* model? Following the spirit of the finite-state compressibility of a sequence, defined by Lempel and Ziv, a "capacity" is defined for such a channel as the highest rate achievable by a designer knowing the particular relation that indeed exists between the input and output for all times, yet is constrained to use a fixed finite-length block communication scheme without feedback, i.e., use the same encoder and decoder over each block. In the case of the modulo additive channel, where the output sequence is obtained by modulo addition of an unknown individual sequence to the input sequence, this capacity is upper bounded by a function of the finite state compressibility of the noise sequence. A universal communication scheme with feedback that attains this capacity universally, without prior knowledge of the noise sequence, is presented.

Index Terms—Arbitrarily varying channels (AVCs), feedback communication, individual sequences, rateless coding, universal communication, unknown channels.

I. INTRODUCTION

CONSIDER the problem of communicating over a channel, where the (possibly stochastic) relation between the input and output is unknown to the transmitter and the receiver and may be, in general, nonstationary. In particular, no assumption is made that the channel behavior up to a certain point in time indicates anything about its expected behavior from this time on. The key characteristic of such a channel is that the channel law cannot be learned, i.e., it is impossible, using an asymptotically short measurement period, to obtain the channel probability law and use it during the rest of the transmission.

Clearly, communication over such an arbitrary channel is challenging. Furthermore, even the question what the limits of such communication are, is not well posed. To emphasize the fact that the relation between input and output is a function of the entire sequences, or vectors, this channel shall be termed a *vector* channel. A simple example of such a channel, which was discussed by Shayevitz and Feder [1] is the modulo-additive channel with an individual noise sequence, defined by

the relation $\mathbf{y} = \mathbf{x} + \mathbf{z}$ where $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}^n$ are n-length vectors, denoting the input, output, and the noise sequence, $\mathcal{X}$ is a finite alphabet, the "+" denotes modulo addition over $\mathcal{X}$, and the sequence $\mathbf{z}$ is arbitrary and unknown. The main focus in this paper is on this channel model. When the alphabet is $\mathcal{X} = \{0, 1\}$ this channel is referred to as the binary additive channel.

In the general vector channel, when the conditional probability of the output vector given the input vector is *known*, the classical Shannon capacity, i.e., the maximum communication rate achievable with an arbitrarily small error probability, is well defined. The Shannon capacity of the general causal vector channel was given by Han and Verdú [2]. When the channel is *unknown*, the same communication rate is in many cases not attainable. In this case, the compound channel or arbitrarily varying channel (AVC) frameworks [3] may be used. In these frameworks, the capacity is defined as the maximum rate of transmission which guarantees robust communication over all possible channels. However, these frameworks do not consider the ability to use feedback to adjust the communication parameters, and are therefore worst case in nature. On the other hand, Shayevitz and Feder [1] have shown that for the modulo-additive channel with an individual noise sequence, by using feedback to adapt the transmission rate to the actual channel occurrence, these worst case assumptions may be alleviated. These results were extended by the authors and others [4], [5].

Since the channel is unknown, the target is to find a universal communication system that operates without knowing the channel. While there are known universal source encoders [6] and universal predictors [7], in the communication problem, the term "universality" had been used mainly with respect to decoders, competing against the maximum likelihood decoder in a compound channel [3], [8], and there is currently no notion of universality with respect to the complete communication system. This is since in the traditional AVC model, feedback is not considered and therefore the encoder is assumed to be fixed. On the other hand, in existing works that consider adaptation of the communication rate using feedback [1], [4], [5], the communication rates achieved do not have a strong justification. For example, these works define the rate using zero-order empirical distributions, and higher rates could be attained by considering empirical distributions with memory.

Let us denote by $P_{Y|X}^{(\theta)}$ a conditional distribution of the channel output given the input defining a vector channel, where θ is an index belonging to a possibly infinite index set Θ. Given a class of vector channels $\{P_{Y|X}^{(\theta)}\}_{\theta \in \Theta}$, the objective is to assign a rate C_θ to each channel, such that on one hand C_θ has an operational meaning, for example, the maximum rate achievable

under certain constraints, and on the other hand, it would be possible to construct a universal system using feedback, that without knowledge of θ, attains a rate of at least C_θ for all θ. The difference from the AVC or compound channel models is that the communication rate depends on θ. As shall be seen, the maximum rate achievable by block encoders and decoders that know θ is a reasonable target that can be used as a definition for C_θ. This target rate is universally achievable for the class of modulo-additive channels, as shown below. More generally, this target is universally achievable for arbitrary channels with fading memory, as shown in a follow up paper [9], using more elaborate tools. On the other hand, this rate is not universally achievable in general.

The main contributions of this paper are as follows. For the general problem of communication over any unknown vector channel, the first definition of competitive universality is given. In particular, the "iterated finite block capacity," $C_\theta = C_{\text{IFB}}(P_{Y|X}^{(\theta)})$ is defined. A significant part of the paper is devoted to exploration of the problem boundaries: For which channel families can the IFB capacity be attained? Which other interesting definitions of the reference class can be given? The other contributions are specific to the modulo-additive channel with an individual noise sequence. A bound on the IFB capacity is given, a universal system that attains the target rate without knowing the channel is presented, and the redundancy in approaching this rate is analyzed. The converse part of the redundancy analysis holds also for larger families of channels that include the modulo-additive channel as a special case.

The paper is organized as follows: in Section II, the motivations for the definition of C_{IFB} are explained. Section III is a high-level overview of the results regarding the modulo-additive channel, and the main ideas behind the proofs. Section IV includes the detailed definitions with some discussion. Section V focuses on the modulo additive channel and includes the upper bound on C_{IFB} and the universal system achieving it. The redundancy, i.e., the convergence rate, in achieving the IFB capacity is explored in Section VI. Section VII is devoted to discussion and comments and suggests some extensions and alternative definitions.

II. MOTIVATION

Let us now discuss the motivations for the definitions of a target rate which is universally achievable. The inherent difficulty of defining the maximal communication rates over arbitrary vector channels can be appreciated by considering even the simple example of a binary additive channel $\mathbf{y} = \mathbf{x} \oplus \mathbf{z}$ with an individual noise sequence $\mathbf{z}$, where "$\oplus$" denotes modulo-2 addition and all vectors are of length n. For every specific individual noise sequence $\mathbf{z}$, the capacity of this channel is 1 bit/use. On the other hand, if the noise sequence is arbitrary and unknown, the AVC capacity [3] is zero. It would initially seem that not much can be done, when the noise sequence is unknown; however it was shown [1] that, using feedback and common randomness, and by adapting the decoding rate, a communication rate of $R = 1 - \hat{H}(\mathbf{z}^n)$ can be achieved, where $\hat{H}(\cdot)$ denotes the empirical entropy of the noise sequence, i.e., the binary entropy of the empirical cross-over probability. The main idea is
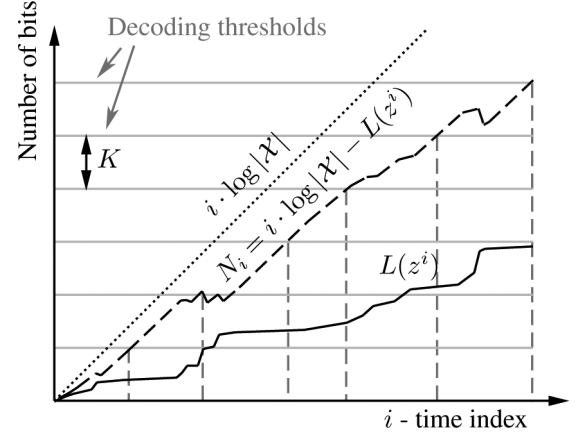

Fig. 1. Illustration of the decoding rule of the rate adaptive system. $L(z^i)$ is the compression length. Decoding thresholds with respect to $N_i = i \cdot \log |\mathcal{X}| - L(z^i)$ are depicted by horizontal lines.

that if the empirical channel can be measured and the communication rate can be adapted, then rather than making *a priori* pessimistic assumptions, one can opportunistically increase the rate when the noise sequence has a low empirical entropy.

A disturbing fact is that some arbitrariness exists in deciding on the rates to achieve per each channel: in the binary additive channel, given a sequence $\mathbf{s}$ of choice, one could also design a system that achieves the rate $1 - \hat{H}(\mathbf{z} \oplus \mathbf{s})$, by adding the sequence $\mathbf{s}$ to the channel output and then applying Shayevitz and Feder's scheme [1]. Doing so, a rate of 1 is obtained for the sequence $\mathbf{z} = \mathbf{s}$, where the original system's rate was $1 - \hat{H}(\mathbf{s})$, and a rate of $1 - \hat{H}(\mathbf{s})$ for the noiseless case $\mathbf{z} = 0$, so one may say that the noise sequence $\mathbf{s}$ is "favored" over $\mathbf{0}$. This demonstrates the arbitrariness in determining which communication rates are possible. To remove this arbitrariness, a reasonable criterion is sought, to decide which channels (noise sequences, in the example) to favor over others.

This issue bears significant resemblance to issues tackled in universal source coding (compression) and in universal prediction. In universal compression, one would like to set a target for the compression rate of an individual sequence. As in the current problem, someone who knows the sequence can design an encoder which compresses it to 1 bit, whereas assuming the sequence is completely unknown and without favoring any sequence over another, no compression can be achieved. There are many possible fixed to variable encoders which are uniquely decodable, and the decision between them may seem arbitrary. One solution proposed by Lempel and Ziv [6] was to set as a target, the compression rates that are achievable by machines with limited capabilities, i.e., finite state machines (FSMs). They defined the notion of *finite state compressibility* for an infinite sequence, as the best compression rate that can be achieved by any information lossless FSM operating over the infinite sequence, and had shown that the LZ78 compression algorithm based on incremental parsing, defined there, achieves this compression rate universally for any sequence. This concept supplies a criterion to decide which sequences to favor over others, without assuming a probability law. A similar notion, i.e., that of comparing against the best machine out of a restricted class, is applied in universal prediction [7], [10].
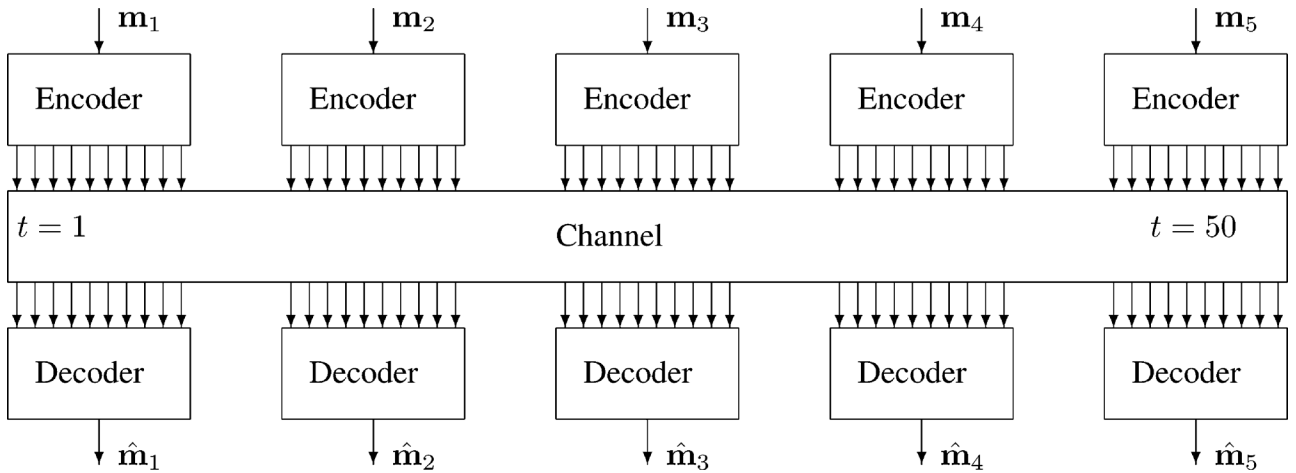
Fig. 2. Illustration of *iterative mapping* used for the definition of average error probability (see Definition 3). The same encoder and decoder are used over each of the $b = 5$ blocks of $k = 10$ channel uses, and the average error probability is computed.

Following this lead, the comparison class is chosen to be the set of fixed finite-length block encoders and decoders, which repeatedly perform the same encoding and decoding operations over blocks of any fixed length (see Fig. 2). This class is a relatively simple one, while still yielding a reasonable criterion to set the communication rate. The *iterated finite block capacity* of an infinite vector channel C_{IFB} is defined as the supremum of all rates which are reliably achievable by encoders and decoders in the comparison class. This capacity value may be smaller, in general, than the Shannon capacity of the vector channel. This definition has operational significance, since many practical communication systems use block encoding, and therefore universally attaining the C_{IFB} means that one can design a system which, without any prior knowledge of the channel, is essentially at least as good as any finite block code. The universal system itself does not belong to the comparison class—it does not operate in fixed blocks, it modifies its behavior based on the past, and it uses feedback. Although achieving C_{IFB} universally is possible for classes of vector channels wider than the modulo-additive channel [9], it is not possible to attain this rate for general unknown vector channels.

The IFB model presented here is simple and intuitive, however it has several drawbacks and alternative definitions and extensions can be proposed. These are discussed in Section VII-A. Most notably, Misra and Weissman [11] generalized the current results to finite-state communication systems with feedback. For the sake of simplicity, the paper focuses on the basic model of reference systems using block coding.

Although the results are currently purely theoretical (see Section VI), they supply motivation for using competitive universality in communication.

III. OVERVIEW OF THE MAIN RESULTS

This section provides an informal review of the main results and rough proof outlines. The purpose is to provide an understanding of the results without diving into mathematical detail. The main results of this paper pertain to the modulo-additive channel with individual noise sequence. For this channel, it is shown in Section V that $C_{\mathrm{IFB}} \leq (1 - \rho(\mathbf{z})) \cdot \log |\mathcal{X}|$, where $\rho(\mathbf{z})$ is the finite state compressibility of the infinite sequence $\mathbf{z}$, as defined by Lempel and Ziv [6]. Assuming that common randomness exists and that there is a feedback link, a universal system employing feedback exists, which asymptotically attains this rate universally without prior knowledge of the noise sequence. In Section VI, upper and lower bounds on the convergence rate are derived. Below, the main ideas in the proofs are described.

Let us begin with the upper bound on C_{IFB}. Suppose the given reference system comprised of an encoder and a decoder, achieves the rate R over b blocks of size k (see Fig. 2). During these b blocks, the reference system "sees" b different noise vectors of length k, namely $\mathbf{z}_{(i-1)k+1}^{ik}$, $i = 1, \ldots, b$. Since the system is fixed during these b blocks, this is equivalent to operating over a stochastic channel, where the noise vector $\tilde{\mathbf{Z}}$ is chosen uniformly from the set of these vectors, with probability $\frac{1}{b}$ for each. This random vector is termed the "collapsed" noise sequence, and the channel generated from it the "collapsed" channel (see Fig. 5). The standard converse of the channel capacity theorem, without the assumption of a memoryless channel, can be applied to the collapsed channel, and yields an upper bound on C_{IFB}, which is roughly $\log |\mathcal{X}| - \frac{1}{k} H(\tilde{\mathbf{Z}})$. The entropy $H(\tilde{\mathbf{Z}})$ is lower bounded using the finite state compressibility of the sequence, since an FSM may achieve a compression rate close to the entropy by standard block-to-variable coding, where the code lengths are tuned to the statistics of the collapsed noise vector. Combining these bounds yields the result $C_{\mathrm{IFB}} \leq (1 - \rho(\mathbf{z})) \cdot \log |\mathcal{X}|$ (see Theorem 1).

Next, a communication scheme is demonstrated, that asymptotically attains the rate $\log |\mathcal{X}| - \frac{1}{n} L(\mathbf{z}^n)$, where $L(\mathbf{z}^n)$ is the compression length of the sequence $\mathbf{z}$ by a given sequential source encoder, and n is the overall block length. The scheme is based on iterative application of rateless coding, sending K bits in each block. Each codeword in the codebook of $\exp(K)$ words is chosen independently and distributed uniformly over $\mathcal{X}^n$. The transmitter sends symbols from the codeword matching the K message bits, until a termination condition occurs at the receiver side. Then, the receiver indicates the end of the block through

the feedback link and a new block begins. The termination condition is based on giving a rank to every possible noise sequence $\mathbf{z}_1^i$. This noise sequence has two parts: the part spanning previous blocks that had already been decoded is known with high probability, as both the channel input and the channel output are known. The part of the noise sequence from the beginning of the current block is unknown, but is related in a one-to-one relation to the unknown message, because the channel output is known. Therefore, there are $\exp(K)$ possible noise sequences, corresponding to the $\exp(K)$ possible codewords sent in the current block. The rank of each noise sequence is the code-length, or the number of bits in its representation by the given source encoder. The decoder terminates the block if for any codeword this length is smaller than a threshold.

The proof of this scheme's performance is roughly as follows. Due to the random coding, most of the hypotheses, except the true one, yield random noise sequences. These sequences are incompressible, and therefore the number of bits representing the last block would be approximately $\log |\mathcal{X}|$ times the number of symbols in the block. It can be shown that setting the threshold approximately K below this value, guarantees a small probability of exceeding the threshold for any of the $\exp(K) - 1$ incorrect codewords, and therefore a small probability of error. It is convenient to define the "porosity" of the sequence up to time i as $N_i = i \cdot \log |\mathcal{X}| - L(\hat{\mathbf{z}}_1^i)$, representing the gap between the compressibility of the hypothetical noise sequence, and the compressibility of a random sequence. The approximate termination condition may be interpreted as decoding when the value of N_i increases by K from the start of the current block. Since when this occurs, the system starts a new block, there is a correspondence between the increase in N_i and the number of blocks and bits that are transmitted, i.e., the termination condition can be approximately interpreted as $N_i \geq K(b+1)$ where b is the number of blocks so far. Therefore, assuming by time n, B blocks were transmitted, the number of transmitted bits is $K \cdot B \approx N_n = n \cdot \log |\mathcal{X}| - L(\hat{\mathbf{z}}_1^n)$. Assuming no errors occurred $\hat{\mathbf{z}}_1^n = \mathbf{z}_1^n$, and dividing by n the desired result is obtained. This is depicted in Fig. 1, where the horizontal axis is the time i. The solid line presents $L(\mathbf{z}_1^i)$, and the dashed line N_i. The decoding thresholds Kb ($b = 1, 2, \ldots$) are depicted as horizontal lines, while the vertical lines depict the decoding times. Decoding occurs whenever N_i crosses a threshold. A random hypothesized sequence in the current block implies that N_i does not increase on average. It can be seen that the number of bits that will be sent is approximately N_n. In the full proof, various overheads that were neglected above are accounted for.

To obtain the universal system attaining C_{IFB} (see Theorem 3), the scheme above is applied with the encoding lengths $L(\mathbf{z}^n)$ determined by the LZ78 source encoder, whose compression ratios asymptotically approach the finite state compressibility: asymptotically $\frac{1}{n}L(\mathbf{z}^n) \leq \rho(\mathbf{z}) \log |\mathcal{X}|$, therefore $\log |\mathcal{X}| - \frac{1}{n}L(\mathbf{z}^n) \geq (1 - \rho(\mathbf{z})) \log |\mathcal{X}|$, where all inequalities are up to asymptotically vanishing factors.

Section VI deals with the question of the redundancy, or how quickly the system converges to the rate attained by the best IFB system with a given block length k. Unfortunately, it is shown that n must grow at least as fast as $|\mathcal{X}|^k$, approximately. The upper bound on redundancy is obtained by using

a similar universal system employing a slightly more refined design: a universal probability assignment based on a mixture of Krichevsky–Trofimov distributions [12] is used instead of the LZ78 encoder. The lower bound is obtained by presenting a design of an IFB system together with a random channel, i.e., a distribution over noise sequences, such that the mutual information over the channel is smaller than the rate obtained by the IFB system. This is possible because the IFB system is designed together with the channel and can use the knowledge of the specific noise sequence. On the other hand, the rate obtained by any universal system with feedback is bounded by the mutual information, and this gap comprises the lower bound on redundancy. The upper bound and the lower bound on the redundancy agree in terms of the asymptotical growth rate of n as function of k (see Fig. 7).

IV. CHANNEL MODEL AND DEFINITIONS

This section begins the formal presentation of the results, by presenting the channel model and the definitions of the capacity C_{IFB}, and discussing their implications.

A. Notation

Vectors are denoted by boldface letters. Subvectors are defined by superscripts and subscripts: $\mathbf{x}_j^i \triangleq [x_j, x_{j+1}, \ldots, x_i]$. $\mathbf{x}_j^i$ 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 or random variable $\mathbf{X}$, $\mathbf{X}_i^{[k]} \triangleq \mathbf{X}_{(i-1)k+1}^{(i-1)k+k}$ denotes the ith 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 as well as information quantities are base 2. Random variables are distinguished from their sample values by capital letters. The indicator function $\text{Ind}(E)$ where E is a set or a probabilistic event is defined as 1 over the set (or when the event occurs) and 0 otherwise.

$h_b(p)$ denotes the binary entropy function, i.e., the entropy of a Bernully-p random variable, and $\rho(\mathbf{z}^\infty)$ denotes the finite state compressibility [6] of $\mathbf{z}$, defined formally in Section V-A, (3)–(5).

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 this 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 $n = 1, 2, \ldots \infty$. A finite length output sequence is considered in order to make the probability well defined. Sometimes, this probability will be informally referred to as $\Pr(\mathbf{Y}_1^\infty | \mathbf{X}_1^\infty)$, and should be understood as the sequence of these distributions for $n = 1, 2, \ldots$.

Definition 1: The channel defined by $\Pr(\mathbf{Y}_1^n | \mathbf{X}_1^\infty)$ is termed *causal* if for all n:

$$\Pr(\mathbf{Y}_1^n | \mathbf{X}_1^\infty) = \Pr(\mathbf{Y}_1^n | \mathbf{X}_1^n). \qquad (1)$$
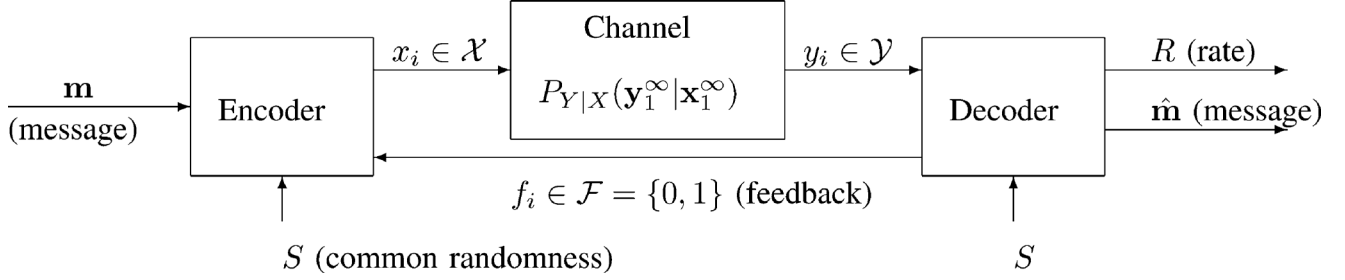
Fig. 3. Rate adaptive encoder-decoder pair with feedback, over an unknown channel.

All the definitions below (including IFB capacity) pertain to causal channels. This characterization of a causal channel is similar to the definition used by Verdú and Han [2] (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 noncausality that consists of bounded negative delays can always be compensated by applying a delay to the output.

C. IFB Capacity

The following definitions lead to the definition of IFB capacity.

Definition 2 (Reference Encoder and Decoder): A finite length encoder E with block length k and a rate R is a mapping $E : \{1, \ldots, M\} \to \mathcal{X}^k$ 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 $D : \mathcal{Y}^k \to \{1, \ldots, M\}$ from the set of output sequences to the set of messages.

Definition 3 (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 $\mathbf{m}_1, \ldots, \mathbf{m}_b$ are chosen as i.i.d. uniformly distributed random variables $\mathbf{m}_i \sim U\{1, \ldots, M\}, i = 1, \ldots, b$. The channel input is set to $\mathbf{X}_i^{[k]} = E(\mathbf{m}_i), i = 1, \ldots, b$, 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 4 (IFB Achievability): A rate R is *iterated-finite-block (IFB) 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 mapping 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 5 (IFB Capacity): The *IFB capacity* of the channel $P_{Y|X}$ is the supremum of the set of IFB achievable rates, and is denoted C_{IFB}.

D. Competitive Universality

In the following, the properties of the adaptive system with feedback, and IFB-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 [13, Sec. 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-universality for rate adaptive systems:

Definition 6 (IFB 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 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 $C_{\mathrm{IFB}}(P_{Y|X}) - \delta$.

Notice that the definitions above (and specifically definitions 4, 6) 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. Discussion on IFB Capacity and Universality

Following are some comments regarding IFB capacity and IFB universality. Note that the use of average error probability over time and messages (expressed in the assumed uniform distribution) rather than maximum error probability (over time or messages) reduces the requirements from E, D and therefore increases C_{IFB}.

As noted, $C_{\mathrm{IFB}} \leq C$, where C is the Shannon capacity [2]. However, for i.i.d. memoryless channels clearly $C_{\mathrm{IFB}} = C$. The difference between C and C_{IFB} relates to the stability of the channel over time, and the ability to utilize channel structure which cannot be observed in finite time. Let us give two examples to sharpen this difference.

Example 1: Consider the binary product channel $y_i = x_i \cdot z_i$, and let the sequence $\mathbf{z}$ alternate between 0 and 1, in blocks of ever growing size, but such that the overall frequency of 0 is $\frac{1}{2}$, and the length of each block is negligible compared to the total length of previous blocks. For example, set z_i to 0 in $i \in$
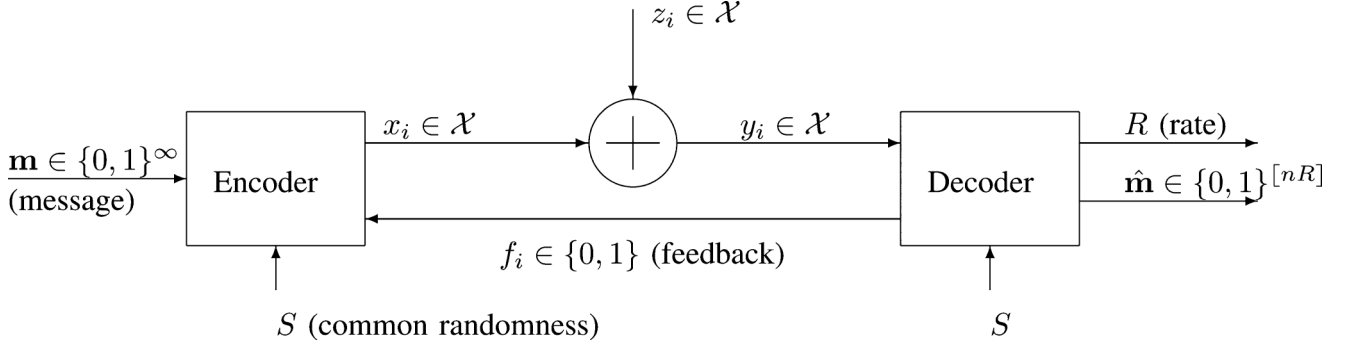
Fig. 4. Adaptive system over the modulo-additive channel with feedback.

$\cup_{k=1}^{\infty}[2k^2, (k+1)^2+k^2]$. For this channel $C_{\text{IFB}} = 0$ while $C = \frac{1}{2}$. The reason is that for every finite length encoder/decoder, ultimately as $m \to \infty$ half the blocks will fall on bursts of $z = 0$ and be in error. Note that if rate adaptation would have been allowed at the IFB decoder, this capacity would not be zero (see Section VII-A)

Example 2: Consider a channel with ever growing delay: suppose that d_i is a sequence of slowly growing delays. For example, $d_i = \lfloor \log i \rfloor$, and the channel is $y_i = x_{i-d_i}$, where x, y are binary. The capacity of this channel is $C = 1$, whereas $C_{\text{IFB}} = 0$. Here, the reason for the gap is the in-ability to utilize the channel structure with a finite block size.

Following these examples the choice of C_{IFB} may be justified by two main reasons: one is its operational significance, i.e., that universally attaining C_{IFB}, means competing with every static block coding system, and the other is the rejection of "pathological" behaviors of the channel, as the ones mentioned in the examples above.

Note that although $C_{\text{IFB}} \le C$, the universal system presented here may opportunistically achieve rates above C. This means the communication rate may exceed C in part of the time. Consider for example the binary nonergodic channel that with probability p has $\mathbf{y} = \mathbf{x}$, and with probability $1 - p$ the output is independent of the input. While the capacity of this channel is $C = 0$ (and $C_{\text{IFB}} = 0$), by adapting the rate, one could attain a rate of 1 with probability p.

An interesting question is whether for a general vector channel, C_{IFB} can be universally attained. Unfortunately, the answer is negative, and the reason is that, because the input sequences used by the reference encoder and by the universal system are different, infinite memory in the channel may cause the channel to get "stuck" in an unfortunate state. This phenomenon is nicknamed a "password" channel, since it is similar to a situation where a password is required at the beginning of transmission, otherwise the channel becomes useless. In this case, a reference system knowing the password may succeed and a universal system, having only one attempt to find the password, is bound to fail. More generally, given an encoder, a channel can be structured such that it will identify the specific encoder's codebook, and fail if any deviation from this codebook is observed. Here is a simple example:

Example 3 (Password Channel): Consider a family of two binary channels. In the first channel, if $x_1 = 0$ then the channel will become clean, i.e., $\forall i \ge 2 : y_i = x_i$, but if $x_1 = 1$, then

it becomes blocked, i.e., $\forall i \ge 2 : y_i = 0$. The second channel is the same, except the roles of $0, 1$ are reversed. Clearly, for both channels $C_{\text{IFB}} = 1$, since the only constraint required to avoid blocking is that the first symbol in each encoded block is constant 0 or 1, and therefore a rate of $\frac{k-1}{k}$ can be obtained with block size k. On the other hand, no universal system can guarantee any rate with a vanishing error probability, since any choice of the first symbol will lead to blocking in one of the two channels.

The conclusion from the above is that the concept of iterated finite block capacity is not as strong as the concept of finite state compressibility, which is truly universally attainable. This problem relates to a fundamental difficulty in universal communication compared to universal compression: in universal compression, the sequence is given and does not depend on the encoder's actions, while in communication, the encoder's actions (the input symbols) affect the channel behavior in an unexpected way.

One may be tempted to think that depriving the IFB class from its blockwise operation and limiting it to i.i.d. distributions would solve the "password" problem. However, it is easy to devise a channel that would identify the input distribution of the reference encoder, while blocking the universal system. See Example 5 in Appendix G. These difficulties exemplify the complexity of the universal communication problem.

V. UNIVERSAL COMMUNICATION OVER THE MODULO-ADDITIVE CHANNEL

This section and the next, focus on the modulo-additive channel with an individual noise sequence (Fig. 4). It is shown that the IFB capacity of this channel is bounded by $(1 - \rho(\mathbf{z})) \log |\mathcal{X}|$ and that this rate is universally achievable. Upper and lower bounds on the convergence rates are given, which show that, unfortunately, the transmission length n required to obtain universal communication grows exponentially with the block length k of the competing system.

The modulo-additive channel is a relatively "easy" case because of two main reasons:

1) It is memoryless in the input, and thus the "password" issue is avoided.
2) There is a single input prior, the uniform i.i.d. distribution, which attains capacity for any noise sequence, since it maximizes the output entropy. Therefore, no adaptation of the prior is needed.
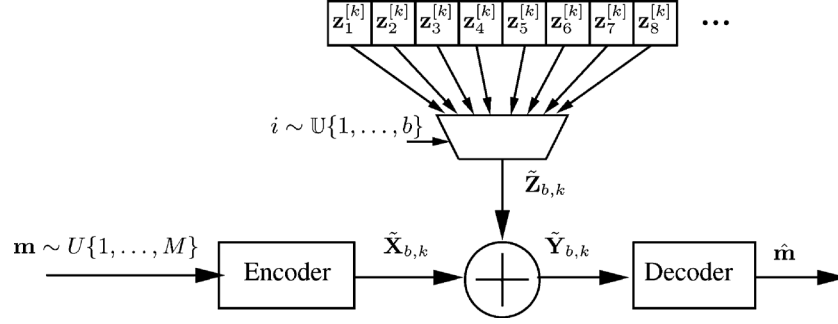
Fig. 5. Collapsed channel: a probabilistic equivalence to iterative mapping.

A. Bound on the IFB Capacity of the Modulo-Additive Channel

In this section, the following theorem is proven.

Theorem 1: The IFB-capacity of the modulo-additive channel $\mathbf{y} = \mathbf{x} + \mathbf{z}$ where $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}^{\infty}$ are infinite sequences denoting the channel input, output, and noise sequence, satisfies

$$C_{\text{IFB}} \leq (1 - \rho(\mathbf{z})) \cdot \log |\mathcal{X}| \qquad (2)$$

where $\rho(\mathbf{z})$ is the finite state compressibility of $\mathbf{z}$.

For the sake of completeness, let us shortly repeat the definition of finite state compressibility. A finite state encoder F with s states is defined by a next state function $g : (\{1, \ldots, s\}, \mathcal{X}) \rightarrow \{1, \ldots, s\}$, and an output function $f : (\{1, \ldots, s\}, \mathcal{X}) \rightarrow \{\{0, 1\}^k\}_{k=0}^{\infty}$, where the output may be a bit sequence of any length, including the empty sequence. The encoder is said to be *information lossless* if for any $\mathbf{z}_1^n$, the input $\mathbf{z}_1^n$ can be uniquely decoded from the output sequence $F(\mathbf{z}_1^n)$, given the initial and final states. Let $\mathcal{F}(s)$ denote the group of all finite state information lossless encoders with at most s states. Let the length of the output sequence of encoder F for an input sequence of length n be denoted $|F(\mathbf{z}_1^n)|$, then the compression ratio of $\mathbf{z}_1^n$ by F is defined as

$$\rho_F(\mathbf{z}_1^n) \triangleq \frac{1}{n \log |\mathcal{X}|} |F(\mathbf{z}_1^n)|. \qquad (3)$$

The compression ratio of the best information lossless finite state encoder with at most s states is denoted

$$\rho_{\mathcal{F}(s)}(\mathbf{z}_1^n) \triangleq \min_{F \in \mathcal{F}(s)} \rho_F(\mathbf{z}). \qquad (4)$$

and finally, the finite state compressibility of the infinite sequence $\mathbf{z} = \mathbf{z}_1^{\infty}$ is defined as

$$\rho(\mathbf{z}) = \lim_{s \to \infty} \limsup_{n \to \infty} \rho_{\mathcal{F}(s)}(\mathbf{z}_1^n). \qquad (5)$$

Note that the order of limits is critical for this definition, since if the number of states is taken to infinity first, any sequence can be compressed to 1 bit by having the state machine "remember" and identify the particular sequence. The outer limit exists, since $\rho_{\mathcal{F}(s)}$ is nonincreasing in s and bounded from below.

Proof of Theorem 1: Define $\tilde{\mathbf{Z}}_{b,k}$ as the random vector of length k formed by selecting one vector from the set of b vectors $(\mathbf{z}_i^{[k]})_{i=1}^b$, with uniform probability of $\frac{1}{b}$ for each. In other words, the probability distribution of $\tilde{\mathbf{Z}}_{b,k}$ equals the empirical distribution of the first b blocks of length k in $\mathbf{z}$. Similarly,

define the random variables $\tilde{\mathbf{X}}_{b,k}$ and $\tilde{\mathbf{Y}}_{b,k}$ derived from the sequences $\mathbf{x}, \mathbf{y}$.

Suppose a given E, D achieve rate R and average error probability ϵ over b blocks of size k. This is equivalent to saying they achieve error probability ϵ when operating on the stochastic channel $\tilde{\mathbf{Y}}_{b,k} = \tilde{\mathbf{X}}_{b,k} + \tilde{\mathbf{Z}}_{b,k}$ (see Fig. 5). Therefore, the standard converse of the channel capacity theorem implies that the rate R can be bounded by $R \overset{\approx}{\leq} \log |\mathcal{X}| - \frac{1}{k} H(\tilde{\mathbf{Z}}_{b,k})$. Then, the limit of $\frac{1}{k} H(\tilde{\mathbf{Z}}_{b,k})$ is related to the finite state compressibility $\rho(\mathbf{z})$. The later relation is a variation of a result by Lempel and Ziv [6, Th. 3] on the convergence of the sliding-window empirical entropy measured over increasing block lengths to the finite state compressibility, whereas here the blockwise empirical entropy is used instead. The full proof is given in Appendix A.

Note that the upper bound of Theorem 1 can sometimes be strict, i.e., there are examples of sequences $\mathbf{z}$ for which $C_{\text{IFB}} < (1 - \rho(\mathbf{z})) \log |\mathcal{X}|$, as shown in the following example. We do not have an expression for the IFB capacity.

Example 4: Consider for the binary additive channel, the sequence $\mathbf{z}$ which consists of blocks with ever increasing size. The first half of each block is 0, and the second half block is chosen randomly $Z_i \sim Ber(\frac{1}{2})$. With high probability, the finite state compressibility of the sequence is $\frac{1}{2}$ (which can be attained, for example, by block-to-variable encoding, using 1 bit to denote the sequence of zeros). However, the IFB capacity of the channel is 0 with high probability, since for any encoder and decoder with large block size, approximately half of the blocks will be received in error. Therefore, there exist sequences for which the inequality is strict.

B. Universally Attaining the IFB Capacity Over the Modulo-Additive Channel

In this section, a universal system for the modulo-additive channel with an unknown state sequence is presented. It is first shown (see also [13, Sec. 10.5], [14]) that for a wide range of sequential source encoders, there is a communication scheme that asymptotically attains the rate $\log |\mathcal{X}| - \frac{1}{n} L(\mathbf{z}^n)$, where $L(\mathbf{z}^n)$ is the compression length of the n-length sequence $\mathbf{z}$ by the source encoder, i.e., the number of bits used to encode the sequence.

Let us first define a class of sequential source encoders, for which Theorem 2 below applies. Only source encoders that have the following structure are considered: the encoding algorithm is unaware and is not a function of the sequence length n. The

encoder receives the sequence for compression letter by letter. From time to time, the encoder emits parts of the compressed sequence. After the last letter is entered into the source encoder, it receives an indication that the sequence has ended and may emit the final part of the compressed sequence.

For each sequence $\mathbf{z}$ of any given length, define $L_S(\mathbf{z})$ as the unterminated coding length of the sequence, i.e., the length of the output of the encoder after the input $\mathbf{z}$ has been fed, but the sequence has not been terminated, i.e., the encoder did not receive an indication that the sequence ended and is expecting additional input. Define $L_T(\mathbf{z}) = L(\mathbf{z})$ as the terminated coding length, i.e., the length of the output after the encoder received the termination indication. The sequence $\mathbf{z}$ is uniquely decodable from the $L_T(\mathbf{z})$ bits of the terminated code, but not necessarily from the $L_S(\mathbf{z})$ bits of the unterminated one. The difference $L_T(\mathbf{z}) - L_S(\mathbf{z}) \geq 0$ is the information stored in the encoder which has not been output yet. The class of source encoders is defined by the two assumptions as follows:

(A) The difference between the terminated and unterminated lengths is bounded by an asymptotically negligible value: $\frac{1}{n}(L_T(\mathbf{z}) - L_S(\mathbf{z})) \leq \frac{1}{n}\Delta_L(n) \xrightarrow[n \to \infty]{} 0$
This can be considered an embodiment of the limitation to "sequential" encoders and precludes encoders that process the entire sequence before producing outputs.

(B) The encoding length does not decrease when the sequence is extended: $L_T(z_1^i) \geq L_T(z_1^{i-1})$. This is a technical requirement intended to simplify the analysis.

Theorem 2: Given a sequential source coding scheme with input symbols from alphabet $\mathcal{X}$ that satisfies assumptions (A), (B), and assigns a codeword length of $L(\mathbf{z})$ to the sequence $\mathbf{z} \in \mathcal{X}^n$, then for any $\epsilon > 0$ there exists a sequence of adaptive-rate encoders and decoders using common randomness and feedback, for increasing block lengths n over the channel $\mathbf{y} = \mathbf{x} + \mathbf{z}$ ($\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}^n$), in which for any individual noise sequence $\mathbf{z}$ with probability at least $1 - \epsilon$, the message is correctly decoded with rate of at least

$$R \geq R_{\mathrm{emp}}(\mathbf{z}) - \delta_n, \qquad (6)$$

where

$$R_{\mathrm{emp}}(\mathbf{z}) = \log|\mathcal{X}| - \frac{1}{n}L(\mathbf{z}), \qquad (7)$$

$$\delta_n = 3\sqrt{\frac{\log|\mathcal{X}|}{n} \cdot \left[\log\left(\frac{n \cdot |\mathcal{X}|}{\epsilon}\right) + \Delta_L^{\max}(n)\right]} \xrightarrow[n \to \infty]{} 0, \qquad (8)$$

and $\Delta_L^{\max}(n) = \max\{\Delta_L(i)\}_{i=1}^n$.

The communication scheme and the proof of Theorem 2 appear in Appendix B (see also the proof outline in Section III). As shown in Appendix C, both assumptions are satisfied by Lempel–Ziv algorithms LZ77 [15] and LZ78 [6]. Note the similarity between the rate expression (7) and the capacity of an ergodic stochastic modulo-additive channel, which is also attained with a uniform prior, $C = \overline{I}(X^\infty; Y^\infty) = \overline{H}(Y^\infty) - \overline{H}(Y^\infty|X^\infty) = \log|\mathcal{X}| - \overline{H}(Z^\infty)$, where $\overline{H}, \overline{I}$ denote entropy rate and mutual information rate, respectively. $\frac{1}{n}L(\mathbf{z})$ can be considered a generalized empirical measure of

the noise entropy rate. In this sense, Theorem 2 is a generalization of Shayevitz and Feder's result [1].

Substituting the compression length of Lempel and Ziv's LZ78 algorithm, the finite state compressibility is obtained. This yields the following theorem.

Theorem 3: When the system of Theorem 2 is used in conjunction with LZ78 source encoder, over the modulo additive channel, then the following holds: for every infinite noise sequence $\mathbf{z}^\infty$ and every $\epsilon, \delta > 0$ there is n large enough so 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 $(1 - \rho(\mathbf{z}))\log|\mathcal{X}| - \delta$.

Corollary 3.1: The system defined above is IFB-universal.

Corollary 3.2: The system attains the Shannon capacity of every modulo-additive channel with a stationary ergodic noise sequence.

The proof of the theorem and its corollaries is given in Appendix C, and its main point is to show that LZ78 satisfies the assumptions of Theorem 2.

Theorems 2 and 3 are finite horizon, i.e., the system is designed for a given transmission length n, and because n needs to grow for the overhead δ to vanish, the asymptotic universality is obtained by a series of systems rather than a single one, as is standard in information theory. However, it is possible to design horizon-free systems in which the transmission length is not limited and redundancy vanishes with time [13, Sec. 8.6].

The results of this section rely on LZ compression algorithm and stress the relations between channel coding rates and compression ratios, and between IFB capacity and finite state compressibility. This relation is intuitively appealing and the resulting system is relatively simple. On the other hand, the modified universal system presented in the next section yields better bounds on the convergence of the overhead terms, which also hold *uniformly* in $\mathbf{z}$.[1]

VI. REDUNDANCY OF THE UNIVERSAL SYSTEM

Let us now consider the redundancy of the universal system and how fast it converges to zero as the block length increases, under the context of the modulo-additive channel. The interesting question is how large the transmission length size n needs to be, in order to successfully compete with an IFB system of a given block size k. Unfortunately, n must grow at least as fast as $|\mathcal{X}|^k$, approximately. Thus, even considering reference systems of relatively small block sizes compared to standard block codes, for instance $k = 100$, the competition becomes infeasible.

A. Definition of Redundancy

Before giving a definition of the redundancy, some considerations for the definition are provided. The finite state compressibility $\rho(\mathbf{z})$ of the infinite sequence $\mathbf{z}$, used in Theorems 1 and 3 is irrelevant for the analysis of convergence. This is because

---

[1]Notice that because Theorem 3 essentially indicates convergence to the IFB *capacity*, the convergence cannot hold uniformly in $\mathbf{z}$, as the IFB capacity may be obtained by competing systems of ever growing complexity, depending on the noise sequence. In the next section, the IFB system and the universal system are compared directly for finite n, k without referring to the asymptotic value of the IFB capacity, thereby making uniform convergence possible.

$\rho(\mathbf{z})$ is an asymptotical value, and the performance of the best block encoder or FSM encoder on any finite block of n symbols, does not indicate anything about the final finite state compressibility. In other words, there is no guarantee on the rate of convergence of the $\limsup$ in (5). Consider as example a sequence $\mathbf{z}$ which is incompressible up to time n_1 and then all zero to infinity, or vice versa. Note that incompressible sequences must exist, by Kraft's inequality. Therefore, instead of considering the convergence of the rates obtained by the best IFB system and the universal system to $(1 - \rho) \log |\mathcal{X}|$, the comparison is between the rate obtained by the best IFB system of block size k, with a universal system, at time n.

While the asymptotic results of Theorems 1 and 3 require the error probability of both systems to tend to zero with n, at a finite block length, a certain nonzero error probability would exist. In the two systems, error probabilities have different meanings: the IFB system's error probability is blockwise and the universal system's error probability is measured on the entire transmission. Therefore, for a fair comparison, and in order to remove the dependence on the error probability from the results, let us consider the following definition of an effective rate, for a system operating over block of size k with rate R and error probability ϵ:

$$R^* = (1 - \epsilon)R - \frac{1}{k}h_b(\epsilon). \qquad (9)$$

This definition is motivated by Fano's inequality; see for example (45). While the first factor is usually termed the good-put, i.e., the number of error free bits, the second factor compensates for the uncertainty in knowing whether there is an error or not. For example, a system delivering $R = 1$ bit per channel use with error probability $\epsilon = \frac{1}{2}$ per block of size $k = 1$, i.e., transmits no information, would have $(1 - \epsilon)R = \frac{1}{2}$ but $R^* = 0$. Equivalently, R^* may be interpreted as a bound on the normalized mutual information between the input message and the decoded message, given the parameters R, ϵ, and k. Notice that $R^* \xrightarrow[\epsilon \to 0]{} R$. Regardless of the interpretation of R^*, the results below yield meaningful bounds on the actual rates R by referring to R^*.

Another issue is how to compare a universal system with transmission length n and an IFB system whose block length k does not divide n. For a worst-case comparison, let us give the IFB system the luxury of using the last block that possibly extends beyond the nth symbol, i.e., $l = \lceil \frac{n}{k} \rceil$ blocks overall, while letting the noise sequence on these symbols $\mathbf{z}_{n+1}^{kl}$ take the values which are best for the IFB system.

A definition of the minimax redundancy is given below. Let E, D define an IFB system with block length k and rate R_{IFB} (Definition 2), which is iteratively mapped to the channel $P_{Y|X}^{(\theta)}, \theta \in \Theta$, over kl symbols, where $l = \lceil \frac{n}{k} \rceil$, and yield average error probability ϵ_{IFB} (Definition 3). Similarly, on the same channel over n symbols, an adaptive system U with feedback and common randomness (Section IV-D), whose design must not depend on θ, guarantees a rate of at least $R_{\text{U}} = R_{\text{U}}(\theta)$ with an error probability of at most ϵ_{U}. As in Definition 6, ϵ_{U} includes both the probability of error and the probability that the system's rate falls below R_{U}. While R_{U}

is allowed to depend on the channel index θ, ϵ_{U} is required to be fixed. Let $R_{\text{IFB}}^* = R_{\text{IFB}} \cdot (1 - \epsilon_{\text{IFB}}) - \frac{1}{k}h_b(\epsilon_{\text{IFB}})$ and $R_{\text{U}}^* = R_{\text{U}} \cdot (1 - \epsilon_{\text{U}}) - \frac{1}{n}h_b(\epsilon_{\text{U}})$. The rate and error probability for each system are defined given the channel and the system. The values related to the IFB system, $R_{\text{IFB}}, \epsilon_{\text{IFB}}$, and R_{IFB}^* depend implicitly on (n, k, E, D, θ), while the values related to the rate adaptive system, R_{U} and R_{U}^* depend implicitly on (n, U, θ).

The minimax redundancy for finite n, k is defined as follows:

$$\Delta^*(n, k) = \min_U \max_{\theta \in \Theta} \left[\max_{E, D} (R_{\text{IFB}}^*) - R_{\text{U}}^* \right]. \qquad (10)$$

In other words, it is the minimal gap $R_{\text{IFB}}^* - R_{\text{U}}^*$ that can be universally guaranteed by a single system U over all channels. Note that the definition allows the universal system to depend on k but this relaxation is not used by the universal system achieving the bounds below. For the special case of the modulo additive channel, the channel index θ is replaced by the noise sequence $\mathbf{z}_1^{kl}$.

B. Minimax Redundancy for the Modulo-Additive Channel Class

The minimax redundancy of a universal system compared to the IFB system over the modulo-additive channel is bounded below. Let us begin with the main asymptotical result which formalizes the notion that the minimum transmission length behaves asymptotically like $|\mathcal{X}|^k$.

Theorem 4: For a given k and $\delta > 0$, let $n^* = n^*(k, \delta)$ be the minimum n such that for the modulo additive channel, $\Delta^*(n, k) \le \delta \log |\mathcal{X}|$, then

$$\lim_{\delta \to 0} \lim_{k \to \infty} \frac{\log n^*(k, \delta)}{k \log |\mathcal{X}|} = 1. \qquad (11)$$

Theorem 4 is an immediate consequence of the explicit bounds given in the remainder of this section. Theorem 5 below specifies bounds on $\Delta^*(n, k)$, and its Corollary 5.1 specifies bounds on the minimum transmission length n^* defined above.

Theorem 5: The minimax redundancy (10) for the channel $\mathbf{y} = \mathbf{x} + \mathbf{z}$ $(\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}^n)$ satisfies

$$\Delta_- \le \Delta^*(n, k) \le \Delta_+ \qquad (12)$$

where

$$\Delta_- = \begin{cases} \left\lfloor \log (k\tau) \frac{1}{\log |\mathcal{X}|} \right\rfloor \frac{\log |\mathcal{X}|}{2k} & \tau > \frac{|\mathcal{X}|}{k} \\ \frac{\log |\mathcal{X}|}{2|\mathcal{X}|} \cdot \tau & \tau \le \frac{|\mathcal{X}|}{k} \end{cases} \qquad (13)$$

and for $\tau \le 1$:

$$\Delta_+ = \frac{\tau}{2} \log \left(\frac{1}{\tau} \right) + \left(\frac{k}{4}\tau^2 + \tau \right) \log e + \delta_n^* + \frac{k}{n} \log(e|\mathcal{X}|). \qquad (14)$$

The parameters are defined as follows:

$$\tau = \frac{|\mathcal{X}|^k}{n}$$

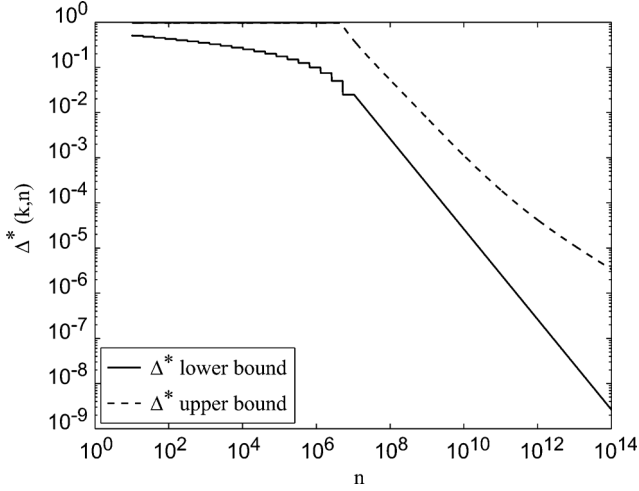$$\delta_n^* = 4\sqrt{\frac{\log |\mathcal{X}| \cdot \log (n^2|\mathcal{X}|)}{n}}. \qquad (15)$$

Fig. 6. Upper and lower bound on the redundancy $\Delta^*(n,k)$ of universal systems given by Theorem 5 for $k = 20, |\mathcal{X}| = 2$.

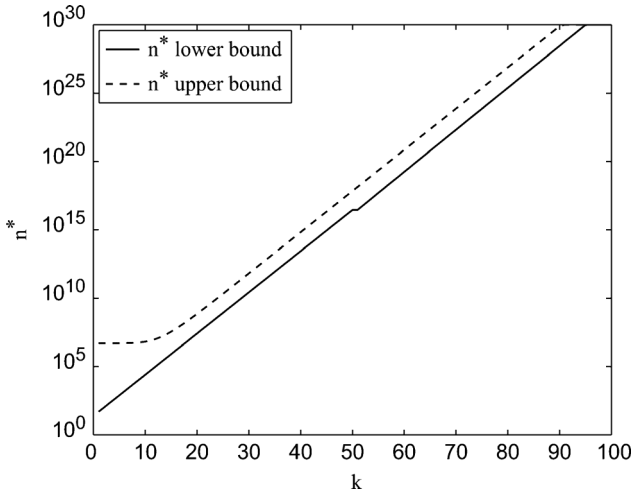Fig. 7. Minimum transmission length n required to obtain a minimax redundancy $\Delta^*(n,k) \leq \delta \cdot \log |\mathcal{X}|$, according to the bounds of Theorem 5 as function of the IFB block size k, for $|\mathcal{X}| = 2, \delta = 0.01$.

Furthermore, the universal system attaining the upper bound Δ_+ does not depend on k.

The theorem is proven in the next section. Note that both bounds require τ to be small, and thus n to be large, in order to achieve a small redundancy. While the lower bound is linear for $\tau \leq \frac{|\mathcal{X}|}{k}$, for large values, it increases significantly more slowly, like $\log \tau$. This is because of the inefficiency of the IFB system used in the lower bound, at high rates. The value of Δ_- in the range $\tau \leq \frac{|\mathcal{X}|}{k}$ is limited to $\frac{\log |\mathcal{X}|}{2k}$, i.e., a rate offset of half a symbol per block. The bound for the range $\tau > \frac{|\mathcal{X}|}{k}$ is useful, in showing that even if one is satisfied with a redundancy of more than $\frac{\log |\mathcal{X}|}{2k}$, τ must be kept small. Fig. 6 illustrates the bounds of Theorem 5 as function of the transmission length n, for a constant value of k. The logarithmic and quantized behavior of the lower bound for small values of n can be observed. Fig. 7 presents $n^*(k, \delta)$, i.e., the minimum n required to obtain $\Delta^*(n,k) \leq \delta \cdot \log |\mathcal{X}|$, according to the bounds of Theorem 5, as a function of k. The gap between the upper and lower bounds is significant: a little more than an order of magnitude.

However, their trend is similar. This observation is formalized by Corollary 5.1 below, concerning the asymptotical behavior of n.

Corollary 5.1:

$$\frac{k}{|\mathcal{X}|} \cdot |\mathcal{X}|^{(1-2\delta)k} \leq n^* \leq \frac{k}{\min\left[T(k,\delta,|\mathcal{X}|),1\right]} \cdot |\mathcal{X}|^k, \quad (16)$$

where $g(\tau) = \tau \log\left(\frac{1}{\tau}\right)$ and

$$T(k,\delta,|\mathcal{X}|) = k \cdot g^{-1}\left(\frac{1}{3}(\delta - 12 \cdot |\mathcal{X}|^{-k/2}) \cdot \log|\mathcal{X}|\right). \quad (17)$$

For large k and fixed δ, $k^{-1} \cdot T(k,\delta,|\mathcal{X}|) \xrightarrow[k\to\infty]{} g^{-1}\left(\frac{1}{3}\delta \cdot \log|\mathcal{X}|\right) = \text{const}$, and thus for large enough k, $T > 1$ and does not dominate the upper bound (16). For a small value of δ both bounds of Corollary 5.1 behave approximately like $|\mathcal{X}|^k$. Corollary 5.1 results from a technical simplification of the bounds of Theorem 5 and is proven in Appendix D. Most important is the lower bound on n^* which indicates the minimum rate at which n^* must grow. Finally, Theorem 4 is an immediate consequence of Corollary 5.1.

Note that the system attaining the upper bound of Theorem 5 yields a stronger type of universality than claimed in Theorem 3, because for each value of n, the overheads are uniformly bounded for any noise sequence $\mathbf{z}$, whereas previously, while the overheads are guaranteed to tend to zero asymptotically with n, this convergence is not necessarily uniform with respect to $\mathbf{z}$.

Unlike other results in this paper where the IFB system is used merely as a converse, in the proof for the lower bound Δ_-, it is required to devise a specific IFB system. Here, the simplicity of the IFB system, which makes the other results intuitive and simple to derive, complicates the proof. The collapsed channel capacity, which upper bounds the IFB system rate, is usually not achievable by a finite block encoder, and a specific channel has to be devised in order for the IFB system to operate provably better than any universal system. It seems that richer classes of reference systems, e.g., systems using feedback as considered in [11], may result in simpler and tighter lower bounds.

C. Proof of Theorem 5

1) Lower Bound (Reverse Part): In order to show that the redundancy must be at least $O\left(\frac{|\mathcal{X}|^k}{n}\right)$ an example random channel is constructed, in the following way. First, the encoder E is defined. Then, a way to generate noise sequences $\mathbf{z}$ is defined, such that the noise sequences belong to a subset of all possible sequences $\mathbf{z} \in \mathbb{Z}_d$, and it is possible to decode the given code with zero error probability for any noise sequence in the set. The IFB decoder D is specified only after the noise sequence has been chosen. The sequence $\mathbf{z}$ is drawn in a randomized way, thus creating a stochastic "test" channel. It is shown that there exists a noise sequence for which the rate of the universal system is bounded by the normalized mutual information over the test channel. Asymptotically, as there are certain constraints on the choice of the noise sequence, this normalized mutual information tends to the rate of the IFB encoder. However, at the beginning of the sequence, the entropy of the sequence is a little higher than its long-term average, and thus the mutual information is a little lower than
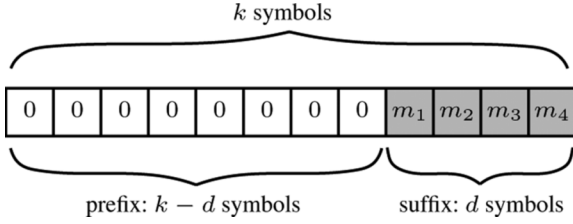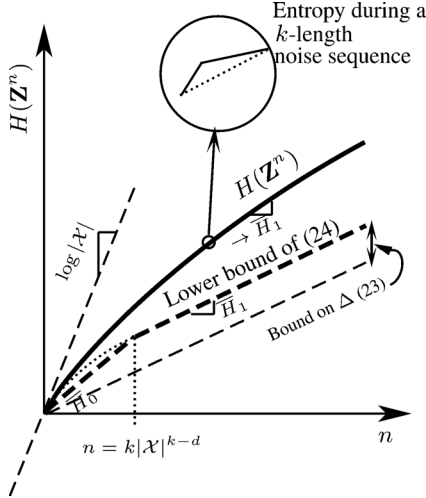
Fig. 8.   Reference encoder for the converse of Theorem 5.

Fig. 9.   Entropy of the noise in the test channel $H(\mathbf{Z}^n)$ over time, and the lower bounds of (24), (80).

its asymptotic value, which equals the rate of the IFB encoder. Thus, the rate of the universal system is bounded by a value lower than the rate of the IFB system.

Let us first describe the IFB encoder. The encoder sends d symbols from the alphabet $\mathcal{X}$ over k channel uses, and therefore has a rate

$$R_{\text{IFB}} = \frac{d}{k} \log |\mathcal{X}|. \tag{18}$$

The encoding is simple (see Fig. 8): the first $k - d$ symbols (prefix) are constant and the rest d symbols (suffix) contain the message. The decoder would be able to know the value of the noise sequence over the prefix symbols, and knows a list of all possible noise sequences. Assuming that there is no more than one noise sequence with any given prefix, then zero error probability is possible: the decoder finds the noise sequence from the prefix symbols, and cancels it on the suffix to find the message.

Next the test channel is defined. The set $\mathbb{Z}_d$ of allowed noise sequences are simply those sequences for which each prefix $\mathbf{z}_{k\cdot(i-1)+1}^{k\cdot(i-1)+k-d}$ $(i = 1, 2, \ldots)$ uniquely determines the respective suffix $\mathbf{z}_{k\cdot(i-1)+k-d+1}^{k\cdot(i-1)+k}$. The random noise sequence is generated as follows: at each block of k symbols, the prefix of $k - d$ symbols is chosen randomly, uniformly over all possible $|\mathcal{X}|^{k-d}$ prefixes, and independently of the past noise sequence. Then, if the prefix had appeared before, the suffix equals the suffix of

the noise sequence that already appeared. Otherwise, the suffix is chosen randomly, uniformly over all possible $|\mathcal{X}|^d$ suffixes.[2]

The choice of the first sequence $\mathbf{z}_1^{[k]}$ is uniform over all possible sequences, and therefore the entropy of the noise sequence in the first block is maximal, $\log(|\mathcal{X}|^k)$. The choice of the noise sequences narrows with time, and after a long while, all possible prefixes would have been chosen, with one noise sequence per prefix. In this case, the choice of the suffix is determined by the prefix, and the entropy per k-block is $\log(|\mathcal{X}|^{k-d})$. This is the minimum entropy per block attained. The behavior of the entropy $H(\mathbf{Z}^n)$ in this channel is shown in Fig. 9.

Now, because $\epsilon_{\text{IFB}} = 0$,

$$R_{\text{IFB}}^* = (1 - \epsilon_{\text{IFB}})R_{\text{IFB}} - \frac{1}{k}h_b(\epsilon_{\text{IFB}}) = R_{\text{IFB}} = \frac{d}{k} \log |\mathcal{X}|. \tag{19}$$

Therefore,

$$\Delta^*(n, k) \overset{(10)}{\geq} \min_U \max_{\mathbf{z}_1^n \in \mathbb{Z}_d} \left[\max_{E, D} (R_{\text{IFB}}^*(E, D)) - R_U^* \right]$$
$$\geq R_{\text{IFB}}^* - \max_U \min_{\mathbf{z}_1^n \in \mathbb{Z}_d} [R_U^*] \tag{20}$$

where R_{IFB}^* denotes the value defined in (19) for the specific reference system described.

The universal system guarantees error probability ϵ_U for any $\mathbf{z}$. By definition, for any $\mathbf{z}_1^n \in \mathbb{Z}_d$, $R_U \geq R_0 = \min_{\mathbf{z}_1^n \in \mathbb{Z}_d} [R_U]$. Therefore, if $\mathbf{z}_1^n$ is drawn randomly in $\mathbb{Z}_d$, then the universal system yields a rate of at least R_0, with error probability at most ϵ_U over the test channel, and can be converted to a fixed-rate system with feedback with rate R_0 over the same channel. Using Fano's inequality, which holds also in the case of feedback[3]

$$R_0(1 - \epsilon_U) - \frac{1}{n}h_b(\epsilon_U) \leq \frac{1}{n}I(\mathbf{X}^n; \mathbf{Y}^n) \leq \log |\mathcal{X}| - \frac{1}{n}H(\mathbf{Z}^n) \tag{21}$$

and therefore for any universal system U:

$$\min_{\mathbf{z}_1^n \in \mathbb{Z}_d} [R_U^*] = (1 - \epsilon_U) \underbrace{\min_{\mathbf{z}_1^n \in \mathbb{Z}_d} [R_U]}_{R_0} - \frac{1}{n}h_b(\epsilon_U)$$
$$\leq \log |\mathcal{X}| - \frac{1}{n}H(\mathbf{Z}^n) \tag{22}$$

which yields the bound

$$\Delta^* \overset{(20),(22)}{\geq} R_{\text{IFB}}^* - \log |\mathcal{X}| + \frac{1}{n}H(\mathbf{Z}^n)$$
$$\overset{(19)}{=} \frac{1}{n}H(\mathbf{Z}^n) - \frac{k-d}{k} \log |\mathcal{X}|. \tag{23}$$

Asymptotically, $\frac{1}{n}H(\mathbf{Z}^n) \xrightarrow[n \to \infty]{} \frac{k-d}{k} \log |\mathcal{X}|$, and thus the bound above tends to zero. The main point of the proof is to bound the convergence rate of $\frac{1}{n}H(\mathbf{Z}^n)$.

It may appear surprising, that while it will be shown that the mutual information over the channel is slightly lower than

---

[2]An alternative way of generating the noise sequence, which yields the maximum entropy, is by uniform drawing over the set of all possible k-length sequences that satisfy the unique prefix condition. However this complicates the bound.

[3]See (45) in the proof of Theorem 1, and (37).

R_{IFB}^*, the IFB system transmits rate R_{IFB}^* with zero error over this channel. This is explained by the fact that the decoder is designed knowing the specific noise sequence, and therefore its effective rate is not limited by the mutual information.

The next step is to bound $H(\mathbf{Z}^n)$. This rather technical derivation is deferred to Appendix E, where the following lemma is proven.

Lemma 1: For the distribution of $\mathbf{Z}^n$ of the test channel defined above, the entropy satisfies

$$H(\mathbf{Z}^n) \geq n \cdot \overline{H}_1 + \min(n, k|\mathcal{X}|^{k-d})(\overline{H}_0 - \overline{H}_1) \qquad (24)$$

where

$$\overline{H}_0 \triangleq \frac{k - d/2}{k} \log |\mathcal{X}|$$
$$\overline{H}_1 \triangleq \frac{k - d}{k} \log |\mathcal{X}| \qquad (25)$$

are the initial slope of the bound for small n, and the asymptotical entropy rate per symbol, respectively.

Substituting the above in (23) yields

$$\Delta^* \geq \frac{1}{n} H(\mathbf{Z}^n) - \frac{k - d}{k} \log |\mathcal{X}|$$
$$\overset{(24)}{\geq} \overline{H}_1 + \min\left(1, \frac{k|\mathcal{X}|^{k-d}}{n}\right)(\overline{H}_0 - \overline{H}_1) - \overline{H}_1 \quad (26)$$
$$\overset{(81)}{=} \min\left(1, \frac{k|\mathcal{X}|^{k-d}}{n}\right) \cdot \frac{d}{2k} \cdot \log |\mathcal{X}|.$$

The bound is true for every $d \in \{1, \ldots, k\}$. Let us find a value of d that approximately maximizes the bound for given n, k. Starting from $d = k$ and decreasing d, each decrease of 1 doubles the first term in the RHS of (26), as long as $\frac{k|\mathcal{X}|^{k-d}}{n} \leq 1$, and only linearly decreases the second term. Therefore, it is beneficial to decrease d as long as $\frac{k|\mathcal{X}|^{k-d}}{n} \leq 1$, and no more than one additional step. For simplicity, let us always take the additional step and determine d as the maximum $d \in \{1, \ldots, k\}$ so that $\frac{k|\mathcal{X}|^{k-d}}{n} \geq 1$, or $d = 1$ if no such d exists, i.e.,

$$d = \max\left(\left\lfloor \log\left(\frac{k|\mathcal{X}|^k}{n}\right) \frac{1}{\log |\mathcal{X}|} \right\rfloor, 1\right). \qquad (27)$$

If $n \geq k|\mathcal{X}|^{k-1}$, then $d = 1$, and $\min\left(1, \frac{k|\mathcal{X}|^{k-d}}{n}\right) = \frac{k|\mathcal{X}|^{k-1}}{n}$. In this case (26) yields

$$\Delta^* \geq \frac{1}{2} \frac{|\mathcal{X}|^{k-1}}{n} \cdot \log |\mathcal{X}|. \qquad (28)$$

Otherwise, $\min\left(1, \frac{k|\mathcal{X}|^{k-d}}{n}\right) = 1$, and (26) yields

$$\Delta^* \geq \frac{1}{2} \left\lfloor \log\left(\frac{k|\mathcal{X}|^k}{n}\right) \frac{1}{\log |\mathcal{X}|} \right\rfloor \frac{\log |\mathcal{X}|}{k} \qquad (29)$$

(28), (29) are represented in a compact form in (13) above. This proves the lower bound of Theorem 5. ∎

Upper Bound (Direct Part): The purpose is to show the existence of a universal system that attains a small redundancy with respect to the reference system, i.e., referring to (10), it is desired to show that there exists a universal system U such that

$$\max_{\mathbf{z}_1^n}\left[\max_{E,D}(R_{\text{IFB}}^*) - R_{\text{U}}^*\right] \leq \Delta_+. \qquad (30)$$

The desired result is similar to the one of Theorem 3, however to reach the desired overheads, a slightly different design of the universal system, and a more careful analysis of the overheads is required.

Following the same logic as the proof of Theorems 1, 3, the difference between the good-put of the two systems is bounded by the following steps:

(a) The relation between R_{U}^* and the ideal R_{emp} target of the rate adaptive system (i.e., the overhead term of Theorem 2).

(b) The relation between R_{emp} and the collapsed channel capacity, or equivalently the collapsed noise entropy $H(\tilde{\mathbf{Z}}_{l,k})$.

(c) The relation between R_{IFB}^* and $H(\tilde{\mathbf{Z}}_{l,k})$ obtained using Fano's inequality (as in the proof of Theorem 1).

Considering the scheme that was described for the achievability result of Theorems 2 and 3, the largest overhead is due to step (b). This large overhead is in some sense unavoidable, as the converse shows, however it is especially large due to the use of LZ78 algorithm which has a slow $O(1/\log n)$ convergence rate. Specifically, using [6, Th. 1.2], this term, i.e., the bound on $\frac{1}{n} L_{78}(\mathbf{z}^n) - \frac{1}{k} H(\tilde{\mathbf{Z}}_{b,k})$ behaves like $O\left(\frac{\log(|\mathcal{X}|^{2k})}{\log n}\right)$, i.e., in order for this term to be small, it is required that $n \gg |\mathcal{X}|^{2k}$, and any small improvement in the overhead requires an ever growing increase in n: improving the overhead by a factor of 2 requires squaring n.

To obtain a tighter bound, a more general result from [13] can be applied. Theorem 8.2 there shows that for every causal probability distribution $P(\mathbf{x}|\mathbf{y})$, i.e., satisfying for all $i \leq n$: $P(\mathbf{x}^i|\mathbf{y}^n) = P(\mathbf{x}^i|\mathbf{y}^i)$, the rate function $R_{\text{emp}} = \frac{1}{n} \log \frac{P(\mathbf{x}^n|\mathbf{y}^n)}{Q(\mathbf{x}^n)}$ is adaptively achievable with overhead of $\delta_n = 3\sqrt{\frac{\log q_{\min}^{-1} \cdot (\log \frac{n}{\epsilon_U} + \log q_{\min}^{-1})}{n}}$, where $q_{\min}$ is the minimum nonzero value of $Q(x_i|\mathbf{x}_{i-1})$.[4]

Substitute as Q the uniform distribution $Q(\mathbf{x}^i) = |\mathcal{X}|^{-i}$ having $q_{\min}^{-1} = |\mathcal{X}|$. Take $P(\mathbf{x}|\mathbf{y}) = P_Z(\mathbf{x} - \mathbf{y})$, for some probability distribution $P_Z(\mathbf{z})$. This choice satisfies the causality condition and yields

$$R_{\text{emp}} = \log |\mathcal{X}| + \frac{1}{n} \log P_Z(\mathbf{x}^n - \mathbf{y}^n), \qquad (31)$$

with $\delta_n = 3\sqrt{\frac{\log |\mathcal{X}| \cdot \log\left(\frac{n|\mathcal{X}|}{\epsilon_U}\right)}{n}}$.

Let us begin by analyzing the relation between R_{emp} and R_{U}^* in step (a) above. While the convergence of $\delta_n \underset{n \to \infty}{\longrightarrow} 0$ requires ϵ_U to decay subexponentially with n, the choice of ϵ_U will lead to a reduction of $\epsilon_U R_{\text{emp}} \leq \epsilon_U \log |\mathcal{X}|$ in rate. For simplicity, let us choose $\epsilon_U = \frac{1}{n}$ as this factor is insignificant. In other

---

[4] Substituting $d_{\text{FB}} = 1$, $D = 0$ in the parameters of the theorem.

words, there exists a system with $\epsilon_U = \frac{1}{n}$ which with probability $1 - \epsilon_U$ transmits a rate $R_{\mathrm{emp}} - \delta_n$ without error over the channel. Therefore,

$$
\begin{aligned}
R_U^* &= R_U(1 - \epsilon_U) - \frac{1}{n} h_b(\epsilon_U) \\
&\geq (R_{\mathrm{emp}} - \delta_n)(1 - \epsilon_U) - \frac{1}{n} h_b(\epsilon_U) \\
&\overset{R_{\mathrm{emp}} \leq \log|\mathcal{X}|,(31)}{\geq} R_{\mathrm{emp}} - \delta_n - \epsilon_U \log|\mathcal{X}| - \frac{1}{n} h_b(\epsilon_U) \\
&\geq R_{\mathrm{emp}} - 3\sqrt{\frac{\log|\mathcal{X}| \cdot \log(n^2|\mathcal{X}|)}{n}} - \frac{1}{n}\log|\mathcal{X}| - \frac{1}{n} \\
&\geq R_{\mathrm{emp}} - \underbrace{4\sqrt{\frac{\log|\mathcal{X}| \cdot \log(n^2|\mathcal{X}|)}{n}}}_{\triangleq \delta_n^*} \qquad (32)
\end{aligned}
$$

where in the last step, for simplification of the bound, it was assumed that $n \geq \log|\mathcal{X}|$—otherwise, δ_n is large. Equation (32) yields the desired relation for step (a) above. Next, the relation between R_{emp} and the collapsed noise sequence entropy in step (b) is considered.

If one is interested in competing with an IFB system with block length k, it would make sense to treat each k symbols of the noise sequence as a single supersymbol, and take as P_Z the universal distribution defined by Krichevsky and Trofimov [12] over these supersymbols. This distribution is universal in the sense that up to a small overhead, $-\frac{1}{n}\log P_Z(\mathbf{z}) \approx \hat{H}(\mathbf{z})$, i.e., the probability matches the empirical entropy of the sequence, which in the current case is $H(\tilde{\mathbf{Z}}_{b,k})$. Furthermore, this holds with a redundancy close to the minimum possible. It is possible to construct a universal distribution P_Z that compares well with all distributions over the n symbols which are i.i.d. over k-length blocks, by a weighted average of Krichevsky–Trofimov distributions.

Let $\pi_k(\mathbf{z}^k)$ denote a distribution over the k-letter $\mathbf{z}^k$, where k is not assumed to divide n. This defines also a distribution on the partial sequence of length $i < k$ by taking the marginal $\pi_k(\mathbf{z}^i) = \sum_{\mathbf{z}_{i+1}^k} \pi_k(\mathbf{z}^k)$. The distribution over n length vectors, associated with π_k is defined as the i.i.d. extension of π_k, where the marginal distribution is used for the remainder that does not divide by k. This n-length distribution will be denoted by the same symbol

$$
\pi_k(\mathbf{z}) \triangleq \prod_{i=1}^{\lfloor n/k \rfloor} \pi_k(\mathbf{z}_{(i-1)k+1}^{(i-1)k+k}) \cdot \pi_k(\mathbf{z}_{\lfloor n/k \rfloor k+1}^n). \qquad (33)
$$

Then, by weighting Krichevsky–Trofimov distributions it is possible to obtain the following result.

Lemma 2: There exists a distribution $P_Z(\mathbf{z}^n)$, $\mathbf{z}^n \in \mathcal{X}^n$, such that for all k for which $\tau \triangleq \frac{|\mathcal{X}|^k}{n} \leq 1$:

$$
\forall \pi_k : \frac{1}{n}\log \pi_k(\mathbf{z}^n) \leq \frac{1}{n}\log P_Z(\mathbf{z}^n) + \Delta_\pi(k,n) \qquad (34)
$$

where

$$
\Delta_\pi = \frac{\tau}{2}\log\left(\frac{1}{\tau}\right) + \left(\frac{k}{4}\tau^2 + \tau + \frac{k}{n}\right)\log e. \qquad (35)
$$

The detailed derivation and proof appears in Appendix F. The next stage is to relate $\pi_k(\mathbf{z}^n)$ to $H(\tilde{\mathbf{Z}}_{l,k})$. Let $\mathbf{z}_i^{[k]} \triangleq \mathbf{z}_{(k-1)i+1}^{(k-1)i+k}$

be the ith k block of $\mathbf{z}^n$. Recall that $l = \lceil \frac{n}{k} \rceil$ is the number of k-blocks that cover the n symbols, and $\tilde{\mathbf{Z}}_{l,k}$ is a random variable generated by uniform selection out of $\mathbf{z}_1^{[k]}, \ldots, \mathbf{z}_l^{[k]}$. Let $P_{\tilde{\mathbf{Z}}_{l,k}}$ be the distribution of $\tilde{\mathbf{Z}}_{l,k}$ which is the empirical distribution of $\mathbf{z}_1^{[k]}, \ldots, \mathbf{z}_l^{[k]}$:

$$
\begin{aligned}
H(\tilde{\mathbf{Z}}_{l,k}) &\triangleq -\sum_{\mathbf{a} \in \mathcal{X}^k} P_{\tilde{\mathbf{Z}}_{l,k}}(\mathbf{a}) \log P_{\tilde{\mathbf{Z}}_{l,k}}(\mathbf{a}) \\
&= -\frac{1}{l}\sum_{i=1}^l \log P_{\tilde{\mathbf{Z}}_{l,k}}(\mathbf{z}_i^{[k]}) \\
&\overset{(a)}{=} -\frac{1}{l}\max_\pi \log \pi_k(\mathbf{z}_1^{k \cdot l}) \qquad (36) \\
&\overset{(b)}{\geq} -\frac{1}{l}\max_\pi \log \pi_k(\mathbf{z}_1^n) \\
&\overset{(34)}{\geq} -\frac{1}{l}\left(\log P_Z(\mathbf{z}) + n\Delta_\pi(k,n)\right)
\end{aligned}
$$

where (a) is because the empirical distribution maximizes the joint distribution of the vector; the expression following (a), where the maximization is over all k-letter distributions π, could be considered an alternative definition of $H(\tilde{\mathbf{Z}}_{l,k})$ (see [13, Sec. 9.14]). Transition (b) holds because extending the vector reduces its probability; see also the definition of $\pi_k(\mathbf{z}^j)$ (33).

Finally, in step (c), let us use Fano's inequality (see (45) in the proof of Theorem 1):

$$
R_{\mathrm{IFB}}^* = R_{\mathrm{IFB}}(1 - \epsilon_{\mathrm{IFB}}) - \frac{1}{k} h_b(\epsilon_{\mathrm{IFB}}) \leq \log|\mathcal{X}| - \frac{1}{k} H(\tilde{\mathbf{Z}}_{l,k}). \qquad (37)
$$

Combining the above yields

$$
\begin{aligned}
R_{\mathrm{IFB}}^* &\overset{(37)}{\leq} \log|\mathcal{X}| - \frac{1}{k}H(\tilde{\mathbf{Z}}_{l,k}) \\
&\overset{(36)}{\leq} \log|\mathcal{X}| + \frac{1}{kl}\left(\log P_Z(\mathbf{z}) + n\Delta_\pi(k,n)\right) \\
&\leq \frac{n}{kl}\left(\log|\mathcal{X}| + \frac{1}{n}\log P_Z(\mathbf{z})\right) \\
&\quad + \frac{kl-n}{kl}\log|\mathcal{X}| + \Delta_\pi(k,n) \qquad (38) \\
&\overset{(31)}{\leq} \frac{n}{kl}R_{\mathrm{emp}} + \frac{k}{n}\log|\mathcal{X}| + \Delta_\pi(k,n) \\
&\overset{(32)}{\leq} \frac{n}{kl}\left(R_U^* + \delta_n^*\right) + \frac{k}{n}\log|\mathcal{X}| + \Delta_\pi(k,n) \\
&\leq R_U^* + \delta_n^* + \frac{k}{n}\log|\mathcal{X}| + \Delta_\pi(k,n).
\end{aligned}
$$

Since this holds for any noise sequence and any pair E, D,

$$
\Delta^*(n,k) \overset{(10)}{\leq} \max_{\mathbf{z},E,D}\left(R_{\mathrm{IFB}}^* - R_U^*\right) \overset{(37)}{\leq} \Delta_\pi(k,n) + \delta_n^* + \frac{k}{n}\log|\mathcal{X}|. \qquad (39)
$$

This proves the upper bound of Theorem 5. ∎

VII. DISCUSSION AND EXTENSIONS

The model presented in this paper supplies the first definition of a "universal communication system," and the results indicate that such universal communication with feedback is possible in the nontrivial example of the modulo additive channel with an individual state sequence.

A. Alternative Definitions of Universality

The IFB comparison class was chosen as perhaps the simplest and most intuitive comparison class for universal communication. However, it has several drawbacks:

1) The reference system is limited in terms of complexity, feedback, etc.
2) On the other hand, universality is only achieved at ultra-high values of the transmission length n. Similar issues exist with Lempel–Ziv universal source coding.
3) The definition motivates learning k-periodic structures in the channel, which is counterintuitive. This may be solved e.g., by starting the reference system at an arbitrary time rather than at time 1, or by using structures that are not periodic such as FSMs [11].
4) While the IFB capacity is limited by the "collapsed channel capacity," it usually falls short of it. Furthermore, had the channel been a stochastic memoryless one, a rather large block size would be needed for the IFB system in order to yield a small error probability. A possible solution is to define the collapsed channel capacity itself as a target rate, but it is not clear how this should be defined for channels with memory.

1) Possible Enhancements of the IFB Class: Since the reference system enjoys the advantage of being designed for the specific noise sequence, this advantage is compensated by imposing some restrictions on the reference system, which are not imposed on the universal system. This is similar to what is done in universal source coding and universal prediction, when the comparison class is too rich. The definition of C_{IFB} limits the reference system in several factors, where the universal system is not restricted. Namely its complexity, the use of feedback, common randomness, and rate adaptivity. Relaxing any of these factors, may generate a higher value of the target rate as an alternative to C_{IFB}, which may still be universally attainable.

Some potential variations are given below:

(a) *Randomness*: allowing the reference system the use of common randomness.
(b) *Rate adaptivity*: allowing rate adaptivity in various levels. Error detection and *automatic repeat request* (ARQ) can be considered a very basic level of adaptivity.
(c) *Complexity*: definition of the encoder/decoder as FSMs rather than block encoders/decoders.
(d) *Feedback*: allowing the use of (a possibly limited amount of) feedback for the reference system.

The first two extensions (a), (b) are trivial, and were not pursued here in order to simplify the presentation. Misra and Weissman presented in [11] a class of FSM encoders and decoders with feedback, termed the *FS class*, that includes all the enhancements above, and had shown that for the modulo-additive channel, the maximum rate achieved by the reference class is at most $R = (1 - \rho(\mathbf{z})) \log |\mathcal{X}|$, so the current result on universality would hold also with respect to this enhanced class. Furthermore, they show that, unlike the IFB class (Section IV-E), the FS class achieves the rate R when the complexity is allowed to grow. Notwithstanding these results, the IFB class is still of interest due to its simplicity, which allows simple analysis and consideration of more complex channel models [9].

Below, these extensions are briefly discussed. Although Misra and Weissman already extended the results in the context of the modulo-additive channel, it is interesting to consider these extensions for more general channel models.

Common randomness: Allowing the reference system the use of common randomness does not change the results, as long as the common randomness is independent of the noise sequence and/or the block number. This is because the IFB capacity would still be upper bounded by the collapsed channel capacity. This holds also for channels with fading memory [9], where the collapsed channel capacity is used as a bound for the IFB rate.

Rate adaptivity: The IFB system may be allowed to choose the transmission rate adaptively at the decoder. A simple form of rate adaptivity is error detection, i.e., the decoder is allowed to choose between rate R and rate 0. In the later case, decoding errors are ignored. On the other hand, the IFB rate is defined in an effective way, considering how many blocks were actually decoded. Under suitable definitions, the effective rate of the IFB system would still be bounded by the collapsed channel capacity, so the results easily extend. Note that allowing error detection effectively models a block coding system using *automatic repeat request* (ARQ). When rate adaptation is considered, for a fair comparison, the decision on the rate must be made at the decoder based on the received sequence alone, rather than be given to the decoder.

Complexity: In order to achieve competitively universal communication, it is essential that *both* the reference encoder and the decoder be limited in some way, assuming they are designed knowing the channel. Consider, for example, the modulo additive channel. If the encoder is not limited, then it can transmit data at the maximum rate $\log |\mathcal{X}|$ bits/channel use, by uncoded transmission and subtraction of the noise sequence at the encoder. In this case, the decoder does nothing essentially, so restrictions on the decoder will not be helpful. Conversely, if the decoder is not limited, the encoder can transmit the message uncoded and the noise sequence can be canceled at the decoder, so limitations on the encoder would not help. As mentioned, an extension to FSMs with feedback (FS-class) has already been shown [11]. An interesting issue for further study is the universality with respect to the FS-class in general channel models.

Feedback: Several types of feedback may be considered:

1) Feedback inside the block, i.e., where the state is reset from block to block. Because the collapsed channel is a channel with memory, feedback can increase its capacity. The increase in capacity is obtained by changing the input distribution (prior) in response to feedback, yielding information on the channel state. Hence, in order to complete in this case, the universal system would also need to adapt its input distribution per symbol based on feedback. Hence, the universal systems presented here and in [9] are not suitable for this setting. However, for the modulo-additive channel, feedback does not increase capacity, because the best input distribution is uniform regardless of any knowledge on channel state (in other words, as easy to see, the bound based on Fano's inequality (45) would hold regardless of feedback), and in this particular case, the results do extend to the case of feedback inside the block (see also [11]).

2) Feedback between blocks, i.e., encoder of block b receives a message from decoder of block $b - 1$. This kind of feedback effectively increases the block size of the IFB system, as it allows it to keep track of the block index to some extent by passing it back and forth between the encoder and the decoder, through the channel in one way and the feedback link in the other way. Of course, this cannot be continued when the number of bits required to represent the block index is larger than $k \log |\mathcal{X}|$. In the modulo-additive channel, knowledge of the block index yields the maximum capacity of $\log |\mathcal{X}|$. It is interesting to note that, while such feedback seems to considerably strengthen the IFB system, Misra and Weissman [11] showed that the rate of the FS-class is limited in spite of feedback. This is because the restriction is on the number of states rather than on the block length.

3) It is possible to allow the reference system the use of asymptotically zero-rate feedback, which does not considerably increase the effective block length and cannot considerably increase the collapsed channel capacity, and is comparable with the amount of feedback used by the universal system.

2) An Alternative Comparison Class: As mentioned, a relatively short block size, limits the IFB class from attaining the collapsed channel capacity. This gap is not utilized in the current bounds. The collapsed channel capacity bound would still hold, if the reference encoder and decoder were allowed to encode multiple blocks together, but treat each block in the same way.

One option to define this class is to limit the encoder to 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). Another similar way is to let the encoder and decoder be general 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. For the modulo-additive channel, it is easy to see that in both cases, the reference rate would be limited to $\log |\mathcal{X}| - H(\tilde{\mathbf{Z}}_{b,k})$. It is more interesting to discuss these classes in the case of general channels—see [9]. Note that although these reference systems would fail for the password channel defined in Example 3, it is possible to devise an alternative example, showing that universal communication with respect to these classes over general channels is not possible (see Appendix G).

B. Other Comments

Theorem 2, connecting the transmission rate to the compression rate of the noise sequence is reminiscent of Ahlswede's channel coding scheme with feedback [16]. This scheme sends information by iteratively compressing the receiver's uncertainty with regard to the transmitted message. Indeed, Ooi [17] used this scheme in order to achieve adaptive communication over compound channels, including compound finite state channels. Ooi assumes a compound channel, i.e., probabilistic with unknown parameters, and varies the rate by changing the transmission length, while here an individual noise sequence is considered and the rate is varied by changing the number of bits

transmitted. Using a variable block length is a simpler, particular case, that can be obtained by transmitting a single block, in the scheme presented here. Adapting Ooi's scheme to the individual noise sequence channel seems complicated while using random coding yields a simple proof for the current result.

The result of Theorem 2 is also closely related to Ziv's result [18] regarding universal decoding over compound finite state channels. If Theorem 2 is particularized to the nonadaptive case, then it can be proven and generalized by the tools used there. The decoder in Ziv's paper uses joint Lempel–Ziv parsing and yields a decoding metric which generalizes in a sense the metric used here, for channels which are not necessarily memoryless. Theorem 2 and particularly Lemma 1 there, relate the size of the error sets M_0, M_u defined there, for the maximum likelihood decoder designed for the finite state channel, and the universal decoder. This relation indicates the rate that can be achieved with a given error probability is asymptotically the same. Furthermore, the only assumption used about the reference maximum likelihood decoder is that it uses a finite state metric (see the proof of Lemma 1 there), and thus the IFB decoder falls into this class.

In a previous paper [5] a different framework, termed "individual channels" was considered, in which no relation between the input and output of the channel is assumed *a priori*, and the communication rate is given as a function of the input and output sequences (see also [13, Part 1]). As an example, the empirical mutual information $\hat{I}(\mathbf{x}, \mathbf{y})$ is shown to be achievable. The current achievability result (Theorem 2) can be stated in these terms by saying that the rate function $R_{\text{emp}}(\mathbf{x}, \mathbf{y}) = \log |\mathcal{X}| - \frac{1}{n} L(\mathbf{y} - \mathbf{x})$ is asymptotically adaptively achievable (i.e., by an adaptive rate system). Note that there is no need to assume that the channel is truly modulo-additive to show this. It is also possible to show [13, Th. 10.2] that all achievable rate functions that depend only on the noise sequence $R_{\text{emp}}(\mathbf{x}, \mathbf{y}) = R(\mathbf{y} - \mathbf{x})$, are asymptotically of this form, i.e., given a system attaining the rate $R_{\text{emp}}(\mathbf{x}, \mathbf{y}) = R(\mathbf{y} - \mathbf{x})$ for each $\mathbf{x}^n, \mathbf{y}^n$ and where the channel input $\mathbf{x}^n$ is uniformly distributed (see the definitions therein), there exists a source encoding scheme with encoding lengths $L(\mathbf{z})$ such that asymptotically $R(\mathbf{z}^n) = \log |\mathcal{X}| - \frac{1}{n} L(\mathbf{z}^n)$.

In previous works [1], [4], rates which reflect the average channel behavior such as $1 - \hat{H}(\mathbf{z}^n)$ were termed "empirical capacity" mainly based on the similarity to the capacity expressions for memoryless channels. The term is not completely justified, since clearly this is not the maximum communication rate. The value C_{IFB} seems to be a better candidate to describe the modulo-additive channel's "empirical capacity," although as discussed above, other interesting definitions can be suggested. Note that there is no fixed order between C_{IFB} and the rate $1 - h_b(\hat{\epsilon})$, where $\hat{\epsilon}$ is the empirical frequency of "1"-s in the sequence (defined in [1]). For example for $\mathbf{z} = 0, 1, 0, 1, 0, \ldots$, the relation is $0 = 1 - h_b(\hat{\epsilon}) < C_{\text{IFB}} = 1$, while in Example 4 the order is inverse $0 = C_{\text{IFB}} < 1 - h_b(\hat{\epsilon}) = 1 - h_b\left(\frac{1}{4}\right)$. On the other hand, the relation $1 - h_b(\hat{\epsilon}) \leq 1 - \rho(\mathbf{z})$ always holds,[5] so the rates achieved by the scheme described here are asymptotically better than the previously achieved rates [1].

---

[5]This can be shown by block to variable encoding to rate $h_b(\hat{\epsilon}_i)$ where $\hat{\epsilon}_i$ is the empirical probability of 1-s in the block, and the convexity of $h_b(\cdot)$

The current results assume the noise sequence is fixed and unknown, and do not extend to the case where the noise sequence is determined by an adversary (i.e., z_i is a function of x_1^{i-1}), and the reference class is aware of the adversary strategy. To see this, it is easy to design an adversary that identifies the codebook used by the reference encoder, and locks the channel (by choosing the noise sequence randomly) once a different channel input appears.

VIII. CONCLUSION

This paper considered target rates for universal systems with feedback and focused on the modulo additive channel. The notion of the *iterated finite block capacity*, denoted C_{IFB}, was defined for a vector channel, as the highest rate achievable by encoders and decoders that may be designed for the particular relation that exists between the input and output, yet are constrained to be of finite block length and use the same scheme over each block. The IFB capacity C_{IFB} was used as a target communication rate to be achieved without any prior knowledge of the channel, using feedback. It was shown that C_{IFB} cannot be achieved universally for completely general input–output relations, however for the modulo-additive channel with an individual noise sequence, it can be achieved universally without knowing the noise sequence. Specifically, it was shown that $C_{\text{IFB}} \leq (1 - \rho) \log |\mathcal{X}|$, where ρ is the finite state compressibility of the noise sequence, and a universal system with feedback attaining a rate of at least $(1 - \rho) \log |\mathcal{X}|$ was presented. This result is relatively simple due to the properties of the modulo additive channel. In a follow-up paper [9] the result is extended to more general channels.

APPENDIX

A) Proof of Theorem 1: Suppose a given E, D achieve rate R and average error probability ϵ over b blocks of size k. Let us adopt the definitions of $\tilde{\mathbf{X}}_{b,k}$, $\tilde{\mathbf{Z}}_{b,k}$, and $\tilde{\mathbf{Y}}_{b,k}$ from Section V-A, and likewise define $\mathbf{m}$ and $\hat{\mathbf{m}}$ to be random variables generated by selecting the block index uniformly over $1, \ldots, b$ and taking the respective encoded/decoded (resp.) messages, i.e., $\mathbf{m} = \mathbf{m}_U$, $\hat{\mathbf{m}} = \hat{\mathbf{m}}_U$, where $U \sim \cup\{1, \ldots, b\}$. See Fig. 5. Then,

$$\frac{1}{b} \sum_{i=1}^{b} \Pr(\hat{\mathbf{m}}_i \neq \mathbf{m}_i) = \sum_{i=1}^{b} \Pr(\hat{\mathbf{m}}_i \neq \mathbf{m}_i)\Pr(U = i)$$
$$= \Pr(\hat{\mathbf{m}} \neq \hat{\mathbf{m}}) \leq \epsilon. \quad (40)$$

The rate R is now bounded by the entropy of $\tilde{\mathbf{Z}}_{b,k}$. By Fano's inequality

$$H(\mathbf{m}|\hat{\mathbf{m}}) \leq h_b(\epsilon) + \epsilon \log M. \quad (41)$$

Therefore, by the information processing inequality

$$I(\tilde{\mathbf{X}}_{b,k}; \tilde{\mathbf{Y}}_{b,k}) \geq I(\mathbf{m}; \hat{\mathbf{m}}) = H(\mathbf{m}) - H(\mathbf{m}|\hat{\mathbf{m}})$$
$$\geq \log M - (h_b(\epsilon) + \epsilon \log M). \quad (42)$$

On the other hand,

$$I(\tilde{\mathbf{X}}_{b,k}; \tilde{\mathbf{Y}}_{b,k}) = H(\tilde{\mathbf{Y}}_{b,k}) - H(\tilde{\mathbf{Y}}_{b,k}|\tilde{\mathbf{X}}_{b,k})$$
$$= H(\tilde{\mathbf{Y}}_{b,k}) - H(\tilde{\mathbf{Z}}_{b,k}) \quad (43)$$
$$\leq \log |\mathcal{X}|^k - H(\tilde{\mathbf{Z}}_{b,k}).$$

Combining the two

$$(1-\epsilon) \log M - h_b(\epsilon) \leq I(\tilde{\mathbf{X}}_{b,k}; \tilde{\mathbf{Y}}_{b,k}) \leq k \log |\mathcal{X}| - H(\tilde{\mathbf{Z}}_{b,k}). \quad (44)$$

Therefore,

$$R \leq \frac{1}{k} \log M \leq (1 - \epsilon)^{-1} \left[\log |\mathcal{X}| - \frac{1}{k} H(\tilde{\mathbf{Z}}_{b,k}) + \frac{1}{k} h_b(\epsilon) \right]. \quad (45)$$

If R is an achievable rate then by Definition 4, for any $\epsilon > 0$ there exist $k > 0$ such that (45) holds for this k and b large enough. Therefore, taking $\liminf_{b \to \infty}$ on both sides yields

$$R \leq (1 - \epsilon)^{-1} \left[\log |\mathcal{X}| - \frac{1}{k} \limsup_{b \to \infty} H(\tilde{\mathbf{Z}}_{b,k}) + \frac{1}{k} h_b(\epsilon) \right]. \quad (46)$$

Next, let us relate $H(\tilde{\mathbf{Z}}_{b,k})$ to the finite state compressibility (see (3)–(5) in Section V-A). There exists an FSM $\tilde{F}$ with $s_k = \mathcal{X}^{k-1} \cdot k$ states that compresses the sequence $\mathbf{z}_1^{kb}$ to at most $b \cdot (H(\tilde{\mathbf{Z}}_{b,k}) + 1)$ bits. This state machine implements a block to variable encoder tuned to the empirical distribution and is structured as follows: its state space includes a counter from 1 to k which counts the index inside the block, and a memory of $k - 1$ input characters. When the counter reaches k the machine outputs an encoded string, and the counter returns to 1. In the other counter states, the machine emits the empty string. The encoded string is generated by a simple block to variable encoder optimized to compress the random variable $Z_{k,b}$ to its minimum average length (e.g., a Huffman encoder, although a simple encoder using lengths $\lceil \log(\Pr(Z_{k,b})^{-1}) \rceil$ is sufficient for this purpose), and therefore its average encoded length for $\tilde{\mathbf{Z}}_{b,k}$ is at most $H(\tilde{\mathbf{Z}}_{b,k}) + 1$ [19, Sec. 5.4]. The encoding length is therefore

$$\sum_{i=1}^{b} |F(\mathbf{z}_i^{[k]})|$$
$$= \sum_{\tilde{\mathbf{z}} \in \mathcal{X}^k} \sum_{i=1}^{b} \text{Ind}\left((\mathbf{z}_i^{[k]})_{i=1}^{b} = \tilde{\mathbf{z}} \right) \cdot |F(\tilde{\mathbf{z}})| \quad (47)$$
$$= \sum_{\tilde{\mathbf{z}} \in \mathcal{X}^k} b \cdot \Pr(Z_{k,b} = \tilde{\mathbf{z}}) \cdot |F(\tilde{\mathbf{z}})|$$
$$\leq b(H(\tilde{\mathbf{Z}}_{b,k}) + 1).$$

Therefore, for $n = bk$

$$\rho_{\mathcal{F}(s_k)}(\mathbf{z}_1^n) \leq \rho_{\tilde{F}}(\mathbf{z}_1^n) = \frac{1}{n \log |\mathcal{X}|} |F(\mathbf{z}_1^n)|$$
$$\leq \frac{1}{n \log |\mathcal{X}|} b(H(\tilde{\mathbf{Z}}_{b,k}) + 1) \quad (48)$$
$$= \frac{1}{k \log |\mathcal{X}|} (H(\tilde{\mathbf{Z}}_{b,k}) + 1).$$

The condition $n = bk$ may be relaxed and the inequality may be applied to any finite n, taking $b = \lfloor \frac{n}{k} \rfloor$ (since if the last block is unfinished it will not contribute to the length, and the normalization by $n > bk$ will only decrease the LHS). Now,

$$
\begin{aligned}
\limsup_{n \to \infty} \rho_{\mathcal{F}(s_k)}(\mathbf{z}_1^n) &\le \limsup_{n \to \infty} \rho_{\tilde{F}}(\mathbf{z}_1^n) \\
&\le \limsup_{b \to \infty} \frac{1}{k \log |\mathcal{X}|} (H(\tilde{\mathbf{Z}}_{b,k}) + 1) \\
&= \frac{1}{k \log |\mathcal{X}|} (\limsup_{b \to \infty} H(\tilde{\mathbf{Z}}_{b,k}) + 1)
\end{aligned}
$$
(49)

and

$$
\begin{aligned}
\rho(\mathbf{z}) &= \lim_{s \to \infty} \limsup_{n \to \infty} \rho_{\mathcal{F}(s)}(\mathbf{z}_1^n) \\
&\le \limsup_{n \to \infty} \rho_{\mathcal{F}(s_k)}(\mathbf{z}_1^n) \le \frac{1}{k \log |\mathcal{X}|} (\limsup_{b \to \infty} H(\tilde{\mathbf{Z}}_{b,k}) + 1).
\end{aligned}
$$
(50)

Combining the above with (46) yields

$$
\forall \epsilon : \exists k :
$$
$$
\begin{aligned}
R &\le (1 - \epsilon)^{-1} \left[\log |\mathcal{X}| - \frac{1}{k} \limsup_{b \to \infty} H(\tilde{\mathbf{Z}}_{b,k}) + \frac{1}{k} h_b(\epsilon) \right] \\
&\le (1 - \epsilon)^{-1} \left[\log |\mathcal{X}| - \log |\mathcal{X}| \rho(\mathbf{z}) + \frac{1}{k} + \frac{1}{k} h_b(\epsilon) \right].
\end{aligned}
$$
(51)

Since the k obtaining the requirements of Definition 4 may be small, the factor $\frac{1}{k}$ on the RHS makes the bound loose. To tighten the bound the following argument is used: choose a number $j > 0$. If there exist E, D with block size k and average error probability ϵ over b large enough which divides by j, then by treating at each consecutive j blocks as a new block (and forming the encoder and decoder with block size $j \cdot k$ by using j times the original encoder and decoder), then by the union bound if ϵ_i denote the error probabilities over the blocks $i \in \{1, \ldots, b\}$, the error probabilities of the aggregate encoder and decoder will satisfy $\epsilon_i' \le \sum_{d=1}^{j} \epsilon_{(i-1)j+d}$, and therefore the average error probability will be $\epsilon' = \frac{1}{b/j} \sum_{i=1}^{b/j} \epsilon_i' \le \frac{j}{b} \sum_{i=1}^{b} \epsilon_i = j \cdot \epsilon$. The conclusion is that if the requirements of Definition 4 are met for a certain ϵ, k, they are also met for $j \cdot \epsilon, j \cdot k$. Therefore,

$$
\forall j, \epsilon : \exists k :
$$
$$
R \le (1 - j\epsilon)^{-1} \left[(1 - \rho(\mathbf{z})) \log |\mathcal{X}| + \frac{1}{jk} (1 + h_b(j\epsilon)) \right].
$$
(52)

Note that Definition 4 requires the rate to be achievable for any $\epsilon > 0$, and therefore it is possible to take $\epsilon \xrightarrow[n \to \infty]{} 0$. By choosing for each j, $\epsilon = \frac{1}{j^2}$, denoting k_j as any k that satisfies (52) for this j, and taking the limit $j \to \infty$ yields

$$
\begin{aligned}
R &\le \lim_{j \to \infty} \left\{ \left(1 - \frac{1}{j} \right)^{-1} \left[(1 - \rho(\mathbf{z})) \log |\mathcal{X}| \right. \right. \\
&\qquad\qquad \left. \left. + \frac{1}{jk_j} (1 + h_b(j^{-1})) \right] \right\} \\
&= (1 - \rho(\mathbf{z})) \log |\mathcal{X}|,
\end{aligned}
$$
(53)

which by Definition 5 proves the theorem. ∎

B) Proof of Theorem 2: In this section, it is shown that the rate $\log |\mathcal{X}| - \frac{1}{n} L(\mathbf{z})$ can be attained for a wide class of source encoders. Notice that this result is derived for a more general case in [13, Sec. 10.5]. A rough proof outline appears in Section III.

The Adaptive Communication Scheme: The scheme applies repeated "rateless" transmissions: fix a value K of the number of information bits per block. Using the common randomness, generate a random codebook of $\exp(K)$ words chosen independently and distributed uniformly over $\mathcal{X}^n$ which is known at the encoder and decoder. In each rateless block $b = 1, 2, \ldots$, the encoder sends K bits to the decoder, by sending the respective symbols from codeword indexed by those K bits. Note that at each block different symbols from the codebook are sent. The block terminates when a termination condition is satisfied at the decoder. Then, the decoder stores the decoded bits and indicates this to the encoder, through the feedback link (a 0–1 feedback is sufficient), and a new block, conveying K new bits, begins. The last block is potentially not decoded, if the termination condition is not satisfied at the last symbol.

The decoding and termination rule are specified next. Suppose that the current symbol number is i and the block number is b. The last symbol of the previous block (number $b - 1$) was sent at symbol j ($j = 0$ if b is the first block). Let $\hat{\mathbf{x}}_1^j$ denote the transmit sequence that follows from the previous decisions made by the decoder (i.e., is composed of the symbols from the codebook matching the decoded bits at each previously decoded block), and let $\mathbf{x}_{j+1}^i(m)$ denote the transmitted symbols matching codeword m ($m = 1, \ldots, \exp(K)$). $\hat{\mathbf{z}}^i(m)$ defined below is the decoder's hypothesis on the noise sequence $\mathbf{z}^i$:

$$
\hat{\mathbf{z}}^i(m) = \mathbf{y}^i - (\hat{\mathbf{x}}_1^j, \mathbf{x}_{j+1}^i(m)).
$$
(54)

Take $\hat{\mathbf{z}}^j = \mathbf{y}^j - \hat{\mathbf{x}}_1^j$ to be the j length prefix of $\hat{\mathbf{z}}^i(m)$ (which is independent of m). The decoder calculates the following condition for all $m = 1, \ldots, \exp(K)$:

$$
L_T(\hat{\mathbf{z}}^i(m)) - L_S(\hat{\mathbf{z}}^j) \le \lfloor (i - j) \cdot \log |\mathcal{X}| - \log \left(\frac{n}{\epsilon} \right) - K \rfloor.
$$
(55)

It announces the end of the block and decodes the bits matching codeword index m if the termination condition is satisfied with respect to codeword m (where ties can be broken arbitrarily), and does not terminate the block if the condition fails for all codewords.

Regarding the termination condition (55) note that the RHS starts from a negative value and increases linearly at a rate of $\log |\mathcal{X}|$ bits per symbol, while the LHS starts from a nonnegative value, but for a compressible noise sequence, it is expected to increase at a rate slower than $\log |\mathcal{X}|$ bits per symbol, therefore if the noise sequence is compressible and the block length n is large enough, the condition will eventually be met.

Proof of the Theorem: In order to prove Theorem 2 it is shown that the scheme above achieves an error probability of at most ϵ, and if an error does not occur, the number of bits decoded (determined by the number of blocks sent), approaches R_{emp} for a suitable choice of K.

Let us begin by bounding the error probability. First let us calculate the probability that the decoder decides in favor of

an incorrect codeword at any given symbol i (where again j denotes the end of the previous block), by using a property of the sequential encoder. Consider a sequence $\mathbf{z}^i$ of length i which is fed into the sequential source encoder in two stages: first, the first j symbols are fed (and the encoder has emitted $L_S(\mathbf{z}^j)$ bits), and then the rest $i - j$ symbols are fed and the encoding is terminated. Between the jth and the ith symbol, the encoder has emitted $L_T(\mathbf{z}^i) - L_S(\mathbf{z}^j)$ additional bits, which can be used to uniquely decode $\mathbf{z}^i_{j+1}$ when $\mathbf{z}^j$ is given (since the entire encoded stream can be generated from the first $L_S(\mathbf{z}^j)$ bits plus these additional bits, and used to decode $\mathbf{z}^i$). Therefore, the number of sequences $\mathbf{z}^i_{j+1}$ for which $L_T(\mathbf{z}^i) - L_S(\mathbf{z}^j) \leq d$ (where $d \in \mathbb{N}$) is upper bounded by $\exp(d)$ (since they are in effect encoded by d bits).

Since the codewords are independent, given the transmitted symbols, the other codewords in the codebook over the period of the current block are independent sequences uniformly drawn from $\mathcal{X}^{i-j}$. Therefore, the hypothesized tail of the sequence $\hat{\mathbf{Z}}^i_{j+1}(m) = \mathbf{Y}^i_{j+1} - \mathbf{X}^i_{j+1}(m)$ for any fixed m is also uniformly distributed (over the common randomness). Since there are at most $\exp(d)$ sequences that satisfy $L_T(\mathbf{z}^i) - L_S(\mathbf{z}^j) \leq d$, the probability that a particular sequence will satisfy the condition is at most

$$\frac{\exp(d)}{|\mathcal{X}|^{i-j}}, \qquad (56)$$

and therefore by the union bound, the probability that any of the competing sequences will satisfy the condition is at most

$$\frac{\exp(d)\exp(K)}{|\mathcal{X}|^{i-j}} = \exp(d + K - (i-j)\log|\mathcal{X}|). \qquad (57)$$

Substituting the value of d given by the termination condition $d = \lfloor (i-j) \cdot \log|\mathcal{X}| - \log\left(\frac{n}{\epsilon}\right) - K \rfloor \leq (i-j) \cdot \log|\mathcal{X}| - \log\left(\frac{n}{\epsilon}\right) - K$, the error probability per symbol is at most $\exp(-\log\left(\frac{n}{\epsilon}\right)) = \frac{\epsilon}{n}$, therefore by the union bound over n symbols, the probability of any error occurring during the decoding process is at most $\frac{\epsilon}{n} \cdot n = \epsilon$.

Next, let us analyze the rate achieved by the scheme. The analysis assumes no decoding errors occur. Denote the number of decoded blocks by B (so potentially there are $B + 1$ blocks, if the last block is not decoded). The proof is based on bounding the value of $L(\mathbf{z})$ based on the number of blocks. $\mathbf{z}$ denotes the true noise sequence.

Suppose a block was decoded in symbol i and the previous block ended at symbol j. By choosing K (or n) large enough it can be guaranteed that decoding never happens at the first symbol of any block, therefore $i > j + 1$. By the assumption that no decoding errors occurred the sequence $\hat{\mathbf{z}}^j$ is identical to $\mathbf{z}^j$. In symbol $i - 1$ the decoding condition was not met for any codeword, including the correct one, for which $\hat{\mathbf{z}}^i(m) = \mathbf{z}^i$. Therefore it holds, with respect to the true noise sequence, that

$$L_T(\mathbf{z}^{i-1}) - L_S(\mathbf{z}^j) > (i-1-j)\log|\mathcal{X}| - \log\left(\frac{n}{\epsilon}\right) - K. \quad (58)$$

This is an inverted version of condition (55). Note that the floor operator $\lfloor \cdot \rfloor$ is not needed here since the LHS is an integer.

Using monotonicity of L_T and the bounded difference $L_T - L_S$ the following telescopic series is lower bounded:

$$\begin{aligned} L_T(\mathbf{z}^i) - L_T(\mathbf{z}^j) &= L_T(\mathbf{z}^i) - L_S(\mathbf{z}^j) - [L_T(\mathbf{z}^j) - L_S(\mathbf{z}^j)] \\ &\geq L_T(\mathbf{z}^i) - L_S(\mathbf{z}^j) - \Delta_L^{\max}(n) \\ &\geq L_T(\mathbf{z}^{i-1}) - L_S(\mathbf{z}^j) - \Delta_L^{\max}(n) \\ &> (i-1-j)\log|\mathcal{X}| - \log\left(\frac{n}{\epsilon}\right) \\ &\quad - K - \Delta_L^{\max}(n) \end{aligned} \qquad (59)$$

where $\Delta_L^{\max}(n) = \max\{\Delta_L(l)\}_{l=1}^n$. By the same argument, this bound is true also for the undecoded block (with $i - 1 = n$). Taking j_b ($b = 1, \dots, B$) to be the symbol in which block b ended, and adding $j_0 = 0$ and $j_{B+1} = n$ the following bound is obtained by summing (59) over $B + 1$ blocks (including the undecoded one, which is taken as a block of length 0 if the last block is decoded)

$$\begin{aligned} L_T(\mathbf{z}) &= L_T(\mathbf{z}^{j_{B+1}}) - L_T(\mathbf{z}^{j_0}) \\ &= \sum_{b=1}^{B+1} \left[L_T(\mathbf{z}^{j_b}) - L_T(\mathbf{z}^{j_{b-1}}) \right] \\ &> \sum_{b=1}^{B+1} \left[(j_b - 1 - j_{b-1})\log|\mathcal{X}| - \log\left(\frac{n}{\epsilon}\right) - K - \Delta_L^{\max}(n) \right] \\ &= n\log|\mathcal{X}| - (B+1)\left(K + \log|\mathcal{X}| + \log\left(\frac{n}{\epsilon}\right) + \Delta_L^{\max}(n)\right). \end{aligned} \qquad (60)$$

The actual rate achieved by the scheme is

$$R_{\mathrm{act}} = \frac{BK}{n}. \qquad (61)$$

Extracting B from (60) and calculating R_{act} yields

$$\begin{aligned} R_{\mathrm{act}} &= \frac{BK}{n} \\ &\geq \frac{K}{n} \cdot \left(\frac{n\log|\mathcal{X}| - L_T(\mathbf{z})}{K + \log|\mathcal{X}| + \log\left(\frac{n}{\epsilon}\right) + \Delta_L^{\max}(n)} - 1 \right) \\ &= \left(1 + \frac{\log(|\mathcal{X}|n/\epsilon) + \Delta_L^{\max}(n)}{K} \right)^{-1} R_{\mathrm{emp}}(\mathbf{z}) - \frac{K}{n} \\ &\overset{(a)}{\geq} \left(1 - \frac{\log(|\mathcal{X}|n/\epsilon) + \Delta_L^{\max}(n)}{K} \right) R_{\mathrm{emp}}(\mathbf{z}) - \frac{K}{n} \\ &\overset{(b)}{\geq} R_{\mathrm{emp}}(\mathbf{z}) \\ &\quad - \left[\frac{\log|\mathcal{X}| \cdot (\log(|\mathcal{X}|n/\epsilon) + \Delta_L^{\max}(n))}{K} + \frac{K}{n} \right], \end{aligned} \qquad (62)$$

where (a) is because $\forall x \geq 0 : (1 + x)^{-1} \geq 1 - x$, and (b) is because $R_{\mathrm{emp}}(\mathbf{z}) \leq \log|\mathcal{X}|$. To choose the value of K that approximately minimizes the overhead term in the lower bound, the following lemma is used.

Lemma 3: For $a > 0, b > 0$ with $b \leq a$

$$r = \min_{k \in \mathbb{N}} \left(\frac{a}{k} + bk \right) \leq 3\sqrt{ab}. \qquad (63)$$

Proof: It is easy to see by derivation that the minimizer over $x \in \mathbb{R}$ of $\frac{a}{x} + bx$ is $x^* = \sqrt{\frac{a}{b}}$. Choosing $k^* = \lceil x^* \rceil$ yields $k^* \in \mathbb{N}$ and since $\sqrt{\frac{a}{b}} \leq k^* \leq \sqrt{\frac{a}{b}} + 1$:

$$\frac{a}{k^*} + bk^* \leq \frac{a}{\sqrt{\frac{a}{b}}} + b\left(\sqrt{\frac{a}{b}} + 1\right)$$
$$= 2\sqrt{ab} + b = 2\sqrt{ab} + \sqrt{b \cdot b} \overset{b \leq a}{\leq} 3\sqrt{ab}. \quad (64)$$

∎

Applying the lemma to the choice of K in (62) yields

$$R_{\text{act}} \geq R_{\text{emp}}(\mathbf{z})$$
$$- \underbrace{3\sqrt{\frac{\log|\mathcal{X}|}{n} \cdot \left[\log(n) + \log\left(\frac{|\mathcal{X}|}{\epsilon}\right) + \Delta_L^{\max}(n)\right]}}_{\delta_n} \quad (65)$$

where by assumption (A) of Theorem 2, $\delta_n \underset{n \to \infty}{\longrightarrow} 0$. ∎

C) Proof of Theorem 3: To prove Theorem 3, it is first shown that LZ77 [15] and LZ78 [6] fulfil the requirements of Theorem 2. Both algorithms operate by creating a dictionary from previous symbols in the string, compressing a new substring to a tuple containing its location in the dictionary, plus, possibly one additional symbol. In LZ77 the dictionary consists of all substrings that begin in a window of specified length before the first symbol that was not encoded yet. LZ78 parses the string $\mathbf{z}$ into phrases. Each phrase is a substring which is not a prefix of any previous phrase, but can be generated from concatenating a previous phrase with one additional symbol. The dictionary contains all phrases.

It is easy to make sure that L_T is monotonous (Assumption (B) of Theorem 2). This depends on the way the last phrase in the string is treated, which does not affect the asymptotical performance. Recall that in LZ compression, in which the new data bits are gathered, and encoded to produce a tuple, once they comprise a phrase that had not appeared before. The last phrase may be an incomplete substring of a string in the dictionary, and therefore does not naturally terminate by this rule. For example, in the following parsed sequence $[1, 0, 11, 110, 00, 110]$, the last phrase 110 had appeared before and therefore would not naturally produce a tuple. There are various ways to treat this last phrase. If, for example, the last phrase is sent without coding, then L_T will not be monotonous, since adding more symbols to $\mathbf{z}$ that will terminate the phrase and may result in a shorter compression. For example, the addition of either 0 or 1 to the sequence above, would generate a phrase that had not appeared before. A simple treatment is to encode the last phrase similarly to other phrases, i.e., refer to one of the phrases in the dictionary which is a prefix of the remaining substring (in the example, refer to the previous appearance of 110), and always give the length of the last substring, or equivalently the length of the block, at the end. This way the compression length associated with the last substring does not decrease when the substring is extended.

In order to bound $L_T(\mathbf{z}) - L_S(\mathbf{z})$ (Assumption (A) of Theorem 2), it is required to bound the tuple which encodes the last phrase. In LZ78 this tuple carries an index to a previous phrase, plus a new symbol. The number of previous phrases is

bounded by n (a coarse bound, but sufficient for the current purpose), and therefore [19, Lemma 13.5.1] its encoding will be of length $\log n + \log \log n + 1$ and the length of the tuple will be $\log n + \log \log n + c$ (where c is a constant accounting also for rounding, encoding of the additional symbol, etc.). Therefore, if the encoder ends the block with an indication of its length then $\Delta_{LZ78}^{\max}(n) = \Delta_{LZ78}(n) \leq 2\log n + 2\log \log n + c$. In LZ77 this tuple carries a pointer to the window and a length (i.e., two numbers bounded to $\{1, \ldots, n\}$). Therefore, after adding an indication of the length at the termination, $\Delta_{LZ77}^{\max}(n) = \Delta_{LZ77}(n) \leq 3\log n + 3\log \log n + c$. In both cases $\Delta_{LZ}^{\max}(n) = O(\log n)$ and the requirement is satisfied. Therefore, the compression length $L_{78}(\mathbf{z})$ may be substituted in Theorem 2.

The rest of the proof deals with analyzing and bounding the overheads related to the achievability of R_{emp}, and the difference between the LZ compression length and the finite state compressibility, in order to show that they tend to 0 with n. Recall the definitions of finite state compressibility (3)–(5) in Section V-A.

A result by Lempel and Ziv [6, Th. 2 (item ii)] shows that for every finite s

$$\rho_{78}(\mathbf{z}_1^n) \triangleq \frac{1}{n \log |\mathcal{X}|} L_{78}(\mathbf{z}_1^n) \leq \rho_{\mathcal{F}(s)}(\mathbf{z}_1^n) + \delta_s(n) \quad (66)$$

where $\delta_s(n) \underset{n \to \infty}{\longrightarrow} 0$. By Theorem 2 for any $\epsilon > 0$, the system attains the rate

$$R \geq R_{\text{emp}}(\mathbf{z}) - \delta_n$$
$$= \log|\mathcal{X}|\left(1 - \frac{1}{n \log|\mathcal{X}|} L_{78}(\mathbf{z}_1^n)\right) - \delta_n$$
$$= (1 - \rho_{78}(\mathbf{z}_1^n))\log|\mathcal{X}| - \delta_n \quad (67)$$
$$\geq (1 - \rho_{\mathcal{F}(s)}(\mathbf{z}_1^n) - \delta_s(n))\log|\mathcal{X}| - \delta_n.$$

Choose a small $\tilde{\delta}$. Since $\lim_{n \to \infty} \delta_n = 0$, it is possible to find n_1^* large enough so that for any $n > n_1^*$, $\delta_n \leq \tilde{\delta}$. By the definition $\rho(\mathbf{z}) = \rho(\mathbf{z}^\infty) = \lim_{s \to \infty} \limsup_{n \to \infty} \rho_{\mathcal{F}(s)}(\mathbf{z}_1^n)$, it is possible to find s large enough such that $\limsup_{n \to \infty} \rho_{\mathcal{F}(s)}(\mathbf{z}_1^n) \leq \rho(\mathbf{z}) + \tilde{\delta}$. For this value of s, because $\lim_{n \to \infty} \delta_s(n) = 0$, it is possible to find n_2^* large enough so that for any $n > n_2^*$, $\delta_s(n) \leq \tilde{\delta}$. For the same s, find $n > n_1^*, n_2^*$ so that $\rho_{\mathcal{F}(s)}(\mathbf{z}_1^n) \leq \limsup_{n' \to \infty} \rho_{\mathcal{F}(s)}(\mathbf{z}_1^{n'}) + \tilde{\delta} \leq \rho(\mathbf{z}) + 2\tilde{\delta}$. Writing (67) for these s, n yields

$$R \geq (1 - \rho_{\mathcal{F}(s)}(\mathbf{z}_1^n) - \tilde{\delta})\log|\mathcal{X}| - \tilde{\delta}$$
$$\geq (1 - \rho(\mathbf{z}) - 3\tilde{\delta})\log|\mathcal{X}| - \tilde{\delta} \quad (68)$$
$$= (1 - \rho(\mathbf{z}))\log|\mathcal{X}| - (3\log|\mathcal{X}| + 1) \cdot \tilde{\delta}.$$

Therefore, the requirements of Theorem 3 are satisfied by substituting $\tilde{\delta} = (3\log|\mathcal{X}| + 1)^{-1}\delta$. ∎

Proof of Corollary 3.1: The corollary follows directly from the definition, by application of Theorems 3 and 1.

Proof of Corollary 3.2: Suppose the sequence $\mathbf{z}$ is drawn by a stationary ergodic source. The mutual information rate is $\overline{I}(\mathbf{X}; \mathbf{Y}) = \overline{H}(\mathbf{Y}) - \overline{H}(\mathbf{Y}|\mathbf{X}) \leq \log|\mathcal{X}| - \overline{H}(\mathbf{Z})$, and to obtain an equality, the capacity is obtained by a uniform i.i.d. prior, which maximizes $\overline{H}(\mathbf{Y})$. Hence, the capacity is $C = \log|\mathcal{X}| - \overline{H}(\mathbf{Z})$. It was shown [6, Th. 4] that the finite state

compressibility equals the entropy rate of the source, with probability one. The proposed communication system would asymptotically attain the communication rate C, without prior knowledge of the noise distribution.

D) Proof of Corollary 5.1: The target is to find the required n such that $\Delta^* \leq \delta \log |\mathcal{X}|$, based on the bounds of Theorem 5. The lower bound of Theorem 5 on Δ^* yields a lower bound on n^* (converse) and the upper bound on Δ^* yields an upper bound on n^* (achievability).

Converse: According to the lower bound (13), either $\tau \leq \frac{|\mathcal{X}|}{k}$, or $\left\lfloor \log\left(k\tau\right) \frac{1}{\log|\mathcal{X}|} \right\rfloor \frac{\log|\mathcal{X}|}{2k} \leq \delta \log|\mathcal{X}|$, which combined with $\lfloor x \rfloor \geq x - 1$ yields, after rearrangement, $\log\left(k\tau\right) \leq (2k\delta + 1)\log|\mathcal{X}|$, i.e., $\tau \leq \frac{1}{k}|\mathcal{X}|^{2k\delta+1}$. This condition on τ is always less strict than the former, and because at least one of the conditions should hold, the second always holds. Translating to a condition on n yields

$$n = \frac{|\mathcal{X}|^k}{\tau} \geq \frac{1}{|\mathcal{X}|} \cdot k|\mathcal{X}|^{(1-2\delta)k}. \tag{69}$$

Achievability: Let us find an n for which the upper bound is at most $\delta \log |\mathcal{X}|$. Define $g(\tau) = \tau \log\left(\frac{1}{\tau}\right)$. Assuming $\tau \leq \frac{1}{2k}$, then $g(\tau)$ is monotonically increasing, $g(\tau) \geq \tau \log(2k)$, and $\frac{k}{4}\tau^2 \leq \frac{1}{8}\tau$. Thus,

$$\begin{aligned} &\frac{\tau}{2}\log\left(\frac{1}{\tau}\right) + \left(\frac{k}{4}\tau^2 + \tau\right)\log e \\ &\leq \frac{1}{2}g(\tau) + \left(\frac{1}{8} + 1\right)\tau \log e \\ &\leq \frac{1}{2}g(\tau) + \frac{9}{8} \cdot \frac{g(\tau)}{\log(2k)}\log e \quad (70) \\ &= \left(\frac{1}{2} + \frac{9\log e}{8\log(2k)}\right)g(\tau) \\ &\leq 3g(\tau). \end{aligned}$$

The same assumption $\tau \leq \frac{1}{k}$ leads to $n \geq k|\mathcal{X}|^k$ and thus

$$\frac{k}{n} \leq |\mathcal{X}|^{-k}, \tag{71}$$

and

$$\begin{aligned} \delta_n^* &\leq 4\sqrt{\frac{\log|\mathcal{X}| \cdot \log\left(k^2|\mathcal{X}|^{2k+1}\right)}{k|\mathcal{X}|^k}} \\ &\overset{(a)}{\leq} 4\sqrt{\frac{5(\log|\mathcal{X}|)^2}{|\mathcal{X}|^k}} \quad (72) \\ &\leq 10\log|\mathcal{X}| \cdot |\mathcal{X}|^{-k/2} \end{aligned}$$

where (a) is because $\log k \leq (k-1)\log e$, so $k^2 \leq \log e^{2(k-1)} \leq |\mathcal{X}|^{3(k-1)}$. Combining (70), (71), (72) with (14) yields

$$\begin{aligned} \Delta_+ &\leq 3g(\tau) + |\mathcal{X}|^{-k}\log(e|\mathcal{X}|) + 10\log|\mathcal{X}| \cdot |\mathcal{X}|^{-k/2} \\ &= 3g(\tau) + \left(|\mathcal{X}|^{-k/2}\left(1 + \frac{\log(e)}{\log|\mathcal{X}|}\right) + 10\right)\log|\mathcal{X}| \cdot |\mathcal{X}|^{-k/2} \\ &\leq 3g(\tau) + \left(2^{-1/2}\left(1 + \frac{\log(e)}{\log 2}\right) + 10\right)\log|\mathcal{X}| \cdot |\mathcal{X}|^{-k/2} \\ &\leq 3g(\tau) + 12\log|\mathcal{X}| \cdot |\mathcal{X}|^{-k/2}. \end{aligned} \tag{73}$$

Thus, to guarantee $\Delta^* \leq \delta \cdot \log |\mathcal{X}|$ is enough if $\tau \leq \frac{1}{k}$ and

$$\tau \leq g^{-1}\left(\frac{1}{3}(\delta - 12 \cdot |\mathcal{X}|^{-k/2}) \cdot \log|\mathcal{X}|\right) \tag{74}$$

i.e., it is required that

$$\tau \leq \min\left[g^{-1}\left(\frac{1}{3}(\delta - 12 \cdot |\mathcal{X}|^{-k/2}) \cdot \log|\mathcal{X}|\right), 1/k\right] \tag{75}$$

and equivalently

$$n = \frac{|\mathcal{X}|^k}{\tau} \geq \frac{|\mathcal{X}|^k}{\min\left[g^{-1}\left(\frac{1}{3}(\delta - 12 \cdot |\mathcal{X}|^{-k/2}) \cdot \log|\mathcal{X}|\right), 1/k\right]}. \tag{76}$$

■

E) Proof of Lemma 1: In this section, Lemma 1 from Section VI-C, regarding the entropy of the noise distribution of the test channel defined there, is proven. See Section VI-C.1 for the definition of the test channel.

The entropy of each prefix, conditioned on the past is $\log|\mathcal{X}|^{k-d} = (k-d)\log|\mathcal{X}| = k \cdot \overline{H}_1$, where $\overline{H}_1$ is the asymptotical entropy rate per symbol. The entropy of the suffix, given the past, changes over time. When choosing the ith noise sequence (of length k), at most $i-1$ different prefixes already appeared. Therefore, the probability that the ith prefix equals one of the previous ones is at most $\frac{i-1}{|\mathcal{X}|^{k-d}}$. Let us define $i_d^* \triangleq |\mathcal{X}|^{k-d}$ and consider first the case $i \leq i_d^*$. In this case, the entropy of the suffix, given all previous symbols, is zero with probability at most $\frac{i-1}{|\mathcal{X}|^{k-d}}$, and $\log|\mathcal{X}|^d$ with probability at least $1 - \frac{i-1}{|\mathcal{X}|^{k-d}}$, and is therefore at least $\left(1 - \frac{i-1}{|\mathcal{X}|^{k-d}}\right) \cdot d \cdot \log|\mathcal{X}|$. Formally, define $P_z^{(i)} \triangleq (\mathbf{Z}_i^{[k]})_1^{k-d}$, $S_z^{(i)} \triangleq (\mathbf{Z}_i^{[k]})_{k-d+1}^k$ as the ith prefix and suffix, and $F_i = \mathrm{Ind}\left[\bigcup_{j=1}^{i-1}\left\{P_z^{(i)} = P_z^{(j)}\right\}\right]$ as a flag indicating whether $P_z^{(i)}$ appeared before. Then, the entropy of the suffix given the past is

$$\begin{aligned} H(S_z^{(i)}|\mathbf{Z}_1^{(i-1)k+k-d}) &= H(S_z^{(i)}|\mathbf{Z}_1^{(i-1)k}, P_z^{(i)}) \\ &\overset{(a)}{=} H(S_z^{(i)}|\mathbf{Z}_1^{(i-1)k}, P_z^{(i)}, F_i) \\ &= H(S_z^{(i)}|\mathbf{Z}_1^{(i-1)k}, P_z^{(i)}, F_i = 0) \cdot \Pr(F_i = 0) \quad (77) \\ &\quad + H(S_z^{(i)}|\mathbf{Z}_1^{(i-1)k}, P_z^{(i)}, F_i = 1) \cdot \Pr(F_i = 1) \\ &= d \cdot \log|\mathcal{X}| \cdot \Pr(F_i = 0) \\ &\geq d \cdot \log|\mathcal{X}| \cdot \left(1 - \frac{i-1}{|\mathcal{X}|^{k-d}}\right) \end{aligned}$$

where (a) is because F_i is a function of $\mathbf{Z}_1^{(i-1)k}, P_z^{(i)}$. Therefore,

$$\begin{aligned} H(\mathbf{Z}_i^{[k]}|\mathbf{Z}^{(i-1)k}) &= H(P_z^{(i)}|\mathbf{Z}^{(i-1)k}) + H(S_z^{(i)}|P_z^{(i)}, \mathbf{Z}^{(i-1)k}) \\ &\geq d \cdot \log|\mathcal{X}| \cdot \left(1 - \frac{i-1}{|\mathcal{X}|^{k-d}}\right) + k \cdot \overline{H}_1. \end{aligned} \tag{78}$$

For $i \geq i_d^*$ simply bound

$$H(\mathbf{Z}_i^{[k]}|\mathbf{Z}^{(i-1)k}) \geq H(P_z^{(i)}|\mathbf{Z}^{(i-1)k}) = k \cdot \overline{H}_1. \tag{79}$$

Notice that H_1 determines the asymptotical entropy rate of the sequence, and matches the bound on the universal system and the rate of the IFB system.

For $i \leq i_d^*$:

$$H(\mathbf{Z}^{ik}) = \sum_{j=1}^{i} H(\mathbf{Z}_j^{[k]} | \mathbf{Z}^{(j-1)k})$$

$$\overset{(78)}{\geq} \sum_{j=1}^{i} \left(d \cdot \log |\mathcal{X}| \cdot \left(1 - \frac{j-1}{|\mathcal{X}|^{k-d}} \right) + k \cdot \overline{H}_1 \right)$$

$$= d \cdot \log |\mathcal{X}| \cdot \left(i - \frac{(i-1)i}{2|\mathcal{X}|^{k-d}} \right) + ik \cdot \overline{H}_1$$

$$= d \cdot \log |\mathcal{X}| \cdot i \underbrace{\left(1 - \frac{i-1}{2|\mathcal{X}|^{k-d}} \right)}_{\geq \frac{1}{2}} + ik \cdot \overline{H}_1$$

$$\geq \frac{1}{2} i \cdot d \cdot \log |\mathcal{X}| + ik \cdot \overline{H}_1. \qquad (80)$$

The above implies that the entropy, at times $n = ik \leq i_d^* k$, is bounded above the straight line with slope

$$\overline{H}_0 = \frac{d}{2k} \cdot \log |\mathcal{X}| + \overline{H}_1. \qquad (81)$$

See Fig. 9. For $i \geq i_d^*$,

$$H(\mathbf{Z}^{ik}) = H(\mathbf{Z}^{i_d^* k}) + \sum_{t=i_d^*+1}^{i} H(\mathbf{Z}_t^{[k]} | \mathbf{Z}^{(t-1)k})$$

$$\overset{(79),(80)}{\geq} k|\mathcal{X}|^{k-d}\overline{H}_0 + k(i - |\mathcal{X}|^{k-d})\overline{H}_1 \qquad (82)$$

and in general

$$H(\mathbf{Z}^{ik}) \geq ik \cdot \overline{H}_1 + \min(i, |\mathcal{X}|^{k-d}) k(\overline{H}_0 - \overline{H}_1). \qquad (83)$$

Consider now $H(\mathbf{Z}^n)$ for n that does not, in general, divide by k. Inside the block of length k, the per-symbol conditional entropy $H(\mathbf{Z}_n | \mathbf{Z}^{n-1})$ is $\log |\mathcal{X}|$ during the prefix, and then increases at a smaller or equal rate during the suffix. Therefore, the entropy $H(\mathbf{Z}^n)$ is concave during the block (see Fig. 9). Because the entropy at block edges is bounded above straight lines (83), the entropy inside the block is bounded by these lines as well, i.e., (83) can be extended to

$$H(\mathbf{Z}^n) \geq n \cdot \overline{H}_1 + \min(n, k|\mathcal{X}|^{k-d})(\overline{H}_0 - \overline{H}_1) \qquad (84)$$

which proves the lemma. ∎

F) Proof of Lemma 2: For the sake of brevity, as long as a single value of k is discussed, let $\mathcal{M} \triangleq \mathcal{X}^k$ denote the super-alphabet of length k and $m = |\mathcal{X}|^k$ denote its size. Let $\pi(\cdot)$ define a distribution over $\mathcal{M}$. The Dirichlet $(\frac{1}{2}, \ldots, \frac{1}{2})$ density over the set of distributions is defined as

$$w_k(\pi) = \exp(-C_m) \prod_{a \in \mathcal{M}} \pi(a)^{-1/2}, \qquad (85)$$

where

$$C_m = \log \left(\Gamma(1/2)^m / \Gamma(m/2) \right), \qquad (86)$$

and for a l-length vector $\mathbf{a} \in \mathcal{M}^l$, let $\pi(\mathbf{a}) = \prod_{i=1}^{l} \pi(a_i)$ be the probability given to $\mathbf{a}$ by the i.i.d. distribution $\pi(\cdot)$. Let

$$P_k(\mathbf{a}) = \int_{\Delta_{\mathcal{M}}} \pi(\mathbf{a}) w_k(\pi) d\pi, \qquad (87)$$

and define the weighted average of all probabilities given to $\mathbf{a}$ by i.i.d. distributions $\pi(\mathbf{a})$, where the integral is over the unit simplex $\Delta_{\mathcal{M}} = \{\pi : \forall a \in \mathcal{M} : \pi(a) \geq 0, \sum_{a \in \mathcal{M}} \pi(a) = 1\}$. By well-known results of Shtarkov, which are detailed in Lemma 1 in Xie and Barron's paper [20], it holds that

$$\log \frac{\max_\pi \pi(\mathbf{a})}{P_k(\mathbf{a})} \leq \frac{m-1}{2} \log \frac{l}{2\pi} + C_m + \left(\frac{m^2}{4l} + \frac{m}{2} \right) \log e$$

$$\triangleq r_{lk}. \qquad (88)$$

Note that the terms that do not scale with n are usually ignored, because m is considered fixed, however here they matter, because the question would be how fast m (equivalently k) may grow with n. Thus, for any π:

$$\pi(\mathbf{a}) \leq P_k(\mathbf{a}) \exp(r_{lk}). \qquad (89)$$

The same inequality would hold when marginalizing the above to any parts of $\mathbf{a}$ (i.e., summing over the remaining elements of $\mathbf{a}$). Using this observation, let us set $l = \lceil n/k \rceil$. Then substituting in (87), $\mathbf{a} = \mathbf{z}_1^{lk}$ yields

$$P_k(\mathbf{z}^{lk}) = \int_{\Delta_{\mathcal{M}}} \pi_k(\mathbf{z}^{lk}) w_k(\pi) d\pi, \qquad (90)$$

and summing both sizes with respect to $\mathbf{z}_{n+1}^{lk}$, yields

$$P_k(\mathbf{z}^n) = \int_{\Delta_{\mathcal{M}}} \pi_k(\mathbf{z}^n) w_k(\pi) d\pi. \qquad (91)$$

Furthermore, by (89)

$$\forall \pi : \pi_k(\mathbf{z}^n) \leq P_k(\mathbf{z}^n) \cdot \exp(r_{lk}). \qquad (92)$$

Let us now bound r_{lk}. Following Xie and Barron's [20, Remark 7], using $\Gamma(1/2) = \sqrt{\pi}$ and Stirling's approximation $\Gamma(m/2) \geq \sqrt{2\pi}(m/2)^{\frac{m-1}{2}} e^{-\frac{m}{2}}$ yields from (86)

$$C_m \leq \frac{m-1}{2} \log \pi - \frac{1}{2} \log 2 - \frac{m-1}{2} \log(m/2) + \frac{m}{2} \log e \qquad (93)$$

and from (88)

$$r_{lk} \leq \frac{m-1}{2} \log \left(\frac{l}{m} \right) + \left(\frac{m^2}{4l} + m \right) \log e - \frac{1}{2} \log 2. \qquad (94)$$

Note that r_{lk} is always positive, even when $l < m$. When $l < m$, the second factor dominates, and the normalized loss $\frac{r_{lk}}{l}$ does not tend to zero. Therefore, it is not useful to consider m in this region. Assuming $l \geq m$ (note that since $m \geq 2$ this also implies $n \geq k$), and substituting $l = \lceil n/k \rceil \leq n$, $m = |\mathcal{X}|^k$, yields

$$r_{lk} \leq \frac{|\mathcal{X}|^k - 1}{2} \log \left(\frac{\lceil n/k \rceil}{|\mathcal{X}|^k} \right) + \left(\frac{k|\mathcal{X}|^{2k}}{4n} + |\mathcal{X}|^k \right) \log e$$

$$- \frac{1}{2} \log 2$$

$$\leq \frac{|\mathcal{X}|^k}{2} \log \left(\frac{n}{|\mathcal{X}|^k} \right) + \left(\frac{k|\mathcal{X}|^{2k}}{4n} + |\mathcal{X}|^k \right) \log e. \qquad (95)$$

Now let

$$P_Z(\mathbf{z}) = \sum_{k=1}^{\infty} 2^{-k} \cdot P_k(\mathbf{z}) \qquad (96)$$

then from (92)

$$\forall \pi : \pi_k(\mathbf{z}^n) \leq P_k(\mathbf{z}^n) \cdot \exp(r_{lk}) \leq \frac{P_Z(\mathbf{z})}{2^{-k}} \exp(r_{lk}) \qquad (97)$$

and thus

$$\forall \pi : \frac{1}{n} \log \pi_k(\mathbf{z}^n) \leq \frac{1}{n} \log P_Z(\mathbf{z}) + \frac{1}{n} \left(k \log(2) + r_{lk} \right). \qquad (98)$$

The factor $\frac{1}{n} \left(k \log(2) + r_{lk} \right)$ can be coarsely bounded by (95):

$$\frac{k \log(2) + r_{lk}}{n} \leq \frac{|\mathcal{X}|^k}{2n} \log \left(\frac{n}{|\mathcal{X}|^k} \right)$$
$$+ \left(\frac{k|\mathcal{X}|^{2k}}{4n^2} + \frac{|\mathcal{X}|^k}{n} + \frac{k}{n} \right) \log e \qquad (99)$$
$$= \frac{\tau}{2} \log \left(\frac{1}{\tau} \right) + \left(\frac{k}{4} \tau^2 + \tau + \frac{k}{n} \right) \log e$$
$$\triangleq \Delta_\pi,$$

with $\tau \triangleq \frac{|\mathcal{X}|^k}{n} \leq 1$. Combining this bound with (98) yields the result of the lemma. ∎

G) Password Channel for i.i.d. Distributions: As noted in Sections IV-E and VII-A, even limiting the reference class it to i.i.d. input distributions would not solve the "password" problem, and therefore universality is not possible even with respect to such encoders, for general channels. To see this, consider the following example, where the channel identifies the input distribution of the encoder. This is a variation of the "password channel" (Example 3).

Example 5: The channel class is a class of binary input–output channels, parameterized by a single a parameter $p \in [0, 1]$. For each value of p, the channel is defined as follows:
1) At each symbol k in time, if the normalized number of ones at the input $\mathbf{x}_1^k$ is not within a range of thresholds $[L_{k,p}, H_{k,p}]$, then from this time on, the channel "locks" and the output is $y_i = 0, \forall i \geq k$. Otherwise, the channel is noise free and the output equals the input $y_k = x_k$.
2) The threshold sequences $L_{k,p}, H_{k,p}$ are computed such that, if the input is i.i.d. $\text{Ber}(p)$, then with high probability $1 - \epsilon_0$, the thresholds will not be crossed during any of the n symbols (i.e., the channel will not lock). Clearly, as k increases, the thresholds will converge to p.

Thus, the channel "identifies" a certain input probability. Notice that all the channels are causal and deterministic, and they allow communication at a rate of approximately $h_b(p)$. The "memoryless" reference schemes mentioned in Section VII-A can communicate over this channel using a $\text{Ber}(p)$ input distribution and approach this rate, with a small error probability. But a universal communication over the class is impossible. Until the channel locks, nothing can be inferred about p from the channel output. Therefore, the transmit distribution of the universal scheme until the lock time is independent of p. On the other hand, any given input sequence, will "lock" some of the channels in the class. Therefore, any operation of the universal system is bound to cause some of the channels to lock, and achieve an asymptotically zero rate.

REFERENCES

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

[2] S. Verdú and T. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, Jul. 1994.

[3] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2148–2177, Oct. 1998.

[4] K. Eswaran, A. Sarwate, A. Sahai, and M. Gastpar, "Zero-rate feedback can achieve the empirical capacity," *IEEE Trans. Inf. Theory*, vol. 58, no. 1, Jan. 2010.

[5] Y. Lomnitz and M. Feder, "Communication over individual channels," *IEEE Trans. Inf. Theory*, vol. 57, no. 11, pp. 7333–7358, Nov. 2011.

[6] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inf. Theory*, vol. 24, no. 5, pp. 530–536, Sep. 1978.

[7] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2124–2147, Oct. 1998.

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

[9] Y. Lomnitz and M. Feder, "Universal communication—Part II: Channels with memory," submitted for publication, Jan. 2012 [Online]. Available: http://arxiv.org/abs/1202.0417

[10] N. Merhav and M. Feder, "Universal schemes for sequential decision from individual data sequences," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1280–1292, Jul. 1993.

[11] V. Misra and T. Weissman, "The porosity of additive noise sequences," in *Proc. IEEE Int. Symp. Inf. Theory*, 2012, pp. 2581–2585.

[12] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inf. Theory*, vol. 27, no. 2, pp. 199–207, Mar. 1981.

[13] Y. Lomnitz, "Universal communication over unknown channels" Ph.D. dissertation, Tel Aviv University, Tel Aviv, Israel, 2012.

[14] Y. Lomnitz and M. Feder, "Communicating over modulo-additive channels with compressible individual noise sequence," in *Proc. IEEE 26th Convent. Elect. Electron. Eng. Israel*, Nov. 2010, pp. 000766–000770.

[15] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inf. Theory*, vol. 23, no. 3, pp. 337–343, Sep. 1977.

[16] R. Ahlswede, "A constructive proof of the coding theorem for discrete memoryless channels with feedback," in *Proc. 6th Prague Conf. Inf. Theory, Statist. Decis. Funct., Random Processes*, 1971, pp. 39–50.

[17] J. Ooi, "A framework for low-complexity communication over channels with feedback" Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1997.

[18] J. Ziv, "Universal decoding for finite-state channels," *IEEE Trans. Inf. Theory*, vol. 31, no. 4, pp. 453–460, Jul. 1985.

[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY: Wiley, 1991.

[20] Q. Xie and A. Barron, "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 431–445, Mar. 2000.

Yuval Lomnitz (S'12–M'13) received the B.Sc. degree from the Technion, Israel in 1995 and the M.Sc. degree from Tel Aviv University, Israel in 2003, both in Electrical engineering. Since 2007 he is a Ph.D. student in the Department of Electrical engineering—Systems in Tel Aviv University. His research focuses on communication over unknown channels, using feedback.

During 2001–2004 he was with Envara working on WiFi, during 2004–2012 he was with Intel working on wireless communication, and from 2012 he is with Vayyar, developing a radio imaging technology.

Mr. Lomnitz is a recipient of the Weinstein prize for a scientific publication in the field of signal processing (2010), the Weinstein prize for excellence in studies (2011), and the Feder Family prize for best student work in communication technology (2011 and 2013).

Meir Feder (S'81–M'87–SM'93–F'99) received the B.Sc. and M.Sc. degrees from Tel-Aviv University, Israel and the Sc.D. degree from the Massachusetts Institute of Technology (MIT) Cambridge, and the Woods Hole Oceanographic Institution, Woods Hole, MA, USA, all in electrical engineering in 1980, 1984 and 1987, respectively.

After being a Research Associate and Lecturer in MIT he joined the Department of Electrical Engineering—Systems, Tel-Aviv University, where he is now a Professor. He had visiting appointments at the Woods Hole Oceanographic Institution, Scripps Institute, Bell laboratories and in 1995/1996 he has been a visiting professor at MIT. He is also extensively involved in the high-tech industry and co-founded several companies including Peach Networks, a developer of a unique server-based interactive TV solution which was acquired on March 2000 by Microsoft, and Amimon, a leading provider of ASIC's for wireless high-definition A/V connectivity at the home.

Prof. Feder is a co-recipient of the 1993 IEEE Information Theory Best Paper Award. He also received the 1978 "creative thinking" award of the Israeli Defense Forces, the 1994 Tel-Aviv University prize for Excellent Young Scientists, the 1995 Research Prize of the Israeli Electronic Industry, and the research prize in applied electronics of the Ex-Serviceman Association, London, awarded by Ben-Gurion University. Between June 1993–June 1996 he served as an Associate Editor for Source Coding of the IEEE TRANSACTIONS ON INFORMATION THEORY.