Achieving the Empirical Capacity Using Feedback: Memoryless Additive Models

Ofer Shayevitz, *Member, IEEE*, and Meir Feder, *Fellow, IEEE*

***Abstract*—We address the problem of universal communications over an unknown channel with an instantaneous noiseless feedback, and show how rates corresponding to the empirical behavior of the channel can be attained, although no rate can be guaranteed in advance. First, we consider a discrete modulo-additive channel with alphabet $\mathcal{X}$, where the noise sequence Z^n is *arbitrary and unknown* and may causally depend on the transmitted and received sequences and on the encoder's message, possibly in an adversarial fashion. Although the classical capacity of this channel is zero, we show that rates approaching the *empirical capacity* $\log|\mathcal{X}| - H_{\rm emp}(Z^n)$ can be universally attained, where $H_{\rm emp}(Z^n)$ is the empirical entropy of Z^n. For the more general setting, where the channel can map its input to an output in an arbitrary unknown fashion subject only to causality, we model the empirical channel actions as the modulo-addition of a realized noise sequence, and show that the same result applies if common randomness is available. The results are proved constructively, by providing a simple sequential transmission scheme approaching the empirical capacity.**

***Index Terms*—Adversarial channels, arbitrarily varying channels, feedback communications, individual sequences, universal communications.**

I. Introduction

THE capacity of a channel is classically defined as the supremum of all rates for which communication with arbitrarily low probability of error can be guaranteed in advance. However, when a noiseless feedback link between the receiver and the transmitter exist, one does not necessarily have to commit to a rate prior to transmission, and communication can take place using some sequential scheme at a variable rate determined by the specific realization of the channel, thus the better the channel realization the higher the rate of transmission. When the channel law is known, this approach cannot yield average rates exceeding those attainable by fixed-rate feedback schemes, and for large classes of channels cannot even exceed the rates of non-feedback schemes [1]–[3]. The variable-rate approach may, however, have the advantages of a better error exponent and a lower complexity. Several transmission schemes possessing such merits were proposed for the binary-symmetric channel (BSC) [4], [5], the Gaussian additive noise channel [6], [7], discrete memoryless channels (DMC) [8], [7], and finite-state channels (FSC) [9].

When the channel law is unknown to some degree, variable-rate feedback schemes become even more attractive, as the realized channel may sometimes be explicitly or implicitly estimated via feedback. In [10], [11], a rate universal scheme for unknown DMC with a random decision time was suggested (later termed *rateless coding*), attaining a rate equal to the mutual information of the channel in use for any selected input distribution. Following this lead, a universal variable-rate transmission schemes for compound BSC and Z-channels with feedback was introduced [12], and shown to attain any fraction of the realized channel's capacity and achieve the Burnashev error exponent [13]. In [14] it was shown that for compound FSC with feedback, it is possible to transmit at a rate approaching the mutual information of the realized channel for any Markov input distribution, by an incremental universal compression of the errors, e.g., via Lempel–Ziv coding.

So far, however, the variable-rate approach was not applied to more stringent channel uncertainty models, where the channel behavior is arbitrary or even adversarial. As a motivating example, consider a binary modulo-additive channel with feedback, where the noise sequence is an individual sequence (i.e., deterministic and unknown). Let us assume (at first) that the fraction of "1's" in the noise sequence (namely, the fraction of errors inserted by the channel) is *a priori* known to be $p \in [0,1]$ at the most. The fixed-rate communication problem in this setting has been addressed before in several different contexts. In a classical work [15], Berlekamp considered this model in the context of error correction capability with feedback, where the receiver is required to correct the errors inserted by the channel and uniquely recover the transmitted message. Since no decoding errors are allowed, the noise sequence in this case can also be thought of as being generated by an adversary that knows the message and the coding scheme, but is "power limited" by p. Berlekamp showed that whenever $p \geq \frac{1}{3}$, there exists an adversarial strategy for error insertion such that the receiver cannot hope to separate even three messages, and so the capacity is zero. For smaller p, he was able to show that the capacity is upper-bounded by[1] $1 - h_B(p)$ and a (tight) straight line tangent to $1 - h_B(p)$, intersecting the horizontal axis at $p = \frac{1}{3}$. The convex part of this bound was later shown to be tight as well [16].

The same communication problem can also be studied in the context of the (discrete memoryless) arbitrarily varying

Manuscript received November 22, 2006; revised August 29, 2008. Current version published February 25, 2009. This work was supported in part by the Israel Science Foundation under Grant 223/05. The work of O. Shayevitz was supported by the Adams Fellowship Program of the Israel Academy of Sciences and Humanities. The material in this paper was presented in part at the IEEE International Symposium on Information Theory (ISIT), Adelaide, Australia, September 2005.

The authors are with the Department of Electrical Engineering–Systems, Tel-Aviv University, Ramat-Aviv 69978, Israel (e-mail: ofersha@eng.tau.ac.il; meir@eng.tau.ac.il).

Communicated by K. Kobayashi, Associated Editor for Shannon Theory.
Digital Object Identifier 10.1109/TIT.2008.2011434

[1] $h_B(\cdot)$ is the binary entropy function.

channel (AVC). In an AVC setting, a memoryless channel law is selected from a given set (state space) at each time point, in an arbitrary unknown manner. The AVC without feedback was studied extensively [17]–[19], and shown to yield different capacities depending on the error criterion (average/maximum error probability) and also on the existence of common randomness (resulting in the so-called *random-code* capacity, which is the same under both error criteria). Within the AVC framework, the channel under discussion is a binary AVC with two states, a clean channel and an inverting channel, where the noise sequence becomes the state sequence and the maximal fraction of channel errors p yields a state constraint. Under the maximum error probability criterion, this AVC with feedback is equivalent to Berlekamp's setting, the capacity of which was given above. Interestingly, it turns out that even *without feedback*, the *random-code* capacity of this channel is given by $1 - h_B(p)$ for any $p < \frac{1}{2}$ (and zero otherwise)[2] and can be attained with merely $\Omega(\log n)$ bits of common randomness [18], [20]. This small amount of randomness can be generated via feedback with a negligible impact on rate, hence the capacity of the discussed binary channel with feedback coincides with its (non-feedback) random-code capacity, yielding a significant gain relative to Berlekamp's capacity through the use of randomness. This approach of extending non-feedback AVC results to the feedback regime was also taken in [21] (albeit without state constraints) where it was shown that the feedback capacity of an AVC is equal to its random-code capacity.

Berlekamp's result and the AVC approach are limited by the requirement to commit to a fixed rate prior to transmission, and so positive rates are obtained only under noise/state sequence constraints. As we shall see, variable-rate coding can be used to obtain a much stronger result that applies to any noise sequence without any constraint (i.e., $p = 1$). As a corollary of our main result, we constructively show that for the binary channel under discussion, rates arbitrarily close to $1 - h_B(p_{\rm emp})$ can be achieved by a simple (deterministic, algorithmic) sequential feedback scheme with probability approaching one and a vanishing (maximum) error probability,[3] where $p_{\rm emp}$ is the *empirical* fraction of "1's" in the individual noise sequence. Thus, although the fixed-rate capacity is zero when there are no constraints on the noise sequence, one can opportunistically attain rates approaching what would have been the capacity of the channel had the fraction of "1's" in the noise sequence been known in advance. It is therefore only appropriate to call the quantity $1 - h_B(p_{\rm emp})$ the (zero-order, modulo-additive) *empirical capacity* of the realized channel.

More generally, in this paper we consider a discrete channel with feedback over a common input/output alphabet $\mathcal{X}$, that maps its input to an output in a modulo-additive fashion, where the corresponding noise sequence is *arbitrary and unknown* and may causality depend on the transmitted and received sequences and on the encoder's message. We constructively show that rates arbitrarily close to the (zero-order, modulo-additive) empirical capacity $\log|\mathcal{X}| - H_{\rm emp}(Z^n)$ can be achieved by a simple sequential scheme with probability approaching one, where $H_{\rm emp}(Z^n)$ is the (zero-order) empirical entropy of the noise sequence Z^n. Furthermore, we consider the more general setting where the channel can map its input to an output in an *arbitrary unknown fashion* (not necessarily modulo-additive), subject only to causality. By modeling the channel actions as the modulo-addition of a realized noise sequence, we show that the corresponding empirical capacity can be achieved, if common randomness is allowed. These channel models can also be interpreted as adversarial, where an adversary (jammer) that knows the transmission scheme and is in possession of the transmitted message, causally listens to the transmitted and received sequences and employs an arbitrary unknown jamming strategy.

The paper is organized as follows. In Section II, some notations and useful lemmas are given. The channel model and the main result of the paper are provided in Section III. A finite-horizon feedback transmission scheme achieving the empirical capacity in a modulo-additive setting is described in Section IV, and its analysis appears in Section V. A short discussion is provided in Section VI. The horizon-free variant of the scheme appears in Appendix B, and its extension to general causal channels under the modulo-additive model using common randomness, is discussed in Appendix C. Part II of this work is dedicated to the investigation of more elaborate models for channel actions, and the corresponding gain in rate that may be achieved by utilizing empirical dependencies in the channel's behavior.

II. Notations and Preliminaries

The following standard asymptotic notations are used:

$$f(n) = \mathrm{O}(g(n)) \iff \limsup_{n\to\infty}\left|\frac{f(n)}{g(n)}\right| < \infty$$

$$f(n) = \mathrm{o}(g(n)) \iff \lim_{n\to\infty}\frac{f(n)}{g(n)} = 0$$

$$f(n) = \Omega(g(n)) \iff \liminf_{n\to\infty}\left|\frac{f(n)}{g(n)}\right| > 0$$

$$f(n) = \omega(g(n)) \iff \lim_{n\to\infty}\frac{f(n)}{g(n)} = \infty.$$

For $n \in \mathbb{N}$, we use the convention $\langle n\rangle \triangleq \{0, 1, \ldots, n-1\}$. All logarithms are taken to the base of 2. For vectors, we write $z_m^n = (z_m, z_{m+1}\ldots, z_n)$ which by convention is the null string if $m > n$, and use $z^n = z_1^n$ for short. For real-valued vectors, $\|\cdot\|_\infty$ is the $\mathcal{L}_\infty$ norm. Random variable (r.v.'s) are usually denoted by upper case letters, with the corresponding lower case letters for realizations. We write $H(\,\cdot\,)$ for the entropy function, $h_B(\,\cdot\,)$ for the binary entropy function, and $D(\cdot\|\cdot)$ for relative entropy. A finite alphabet $\mathcal{X}$ in this paper is taken to be the set $\mathcal{X} = \langle|\mathcal{X}|\rangle$ associated with the modulo-addition operator $+$, unless otherwise stated.

Lemma 1 (Entropy $\mathcal{L}_\infty$ Bound): Let $\boldsymbol{p}$ be a probability distribution over a finite alphabet $\mathcal{X}$. Then

$$H(\boldsymbol{p}) \geq \log|\mathcal{X}|(1 - |\mathcal{X}|\|\boldsymbol{p} - \boldsymbol{p}_u\|_\infty)$$

where $\boldsymbol{p}_u$ is the uniform distribution over $\mathcal{X}$.

Proof: See Appendix A. □

[2] In fact, this is also the deterministic coding capacity without feedback, under the *average* error probability criterion.

[3] The probabilities in this case are taken with respect to (w.r.t.) randomness created via the feedback link.

For a sequence $z^n \in \mathcal{X}^n$, the number of occurrences of the symbol $i \in \mathcal{X}$ is denoted by $n_i(z^n)$. The *empirical distribution* of z^n is the vector of relative symbol occurrences in z^n

$$\boldsymbol{p}_{\rm emp}(z^n) \triangleq \left(\frac{n_0(z^n)}{n}\,\frac{n_1(z^n)}{n},\ldots,\frac{n_{|\mathcal{X}|-1}(z^n)}{n}\right)$$

where by convention, the empirical distribution of a null string is taken to be uniform. When z^n is a binary sequence, we write $p_{\rm emp}(z^n)$ for its empirical fractions of "1's," and loosely refer to $p_{\rm emp}(z^n)$ as the empirical distribution of z^n. The (zero-order) *empirical entropy* of z^n is $H(\boldsymbol{p}_{\rm emp}(z^n))$, the entropy pertaining to the empirical distribution, and is denoted by $H_{\rm emp}(z^n)$ for short. For a binary sequence, the empirical entropy is written $h_B(p_{\rm emp}(z^n))$.

A *sequential probability estimator* over a finite alphabet $\mathcal{X}$ is a sequence of nonnegative functions $\{\widehat{p}_k(\cdot|z^{k-1})\}_{k=1}^{\infty}$ which sum to unity for any $k \in \mathbb{N}, z^{k-1} \in \mathcal{X}^{k-1}$. As usual, the function $\widehat{p}_k(\cdot|z^{k-1})$ is thought of as a *probability assignment* for the next symbol z_k given past observations z^{k-1}. The probability assigned by the sequential estimator to any finite individual sequence z^n is therefore defined as

$$\widehat{p}(z^n) \triangleq \prod_{k=1}^{n} \widehat{p}_k(z_k|z^{k-1}), \quad z^n \in \mathcal{X}^n.$$

We would also be interested in the following quantity:

$$\widehat{p}(z^n\|w^n) \triangleq \prod_{k=1}^{n} \widehat{p}_k(z_k|w^{k-1}), \quad z^n, w^n \in \mathcal{X}^n$$

which is the probability assigned to the individual sequence z^n by a sequential estimator matched to a different individual sequence w^n, namely, the case of sequential estimation from noisy observations.

A well known probability estimator is the Krichevsky–Trofimov (KT) estimator [22] given by

$$\widehat{p}_k^{\rm KT}(i|z^{k-1}) = \frac{n_i(z^{k-1})+\frac{1}{2}}{k-1+\frac{|\mathcal{X}|}{2}}, \quad i \in \mathcal{X}.$$

The following lemma shows that the per-symbol code-length assigned by the KT estimator to any individual sequence is close to its empirical entropy.

Lemma 2 (KT Redundancy [23]): For any individual sequence $z^n \in \mathcal{X}^n$

$$-\frac{1}{n}\log\widehat{p}^{\rm KT}(z^n) \le H_{\rm emp}(z^n) + \frac{(|\mathcal{X}|-1)\log n + {\rm o}(1)}{2n}.$$

A sequential probability estimator is said to be a $\mathrm{KT}(b)$ *estimator* if it can be obtained from a KT estimator by updating the estimates *at least* once per b symbols. Such an estimator is given by

$$\widehat{p}_k^{{\rm KT}(b)}(i|z^{k-1}) = \widehat{p}_k^{\rm KT}(i|z^{\nu_k}), \quad i \in \mathcal{X}$$

where $\{\nu_k\}_{k=1}^{\infty}$ is a nondecreasing index sequence determining the positions where the KT estimates are updated, thus satisfying $k-b \le \nu_k \le k-1$. In the sequel, we will be interested in the excess redundancy incurred when using a $\mathrm{KT}(b)$ estimator in lieu of the KT estimator, and possibly when the estimator is matched to a different individual sequence, i.e., the case of noisy observations.

Lemma 3 (Noisy $\mathrm{KT}(b)$ Excess Redundancy): Let $\widehat{p}^{{\rm KT}(b)}$ be a $\mathrm{KT}(b)$ estimator for some $b \in \mathbb{N}$. Then for any pair of individual sequences $z^n, w^n \in \mathcal{X}^n$

$$-\log\frac{\widehat{p}^{{\rm KT}(b)}(z^n\|w^n)}{\widehat{p}^{\rm KT}(z^n)} \le 2|\mathcal{X}|(b+d(z^n,w^n)-1)\log 2ne \tag{1}$$

where $d(\cdot,\cdot)$ is the Hamming distance operator.

Proof: See Appendix A. □

Let $z^n \in \mathcal{X}^n$, $b^n \in \{0,1\}^n$, and let $\sigma_k(b^n)$ be the index of the kth nonzero element in b^n. Define $z^n \downarrow b^n \in \mathcal{X}^{n_1(b^n)}$ to be the vector whose kth element is $z_{\sigma_k(b^n)}$, i.e., $z^n \downarrow b^n$ is a sample of size $n_1(b^n)$ of z^n where sampling positions are determined by the "1's" in b^n. The next lemma bounds the probability that the deviation (measured in the $\mathcal{L}_\infty$ norm) between the empirical distributions of an r.v. sequence, and that of a fixed-size random sample without replacement from that sequence, exceeds some threshold. It is a direct consequence of a result by Hoeffding [24]

Lemma 4 (Sampling Without Replacement): Let Z^n, B^n be a pair of statistically independent r.v. sequences, where Z^n takes values in a finite alphabet $\mathcal{X}$, and B^n is uniformly distributed over the set $\{b^n \in \{0,1\}^n : n_1(b^n) = m\}$. Then for any $\tau > 0$

$$\mathbb{P}(\|\boldsymbol{p}_{\rm emp}(Z^n) - \boldsymbol{p}_{\rm emp}(Z^n \downarrow B^n)\|_\infty > \tau) \le 2|\mathcal{X}|\exp(-2m\tau^2) \tag{2}$$

Proof: See Appendix A. □

The following lemma is an analogue of Lemma 4 for causally independent sampling, and is a direct consequence of the Azuma–Hoeffding inequality for bounded-difference martingales [25].

Lemma 5 (Causally Independent Sampling): Let Z^n, B^n be a pair of r.v. sequences, where Z^n takes values in a finite alphabet $\mathcal{X}$. Suppose $B_{k+1} \sim \mathrm{Ber}(q)$ and is statistically independent of (B^k, Z^{k+1}) for any $k \in \langle n \rangle$. Then for any $\tau > 0$

$$\mathbb{P}(\|\boldsymbol{p}_{\rm emp}(Z^n) - \alpha(B^n)\boldsymbol{p}_{\rm emp}(Z^n \downarrow B^n)\|_\infty > \tau) \le 2|\mathcal{X}|\exp\left(-\frac{n\tau^2q^2}{2}\right) \tag{3}$$

where

$$\alpha(B^n) \triangleq \frac{n_1(B^n)}{\mathbb{E}\{n_1(B^n)\}}. \tag{4}$$

Proof: See Appendix A. □

In the context of Lemma 5 above, B^n is said to be an independent and identically distributed (i.i.d.) *causal sampling sequence*[4] for Z^n. The multiplication of the sample's empirical distribution by the factor $\alpha(\,\cdot\,)$ is referred to as α-*normalization*. Note that $\alpha(B^n)\boldsymbol{p}_{\rm emp}(Z^n \downarrow B^n)$ is in fact the vector of symbol occurrences in $Z^n \downarrow B^n$, normalized by $\mathbb{E}(n_1(B^n)) = nq$ instead of by $n_1(B^n)$. Moreover, $\alpha(B^n) \to 1$ almost surely (a.s.), hence this vector converges a.s. to a probability distribution as $n \to \infty$.

III. Channel Model and Main Result

A (causal) *channel* over a common input and output finite alphabet $\mathcal{X}$ is a sequence of conditional probability distributions $\mathcal{W} = \{W_k(\cdot|x^k, y^{k-1})\}_{k=1}^{\infty}$ over $\mathcal{X}$, where $x^k \in \mathcal{X}^k, y^{k-1} \in \mathcal{X}^{k-1}$. Two sequences of r.v.'s (X^∞, Y^∞) taking values in $\mathcal{X}$ are said to be a pair of *input/output sequences* for the channel, respectively, if for any $k \in \mathbb{N}, x^k \in \mathcal{X}^k, y^k \in \mathcal{X}^k$

$$\mathbb{P}_{Y_k|X^kY^{k-1}}(y_k|x^k, y^{k-1}) = W_k(y_k|x^k, y^{k-1}).$$

We will find it convenient to *model* the channel's action on its input as the modulo-addition of a *realized noise sequence* Z^∞ corresponding to (X^∞, Y^∞), implicitly defined by

$$Y_k = X_k + Z_k, \quad k \in \mathbb{N} \tag{5}$$

In fact, a channel can be equivalently defined by the conditional distribution of the noise sequence given past and present inputs and past outputs $\mathbb{P}_{Z_k|X^kY^{k-1}}(z_k|x^k, y^{k-1})$. A channel $\mathcal{W}$ is called *modulo-additive* if the following Markov relation is satisfied for any pair of input/output sequences and any $k \in \mathbb{N}$:

$$Z_k \leftrightarrow X^{k-1}Y^{k-1} \leftrightarrow X_k \tag{6}$$

or, equivalently, if $W_k(y_k + z|x_k + z, x^{k-1}, y^{k-1})$ is independent of $z \in \mathcal{X}$ for any $x^k, y^k \in \mathcal{X}^k$. Note that this definition of a modulo-additive channel allows the noise sequence to depend on previous inputs and outputs in a general way. The more restricted class of modulo-additive channels, where the channel is completely defined by the noise distribution itself, is discussed below.

The family of all causal channels over $\mathcal{X}$ is denoted by $\mathsf{C}_\mathcal{X}$, and the family of all modulo-additive channels over $\mathcal{X}$ is denoted by $\mathsf{M}_\mathcal{X} \subset \mathsf{C}_\mathcal{X}$. The families $\mathsf{C}_\mathcal{X}$ and $\mathsf{M}_\mathcal{X}$ are broad, including also nonstationary and nonergodic channels. In particular, $\mathsf{C}_\mathcal{X}$ includes the following families of channels sometimes used for modeling channel uncertainty [26].

- The *Compound Memoryless Channel*, which is a family of time-invariant memoryless channels, or in our notation all channels for which (informally) $W_k(\cdot|x^k, y^{k-1}) = W(\cdot|x_k)$, where $W(\cdot|\cdot) \in S$ for some set S of conditional probability distributions over $\mathcal{X}$. Specifically, $\mathsf{M}_\mathcal{X}$ includes compound channels for which S consists only of (memoryless) modulo-additive mappings.
- The *Arbitrarily Varying Channel (AVC)*, which is a family of time-varying memoryless channels, or in our notation all channels for which (informally) $W_k(\cdot|x^k, y^{k-1}) = W_k(\cdot|x_k)$, where each $W_k(\cdot|\cdot) \in S$ for some set S (state space) of conditional probability distributions over $\mathcal{X}$. Alternatively, an AVC can also be defined via the noise sequence by requiring $Z_k \leftrightarrow X_k \leftrightarrow X^{k-1}Y^{k-1}$. Once again, $\mathsf{M}_\mathcal{X}$ includes all AVCs for which S consists of only (memoryless) modulo-additive mappings.
- *Noise Sequence Channels*: This family is denoted by $\mathsf{N}_\mathcal{X}$, and consists of all (modulo-additive) channels that are completely defined by the noise sequence itself, i.e., for which the (stricter) Markov relation

$$Z_k \leftrightarrow Z^{k-1} \leftrightarrow X^kY^{k-1} \tag{7}$$

holds for any $k \in \mathbb{N}$. Note that some texts use noise sequence channels as the standard definition for a modulo-additive channel. Our definition for a modulo-additive channel is broader, allowing a general coupling between the noise sequence and previous inputs/outputs, and so $\mathsf{N}_\mathcal{X} \subset \mathsf{M}_\mathcal{X}$ with the inclusion being strict.
- *Individual Noise Sequence Channels*: This family consists of all noise sequence channels for which the noise sequence is an individual sequence $Z^\infty = z^\infty$, i.e., $W_k(y_k|x^k, y^{k-1}) = \delta_{y_k, x_k+z_k}$, where $\delta_{i,j}$ is Kronecker's delta. It is a subfamily of $\mathsf{N}_\mathcal{X}$ defined above, and may be viewed as an AVC with S being the set of all deterministic modulo-additive mappings. The example of the binary channel given in the Introduction falls into this category.
- *Causal Adversarial Channels*: Loosely speaking, a causal adversarial channel is one for which at each time point an adversary (jammer) chooses a (possibly random) input–output mapping according to some (possibly random) *strategy*, that may arbitrarily depend on previous channel inputs and outputs. It is easy to see that the family of causal adversarial channels is in fact equivalent to $\mathsf{C}_\mathcal{X}$, since any strategy employed by the adversary can be equivalently described by the sequence $\{W_k(\cdot|x^k, y^{k-1})\}_{k=1}^{\infty}$. If the adversary is limited to use only modulo-additive mapping strategies, then this is equivalent to the family $\mathsf{M}_\mathcal{X}$. In the sequel, we will sometimes find it convenient to use the adversarial point of view.

The communication problem with feedback over a channel $\mathcal{W} \in \mathsf{C}_\mathcal{X}$ is now described. Without loss of generality, we assume a *transmitter* is in possession of a *message point* $\theta_0 \in [0, 1)$, its binary expansion representing an infinite bit string to be reliably conveyed to a *receiver* over the channel $\mathcal{W}$ (later assumed to be unknown). A (sequential) *feedback transmission scheme* is described by a triplet $(G, \mathcal{S}, \Delta)$, where $G = \{g_k : [0,1)\times\mathcal{X}^{k-1} \mapsto \mathcal{X}\}_{k=1}^{\infty}$ is a sequence of *transmission functions*, $\mathcal{S} \in \mathsf{C}_\mathcal{X}$ is a *feedback strategy*, $\Delta = \{\Delta_k : \mathcal{X}^k \times \mathcal{X}^{k-1} \mapsto \mathfrak{J}\}_{k=1}^{\infty}$ is a sequence of *decoding rules*, and $\mathfrak{J}$ is the set of all binary subintervals of the unit interval.[5] A scheme is said to use *passive feedback* if $\mathcal{S}$ consists of only deterministic conditional distributions, and is otherwise said to use *active feedback*. A scheme is said to use *asymptotically passive feedback* if the portion of nondeterministic conditional distributions within the first n elements of $\mathcal{S}$ tends to zero with n.

[4]The i.i.d. sequence B^n is only causally independent of Z^n, but the two sequences may generally be dependent. For instance, setting Z_1 constant and $Z_k = B_{k-1}$ satisfies the conditions of the lemma.

[5]There is a one-to-one correspondence between any finite binary string $b_1b_2\ldots b_k$ and a *binary subinterval* $[\alpha, \beta) \subseteq [0, 1)$ where $\alpha = 0.b_1b_2\ldots b_k$ and $\beta = \alpha + 2^{-k}$ (similar to arithmetic coding).

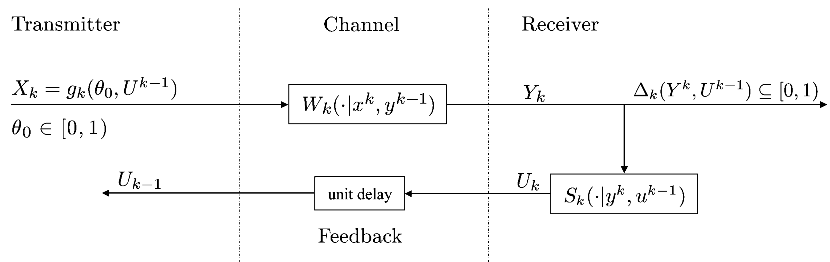

Fig. 1. Channel model and a feedback transmission scheme.

A feedback transmission scheme $(G, \mathcal{S}, \Delta)$ used over the channel $\mathcal{W} \in C_{\mathcal{X}}$ with a message point $\theta_0 \in [0,1)$ is described by the following construction, also depicted in Fig. 1.

- (X^∞, Y^∞) is an input/output pair for the channel $\mathcal{W}$. The channel input sequence is said to be *generated* by the transmitter, while the channel output sequence is said to be *observed* by the receiver.
- (Y^∞, U^∞) is an input/output pair for the feedback strategy $\mathcal{S}$.
- The channel input sequence is generated by the transmitter for any $k \in \mathbb{N}$, as follows:
$$X_k = g_k(\theta_0, U^{k-1}). \tag{8}$$
The existence of an instantaneous noiseless feedback link is manifested through the fact that the *feedback sequence* U^∞, which is causally generated from Y^∞ by the receiver via the feedback strategy $\mathcal{S}$, is causally available to the transmitter. Note that for passive feedback this means that U_k is a deterministic function of Y^k, with the most common example being when the channel output is fed back to the transmitter, i.e., $U_k = Y_k$.
- The following Markov relation is satisfied for any $k \in \mathbb{N}$:
$$Y_k \leftrightarrow X^k Y^{k-1} \leftrightarrow U^{k-1}. \tag{9}$$
Loosely speaking, this relation guarantees that any randomness generated by the receiver (and shared with the transmitter via feedback) is "private," i.e., the channel/adversary has no direct access to it and its actions are based on observing channel inputs/outputs only.
- $\Delta_k(Y^k, U^{k-1})$ is the receiver's *decoded interval* at time k.

The construction above uniquely determines the joint distribution of $(X^\infty, Y^\infty, U^\infty)$. If transmission is terminated at time n, the receiver decodes bits that correspond to the decoded interval $\Delta_n(Y^n, U^{n-1})$ as being the leading bits in the message point's binary expansion. Accordingly, the associated *rate* and (pointwise) *error probability* at time n are defined as

$$\begin{aligned} R_n(\mathcal{W}, \theta_0) &\triangleq -\frac{1}{n}\log|\Delta_n(Y^n, U^{n-1})| \\ p_e(n, \mathcal{W}, \theta_0) &\triangleq \mathbb{P}(\theta_0 \notin \Delta_n(Y^n, U^{n-1})). \end{aligned} \tag{10}$$

Modeling the channel actions as the modulo-addition of a realized noise sequence Z^∞ as in (5) allows us to define the (modulo-additive, zero-order) *empirical capacity* at time n as

$$C_n^{\rm emp}(\mathcal{W}, \theta_0) \triangleq \log|\mathcal{X}| - H_{\rm emp}(Z^n) \tag{11}$$

where the r.v. $H_{\rm emp}(Z^n)$ is the zero-order empirical entropy of Z^n. Hence, the empirical capacity is the capacity of a corresponding memoryless modulo-additive channel, with a marginal noise distribution that coincides with the empirical distribution of Z^n. In general, both the instantaneous rate and the empirical capacity are r.v.'s with distributions that depend on the channel, the message point, and even the transmission scheme itself (the latter dependency is suppressed). Note however that in the special case where communications take place over a noise sequence channel $\mathcal{W} \in N_{\mathcal{X}}$, the empirical capacity depends only on the channel, and for an individual noise sequence channel, it is deterministic.

The universal communication problem over a family of channels $F \subseteq C_{\mathcal{X}}$ is now described. Suppose a feedback transmission scheme $(G, \mathcal{S}, \Delta)$ is used for communication over an <u>unknown</u> channel $\mathcal{W} \in F$. Regarding the empirical capacity as a measure for how well the channel behaves, a desirable property would be for the scheme, although being fixed and independent of the actual channel in use, to achieve rates close to the empirical capacity with a low error probability. Making this notion precise, a scheme $(G, \mathcal{S}, \Delta)$ is said to (uniformly) *achieve the empirical capacity* over the family F, if

$$\begin{aligned} &\sup_{\mathcal{W}\in F, \theta_0\in[0,1)} p_e(n, \mathcal{W}, \theta_0) < \varepsilon_1(n) \\ &\inf_{\mathcal{W}\in F, \theta_0\in[0,1)} \mathbb{P}\left(R_n(\mathcal{W}, \theta_0)\right. \\ &\qquad \left. > C_n^{\rm emp}(\mathcal{W}, \theta_0) - \varepsilon_2(n)\right) > 1 - \varepsilon_3(n) \end{aligned} \tag{12}$$

where all $\varepsilon_1(n), \varepsilon_2(n), \varepsilon_3(n) \to 0$. Such a scheme is also called *universal* for the family $\mathcal{F}$.

In the discussion so far we have considered *horizon-free* transmission schemes, namely, schemes that do not depend on any decoding deadline and can be terminated at any time. In the sequel, we also consider *finite-horizon* schemes, which are schemes that must terminate at some given time n (horizon). The horizon-free construction and the subsequent definitions of rate and error probability immediately carry over to the finite-horizon setting, via simple truncation. A sequence $\{(G, \mathcal{S}, \Delta)_n\}_{n=1}^\infty$ of finite-horizon transmission schemes, with $(G, \mathcal{S}, \Delta)_n$ having a horizon n, is said to achieve the empirical capacity over a family $\mathcal{F}$ if for any $n \in \mathbb{N}$ the scheme $(G, \mathcal{S}, \Delta)_n$ satisfies (12), and $\varepsilon_1(n), \varepsilon_2(n), \varepsilon_3(n) \to 0$. A finite-horizon scheme is loosely said to achieve the empirical capacity, whenever it is clear that a suitable sequence of such schemes with an arbitrarily large horizon can be constructed. We now state our main result.

Theorem 1: There exists a horizon-free feedback transmission scheme $(G, \mathcal{S}, \Delta)$ using asymptotically passive feedback, that achieves the empirical capacity over the family $\mathsf{M}_{\mathcal{X}}$. Such a universal scheme is constructed explicitly below. Furthermore, the scheme can be adapted to achieve the empirical capacity over the larger family $\mathsf{C}_{\mathcal{X}}$, if common randomness is available.

Proof: The rest of the paper is dedicated to the construction of the universal scheme and hence to the proof of this theorem. The discussion in the body of the paper focuses on a finite-horizon feedback transmission scheme, which is introduced in Section IV, and shown to be universal for the family $\mathsf{M}_{\mathcal{X}}$ of modulo-additive channels in Section V. The horizon-free variant of this scheme is discussed in Appendix B, and the adaptations (via common randomness) required to obtain universality for the family $\mathsf{C}_{\mathcal{X}}$ of all causal channels, are relegated to Appendix C. □

The following remarks are now in order.

1) The probabilities in (10) and (12) are taken over the randomness created both by the feedback strategy, and by the channel. The randomness due to feedback is negligible yet essential as manifested by the special case of an individual noise sequence channel, where without randomness one is limited by Berlekamp's results [15] and the empirical capacity cannot be attained, even being known in advance. Note also that the definition in (12) requires uniform convergence over the set of message points. This is the variable-rate counterpart of a maximum error probability criterion, and from an adversarial viewpoint is equivalent to the assumption that the adversary knows the message point.
2) As already mentioned, when communicating over an unknown member of $\mathsf{M}_{\mathcal{X}}$ or $\mathsf{C}_{\mathcal{X}}$, no rate can be guaranteed in advance since both families include (many) channels with zero capacity. Furthermore, since the channel law may vary arbitrarily there is no hope to identify the actual channel in use and attain its capacity, even with feedback. Our approach is more "optimistic": We disregard the complex nature of $\mathsf{M}_{\mathcal{X}}$, $\mathsf{C}_{\mathcal{X}}$ and model the channel actions as being memoryless modulo-additive, although these are usually not. This simple model allows us to opportunistically attain rates that correspond to the empirical goodness of the realized channel (measured relative to our model), no matter what the true channel law is. In the special case of noise sequence channels, we obtain universality w.r.t. any competing scheme that is informed of the empirical distribution of the noise sequence in advance. Of course, more complex models for channel actions can be considered. For instance, one can model the actions as being modulo-additive with some Markovian statistical dependence, or as being memoryless but input dependent. These more elaborate models allow to universally approach suitably defined (and possibly higher) empirical capacities.
3) The empirical capacity over the family $\mathsf{M}_{\mathcal{X}}$ is achieved essentially without common randomness, since the negligible amount nevertheless required can be generated via feedback. However, when communicating over the family $\mathsf{C}_{\mathcal{X}}$, we require common randomness that cannot be accommodated by feedback. As described in Appendix C, this randomness is chiefly used for dithering, i.e., making the input distribution uniform. This is not merely an artifact, but has to do with the fact that the empirical capacity is defined in terms of a noise sequence, and for channels in $\mathsf{C}_{\mathcal{X}}$ the empirical distribution of the realized noise sequence depends on the empirical input distribution, a dependence which a modulo-additive model cannot capture. For instance, consider the extreme case of a binary channel where the channel's output at each time point is randomly chosen in an i.i.d. fashion to be $\sim \mathrm{Ber}(\varepsilon)$, independently of the inputs. This memoryless channel is in $\mathsf{C}_{\{0,1\}}$ (but not in $\mathsf{M}_{\{0,1\}}$), and its capacity is of course zero. Suppose one tries to communicate over this channel using some transmission scheme, and at the end of transmission the empirical distribution of the inputs turns out to be q. Then with high probability, the empirical distribution of the realized noise sequence will be close to $q * \varepsilon \triangleq q(1-\varepsilon)+(1-q)\varepsilon$, and the empirical capacity will therefore be close to $1 - h_B(q * \varepsilon)$ which is positive for $q \neq \frac{1}{2}$. This example demonstrates that for the family $\mathsf{C}_{\mathcal{X}}$, unless the input distribution is guaranteed to be close to uniform with high probability, the empirical capacity as defined may not be the right quantity to look at.
4) When $\mathcal{W} \in \mathsf{M}_{\mathcal{X}}$ happens to be a memoryless channel, the empirical capacity converges a.s. to the classical capacity of the channel. This is of course generally untrue for memoryless channels $\mathcal{W} \in \mathsf{C}_{\mathcal{X}}$ (with dithering). It is straightforward that due to the modulo-additive modeling, the empirical capacity cannot exceed the mutual information of $\mathcal{W}$ with a uniform input, but in fact the penalty may even be larger. For example, consider a general binary memoryless channel $\mathcal{W} \in \mathsf{C}_{\{0,1\}}$, described by

$$W_k(j \,|\, x_k = i, x^{k-1}, y^{k-1}) = p_{ij} \quad i, j \in \{0, 1\}.$$

With a uniform input (obtained via dithering), the empirical capacity will converge a.s. to $1 - h_B(\frac{1}{2}(p_{00} + p_{11}))$. This quantity is the capacity of a BSC obtained by averaging the channel $\mathcal{W}$ with its "cyclicly shifted" counterpart, which is a binary memoryless channel characterized by the transition probabilities $q_{ij} = p_{i+1,j+1}$ (modulo-addition). By the convexity of the mutual information in the transition matrix, and due to the symmetry between $\mathcal{W}$ and its cyclically shifted counterpart, the capacity of this BSC is upper-bounded by the mutual information of $\mathcal{W}$ with a uniform input. Furthermore, unless $\mathcal{W}$ happens to be a BSC to begin with, this inequality is strict and the empirical capacity is a.s. strictly smaller than the mutual information of $\mathcal{W}$ with a uniform input. The discussion is easily extended to larger alphabets.

IV. The Universal Scheme

In this section, we introduce a finite-horizon transmission scheme achieving the empirical capacity over the family $\mathsf{M}_{\mathcal{X}}$. We find it instructive to focus our discussion on this setting, as it is simpler yet includes all the core ideas. The more exhaustive horizon-free scheme and its extension to the larger family $\mathcal{C}_{\mathcal{X}}$ using common randomness, are discussed in Appendices B and C. We start by building intuition for the binary alphabet case, followed by a step-by-step construction of a binary alphabet universal scheme. This scheme is then generalized to a finite alphabet setting via some simple modifications. The rate and error probability analysis of the scheme appears in Section V. In this section, transmission is assumed to take place over a fixed period of n channel uses.

A. The Horstein Scheme for the BSC

We first discuss the simple case where the channel in use is known to be a BSC with a given crossover probability p, which in our terminology means a noise sequence channel (i.e., one satisfying the Markov relation (7)) with an i.i.d. $\sim \mathrm{Ber}(p)$ noise sequence Z^∞. For this setting, we describe the well-known passive feedback transmission scheme proposed by Horstein [4]. In that scheme, the message point is assumed to be selected at random uniformly over the unit interval. The receiver constantly calculates the *a posteriori* probability distribution of the message point given the bits it has seen so far. These bits are passively fed back to the transmitter (namely, $U_k = Y_k$ in our terminology), which can therefore calculate the posterior as well. A zero or one is transmitted according to whether the message point currently lies to the left or to the right of the posterior's median point. Thus, the transmitter always answers the most informative yes/no question that can be posed by the receiver.

Specifically, let Θ_0 be the random message point and denote its posterior density at time k (given the observed outputs) by $f_k(\theta) \triangleq f_{\Theta_0|Y^k}(\theta|y^k)$ for $\theta \in [0,1)$. Denote the median point corresponding to $f_k(\theta)$ by μ_k. Since Θ_0 is uniform over the unit interval, we have $f_0(\theta) = \mathbb{1}_{[0,1)}(\theta)$ and $\mu_0 = \frac{1}{2}$. The transmission functions are hence given by

$$g_k(\theta_0, y^{k-1}) = \begin{cases} 0, & \theta_0 < \mu_{k-1} \\ 1, & \theta_0 > \mu_{k-1} \end{cases}$$

and the transition from $f_k(\theta)$ to $f_{k+1}(\theta)$ is given by

$$f_{k+1}(\theta) = \begin{cases} 2(py_{k+1} + q(1-y_{k+1}))f_k(\theta), & \theta < \mu_k \\ 2(qy_{k+1} + p(1-y_{k+1}))f_k(\theta), & \theta > \mu_k \end{cases}$$

where $q = 1-p$. The transition from $f_k(\theta)$ to $f_{k+1}(\theta)$ and the corresponding transmission of $X_k = g_k(\theta_0, Y^{k-1})$ are referred to in the sequel as a *Horstein iteration*. Several optimal decoding rules are associated with the Horstein scheme. A fixed rate R rule is to decode the binary interval of size $2^{-\lfloor nR \rfloor}$ with the maximal posterior probability, which is our notations reads

$$\begin{aligned}\Delta_n(y^n, u^{n-1}) &= \Delta_n(y^n) \\ &= \mathrm{argmax}_{I \in \mathfrak{J}, |I| = 2^{-\lfloor nR \rfloor}} \int_I f_{\Theta_0|Y^n}(\theta|y^n)d\theta\end{aligned}$$

A variable rate rule with a target error probability p_e is to decode the smallest binary interval with a posterior probability exceeding a threshold $1-p_e$. There is also the bit-level decoding rule in which a bit is decoded whenever its corresponding binary interval has accumulated a posterior probability greater than $1-p_e$, where p_e is a target probability of bit error.[6] The Horstein scheme has been long conjectured to achieve the capacity of the BSC with either a fixed or a variable rate decoding rule, but this fact was proved in rigor only recently [27].

B. Binary Channels With Noise Constraints

Let us now take a step towards the unknown channel setting by considering a subfamily of $M_{\{0,1\}}$, where the empirical distribution of the noise sequence is known in advance to a.s. satisfy $p_{\mathrm{emp}}(Z^n) < p < \frac{1}{2}$ (e.g., an individual noise sequence with a fractions of "1's" smaller than p). From an adversarial point of view, this can be thought of as imposing a power constraint on the adversary. A plausible idea would be to communicate by performing Horstein iterations using p in lieu of the crossover probability, hoping that the average performance of the scheme in the BSC setting will carry over to this more stringent setting, i.e., enable to achieve $1-h_B(p)$ uniformly over the noise-constrained family. Unfortunately, this is not the case since Berlekamp's results [15] imply that for many values of p there exist pairs of message points and individual noise sequences (satisfying the constraint) for which decoding will surely fail. Nevertheless, as we now show, there is little information missing at the receiver to get it right.

We now make two key observations regarding the Horstein transmission process. First, we notice that $f_k(\theta)$ is a quasi-constant function, over at most $k+1$ distinct intervals whose union is the unit interval. Second, when transmission is terminated after n channel uses, we have

$$\begin{aligned}f_n(\theta = \theta_0) &= 2^n(1-p)^{n_0(Z^n)}p^{n_1(Z^n)} \\ &= 2^n(1-p)^{n(1-p_{\mathrm{emp}}(Z^n))}p^{np_{\mathrm{emp}}(Z^n)} \quad (13)\end{aligned}$$

where θ_0 is the message point. This stems directly from the fact that θ_0 is always on the correct side of the median, so that its density is multiplied by $2(1-p)$ when there is no error, and by $2p$ otherwise. Now, let the *message interval* at time k be the interval containing θ_0 over which $f_k(\theta)$ is constant, and let $2^{-\ell}$ be its length at the end of transmission, for some $\ell > 0$. Using (13) we have that

$$\begin{aligned}2^{-\ell} \cdot f_n(\theta = \theta_0) \le 1 \Rightarrow \ell \\ \ge n(1 - h_B(p_{\mathrm{emp}}(Z^n)) - D(p_{\mathrm{emp}}(Z^n)\|p).\end{aligned}$$

Now, assume the decoder could identify with certainty which is the message interval at the end of transmission. In that case, the common most significant bits in the binary expansion of points inside the message interval (which also correspond to the message point itself) could be decoded, error free! This means that an instantaneous decoding rate of

$$R_n = \frac{\lfloor \ell \rfloor}{n} \ge 1 - h_B(p_{\mathrm{emp}}(Z^n)) - D(p_{\mathrm{emp}}(Z^n)\|p) - \frac{1}{n} \quad (14)$$

could be attained. Actually, another information bit is required to allow the above rate, as the message interval may sometimes be inconveniently located over the binary grid and not enough bits (if any) can be decoded. We further elaborate on this point in

[6] For instance, when the posterior probability (w.r.t. $f_k(\theta)$) of either $[0, \frac{1}{2})$ or $[\frac{1}{2}, 1)$ exceeds $1-p_e$, the most significan bit (MSB) of the message point is decoded as either 0 or 1, respectively.

Section IV-D, when the full scheme is presented. Notice that the expression in (14) can be divided into two parts: $1-h_B(p_{\mathrm{emp}})$ is the empirical capacity of the channel, and $D(p_{\mathrm{emp}}\|p)$ is a penalty term for using the maximal value p of p_{emp} instead of p_{emp} itself. Note also that

$$\inf_{p_{\mathrm{emp}}<p} R_n \geq 1-h_B(p)-\frac{1}{n} \tag{15}$$

and therefore the rate attained by this variable rate scheme (had the message interval been known at the end of transmission) is guaranteed to be asymptotically no less than $1-h_B(p)$, for *any* message point $\theta_0\in[0,1)$.

As observed before, there are at most $n+1$ distinct intervals over which $f_n(\theta)$ is constant, therefore, no more than $\lceil\log(n+1)\rceil$ bits of side information are required in order to identify the message interval at the end of transmission. This means that the decoding rate in (14) is achievable (error free) if only $1+\lceil\log(n+1)\rceil$ bits could be reliably conveyed to the receiver at the end of transmission. Thus, while [15] determines that it is generally impossible to communicate at such rate with no errors, the size of the decoding uncertainty preventing us from attaining it is very small. For instance, if list decoding is allowed then (15) could be attained using a list whose size grows only linearly (and not exponentially) with n.

In the following subsections, we present a simple randomization technique by which these extra bits can be reliably conveyed to the receiver, with no asymptotic decrease in the data rate, so that the decoder can determine with high probability the correct message from the list. This technique requires a sublinear number of random bits shared by the transmitter and the receiver (obtained via feedback, or possibly via a common random source). Moreover, through a sequential use of randomness we will be able to present a feasible transmission scheme that tracks the empirical distribution of the realized noise sequence, so that a significantly higher rate approaching the *empirical capacity* $1-h_B(p_{\mathrm{emp}}(Z^n))$ is attained, avoiding the penalty term in (14).

C. Sequential Probability Estimation

We now turn to the general case where the channel in use is an arbitrary unknown member of $M_{\{0,1\}}$, i.e., from the adversarial point of view there are no constraints (besides causality) on the way the noise sequence is generated by the adversary (e.g., the noise may be some unknown individual sequence). A reasonable idea could be to plug in a sequential estimator for the empirical distribution of the noise $p_{\mathrm{emp}}(Z^k)$ into the Horstein iterations, that is, to let the "crossover probability" used by the receiver to calculate the so-called "posteriori distribution" of the message point, vary with time. Specifically, this amounts to

$$f_{k+1}(\theta)=\begin{cases}2(\widehat{p}_{k+1}(Z^k)y_{k+1}+\widehat{q}_{k+1}(Z^k)(1-y_{k+1}))f_k(\theta), & \theta<\mu_k\\ 2(\widehat{q}_{k+1}(Z^k)y_{k+1}+\widehat{p}_{k+1}(Z^k)(1-y_{k+1}))f_k(\theta), & \theta>\mu_k\end{cases}$$

where $\widehat{p}_{k+1}$ is a sequential estimator applied to the noise sequence Z^k and $\widehat{q}_{k+1}\triangleq 1-\widehat{p}_{k+1}$. Note that $f_k(\theta)$ is still a probability density function, but looses the meaning of a true posterior in this unknown channel setting. We henceforth loosely refer to $f_k(\theta)$ as the *empirical posterior* of the message point (relative to the estimator in use).

This idea is of course problematic, since the noise sequence is causally known only to the transmitter and not to the receiver, but for the moment let us assume that the estimates can be somehow made known to the receiver, and take care of this point later. The first of the two key observations from the previous subsection still holds, i.e., $f_k(\theta)$ is quasi-constant over $k+1$ disjoint intervals whose union is the unit interval. The empirical posterior evaluated at the message point at the end of transmission is now equal to

$$f_n(\theta=\theta_0)=2^n\prod_{k=1}^{n}\widehat{p}_k^{Z_k}(1-\widehat{p}_k)^{1-Z_k}=2^n\widehat{p}(Z^n)$$

where $\widehat{p}(Z^n)$ is the probability assigned to the entire noise sequence Z^n by the estimator in use. Using this fact and assuming again that at the end of transmission we know which one of the intervals is the message interval (which is then set to be the decoded interval), the instantaneous decoding rate attained is given by

$$R_n\geq\frac{1}{n}\lfloor\log 2^n\widehat{p}(Z^n)\rfloor\geq 1-\frac{1}{n}\log\widehat{p}(Z^n)-\frac{1}{n}$$

so the shorter code-length assigned by the estimator to the noise sequence, or the more *compressible* the strategy of the adversary is (w.r.t. a memoryless modulo-additive model), the higher the achieved rate. It is therefore only reasonable to make use of the KT estimator, or more generally an intermittently updated $\mathrm{KT}(b)$ estimator. Applying Lemmas 2 and 3, the instantaneous decoding rate achieved when using a $\mathrm{KT}(b)$ estimator is

$$\begin{aligned}R_n&\geq 1-\frac{1}{n}\log\widehat{p}^{\mathrm{KT}(b)}(Z^n)-\frac{1}{n}\\&\geq 1-h_B(p_{\mathrm{emp}}(Z^n))-K_1\frac{b\log n}{n}\\&=C_n^{\mathrm{emp}}(\mathcal{W},\theta_0)-K_1\frac{b\log n}{n}\end{aligned} \tag{16}$$

where $K_1>0$ is constant. Thus, if $bn^{-1}\log n=\mathrm{o}(1)$ the empirical capacity is asymptotically attained. This holds however only under the assumptions that the receiver knows the $\mathrm{KT}(b)$ estimates online, and can also recognize the message interval with certainty at the end of transmission.

There are two key elements that make this approach work. First, the update information required by the receiver so that the assumptions above are satisfied can be made negligible, namely, have rate zero. The message interval is one of at most $n+1$ possible intervals hence requires only $\lceil\log(n+1)\rceil$ bits, and using a $\mathrm{KT}(b)$ estimator only requires to communicate the number of "1's" in the last b channel uses, which requires $\lceil\log b\rceil$ bits per b channel uses and is negligible if $b^{-1}\log b=\mathrm{o}(1)$. Note the core tradeoff between a small b required to obtain a small redundancy term in (16), and a large b required to make the update information rate negligible. Second, as we shall see, it is possible to obtain reliable zero rate communications over an unknown member of $M_{\{0,1\}}$ as long as the empirical capacity is not too small, and that the latter condition can be identified with high probability. These two observations allow us to make the seemingly unfeasible approach described so far into a practical scheme that achieves the empirical capacity.

D. A Universal Binary Alphabet Scheme

In this subsection, we introduce the universal scheme achieving the empirical capacity for the binary alphabet, finite-horizon case. Let us first provide a rough outline of the scheme. Transmission takes place over a period of n channel uses, which is divided into blocks of equal length $b = b(n)$. Inside each block, *Horstein iterations* are performed over the majority of channel uses, always using the most updated KT estimate. *Update information* containing the number of “1’s” in the previously accepted block together with the index of the current message interval, is coded using a repetition code and passed to the receiver over randomly selected positions inside the block, which are selected via feedback. The idea is that since positions are random, the “effective” channel for the update information transmission is roughly a BSC with transition probabilities close to the empirical distribution of the noise sequence inside the block. This distribution is estimated using a randomly positioned training sequence, and if the estimation is too close to being uniform, the block is discarded. Otherwise, the update information can be reliably decoded with high probability. Loosely speaking, the discarding process partitions the noise sequence into a “good” part and a “bad” part, and with high probability the empirical capacity of the latter part is small. Therefore, discarding the “bad” part increases the rate with high probability, due to the concavity of the entropy.

If a block is accepted then the polarity of the “effective” crossover probability (i.e., above/below $\frac{1}{2}$) can be reliably determined, and hence the update information (which has a negligible rate) can be reliably decoded. Once the update information is successfully decoded, the receiver uses the number of “1’s” in the noise sequence from the previous block to update the KT estimate, which is then used for communications in the next block. At the end of transmission, the last known message interval is set to be the decoded interval.

What takes place inside each block is now described in detail. We define four types of positions within the block—*regular positions* over which Horstein iterations are performed, *training positions* over which a training sequence is transmitted, *update positions* over which update information is transmitted, and *active feedback positions* used to select the random positions for the other types. The non-active feedback positions (regular, training, and update) are referred to as *passive feedback positions*. Apart from active feedback positions, the receiver passively feeds back what it receives (i.e., $U_k = Y_k$ over these positions).

(A) **Random positions generation (active feedback)**: We set a parameter $m = m(n)$ which will indirectly determine the number of nonregular positions. The active feedback positions are always at the beginning of the block, and occupy exactly $b_a = b_a(n)$ positions where b_a is determined in the sequel as a function of m, b. The active positions are used in order to synchronize the terminals regarding the type of each passive position that follows. The number of passive positions is fixed and given by $b_p \triangleq b - b_a$. The type of each of the b_p passive positions is determined by an i.i.d. sequence Λ^{b_p} over the alphabet $\{\mathrm{training}, \mathrm{update}, \mathrm{regular}\}$ with a marginal distribution given by $(\frac{m}{b_p}, \frac{m}{b_p}, 1-\frac{2m}{b_p})$. The selection of Λ^{b_p} is synchronized between the transmitter and the receiver as follows:

- (A1) The receiver randomly selects the sequence Λ^{b_p} according to the i.i.d. distribution above. Let the r.v.’s (M_t, M_u, M_r) denote the number of occurrences of the corresponding symbols in Λ^{b_p}, and note that $\mathbb{E}(M_t, M_u, M_r) = (m, m, b_p - 2m)$.
- (A2) The type of the sequence[7] Λ^{b_p} is then binary encoded and sent via feedback over active positions, which requires no more than $2\lceil \log b \rceil$ bits.
- (A3) If both $\frac{m}{2} \leq M_t, M_u \leq 2m$ (not too little or too many training and update positions) then the index of the sequence Λ^{b_p} within its type[7] is communicated via the feedback (active positions), which requires no more than $4m\lceil \log b \rceil$ bits. Otherwise, the transmitter overrides the receiver’s selection, and randomly selects the sequences Λ^{b_p} itself.

Now, another sequence Γ^{M_u} is selected, determining which of the update bits is to be transmitted over which update position (with repetitions). The number of update bits (see step (C) below) is $2\lceil \log(n+1) \rceil$, hence Γ^{M_u} is selected in a uniform i.i.d. fashion over the alphabet $\langle 2\lceil \log(n+1) \rceil \rangle$. If both $\frac{m}{2} \leq M_t, M_u \leq 2m$, then Γ^{M_u} is selected by the receiver, binary encoded using no more than $2m\lceil 1 + \log\lceil \log(n+1) \rceil) \rceil$ bits, and sent via feedback over active positions. Otherwise, Γ^{M_u} is selected by the transmitter.

Let us now assume that $b > \log(n+1)$, and set the total number of active positions to $b_a \triangleq 8m\lceil \log b \rceil$, which is sufficient to accommodate the synchronization process described above. If $b^{-1} m \log b = \mathrm{o}(1)$, then this amount becomes negligible and the feedback strategy is asymptotically passive. Fig. 2 depicts a “typical” position assignment within a block.

(B) **Training transmission**: A training sequence is transmitted over the M_t random positions as determined by Λ^{b_p}. At the end of the block, the receiver calculates the *training estimate* p^{train} for the empirical distribution of the noise in the block, which is a coarse estimate later used for block discarding. Let $\tilde{Z}^{b_p}$ denote the noise sequence over passive positions within the current block, i.e., if this is the kth block then $\tilde{Z}^{b_p} = Z_{bk-b_p+1}^{bk}$. Let B^{b_p} be the corresponding *training pattern sequence*, i.e., $B_k = \mathbb{1}_{\mathrm{training}}(\Lambda_k)$. The training estimate is set to

$$\begin{aligned} \boldsymbol{p}^{\mathrm{train}} &\triangleq \alpha(B^{b_p})\boldsymbol{p}_{\mathrm{emp}}(\tilde{Z}^{b_p} \downarrow B^{b_p}) \\ &= \left(\frac{M_t}{m}\right) \cdot \boldsymbol{p}_{\mathrm{emp}}(\tilde{Z}^{b_p} \downarrow B^{b_p}) \end{aligned} \tag{17}$$

where the α-normalization factor is defined in (4).

(C) **Update transmission**: Update information is transmitted over the M_u random positions determined by Λ^{b_p}. The

[7] Note that by *type* of a sequence we refer to the vector of symbols occurrences, not to be confused with position types.

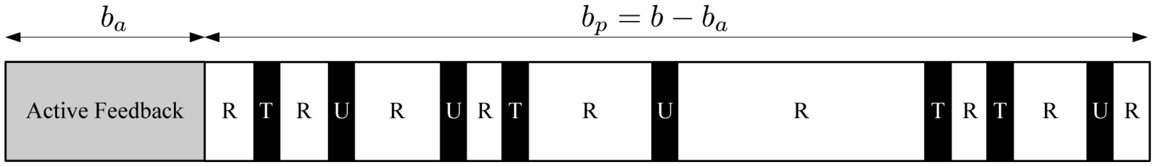

Fig. 2. A block always begins with b_a active feedback positions, followed by a bulk of **R**egular positions that are randomly replaced with (on the average) m **T**raining positions and m **U**pdate positions.

uncoded update information includes the following quantities, all binary encoded:

(C1) The number of "1's" in the noise sequence over regular positions in the previously accepted block ($\lceil \log b \rceil$ bits).

(C2) The index of the message interval w.r.t. the interval partitioning of the empirical posterior at the end of the previously accepted block ($\lceil \log{(n+1)} \rceil$ bits at the most).

(C3) One *ambiguity resolving bit* which is discussed later on.

The number of uncoded update bits in total is therefore no more than $2\lceil \log(n+1) \rceil$, and for simplicity we assume that exactly $2\lceil \log(n+1) \rceil$ uncoded update bits are to be transmitted (and e.g., zero pad if necessary). Now, on the kth update position (determined by Λ^{b_p}) the transmitter sends the Γ_kth uncoded update bit. By properly tuning the scheme parameters we typically have $M_u \gg \log n$, hence each update bit is coded using a repetition code with a random number of repetitions.

(D) **Horstein iterations with $\mathrm{KT}(b)$ estimates**: Horstein iterations are performed over the (random) M_r regular positions as determined by Λ^{b_p}. The "crossover probability" used is the most updated KT estimate of the empirical noise distribution available to the receiver. On the kth block, this estimate is given by

$$\widehat{p}^{(k)} = \frac{\frac{1}{2} + \sum_{j=1}^{k-2} \widehat{n}_1(j) I(j)}{1 + \sum_{j=1}^{k-2} M_r(j) I(j)},$$

where $\widehat{n}_1(j)$ is the number of "1's" the receiver assumes appeared in the noise sequence over regular positions in the jth block, as communicated by the update information so far (may be different than the actual number due to errors), $M_r(j)$ is the value of M_r (number of regular positions) on the jth block, and $I(j)$ is an indicator function that evaluates to one if the jth block was accepted, and to zero if it was discarded. Note that the estimator works on the sequence of accepted positions, and is always two accepted blocks behind.

(E) **Block discarding**: The block is discarded if either M_t, M_u are out of range (in the sense of (A3)), or if

$$\|\boldsymbol{p}^{\mathrm{train}} - \boldsymbol{p}_u\|_\infty < \tau_d \tag{18}$$

for some *discarding threshold* $\tau_d(n) = \mathrm{o}(1)$, where $\boldsymbol{p}_u$ is the uniform distribution over $\{0, 1\}$. Otherwise, the block is accepted. When a block is discarded, the transmitter and receiver return to the state they were in before the block has started.

(F) **Update information decoding**: For an accepted block, the update information is decoded according to the estimated noise probability, as follows. Let $\tilde{Y}^{b_p}$ denote the channel output sequence over passive positions within the current block, i.e., if this is the kth block then $\tilde{Y}^{b_p} = Y_{bk-b_p+1}^{bk}$. Let $B_{(i)}^{b_p}$ be the *repetition pattern sequence* of the ith update bit (as determined by Λ^b, Γ^{M_u}), i.e., a binary sequence with "1's" only in update positions that correspond to a repetition of that bit. For any $i \in \langle 2\lceil \log n \rceil \rangle$, the receiver calculates the following *update estimate* for the ith bit:

$$\begin{aligned}\boldsymbol{p}^{\mathrm{upd},i} &\triangleq \alpha\left(B_{(i)}^{b_p}\right) \boldsymbol{p}_{\mathrm{emp}}\left(\tilde{Y}^{b_p} \downarrow B_{(i)}^{b_p}\right) \\ &= \left(\frac{n_i(\Gamma^{M_u})}{m/(2\lceil \log(n+1) \rceil)}\right) \cdot \boldsymbol{p}_{\mathrm{emp}}\left(\tilde{Y}^{b_p} \downarrow B_{(i)}^{b_p}\right)\end{aligned}$$

where α-normalization is used again. Intuitively, we expect the update estimate to be close to the training estimate only when the corresponding update bit was a "0," unless the noise sequence within the block is close to uniform in which case it is likely to be discarded anyway. Accordingly, the decision rule for the ith update bit is given by

$$\|\boldsymbol{p}^{\mathrm{upd},i} - \boldsymbol{p}^{\mathrm{train}}\|_\infty \underset{1}{\overset{0}{\lesseqgtr}} \tau_u \tag{19}$$

for some *update decision threshold* $\tau_u(n) = \mathrm{o}(1)$, where in case of an equality a "0" is decoded. The decoded information is used to update the KT estimate, and to store the new identity of the message interval.

Decoding Rule: Ideally, when transmission ends one would like to decode the minimal binary interval (i.e., its corresponding MSBs) containing the last message interval given by the update information, which (if no error has occurred) contains the message point. However, as happens in arithmetic coding, sometimes this minimal binary interval is much larger than the message interval itself (for instance, if the message interval contains the point $\frac{1}{2}$ then we cannot decode even a single bit). To solve this, note that it is possible to divide the interval $[0, 1)$ into binary intervals of size corresponding to the message interval's size, such that the message interval intersects no more than two of those, and the only uncertainty that may be left is which one. The *ambiguity resolving bit* mentioned in step (C) above is used to that end. The receiver thus uses the following decoding rule:[8] Seek the two smallest adjacent binary intervals (of the same size) whose union contains the last known message interval, and decode one of them according to the last known ambiguity resolving bit. This rule guarantees that the decoded interval is less than twice the size of the last message interval.

As we show in Section V, by properly selecting the dependence of the scheme parameters on n, the update information is guaranteed to be correctly decoded with probability approaching one without causing any asymptotic decrease in the data rate, thus allowing the empirical capacity to be approached.

E. A Universal Finite-Alphabet Scheme

We now describe the (finite-horizon) finite alphabet $\mathcal{X}$ variant of the universal scheme. Suppose the transmitter and the receiver can agree on a sequential $\mathrm{KT}(b)$ estimator $\widehat{p}_k^{\mathrm{KT}(b)}(\cdot|Z^{k-1})$ for the noise sequence at each time k (as in Section IV-C). Horstein iterations using this estimate are performed as follows. The empirical posterior is initialized as before to a uniform distribution over the unit interval $f_0(\theta) = \mathbb{1}_{[0,1)}(\theta)$. At each time point k, unit interval is divided into $|\mathcal{X}|$ consecutive subintervals with identical probability $|\mathcal{X}|^{-1}$ under the empirical posterior $f_{k-1}(\theta)$, and the transmitter sends a symbol that corresponds to the subinterval containing the message point θ_0. Upon receiving $Y_k \in \mathcal{X}$, the receiver generates the new empirical posterior $f_k(\theta)$ by multiplying $f_{k-1}(\theta)$ in the interval corresponding to the symbol $i \in \mathcal{X}$ by the factor $|\mathcal{X}|\widehat{p}_k^{\mathrm{KT}(b)}(Y_k - i|Z^{k-1})$, where the minus sign is the modulo-subtraction operator.[9] Thus, the message point is always in the interval multiplied by the estimate corresponding to the next value of the noise, and hence we have

$$f_n(\theta = \theta_0) = |\mathcal{X}|^n \widehat{p}^{\mathrm{KT}(b)}(Z^n)$$

where $\widehat{p}^{\mathrm{KT}(b)}(Z^n)$ is the probability assigned to the entire noise sequence by the $\mathrm{KT}(b)$ estimator. Similarly to the binary case, there are no more than $n(|\mathcal{X}|-1)+1$ subintervals over which $f_n(\theta)$ is constant. Assuming further that the message interval index is known to the receiver at the end of transmission, then using Lemmas 2 and 3 for the $\mathrm{KT}(b)$ redundancy, the achieved rate is given by

$$\begin{aligned} R_n &\geq \log|\mathcal{X}| - H_{\mathrm{emp}}(Z^n) - K_2\frac{|\mathcal{X}|b\log n}{n} \\ &= C_n^{\mathrm{emp}}(\mathcal{W},\theta_0) - K_2\frac{|\mathcal{X}|b\log n}{n} \end{aligned} \tag{20}$$

for some $K_2 > 0$. Once again, the update information required by the receiver for the above assumptions to hold incurs in a vanishing rate penalty, and can be reliably transmitted over random positions if the empirical capacity is not too small.

The finite alphabet scheme now follows a similar path to that of the binary alphabet scheme. The transmission period is divided into blocks of equal size $b = b(n)$, and inside each block we again have the active feedback, training, update, and regular positions. However, the update information transmission and decoding is somewhat more involved in this case. We now describe what happens inside each block.

(A) **Random position generation (active feedback)**: We use the same parameters b, m, b_a, generate the corresponding r.v.'s M_t, M_u, M_r, and pick the sequence Λ^{b_p} the same way.[10] However, the sequence Γ^{M_u} is now selected uniformly over an alphabet $\langle(|\mathcal{X}|-1)s\rangle$, where $s = s(n)$ corresponds to a larger number of update bits, and is defined in step (C) below. Again, apart from active feedback positions the receiver passively feeds back what it receives.

(B) **Training transmission**: The training estimate $\boldsymbol{p}^{\mathrm{train}}$ is given by (17).

(C) **Update transmission**: Update information is transmitted over the M_u random positions determined by Λ^{b_p}. The uncoded update information includes the type (symbol occurrences) of the noise sequence over regular positions in the previously accepted block, the index of the message interval w.r.t. the interval partitioning of the empirical posterior at the end of that block, and one ambiguity resolving symbol. Using a binary representation, the total number of uncoded update bits is no more than

$$\begin{aligned} &\left\lceil \log b^{(|\mathcal{X}|-1)} \right\rceil + \lceil \log\left(n(|\mathcal{X}|-1)+1\right)\rceil \\ &\qquad +1 \leq 2(|\mathcal{X}|-1)\lceil \log(n+1)\rceil \triangleq s(n) \end{aligned} \tag{21}$$

and again we zero-pad the uncoded update bits up to the length s above, for simplicity. For a nonbinary alphabet, using a random "repetition code" method similar to the one used in the binary alphabet case may result in a decoding ambiguity,[11] and thus a different method must be used. The sequence Γ^{M_u} takes values over an alphabet $\langle(|\mathcal{X}|-1)s\rangle$, so we can write for any k, $\Gamma_k = i + js$ for some $i \in \langle s\rangle$, and $j \in \langle|\mathcal{X}|-1\rangle$. Following this representation, in the kth update position we send one of the symbols from the pair $\{0, j+1\}$, where which one is determined by the ith uncoded update bit.[12] For a suitable selection of parameters, this procedure guarantees (with high probability) that any of the uncoded update bits is sent several times using pairs of channel inputs with any possible modulo-additive distance. This in turn guarantees that any bit is resolvable with high probability via at least one of the pairs, unless the empirical capacity of that block is close to zero.

(D) **Horstein iterations with $\mathbf{KT}(b)$ estimates**: Similar to the binary alphabet case. Horstein iterations are performed over the (random) M_r regular positions determined by Λ^{b_p}, using the most updated KT estimate of

[8]If all blocks were discarded, the decoded interval is trivially taken to be $[0,1)$.

[9]The probability of the ith interval under $f_k(\theta)$ is exactly $\widehat{p}_k^{\mathrm{KT}(b)}(Y_k - i|Z^{k-1})$, hence $f_k(\theta)$ is a probability distribution.

[10]Note that it is possible to reduce the number of active positions b_a since the feedback has a larger capacity now, but this has a negligible effect, and for consistency we refrain from doing so.

[11]This occurs when the empirical distribution of the noise inside the block is invariant under some cyclic shift. Take for example the distribution $(0.4, 0.1, 0.4, 0.1)$ over a quaternary alphabet, in which case one cannot separate, say, the all-"0's" and the all-"2's" repetition words, but the empirical capacity is nevertheless positive. Even in the simple modulo-additive DMC setting with the above noise distribution, one would use only two inputs to attain capacity, say the first and the second.

[12]We use "0's"$\rightarrow \{0\}$,"1's" $\rightarrow \{j+1\}$.

the empirical noise distribution (over accepted blocks) available to the receiver.

(E) **Block discarding**: Same criterion as in the binary alphabet case, where now $\boldsymbol{p}_u$ in (18) is taken to be the uniform distribution over $\mathcal{X}$.

(F) **Update information decoding**: For an accepted block, the update information is decoded using $\boldsymbol{p}^{\mathrm{train}}$ as follows. Let $B^{b_p}_{(i,j)}$ be the repetition pattern sequence of the ith update bit using the inputs $\{0, j+1\}$, as determined by Λ^b, Γ^{M_u}. For any $i \in \langle s \rangle$ and $j \in \langle |\mathcal{X}| - 1 \rangle$, the receiver calculates the following update estimate:

$$\begin{aligned}\boldsymbol{p}^{\mathrm{upd}}_{(i,j)} &\triangleq \alpha\left(B^{b_p}_{(i,j)}\right)\boldsymbol{p}_{\mathrm{emp}}\left(\tilde{Y}^{b_p} \downarrow B^{b_p}_{(i,j)}\right)\\ &= \left(\frac{n_{i+js}(\Gamma^{M_u})}{m/((|\mathcal{X}|-1)s)}\right) \cdot \boldsymbol{p}_{\mathrm{emp}}\left(\tilde{Y}^{b_p} \downarrow B^{b_p}_{(i,j)}\right).\end{aligned}$$

The decoding rule for the ith update bit is given by

$$\max_{j \in \langle |\mathcal{X}|-1 \rangle} \left\| \boldsymbol{p}^{\mathrm{train}} - \boldsymbol{p}^{\mathrm{upd}}_{(i,j)} \right\|_\infty \underset{1}{\overset{0}{\lesseqgtr}} \tau_u \tag{22}$$

where in cases of an equality a "0" is decoded. The decoded information is used to update the KT estimate, and to store the new identity of the message interval.

The decoding rule used is the same one as in the binary setting, see Section IV-D. In the following section, we analyze the performance of the described scheme, proving it is universal for the family $M_{\mathcal{X}}$.

V. Analysis

In this section, we analyze the performance of the finite-alphabet finite-horizon scheme presented in Section IV-E, and show it achieves the empirical capacity in the limit of infinite horizon, for a suitable selection of the parameters $m(n), b(n), \tau_d(n), \tau_u(n)$. For brevity, the dependence of the parameters on n will usually be omitted. In the sequel, we also show how faster convergence to the empirical capacity is obtained when operating over noise sequence channels, and discuss the amount of randomness generated by the scheme. The following observation plays a key role in our subsequent derivations:

Lemma 6: For any specific block, let $\widetilde{Z}^{b_p}$, B^{b_p}, and $B^{b_p}_{(i,j)}$ be the corresponding noise sequence over passive positions, training pattern sequence, and repetition pattern sequence for the ith update bit with the input pair $\{0, j+1\}$, respectively. Then B^{b_p} and $B^{b_p}_{(i,j)}$ each constitute a causal sampling sequence for $\widetilde{Z}^{b_p}$.

Proof: See Appendix A. In short, the training/update positions are i.i.d. by construction, and causal independence is established by combining that with the Markov relations (9) and (6). □

A. Error Probability

The only source for error in our scheme lies in the incorrect decoding of update information in the last accepted block before transmission is terminated, which causes the wrong message interval to be decoded. However, for simplicity of the exposition, we leniently assume that an error is declared whenever the update information in any of the blocks is erroneously decoded.

The error probability is hence upper-bounded by the probability of erroneous update decoding in any of the blocks. Therefore, we now focus on a specific block and find the corresponding update decoding error probability, where it is emphasized that while discarding a block has impact on the rate, it does not constitute an error event. The noise sequence over passive positions in the block is denoted as before by $\widetilde{Z}^{b_p}$. For any $i \in \langle s \rangle$ and $j \in \langle |\mathcal{X}| - 1 \rangle$, define

$$\boldsymbol{p}_{(i,j)} \triangleq \alpha\left(B^{b_p}_{(i,j)}\right)\boldsymbol{p}_{\mathrm{emp}}\left(\tilde{Z}^{b_p} \downarrow B^{b_p}_{(i,j)}\right) \tag{23}$$

which is the counterpart of $\boldsymbol{p}^{\mathrm{upd}}_{(i,j)}$, yet sampling the noise sequence rather than the output sequence over the update positions corresponding to the ith update bit and the input pair $\{0, j+1\}$. Define the following two events:

$$\begin{aligned}E_1 &\triangleq \left\{ \left\| \boldsymbol{p}^{\mathrm{train}} - \boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p}) \right\|_\infty > \tau \right\}\\ E_2 &\triangleq \left\{ \max_{i \in \langle s \rangle, j \in \langle |\mathcal{X}|-1 \rangle} \left\| \boldsymbol{p}_{(i,j)} - \boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p}) \right\|_\infty > \tau \right\}.\end{aligned}$$

For some $\tau(n) = \mathrm{o}(1)$. We assert that for a suitable selection of the thresholds (τ_d, τ_u, τ), a necessary condition for an update decoding error in the block is given by the event $E_1 \cup E_2$. To see why this holds, let us assume the complementary event $E_1^c \cap E_2^c$ and show it implies no update decoding errors for a suitable thresholds selection. If the block was discarded then surely there is no error, so assume further the block was not discarded. Now consider the ith update bit. If this bit is a "0" then the channel input at the corresponding update positions (determined by $B^{b_p}_{(i,j)}$) is $0 \in \mathcal{X}$, and therefore $\boldsymbol{p}_{(i,j)} = \boldsymbol{p}^{\mathrm{upd}}_{(i,j)}$ for any $j \in \langle |\mathcal{X}| - 1 \rangle$. Thus, in this case we have

$$\begin{aligned}\left\| \boldsymbol{p}^{\mathrm{upd}}_{(i,j)} - \boldsymbol{p}^{\mathrm{train}} \right\|_\infty &= \left\| \boldsymbol{p}_{(i,j)} - \boldsymbol{p}^{\mathrm{train}} \right\|_\infty\\ &\le \left\| \boldsymbol{p}_{(i,j)} - \boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p}) \right\|_\infty\\ &\quad + \left\| \boldsymbol{p}^{\mathrm{train}} - \boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p}) \right\|_\infty \le 2\tau\end{aligned}$$

which holds for any $j \in \langle |\mathcal{X}| - 1 \rangle$. Therefore, if we set $2\tau \le \tau_u$ then the above together with the update decoding rule (22) imply that the ith update bit is correctly decoded. Now suppose the ith update bit is a "1," in which case the channel input at the corresponding update positions is $(j+1) \in \mathcal{X}$, and thus $\boldsymbol{p}^{\mathrm{upd}}_{(i,j)}$ is a *cyclic right-shift* of $\boldsymbol{p}_{(i,j)}$ by $j+1$ positions. Writing $\boldsymbol{p}^{(-j)}_{\mathrm{emp}}(\widetilde{Z}^{b_p})$ for a *cyclic left-shift* of $\boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p})$ by $j+1$ positions, we have the following chain of inequalities:

$$\begin{aligned}&\left\| \boldsymbol{p}^{\mathrm{upd}}_{(i,j)} - \boldsymbol{p}^{\mathrm{train}} \right\|_\infty\\ &\overset{(a)}{\ge} \left\| \boldsymbol{p}^{\mathrm{upd}}_{(i,j)} - \boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p}) \right\|_\infty - \left\| \boldsymbol{p}^{\mathrm{train}} - \boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p}) \right\|_\infty\\ &\overset{(b)}{\ge} \left\| \boldsymbol{p}^{\mathrm{upd}}_{(i,j)} - \boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p}) \right\|_\infty - \tau\\ &= \left(\left\| \boldsymbol{p}^{\mathrm{upd}}_{(i,j)} - \boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p}) \right\|_\infty + \left\| \boldsymbol{p}_{(i,j)} - \boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p}) \right\|_\infty \right)\\ &\quad - \left\| \boldsymbol{p}_{(i,j)} - \boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p}) \right\|_\infty - \tau\\ &\overset{(c)}{\ge} \left(\left\| \boldsymbol{p}_{(i,j)} - \boldsymbol{p}^{(-j)}_{\mathrm{emp}}(\widetilde{Z}^{b_p}) \right\|_\infty \right.\end{aligned}$$

$$+\left\|\boldsymbol{p}_{(i,j)}-\boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p})\right\|_{\infty}\Big)-2\tau$$
$$\stackrel{(\mathrm{d})}{\geq}\left\|\boldsymbol{p}_{\mathrm{emp}}^{(-j)}(\widetilde{Z}^{b_p})-\boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p})\right\|_{\infty}-2\tau. \tag{24}$$

In transition (a) we used the triangle inequality for the $\mathcal{L}_\infty$ norm, transition (b) holds since we assume E_1^c, in (c) we use E_2^c, and also replace a cyclic right-shift of one vector with a corresponding cyclic left-shift of the other vector inside the first $\mathcal{L}_\infty$ norm term. Finally, the triangle inequality is used once again in (d).

We can now maximize both sides of (24) over $j\in\langle|\mathcal{X}|-1\rangle$, to obtain

$$\begin{aligned}&\max_{j\in\langle|\mathcal{X}|-1\rangle}\left\|\boldsymbol{p}_{(i,j)}^{\mathrm{upd}}-\boldsymbol{p}^{\mathrm{train}}\right\|_{\infty}\\ &\quad\geq\max_{j\in\langle|\mathcal{X}|-1\rangle}\left\|\boldsymbol{p}_{\mathrm{emp}}^{(-j)}(\widetilde{Z}^{b_p})-\boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p})\right\|_{\infty}-2\tau\\ &\quad\stackrel{(\mathrm{a})}{=}\max\left(\boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p})\right)-\min\left(\boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p})\right)-2\tau\\ &\quad\stackrel{(\mathrm{b})}{>}\left\|\boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p})-\boldsymbol{p}_u\right\|_{\infty}-2\tau\end{aligned} \tag{25}$$

where $\max(\cdot),\min(\cdot)$ return the maximal and minimal element of the vector argument, respectively. Transition (a) holds since the maximization is over the $\mathcal{L}_\infty$ distance between a vector and all its cyclic shifts, and for (b) to hold with a strict inequality we further assume that $\boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p})$ is not precisely uniform, which is satisfied by setting $\tau<\tau_d$, since

$$\begin{aligned}\tau_d&\leq\left\|\boldsymbol{p}^{\mathrm{train}}-\boldsymbol{p}_u\right\|_{\infty}\\ &\leq\left\|\boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p})-\boldsymbol{p}_u\right\|_{\infty}+\left\|\boldsymbol{p}^{\mathrm{train}}-\boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p})\right\|_{\infty}\\ &\leq\left\|\boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p})-\boldsymbol{p}_u\right\|_{\infty}+\tau.\end{aligned}$$

Finally, combining the above with (25) we obtain

$$\max_{j\in\langle|\mathcal{X}|-1\rangle}\left\|\boldsymbol{p}_{(i,j)}^{\mathrm{upd}}-\boldsymbol{p}^{\mathrm{train}}\right\|_{\infty}>\tau_d-3\tau.$$

If we set $\tau_d-3\tau\geq\tau_u$ then the above together with the update decoding rule (22) imply that the ith update bit is correctly decoded in this case as well. Therefore, we now set

$$\tau_u=2\tau\quad\tau_d=5\tau \tag{26}$$

and continue our analysis henceforth depending on the parameter $\tau=\tau(n)$. As we have just seen, this selection guarantees that the event $E_1\cup E_2$ is indeed a necessary condition for an update decoding error within the block.

Let us now bound the probability of the event E_1. To that end, we note that Lemma 6 together with the α-normalization used in the definition of the training estimate, facilitate the use of Lemma 5. Since the training pattern sequence has a marginal distribution $\sim\mathrm{Ber}(q)$ with $q=\frac{m}{b_p}$, we obtain

$$\begin{aligned}\mathbb{P}(E_1)&=\mathbb{P}\left(\left\|\boldsymbol{p}^{\mathrm{train}}-\boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p})\right\|_{\infty}>\tau\right)\\ &\leq 2|\mathcal{X}|\exp\left(-\frac{b_p\tau^2\left(\frac{m}{b_p}\right)^2}{2}\right)\\ &\leq 2|\mathcal{X}|\exp\left(-\frac{1}{2}\tau^2m^2b^{-1}\right)\triangleq\varepsilon_1^{(1)}(n).\end{aligned}$$

Bounding the probability of the event E_2 is similar, and Lemma 6 together with (23) facilitate the use of Lemma 5 for any of the repetition pattern sequences. These sequences all have a marginal distribution $\sim\mathrm{Ber}(q)$ with $q=\frac{m}{b_p(|\mathcal{X}|-1)s}$ where $s=2(|\mathcal{X}|-1)\lceil\log(n+1)\rceil$ was given in (21). Using Lemma 5 and applying the union bound over all update bits and input pairs (i.e., all repetition pattern sequences) leads to

$$\begin{aligned}\mathbb{P}(E_2)&\leq\sum_{i\in\langle s\rangle,j\in\langle|\mathcal{X}|-1\rangle}\mathbb{P}\left(\left\|\boldsymbol{p}_{(i,j)}-\boldsymbol{p}_{\mathrm{emp}}(\widetilde{Z}^{b_p})\right\|_{\infty}>\tau\right)\\ &\leq(|\mathcal{X}|-1)s\cdot 2|\mathcal{X}|\exp\left(-\frac{\tau^2m^2b^{-1}}{2(|\mathcal{X}|-1)^2s^2}\right)\\ &\leq 4|\mathcal{X}|(|\mathcal{X}|-1)^2\lceil\log(n+1)\rceil\\ &\qquad\times\exp\left(-\frac{\tau^2m^2b^{-1}}{8(|\mathcal{X}|-1)^4\lceil\log(n+1)\rceil^2}\right)\triangleq\varepsilon_1^{(2)}(n).\end{aligned}$$

So far, we have established that the probability of an update decoding error in any given block is upper-bounded by $\mathbb{P}(E_1\cup E_2)\leq\varepsilon_1^{(1)}(n)+\varepsilon_1^{(2)}(n)$. Using the union bound over the blocks and the fact that $\varepsilon_1^{(1)}(n),\varepsilon_1^{(2)}(n)$ do not depend on the message point or the channel, we obtain a uniform upper bound for the error probability achieved by the scheme

$$\begin{aligned}&\sup_{\mathcal{W}\in M_{\mathcal{X}},\theta_0\in[0,1)}p_e(n,\mathcal{W},\theta_0)\\ &\qquad\leq nb^{-1}\left(\varepsilon_1^{(1)}(n)+\varepsilon_1^{(2)}(n)\right)\triangleq\varepsilon_1(n).\end{aligned} \tag{27}$$

From the expression above it is easily verified that if $b(n),m(n),\tau(n)$ are selected such that $\tau^2m^2b^{-1}=\omega(\log^2 n\log(b^{-1}n\log n))$, then $\varepsilon_1(n)\to 0$ and the error probability tends to zero uniformly as desired. However, since this is a variable rate scheme, a low probability of error is not enough since while not making an error indicates we have correctly decoded bits, it does not indicate how many.

B. Rate

Due to the inherent randomness generated by the transmission scheme and the possibly random actions of the channel, the rate achieved by the scheme is random.[13] In this subsection, we show that this rate is arbitrarily close to the empirical capacity of the channel, with probability that tends to one.

At the first stage of the proof, we look only at regular positions (which are used for Horstein iterations), and analyze the rate w.r.t. these channel uses only. Later, we make the necessary adjustments taking into account the negligible effect of nonregular positions as well. For an accepted block, the number of regular positions is in the range $(b_{\min},b_{\max})$, where $b_{\min}=b_p-4m,b_{\max}=b_p-m$. Let n^{reg} be the (random) total number of regular positions over the entire transmission period, and let $\beta\in[0,1]$ denote the (random) fraction of these positions that are accepted (namely, reside in accepted blocks). Communications (via Horstein iterations) take place only on accepted regular positions, namely, over βn^{reg} channel uses. The KT estimates used by the decoder are updated at varying intervals, but these intervals do not exceed $2b_{\max}$ (measured relative to the sequence of accepted regular positions). Hence, the estimator used

[13]Note that even in the case of an individual noise sequence, the rate is still random due to training/update randomization.

in effect is a $\mathrm{KT}(2b_{\max})$ estimator over a sequence of length βn^{reg}.

Define V_0 to be the event where no update decoding errors have occurred, and let V_1 be the event where none of the blocks were discarded due to a too small or too large selection of M_t, M_u made by the receiver. Given V_0, the KT estimates are based on noiseless observations of the noise sequence. Given V_1, we have $nb^{-1}b_{\min} \le n^{\mathrm{reg}} \le nb^{-1}b_{\max}$. Let R^{reg} be the (random) decoding rate measured over regular channel positions only (including both accepted and discarded blocks), and denote by $\boldsymbol{p}_a^{\mathrm{reg}}$ the (random) empirical distribution of the noise sequence over accepted regular positions (entire transmission period). Using (20) and substituting $n \to \beta n^{\mathrm{reg}}$ and $b \to 2b_{\max}$, we have that given $V_0 \cap V_1$

$$\begin{aligned} R^{\mathrm{reg}} &\ge \beta\left(\log|\mathcal{X}| - H\left(\boldsymbol{p}_a^{\mathrm{reg}}\right) - K_2\frac{2|\mathcal{X}|b_{\max}\log\left(\beta n^{\mathrm{reg}}\right)}{\beta n^{\mathrm{reg}}}\right) \\ &= \beta(\log|\mathcal{X}| - H(\boldsymbol{p}_a^{\mathrm{reg}})) - K_2\frac{2|\mathcal{X}|b_{\max}\log\left(\beta n^{\mathrm{reg}}\right)}{n^{\mathrm{reg}}} \\ &\ge R_\beta - 2|\mathcal{X}|K_2\cdot\frac{b}{b_{\min}}\cdot\frac{b\log n}{n} \end{aligned} \tag{28}$$

where $R_\beta \triangleq \beta(\log|\mathcal{X}| - H(\boldsymbol{p}_a^{\mathrm{reg}}))$.

We now focus on the principal rate term R_β. As already mentioned, due to the concavity of the entropy it is expected that discarding blocks will only increase the achieved rate with high probability, as discarded blocks usually have a low empirical capacity. Therefore, we would like to seek conditions under which R_β is minimized by $\beta = 1$ (no discarded blocks), and later show that these conditions are satisfied with high probability. Denote by $\boldsymbol{p}^{\mathrm{reg}}$ and $\boldsymbol{p}_d^{\mathrm{reg}}$ the (random) empirical distributions of noise sequence over all regular positions, and over regular positions inside discarded blocks only, respectively. These distributions together with $\boldsymbol{p}_a^{\mathrm{reg}}$ satisfy

$$\boldsymbol{p}^{\mathrm{reg}} = \beta\boldsymbol{p}_a^{\mathrm{reg}} + (1-\beta)\boldsymbol{p}_d^{\mathrm{reg}}.$$

Extracting $\boldsymbol{p}_a^{\mathrm{reg}}$ and substituting into the expression for R_β yields

$$R_\beta = \beta\left(\log|\mathcal{X}| - H\left(\boldsymbol{p}_d^{\mathrm{reg}} + \beta^{-1}\left(\boldsymbol{p}^{\mathrm{reg}} - \boldsymbol{p}_d^{\mathrm{reg}}\right)\right)\right).$$

Note that for any given values of $\boldsymbol{p}^{\mathrm{reg}}$ and $\boldsymbol{p}_d^{\mathrm{reg}}$, R_β is defined only for values of β large enough such that $\boldsymbol{p}_d^{\mathrm{reg}} + \beta^{-1}(\boldsymbol{p}^{\mathrm{reg}} - \boldsymbol{p}_d^{\mathrm{reg}})$ is still a probability vector. Now, if the derivative of R_β w.r.t. β were to be nonpositive over all the range of permissible β, then R_β would be minimized by $\beta = 1$. We would therefore like to derive a condition for the nonpositivity of the derivative.

Lemma 7: For any given $\boldsymbol{p}^{\mathrm{reg}}, \boldsymbol{p}_d^{\mathrm{reg}}$ and corresponding permissible β

$$\frac{\partial R_\beta}{\partial \beta} \le \log|\mathcal{X}| - H\left(\boldsymbol{p}_d^{\mathrm{reg}}\right) - D\left(\boldsymbol{p}_d^{\mathrm{reg}}\,\|\,\boldsymbol{p}^{\mathrm{reg}}\right).$$

Proof: See Appendix A. □

Based on the preceding lemma, the following chain of inequalities provide a $\mathcal{L}_\infty$-type upper bound on the derivative:

$$\begin{aligned} \frac{\partial R_\beta}{\partial \beta} &\le \log|\mathcal{X}| - H\left(\boldsymbol{p}_d^{\mathrm{reg}}\right) - D\left(\boldsymbol{p}_d^{\mathrm{reg}}\,\|\,\boldsymbol{p}^{\mathrm{reg}}\right) \\ &\overset{(a)}{\le} |\mathcal{X}|\log|\mathcal{X}|\,\|\boldsymbol{p}_d^{\mathrm{reg}} - \boldsymbol{p}_u\|_\infty - \frac{1}{2\ln 2}\|\boldsymbol{p}_d^{\mathrm{reg}} - \boldsymbol{p}^{\mathrm{reg}}\|_1^2 \\ &\overset{(b)}{\le} |\mathcal{X}|\log|\mathcal{X}|\,\|\boldsymbol{p}_d^{\mathrm{reg}} - \boldsymbol{p}_u\|_\infty - \frac{1}{2\ln 2}\|\boldsymbol{p}_d^{\mathrm{reg}} - \boldsymbol{p}^{\mathrm{reg}}\|_\infty^2 \\ &\overset{(c)}{\le} |\mathcal{X}|\log|\mathcal{X}|\,\|\boldsymbol{p}_d^{\mathrm{reg}} - \boldsymbol{p}_u\|_\infty \\ &\quad - \frac{1}{2\ln 2}\left(\|\boldsymbol{p}_d^{\mathrm{reg}} - \boldsymbol{p}_u\|_\infty - \|\boldsymbol{p}^{\mathrm{reg}} - \boldsymbol{p}_u\|_\infty\right)^2 \end{aligned}$$

where $\|\cdot\|_1$ is the $\mathcal{L}_1$ norm. Transition (a) is due to Pinsker's inequality for the relative entropy [28] and the $\mathcal{L}_\infty$ bound for the entropy (Lemma 1), transition (b) holds since the $\mathcal{L}_1$ norm dominates the $\mathcal{L}_\infty$ norm, and in (c) we used the triangle inequality. Thus, a sufficient condition for $\frac{\partial R_\beta}{\partial \beta} \le 0$ is given by

$$\begin{aligned} &\left|\|\boldsymbol{p}^{\mathrm{reg}} - \boldsymbol{p}_u\|_\infty - \|\boldsymbol{p}_d^{\mathrm{reg}} - \boldsymbol{p}_u\|_\infty\right| \\ &\qquad \ge \sqrt{2\ln 2\cdot|\mathcal{X}|\log|\mathcal{X}|}\cdot\|\boldsymbol{p}_d^{\mathrm{reg}} - \boldsymbol{p}_u\|_\infty^{\frac{1}{2}}. \end{aligned} \tag{29}$$

Practicing some algebra, it is easily verified[14] that a sufficient condition for (29) to hold is given by

$$\|\boldsymbol{p}^{\mathrm{reg}} - \boldsymbol{p}_u\|_\infty \ge 2|\mathcal{X}|\cdot\|\boldsymbol{p}_d^{\mathrm{reg}} - \boldsymbol{p}_u\|_\infty^{\frac{1}{2}} \tag{30}$$

and so given (30) we have $R_\beta \ge R_{\beta=1} = \log|\mathcal{X}| - H(\boldsymbol{p}^{\mathrm{reg}})$. Using (28), it is therefore established that (30) together with $V_0 \cap V_1$ imply that

$$R^{\mathrm{reg}} \ge \log|\mathcal{X}| - H(\boldsymbol{p}^{\mathrm{reg}}) - 2|\mathcal{X}|K_2\cdot\frac{b}{b_{\min}}\cdot\frac{b\log n}{n}. \tag{31}$$

We would like to obtain a similar result involving $\boldsymbol{p}_{\mathrm{emp}}(Z^n)$ and the rate R_n. To that end, let $\eta \in [0,1]$ be the fraction of regular positions (out of n), and so $R_n = \eta R^{\mathrm{reg}}$. Let $\boldsymbol{p}^{\mathrm{nreg}}$ denote the distribution of the noise sequence over nonregular positions. Given the event V_1 we have $\eta \ge \frac{b_{\min}}{b}$, and so

$$\begin{aligned} \|\boldsymbol{p}_{\mathrm{emp}}(Z^n) - \boldsymbol{p}_u\|_\infty &= \|\eta(\boldsymbol{p}^{\mathrm{reg}} - \boldsymbol{p}_u) + (1-\eta)(\boldsymbol{p}^{\mathrm{nreg}} - \boldsymbol{p}_u)\|_\infty \\ &\le \eta\|\boldsymbol{p}^{\mathrm{reg}} - \boldsymbol{p}_u\|_\infty + (1-\eta) \\ &\le \|\boldsymbol{p}^{\mathrm{reg}} - \boldsymbol{p}_u\|_\infty + \frac{b - b_{\min}}{b} \\ &\le \|\boldsymbol{p}^{\mathrm{reg}} - \boldsymbol{p}_u\|_\infty + 4mb^{-1}(1 + 2\log b) \\ &\le \|\boldsymbol{p}^{\mathrm{reg}} - \boldsymbol{p}_u\|_\infty + K_3 mb^{-1}\log b \end{aligned} \tag{32}$$

for some $K_3 > 0$, where we have used the convexity of the norm, and the fact that $\|\boldsymbol{p}^{\mathrm{nreg}} - \boldsymbol{p}_u\|_\infty \le 1$. Furthermore, using the concavity and nonnegativity of the entropy we have that given V_1

$$\begin{aligned} H_{\mathrm{emp}}(Z^n) &= H(\boldsymbol{p}_{\mathrm{emp}}(Z^n)) \\ &= H(\eta\boldsymbol{p}^{\mathrm{reg}} + (1-\eta)\boldsymbol{p}^{\mathrm{nreg}}) \\ &\ge \eta H(\boldsymbol{p}^{\mathrm{reg}}) \ge \frac{b_{\min}}{b}H(\boldsymbol{p}^{\mathrm{reg}}). \end{aligned} \tag{33}$$

Now, introduce an auxiliary parameter $\gamma(n) = \mathrm{o}(1)$ chosen to satisfy $\frac{\gamma}{2} > K_3 mb^{-1}\log b$, which is made feasible by requiring $mb^{-1}\log b = \mathrm{o}(1)$. Define the events

$$\begin{aligned} V_2 &\triangleq \left\{2|\mathcal{X}|\,\|\boldsymbol{p}_d^{\mathrm{reg}} - \boldsymbol{p}_u\|_\infty^{\frac{1}{2}} \le \frac{\gamma}{2}\right\} \\ V_3 &\triangleq \{\|\boldsymbol{p}_{\mathrm{emp}}(Z^n) - \boldsymbol{p}_u\|_\infty \ge \gamma\}. \end{aligned} \tag{34}$$

[14] Using $\|\boldsymbol{p}_d^{\mathrm{reg}} - \boldsymbol{p}_u\|_\infty \le 1$, $\ln 2 < 1$, $|\mathcal{X}| \ge 2$, and $x\log x < \frac{x^2}{2}$ for $x = 2, 3, \ldots$.

Now, using (32) it is readily verified that $V_1 \cap V_2 \cap V_3$ implies (30), and therefore $\bigcap_{i=0}^{3} V_i$ implies (31). Furthermore, V_1 implies (33). Putting (31) and (33) together, we establish that given $\bigcap_{i=0}^{3} V_i$

$$\begin{aligned} R_n = \eta R^{\text{reg}} &\geq \frac{b_{\min}}{b} \cdot R^{\text{reg}} \\ &\geq \frac{b_{\min}}{b} \left(\log |\mathcal{X}| - \frac{b}{b_{\min}} H_{\text{emp}}(Z^n) \right. \\ &\qquad \left. -2|\mathcal{X}| K_2 \cdot \frac{b}{b_{\min}} \cdot \frac{b \log n}{n} \right) \\ &\geq \log |\mathcal{X}| - H_{\text{emp}}(Z^n) - \log |\mathcal{X}| \\ &\qquad \cdot K_4 \frac{m \log b}{b} - 2|\mathcal{X}| K_2 \frac{b \log n}{n} \end{aligned} \tag{35}$$

for some $K_4 > 0$. Moreover, given V_3^c, the $\mathcal{L}_\infty$ bound for the entropy (Lemma 1) yields

$$\log |\mathcal{X}| - H_{\text{emp}}(Z^n) \leq \gamma |\mathcal{X}| \log |\mathcal{X}| \tag{36}$$

which enables us to remove the dependence on the event V_3 by incorporating the above into the redundancy term. Namely, since $\{\bigcap_{i=0}^{3} V_i \cup V_3^c\} \supseteq \bigcap_{0=1}^{2} V_i$, we can combine (35) and (36) and the definition of the empirical capacity, to obtain

$$\mathbb{P}\left(R_n(\mathcal{W}, \theta_0) \geq C_n^{\text{emp}}(\mathcal{W}, \theta_0) - \varepsilon_2(n)\right) \geq \mathbb{P}\left(\bigcap_{i=0}^{2} V_i \right) \tag{37}$$

where

$$\begin{aligned} \varepsilon_2(n) \triangleq \log |\mathcal{X}| \cdot K_4 \frac{m \log b}{b} \\ +2|\mathcal{X}| K_2 \frac{b \log n}{n} + \gamma |\mathcal{X}| \log |\mathcal{X}|. \end{aligned} \tag{38}$$

To conclude the rate analysis, we need to lower-bound the probability of $\bigcap_{0=1}^{2} V_i$ and set γ as a function of the scheme parameters. While analyzing the error probability, we have already established that $\mathbb{P}(V_0^c) \leq \varepsilon_1(n)$. We note that $\mathbb{P}(V_1^c)$ is simply upper-bounded by the event where at least one of $2nb^{-1}$ Binomial r.v.'s $\sim B(b_p, \frac{m}{b_p})$ (the M_t, M_u of all the blocks) deviates by more than $\frac{m}{2}$ from its expected value. Applying (say) the Hoeffding inequality [24] and then the union bound, we obtain

$$\begin{aligned} \mathbb{P}(V_1^c) &\leq 4nb^{-1} \exp\left(-\frac{2(m/2)^2}{b_p} \right) \\ &\leq 4nb^{-1} \exp\left(-\frac{1}{2} m^2 b^{-1} \right) \triangleq \varepsilon_3^{(1)}(n). \end{aligned} \tag{39}$$

For V_2, we have $\widetilde{V}_2 \subseteq V_2$ where $\widetilde{V}_2^c$ is the event where at least one discarded block has an empirical distribution $\boldsymbol{q}^{\text{reg}}$ over regular positions which does not satisfy the condition defining the event V_2, namely

$$2|\mathcal{X}| \|\boldsymbol{q}^{\text{reg}} - \boldsymbol{p}_u\|_\infty^{\frac{1}{2}} > \frac{\gamma}{2}.$$

We would like to obtain a corresponding necessary condition on the deviation from uniformity of the training estimate, the probability of which we can then bound. Using norm properties, it is easily verified that

$$\begin{aligned} &\left| \|\boldsymbol{p}_{\text{emp}}(\widetilde{Z}^{b_p}) - \boldsymbol{p}_u\|_\infty - \|\boldsymbol{q}^{\text{reg}} - \boldsymbol{p}_u\|_\infty \right| \\ &\leq \frac{2 \cdot 4m}{b_{\min}} = \frac{8m}{b - 8m\lceil \log b \rceil - 4m} \\ &\leq K_5 m b^{-1} \end{aligned}$$

for some $K_5 > 0$, where $\boldsymbol{p}_{\text{emp}}(\widetilde{Z}^{b_p})$ is the corresponding empirical distribution over all passive positions in the block. Hence, we have $\widehat{V}_2 \subseteq \widetilde{V}_2 \subseteq V_2$ where $\widehat{V}_2^c$ is the event where for at least one block with an empirical distribution over passive positions $\boldsymbol{p}_{\text{emp}}(\widetilde{Z}^{b_p})$ and training estimate $\boldsymbol{p}^{\text{train}}$, simultaneously satisfies

$$\begin{aligned} \|\boldsymbol{p}_{\text{emp}}(\widetilde{Z}^{b_p}) - \boldsymbol{p}_u\|_\infty &> \left(\frac{\gamma}{4|\mathcal{X}|} \right)^2 - K_5 m b^{-1} \\ \|\boldsymbol{p}^{\text{train}} - \boldsymbol{p}_u\|_\infty &< \tau_d = 5\tau. \end{aligned}$$

Hence, using the triangle inequality, a necessary condition for $\widehat{V}_2^c$ is for the training estimate in some block to deviate by at least

$$\|\boldsymbol{p}^{\text{train}} - \boldsymbol{p}_{\text{emp}}(\widetilde{Z}^{b_p})\|_\infty > \left(\frac{\gamma}{4|\mathcal{X}|} \right)^2 - K_5 m b^{-1} - 5\tau. \tag{40}$$

We would now like to set γ so that the right-hand side of the above is strictly positive, and such that $\frac{\gamma}{2} > K_3 m b^{-1} \log b$ which was previously required, is also satisfied. This is obtained by setting γ to

$$\left(\frac{\gamma}{4|\mathcal{X}|} \right)^2 - K_5 m b^{-1} - 5\tau = K_6(\tau + m b^{-1} \log b) \tag{41}$$

with $K_6 > 0$ large enough. Using Lemma 5 and the union bound over blocks, we get

$$\begin{aligned} \mathbb{P}(V_2^c) &\leq \mathbb{P}\left(\widehat{V}_2^c \right) \\ &\leq nb^{-1} \mathbb{P}(\|\boldsymbol{p}^{\text{train}} - \boldsymbol{p}_{\text{emp}}(\widetilde{Z}^{b_p})\|_\infty \\ &\qquad > K_6(\tau + m b^{-1} \log b)) \\ &\leq 2|\mathcal{X}| n b^{-1} \exp\left(-\frac{1}{2} K_6^2 (m b^{-1} \log b + \tau)^2 m^2 b^{-1} \right) \\ &\triangleq \varepsilon_3^{(2)}(n) \end{aligned}$$

and finally

$$\begin{aligned} \mathbb{P}\left(\bigcap_{i=0}^{2} V_i \right) &\geq 1 - \sum_{i=0}^{2} \mathbb{P}(V_i^c) \\ &\geq 1 - \left(\varepsilon_1(n) + \varepsilon_3^{(1)}(n) + \varepsilon_3^{(2)}(n) \right) \\ &\triangleq 1 - \varepsilon_3(n). \end{aligned} \tag{42}$$

Note that $\varepsilon_3(n) = \mathrm{O}(\varepsilon_1(n))$ and, hence, $\varepsilon_3(n) \to 0$ under the same condition provided for the error probability in Section V-A.

Summarizing, we have found that a rate $\varepsilon_2(n)$ close to the empirical capacity (see (38) with γ given in (41)) is achieved with probability at least $1 - \varepsilon_3(n)$ (see (42)), and an error probability no larger than $\varepsilon_1(n)$ (see (27)). In passing, we have also described a set of constraints on the asymptotic behavior of the scheme parameters that are sufficient to guarantee that $\varepsilon_1(n), \varepsilon_2(n), \varepsilon_3(n) \to 0$. In the next subsection, we summarize these constraints, and show that there exist (many) selections of scheme parameters for which they are satisfied.

TABLE I
THE ASYMPTOTIC BEHAVIOR OF CONVERGENCE TERMS FOR THE UNIVERSAL SCHEME

Error probability $\varepsilon_1(n)$	$-\log\varepsilon_1(n) = \Omega\left(\frac{\tau^2 m^2 b^{-1}}{\log^2 n}\right)$
Target redundancy $\varepsilon_2(n)$	$\varepsilon_2(n) = \mathrm{O}(bn^{-1}\log n) + \mathrm{O}\left(\sqrt{\tau + mb^{-1}\log b}\right)$
Redundancy exceeding probability $\varepsilon_3(n)$	$-\log\varepsilon_3(n) = \Omega\left(\frac{\tau^2 m^2 b^{-1}}{\log^2 n}\right)$

C. Parameter Selection and Asymptotic Behavior

There are many different selections of the scheme parameters[15] $b(n), m(n), \tau(n)$ which allow all the convergence parameters $\varepsilon_1(n), \varepsilon_2(n), \varepsilon_3(n)$ to become asymptotically negligible, and result in various tradeoffs between them. The following is the set of sufficient asymptotic conditions to that end, derived directly from the discussion in the two previous subsections:

$$\begin{aligned} &(\mathrm{I})\ \tau = \mathrm{o}(1)\\ &(\mathrm{II})\ mb^{-1}\log b = \mathrm{o}(1)\\ &(\mathrm{IV})\ bn^{-1}\log n = \mathrm{o}(1)\\ &(\mathrm{V})\ \tau^2 m^2 b^{-1} = \omega(\log^2 n \log(b^{-1} n\log n)). \end{aligned} \tag{43}$$

Note that the above conditions also imply that $b = \omega(\log n)$, which was an assumption made when computing the number of update bits. Under (43), the asymptotical behavior for the convergence parameters shown in Table I is achievable.

To demonstrate that the sufficient conditions (43) can be met, let us specifically set the parameters to

$$b(n) = n^{a_0}, m(n) = n^{a_1}, \tau(n) = n^{-a_2} \tag{44}$$

for some positive constants a_0, a_1, a_2. The conditions then translate into

$$a_1 < a_0 < 1, \quad a_0 < 2(a_1 - a_2).$$

It is easy to find many parameter selections satisfying the above conditions, and one possible selection is given by $(a_0, a_1, a_2) = (\frac{3}{4}, \frac{1}{2}, \frac{1}{16})$.

D. Noise Sequence Channels

As already mentioned, the family of noise sequence channels $\mathsf{N}_{\mathcal{X}}$ is a subfamily of the family of modulo-additive channels $\mathsf{M}_{\mathcal{X}}$, and therefore the analysis presented thus far specifically holds when communications take place over an unknown member of $\mathsf{N}_{\mathcal{X}}$. Specifically, this is also true for the special case of an individual noise sequence, which was given as a motivating example in Section I. However, it turns out that when transmission takes place over the family $\mathsf{N}_{\mathcal{X}}$, it is possible to obtain better convergence tradeoffs than when operating over $\mathsf{M}_{\mathcal{X}}$. This is achieved by the following simple modification within each block: The sequence Λ^{b_p} is drawn uniformly over the type of all sequences with *exactly* m training positions and *exactly* m update positions. The sequence Γ^m (note that now $M_u = m$) is drawn uniformly over the type of all sequences with a uniform composition[16] over the alphabet $\langle(|\mathcal{X}| - 1)s\rangle$.

[15]Recall that the thresholds $\tau_d(n), \tau_u(n)$ are determined by $\tau(n)$, as given in (26).

[16]We assume m divides the size of the alphabet, however using a close to uniform composition works as well.

These changes amount to using a fixed number of training/update positions, and using a fixed repetition code for each update bit with each input pair, which means that position types are not selected in an i.i.d. fashion anymore. Thus, a fully informed adversary can now predict the type of the next position with some accuracy, and possibly exhibit "atypical behavior" accordingly (say over training positions), rendering the scheme useless. For noise sequence channels, however, the noise sequence is "generated separately" from the input/output sequence (in the sense described by the Markov relation (7)), hence, the adversary cannot change its strategy based in its ability to predict, which is why the scheme can still work. The most basic example for that is the case where the noise is an individual sequence, which is fixed at the beginning of time and cannot adapt according to the observed inputs/outputs.

The only derivation in the achievability proof that needs to be modified is that of the deviation probability of a sample's empirical distribution from the true empirical distribution, where this sample is now uniform over a type. To this end, we can use Lemma 4 (sampling without replacement) in lieu of Lemma 5, as the noise sequence is now statistically independent of the (training, update) sampling sequences. Interestingly, the exponential decay of the deviation probability is linear in the *number of samples* (either m or $\frac{m}{2(|\mathcal{X}|-1)^2\lceil\log n+1\rceil}$ in our case) and does not involve the length of the sequence sampled from (b_p in our case). Therefore, the expressions for $\varepsilon_i(n)$ for noise sequence channels is essentially given by exchanging the expressions $b^{-1}m^2 \to m$ and $\lceil\log(n+1)\rceil^2 \to \lceil\log(n+1)\rceil$ in all the exponents (up to constant factors). A set of sufficient conditions for communications over $\mathsf{N}_{\mathcal{X}}$ is given by

$$\begin{aligned} &(\mathrm{I})\ \tau = \mathrm{o}(1)\\ &(\mathrm{II})\ mb^{-1}\log b = \mathrm{o}(1)\\ &(\mathrm{III})\ bn^{-1}\log n = \mathrm{o}(1)\\ &(\mathrm{VI})\ b = \omega(\log n)\\ &(\mathrm{V})\ \tau^2 m = \omega(\log n\log(b^{-1}n\log n)) \end{aligned} \tag{45}$$

and the corresponding asymptotical behavior of the convergence parameters is given in Table II at the top of the following page.

Finally, note that moving from i.i.d. sampling to fixed-size sampling without replacement is fundamental for reaping this performance gain in noise sequence channels, since even when the noise is independent of the sampling sequence, the tail of the binomial distribution renders i.i.d. sampling inferior.

E. Randomness Resources

Randomness is a key element in achieving the empirical capacity. Let us examine just how many common random bits are consumed by our scheme. The receiver generates

TABLE II
THE ASYMPTOTIC BEHAVIOR OF CONVERGENCE TERMS FOR THE MODIFIED UNIVERSAL SCHEME OVER NOISE SEQUENCE CHANNELS

Error probability $\varepsilon_1(n)$	$-\log\varepsilon_1(n) = \Omega\left(\frac{\tau^2 m}{\log n}\right)$
Target redundancy $\varepsilon_2(n)$	$\varepsilon_2(n) = \mathrm{O}(bn^{-1}\log n) + \mathrm{O}\left(\sqrt{\tau + mb^{-1}\log b}\right)$
Redundancy exceeding probability $\varepsilon_3(n)$	$-\log\varepsilon_3(n) = \Omega\left(\frac{\tau^2 m}{\log n}\right)$

TABLE III
COMMON RANDOMNESS RESOURCES AND CORRESPONDING SCHEME PARAMETERS SUFFICIENT FOR ACHIEVING THE EMPIRICAL CAPACITY OVER THE VARIOUS FAMILIES OF CHANNELS

Family of Channels	Random Bits	Generation Mechanism	Parameters
Noise Sequence $\mathscr{N}_\mathcal{X}$	$\mathrm{O}(\log^{3+\varepsilon} n)$	Feedback	$b = \Omega\left(\frac{n}{(\log n)^{1+\frac{\varepsilon}{2}}}\right), m = \mathrm{O}\left((\log n)^{1+\frac{\varepsilon}{2}}\right)$
Modulo-Additive $\mathscr{M}_\mathcal{X}$	$\mathrm{O}(\sqrt{n}\log^{2+\varepsilon} n)$	Feedback	$b = \Omega\left(\frac{n}{(\log n)^{1+\frac{\varepsilon}{2}}}\right), m = \mathrm{O}\left(\sqrt{n(\log n)^{1+\varepsilon}}\right)$
General Causal $\mathscr{C}_\mathcal{X}$	$\mathrm{O}(n)$	Common Randomness	Any feasible selection

$\mathrm{O}(nb^{-1}m\log b)$ random bits over the entire transmission period. Under the parameter constraints (43) or (45), this amount is sublinear in n, as otherwise it could not be accommodated by feedback. It is easily verified that these conditions imply that for any $\varepsilon > 0$, the following amount of randomness is sufficient for achieving the empirical capacity, for the different channel families (see Table III).[17]

Interestingly, the randomness resources consumed are significantly reduced when operating over $N_\mathcal{X}$, i.e., over noise sequence channels. Thus, in a sense it seems that when working against an $M_\mathcal{X}$ adversary most of the randomness resources are dedicated to "decoupling" its actions from the channel inputs/outputs, and only a negligible amount of randomness is used for combating the "noise effect" itself. In the most general case of communication over $C_\mathcal{X}$, the much larger amount of random bits (mainly used for dithering) cannot be accommodated by the feedback link, and an external common randomness source is required.

VI. SUMMARY AND DISCUSSION

The universal communication problem over an unknown discrete channel with noiseless feedback was addressed. An extreme channel uncertainty model was considered, where the channel law is unknown to both transmitter and receiver, and may vary arbitrarily from symbol to symbol depending on previous inputs and outputs, possibly in an adversarial fashion. Although in such a general setting no positive rate can be guaranteed in advance, it was constructively shown that reliable communications at a variable rate that corresponds to the empirical goodness of the channel, can be attained. As a measure for this empirical goodness, the *empirical capacity* of the channel was defined as the capacity of an equivalent memoryless modulo-additive channel, with an additive noise marginal distribution given by the empirical distribution of a noise sequence realized by channel actions throughout transmission. An explicit sequential transmission scheme was then described, and shown to achieve rates arbitrarily close to the empirical capacity with probability approaching one, independent of the actual channel in use and uniformly over the message set. For the special case of individual noise sequence channels, the scheme is universal in the sense of successfully competing with any fixed-rate transmission scheme that knows the empirical distribution of the noise sequence in advance.

Achieving the empirical capacity requires randomization. This is especially evident in the case of an individual noise sequence channel, where it is well known that deterministic coding schemes cannot attain the empirical capacity uniformly over the message set in general, even if the empirical distribution of the noise sequence is given in advance. Consequently, the described scheme requires the generation of common random bits shared by the transmitter and the receiver. In the most general setting, $\mathrm{O}(n)$ random bits are used by the scheme, a quantity requiring an external source of common randomness available to the terminals. However, if the channel law is known to be modulo-additive at any time instant (but otherwise arbitrary varying, depending on previous inputs/outputs), only $\mathrm{O}(\sqrt{n}\log^{2+\varepsilon} n)$ random bits are sufficient for any $\varepsilon > 0$, an amount that can be generated exclusively via feedback at no asymptotical cost. Furthermore, in the special case of noise sequence channels (where the channel is completely defined by the noise sequence), the scheme exhibits improved performance in terms of error probability and redundancy, and the amount of common randomness is further reduced to merely $\mathrm{O}(\log^{3+\varepsilon} n)$ random bits, which again can be produced by feedback alone.

The tradeoff between error probability and transmission time attained by the scheme is subexponential in n. This is to be expected, since the actual channel over which communications take place might be (say) a BSC, in which case the empirical capacity converges to the channel capacity a.s., so one cannot hope to universally obtain a positive error exponent when operating at the empirical capacity. However, if one is willing to give up a constant portion of the empirical capacity then it is plausible that a positive error exponent could be universally attained, yet we were unable to adapt our scheme to that end. In order to make the errors due to training estimate deviations vanish exponentially with n, a linear number of training positions must be set, which in the finite-horizon setting implies a constant number of

[17]The parameters b, m are provided, it is readily verified that τ can be set to satisfy the required conditions.

blocks. This however results in an excessively slow update rate for the KT estimate which prohibits any positive rate from being attained, and it therefore seems that an altogether different approach is required.

In this paper, the discussion was limited to a memoryless modulo-additive model, where universality is sought w.r.t. the marginal empirical distribution of the realized noise sequence. In an extension of this work, the concepts presented here are further developed to encompass more general models for channel actions. Specifically, we discuss models that take into account *empirical dependencies* between the channel actions and the input, and exploit *empirical memory* within consecutive channel actions. This approach will facilitate universality w.r.t. higher order empirical statistics, achieving the empirical capacity corresponding to more elaborate models.

Appendix A
Proofs of Lemmas

Proof of Lemma 1: Let $\boldsymbol{p} = (p_0, p_1, \ldots, p_{|\mathcal{X}|-1})$, and assume without loss of generality that $\min(\boldsymbol{p}) = p_0$. For any $i \in \langle|\mathcal{X}|\rangle$, define

$$\boldsymbol{p}^{(i)} = (\underbrace{0,0,\ldots,0}_{i}, 1, \underbrace{0,\ldots,0}_{|\mathcal{X}|-i-1})$$

and let $\boldsymbol{p}_u$ be the uniform distribution over $\mathcal{X}$. Express $\boldsymbol{p}$ as the following convex combination:

$$\boldsymbol{p} = p_0 \cdot |\mathcal{X}| \cdot \boldsymbol{p}_u + \sum_{i=1}^{|\mathcal{X}|-1} (p_i - p_0) \cdot \boldsymbol{p}^{(i)}.$$

Using Jensen's inequality we get

$$\begin{aligned} H(\boldsymbol{p}) &\geq p_0 \cdot |\mathcal{X}| \cdot H(\boldsymbol{p}_u) \\ &\quad + \sum_{i=1}^{|\mathcal{X}|-1} (p_i - p_0) \cdot H(\boldsymbol{p}^{(i)}) = p_0 \cdot |\mathcal{X}| \log |\mathcal{X}|. \end{aligned} \tag{46}$$

Now, by definition we have $\|\boldsymbol{p} - \boldsymbol{p}_u\|_\infty \geq \frac{1}{|\mathcal{X}|} - p_0$, and hence, $p_0 \geq \frac{1}{|\mathcal{X}|} - \|\boldsymbol{p} - \boldsymbol{p}_u\|_\infty$. Substituting this into (46), we obtain

$$H(\boldsymbol{p}) \geq \log |\mathcal{X}|(1 - |\mathcal{X}|\|\boldsymbol{p} - \boldsymbol{p}_u\|_\infty)$$

as desired. Note that (46) is in fact a uniformly better bound, but the $\mathcal{L}_\infty$ bound is sufficient for our needs and easier to work with. □

Proof of Lemma 3: Let ℓ_k^i be the position within the sequence z^n of the kth appearance of the ith symbol, and denote $d = d(z^n, w^n)$. We bound the excess redundancy incurred by using the $\mathrm{KT}(b)$ estimator with noisy observations, assuming that $b + d \leq n + 1$, as follows:

$$\begin{aligned} &-\log \frac{\widehat{p}^{\mathrm{KT}(b)}(z^n \| w^n)}{\widehat{p}^{\mathrm{KT}}(z^n)} \\ &= \sum_{i\in\mathcal{X}} \sum_{k=1}^{n_i(z^n)} \log \frac{\widehat{p}^{\mathrm{KT}}(i|z^{\ell_k^i - 1})}{\widehat{p}^{\mathrm{KT}}(i|w^{\nu_{\ell_k^i}})} \\ &= \sum_{i\in\mathcal{X}} \sum_{k=1}^{n_i(z^n)} \log \left(\frac{\nu_{\ell_k^i} + \frac{|\mathcal{X}|}{2}}{n_i(w^{\nu_{\ell_k^i}}) + \frac{1}{2}} \cdot \frac{(k-1) + \frac{1}{2}}{(\ell_k^i - 1) + \frac{|\mathcal{X}|}{2}} \right) \\ &\leq \sum_{i\in\mathcal{X}} \sum_{k=1}^{n_i(z^n)} \log \left(\frac{\ell_k^i - 1 + \frac{|\mathcal{X}|}{2}}{|k-b-d|^+ + \frac{1}{2}} \cdot \frac{k - \frac{1}{2}}{\ell_k^i - 1 + \frac{|\mathcal{X}|}{2}} \right) \\ &\leq |\mathcal{X}| \sum_{k=1}^{n} \log \left(\frac{k - \frac{1}{2}}{|k-b-d|^+ + \frac{1}{2}} \right) \\ &\leq |\mathcal{X}|(b+d-1)\log(2(b+d-1)) \\ &\quad + |\mathcal{X}| \sum_{k=b+d}^{n} \log \left(1 + \frac{b+d-1}{k-b-d+\frac{1}{2}} \right) \\ &\leq |\mathcal{X}|(b+d-1)\log(2(b+d-1)) \\ &\quad + |\mathcal{X}|(b+d-1)\log e \sum_{k=0}^{n-1} \frac{1}{k+\frac{1}{2}} \\ &\leq |\mathcal{X}|(b+d-1)\log(2(b+d-1)) \\ &\quad + |\mathcal{X}|(b+d-1)(2\log e + \log(2n-1)) \\ &\leq 2|\mathcal{X}|(b+d-1)\log 2ne. \end{aligned}$$

This completes the proof for $b + d \leq n + 1$. For $b + d > n + 1$, we note that the maximal possible excess redundancy per symbol at each step is upper-bounded by $\log(2n + |\mathcal{X}| - 2)$, hence, the total excess redundancy is upper-bounded by

$$\begin{aligned} &n \log(2n + |\mathcal{X}| - 2) \\ &\quad \leq (b+d-1)\log(2n + |\mathcal{X}| - 2) \\ &\quad \leq 2|\mathcal{X}|(b+d-1)\log 2ne \end{aligned}$$

where the last transition is a simple exercise using the fact that $|\mathcal{X}| \geq 2$. □

Proof of Lemma 4: We first prove the lemma for $\mathcal{X} = \{0,1\}$ and a deterministic $Z^n = z^n$, the general case then results as a simple corollary. Under these assumptions, we have

$$\begin{aligned} &\mathbb{P}(\|\boldsymbol{p}_{\mathrm{emp}}(z^n) - \boldsymbol{p}_{\mathrm{emp}}(z^n \downarrow B^n)\|_\infty > \tau) \\ &\quad = \mathbb{P}(|p_{\mathrm{emp}}(z^n) - p_{\mathrm{emp}}(z^n \downarrow B^n)| > \tau) \\ &\quad = \mathbb{P}\left(\left| \frac{1}{n} \sum_{k=1}^{n} z_k - \frac{1}{m} \sum_{k=1}^{m} z_{\sigma_k(B^n)} \right| > \tau \right). \end{aligned} \tag{47}$$

Hoeffding showed that the distribution of the sample mean for sampling without replacement from a finite (deterministic) population, obeys the same bounds for deviation from the mean as the ones he obtained for an i.i.d. sample with mean equal to the empirical mean of the population [24, Sec. 6]. Following this, let A^m be an i.i.d. $\sim \mathrm{Ber}(\frac{1}{n}\sum_{k=1}^n z_k)$ sequence. Using the standard Hoeffding inequality [24, Sec. 2], we obtain

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{k=1}^{n} z_k - \frac{1}{m} \sum_{k=1}^{m} A_k \right| > \tau \right) \leq 2\exp(-2m\tau^2)$$

and by Hoeffding's claim above, this bound also holds for (47). This concludes the proof for the binary alphabet[18] with a deterministic Z^n. For a stochastic Z^n, the result stands since B^n and Z^n are independent, and the bound holds for any realization $Z^n = z^n$. For a larger alphabet, one can define an indicator sequence $Z_{(i)}^n$ for any symbol $i \in \mathcal{X}$, namely, a sequence whose kth element is given by $Z_{k,(i)} \triangleq \mathbb{1}_i(Z_k)$. The proof then follows from the binary alphabet analysis, and the union bound over all the symbols. □

[18] In the binary case we get a coefficient of 2 instead of 4 multiplying the exponent.

Proof of Lemma 5: For any $i \in \mathcal{X}$, define the r.v. sequence $A_{(i)}^n$ of length n, whose kth element is given by

$$A_{k,(i)} = \sum_{j=1}^{k} \mathbb{1}_i(Z_j) - q^{-1}\mathbb{1}_i(Z_j) \cdot B_j.$$

We have that

$$\begin{aligned}
\mathbb{E}&\left(A_{k+1,(i)} | A_{(i)}^k = a^k\right) \\
&= a_k + \mathbb{E}\left(\mathbb{1}_i(Z_{k+1}) \cdot (1 - q^{-1}B_{k+1}) | A_{(i)}^k = a^k\right) \\
&= a_k + \left[1 - q^{-1}\mathbb{E}\left(B_{k+1} | A_{(i)}^k = a^k\right)\right] \\
&\quad \times \mathbb{E}\left(\mathbb{1}_i(Z_{k+1}) | A_{(i)}^k = a^k\right) \\
&= a_k
\end{aligned}$$

where in the last two transitions we used the causal independence assumption, the fact that $B_k \sim \mathrm{Ber}(q)$. Similarly, we also have that $\mathbb{E}(A_{1,(i)}) = 0$. Therefore, for any $i \in \mathcal{X}$, $A_{(i)}^n$ is a zero mean martingale with differences that are bounded by $|A_{k+1,(i)} - A_{k,(i)}| \le \max\{1, q^{-1} - 1\} \le q^{-1}$. By the Azuma–Hoeffding inequality for bounded-difference martingales [25], for any $\tau > 0$

$$\mathbb{P}(|A_{n,(i)}| > n\tau) \le 2\exp\left(-\frac{n\tau^2 q^2}{2}\right).$$

The result is now established as follows:

$$\begin{aligned}
\mathbb{P}&(\|\boldsymbol{p}_{\mathrm{emp}}(Z^n) - \alpha(B^n)\boldsymbol{p}_{\mathrm{emp}}(Z^n \downarrow B^n)\|_\infty > \tau) \\
&= \mathbb{P}\left(\bigcup_{i\in\mathcal{X}} \left\{ \left| \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_i(Z_k) - \frac{1}{nq}\sum_{k=1}^{n} \mathbb{1}_i(Z_k) \cdot B_k \right| > \tau \right\}\right) \\
&\le \sum_{i\in\mathcal{X}} \mathbb{P}(|A_{n,(i)}| > n\tau) \\
&\le 2|\mathcal{X}| \exp\left(-\frac{n\tau^2 q^2}{2}\right). \qquad \square
\end{aligned}$$

Proof of Lemma 6: To avoid heavy indexing, we prove the result for the first block and the training pattern sequence, this is easily extended to any other block and to the repetition pattern sequences. For $1 \le k \le b$, let A_k be an r.v. denoting the exact usage of each position, thus taking values over a generic alphabet $\mathcal{A} = \{\mathit{active}, \mathit{regular}, \mathit{training}\} \cup \langle(|\mathcal{X}| - 1)s\rangle$, where the numerical values correspond to update positions for a specific update bit using a specific input pair, as described in Section IV-E. For brevity, we omit subscripts and write $\mathbb{P}(x|y)$ for $\mathbb{P}_{X|Y}(x|y)$. Summations are understood to be taken over all feasible values of the summation variables. Let us now prove that A_{k+1} and (Z^{k+1}, A^k) are statistically independent for $b_a \le k < b$, from which the result then follows immediately, since the training pattern is simply an indicator sequence $B_k = \mathbb{1}_{\mathrm{training}}(A_{k+b_a})$. We have

$$\begin{aligned}
\mathbb{P}(a_{k+1}|z^{k+1}, a^k) &= \frac{\mathbb{P}(a^{k+1})\mathbb{P}(z^{k+1}|a^{k+1})}{\mathbb{P}(a^k)\mathbb{P}(z^{k+1}|a^k)} \\
&= \mathbb{P}(a_{k+1})\frac{\mathbb{P}(z^{k+1}|a^{k+1})}{\mathbb{P}(z^{k+1}|a^k)}
\end{aligned}$$

where in the second transition we used the fact that by construction, A^{b_a} is a constant deterministic sequence and $A_{b_a+1}^b$ is an i.i.d. sequence. It is therefore sufficient to show that $Z^{k+1} \leftrightarrow A^k \leftrightarrow A_{k+1}$. To that end

$$\begin{aligned}
\mathbb{P}&(z^{k+1}|a^{k+1}) \\
&= \sum_{x^k y^k} \mathbb{P}(z^{k+1}|x^k, y^k, a^{k+1})\mathbb{P}(x^k, y^k|a^{k+1}) \\
&= \sum_{x^k y^k} \mathbb{P}(z^k|x^k, y^k, a^{k+1}) \\
&\quad \times \mathbb{P}(z_{k+1}|z^k, x^k, y^k, a^{k+1})\mathbb{P}(x^k, y^k|a^{k+1}) \\
&= \sum_{x^k y^k} \mathbb{P}(z^k|x^k, y^k)\mathbb{P}(z_{k+1}|x^k, y^k, a^{k+1})\mathbb{P}(x^k, y^k|a^{k+1}) \\
&\overset{(a)}{=} \sum_{x^k y^k} \mathbb{P}(z^k|x^k, y^k)\mathbb{P}(z_{k+1}|x^k, y^k, a^k) \\
&\quad \times \mathbb{P}(x^k|y^k, a^{k+1})\mathbb{P}(y^k|a^{k+1}) \\
&\overset{(b)}{=} \sum_{x^k y^k} \mathbb{P}(z^k|x^k, y^k)\mathbb{P}(z_{k+1}|x^k, y^k, a^k) \\
&\quad \times \mathbb{P}(x^k|y^{k-1}, a^k)\mathbb{P}(y^k|a^{k+1}) \\
&= \sum_{x^k y^k} \mathbb{P}(z^k|x^k, y^k)\mathbb{P}(z_{k+1}|x^k, y^k, a^k)\mathbb{P}(x^k|y^{k-1}, a^k) \\
&\quad \times \prod_{j=1}^{k} \mathbb{P}(y_j|y^{j-1}, a^{k+1}) \\
&\overset{(c)}{=} \sum_{x^k y^k} \mathbb{P}(z^k|x^k, y^k)\mathbb{P}(z_{k+1}|x^k, y^k, a^k)\mathbb{P}(x^k|y^{k-1}, a^k) \\
&\quad \times \prod_{j=1}^{k} \mathbb{P}(y_j|y^{j-1}, a^j) \\
&\overset{(d)}{=} \mathbb{P}(z^{k+1}|a^k)
\end{aligned} \tag{48}$$

Where the transitions are justified as follows ($b_a < k \le b$):

(a) $Z_k \leftrightarrow X^{k-1}Y^{k-1} \leftrightarrow A^k$
Proof: We easily find that $Z_k \leftrightarrow X^{k-1}Y^{k-1} \leftrightarrow U^{k-1}$ by combining $Z_k \leftrightarrow X^{k-1}Y^{k-1} \leftrightarrow X_k$ given in (6), with $Z_k \leftrightarrow X^kY^{k-1} \leftrightarrow U^{k-1}$ which is equivalent to (9). The relation now follows since by construction $A^k = \mathrm{func}(U^{b_a})$.

(b) $X^k = \mathrm{func}(\theta_0, Y^{k-1}, A^k)$, by construction.

(c) $Y_k \leftrightarrow Y^{k-1}A^k \leftrightarrow A_{k+1}^b$
Proof:

$$\begin{aligned}
\mathbb{P}&(y_k|y^{k-1}, a^b) \\
&= \sum_{x^k} \mathbb{P}(y_k|y^{k-1}, x^k, a^b)\mathbb{P}(x^k|y^{k-1}, a^b) \\
&= \sum_{x^k} \mathbb{P}(y_k|y^{k-1}, x^k, a^k)\mathbb{P}(x^k|y^{k-1}, a^k) \\
&= \mathbb{P}(y_k|y^{k-1}, a^k)
\end{aligned}$$

where in the second transition we used (b) above together with $Y_k \leftrightarrow X^kY^{k-1} \leftrightarrow A^b$, which stems from (9) using the fact that by construction $A^b = \mathrm{func}(U^{b_a})$.

(d) The dependence of the expression on a_{k+1} has been removed. $\square$

Proof of Lemma 7: Let $\boldsymbol{p}, \boldsymbol{q}$ be two distributions over $\mathcal{X}$. Then for any $a \in \mathbb{R}^+$ such that $\boldsymbol{p} + a(\boldsymbol{q} - \boldsymbol{p})$ is also a probability distribution

$$\begin{aligned}
&\frac{\partial}{\partial a} H(\boldsymbol{p} + a(\boldsymbol{q} - \boldsymbol{p})) \\
&\quad = -\sum_{i \in \mathcal{X}} (q_i - p_i) \log(p_i + a(q_i - p_i)) \\
&\qquad - \sum_{i \in \mathcal{X}} (q_i - p_i) \frac{p_i + a(q_i - p_i)}{p_i + a(q_i - p_i)} \log e \\
&\quad = -\sum_{i \in \mathcal{X}} (q_i - p_i) \log(p_i + a(q_i - p_i)) \\
&\quad = -\frac{1}{a} \left(\sum_{i \in \mathcal{X}} (p_i + a(q_i - p_i)) \log(p_i + a(q_i - p_i)) \right. \\
&\qquad\qquad \left. - \sum_{i \in \mathcal{X}} p_i \log(p_i + a(q_i - p_i)) \right) \\
&\quad = \frac{1}{a} (H(\boldsymbol{p} + a(\boldsymbol{q} - \boldsymbol{p})) - H(\boldsymbol{p}) - D(\boldsymbol{p} \| \boldsymbol{p} + a(\boldsymbol{q} - \boldsymbol{p}))).
\end{aligned}$$

Using the above, we have that for any $\beta \in [0, 1]$ such that $\boldsymbol{p} + \beta^{-1}(\boldsymbol{q} - \boldsymbol{p})$ is a probability distribution

$$\begin{aligned}
&\frac{\partial}{\partial \beta} \{\beta(\log |\mathcal{X}| - H(\boldsymbol{p} + \beta^{-1}(\boldsymbol{q} - \boldsymbol{p})))\} \\
&\quad = \log |\mathcal{X}| - H(\boldsymbol{p} + \beta^{-1}(\boldsymbol{q} - \boldsymbol{p})) \\
&\qquad + (H(\boldsymbol{p} + \beta^{-1}(\boldsymbol{q} - \boldsymbol{p})) - H(\boldsymbol{p}) \\
&\qquad - D(\boldsymbol{p} \| \boldsymbol{p} + \beta^{-1}(\boldsymbol{q} - \boldsymbol{p}))) \\
&\quad = \log |\mathcal{X}| - H(\boldsymbol{p}) - D(\boldsymbol{p} \| \boldsymbol{p} + \beta^{-1}(\boldsymbol{q} - \boldsymbol{p})) \\
&\quad \le \log |\mathcal{X}| - H(\boldsymbol{p}) - \beta D(\boldsymbol{p} \| \boldsymbol{p} \\
&\qquad + \beta^{-1}(\boldsymbol{q} - \boldsymbol{p})) - (1 - \beta) D(\boldsymbol{p} \| \boldsymbol{p}) \\
&\quad \overset{\text{(a)}}{\le} \log |\mathcal{X}| - H(\boldsymbol{p}) - D(\boldsymbol{p} \| \beta(\boldsymbol{p} \\
&\qquad + \beta^{-1}(\boldsymbol{q} - \boldsymbol{p})) + (1 - \beta)\boldsymbol{p}) \\
&\quad = \log |\mathcal{X}| - H(\boldsymbol{p}) - D(\boldsymbol{p} \| \boldsymbol{q})
\end{aligned}$$

where in (a) we have used the convexity of the relative entropy. The proof is concluded by substituting $\boldsymbol{p}, \boldsymbol{q}$ with $\boldsymbol{p}_d^{\mathrm{reg}}, \boldsymbol{p}^{\mathrm{reg}}$, respectively. □

APPENDIX B
A HORIZON-FREE UNIVERSAL SCHEME

In this section, we show how the presented finite-horizon feedback transmission scheme can be transformed into a horizon-free scheme, with an instantaneous rate approaching the empirical capacity. To motivate this generalization, suppose one wishes to transmit a fixed number of bits using the finite-horizon scheme. In this case, it may be that capacity-wise, the receiver could have potentially decoded enough bits half way throughout transmission, and even worse —could not do so when transmission ends due to a deterioration in channel conditions. In this case, it is therefore critical that the transmission can be stopped at any given time, while achieving the instantaneous empirical capacity.

The idea is that instead of taking a fixed transmission period n and dividing it into blocks of a fixed size $b(n)$, a variable block size is used, growing with time. The apparent difficulty with this approach is that, in contrast to the finite-horizon case and although the size of the last block is increasing, the size of any specific block is constant, and a nonnegligible update decoding error probability in each specific block is incurred. This in turn results in two problems. First, bounding the error probability as before using a union bound over update decoding error events provides a nonvanishing bound dominated by the first block. Second, the resulting KT estimates use noisy observations, which incurs a redundancy penalty. The first problem is essentially solved by making sure that in the event where the last accepted block is not "recent enough," no bits are decoded. Loosely speaking, this event implies the empirical capacity is small anyway with high probability, hence the resulting excess redundancy is negligible. As for the second problem, for a suitable selection of scheme parameters we can show that with high probability, the Hamming distance between the noise sequence and the corresponding noisy observations sequence increases slowly enough, so that the excess redundancy becomes negligible.

Following this discussion, the horizon-free scheme is obtained via the following modifications of the finite-horizon scheme.

(A) The size of the kth block is set to b_k, where $\{b_k \in \mathbb{N}\}_{k=1}^{\infty}$ is strictly increasing. In our proof, we use an arithmetically growing block size,[19] i.e., $b_k = b_0 + k$ for some $b_0 \in \mathbb{N}$.

(B) The parameters of the kth block are fixed functions of its size b_k, i.e., $m_k = m(b_k), \tau_k = \tau(b_k)$, etc.

(C) On the kth block, the update information consists of the type of the noise sequence (symbols occurrences vector) over regular positions in the previously accepted block, and the index of the corresponding message interval. Let $n_k \triangleq \sum_{j=1}^{k} b_j$ be the number of channel uses in the first k blocks. The number of uncoded update bits in the kth block is therefore given by replacing $b \to b_{k-1}, n \to n_{k-1}$ in the left-hand side of (21).

(D) Transmission can be terminated at any point. When terminated, the receiver normally decodes the binary interval pertaining to the last known message interval (from the last accepted block) using the corresponding ambiguity resolving bit. However, if transmission ended during the kth block and the last accepted block k_{acc} is not recent enough, namely $k_{\mathrm{acc}} < \rho_k$ for some predetermined *recency threshold* ρ_k, then the decoded interval is $[0, 1)$, i.e., no bits are decoded.

We now turn to prove that this modified scheme achieves the empirical capacity for a suitable selection of parameters b_k, m_k, τ_k, ρ_k, where the thresholds $\tau_{d,k}$ and $\tau_{u,k}$ are determined by τ_k as before. For simplicity of the exposition, assume the length of the kth block is $b_k = k$, and choose m_k and τ_k to be

$$m_k \triangleq b_k^{a_1} = k^{a_1}, \quad \tau_k \triangleq b_k^{-a_2} = k^{-a_2}$$

for some constants $a_1, a_2 \in (0, 1)$. For brevity, we mostly disregard noninteger issues throughout this section, as these have no

[19]Other block size increments are possible, resulting in different tradeoffs between error probability and convergence rate. For instance, one can use the recursion $b_k = (\sum_{j=1}^{k-1} b_j)^{\nu}$ for some $\nu \in (0, 1)$.

asymptotic effect. Let $n_k = \sum_{j=1}^k b_j = \frac{k(k+1)}{2}$. The number of uncoded update bits in the kth block is zero padded up to s_k, which is given by

$$\begin{aligned}\left\lceil \log b_{k-1}^{|\mathcal{X}|-1} \right\rceil &+ \lceil \log(1+(|\mathcal{X}|-1)n_{k-1})\rceil \\ &\le (|\mathcal{X}|-1)\log k + \log((|\mathcal{X}|-1)k(k+1)) \\ &\le 2|\mathcal{X}|\log(k+1) \triangleq s_k. \end{aligned} \tag{49}$$

Following the same derivations as in Section V, we have that

$$\begin{aligned}\mathbb{P}(E_1^{(k)}) &\le 2|\mathcal{X}|\exp\left(-\frac{1}{2}\tau_k^2 m_k^2 b_k^{-1}\right) \\ &= 2|\mathcal{X}|\exp\left(-\frac{1}{2}k^{2(a_1-a_2)-1}\right) \\ \mathbb{P}(E_2^{(k)}) &\le 4|\mathcal{X}|^3\lceil\log(k+1)\rceil\exp\left(-\frac{\tau_k^2 m_k^2 b_k^{-1}}{8|\mathcal{X}|^4\lceil\log(k+1)\rceil^2}\right) \\ &= 4|\mathcal{X}|^3\lceil\log(k+1)\rceil\exp\left(-\frac{k^{2(a_1-a_2)-1}}{8|\mathcal{X}|^4\lceil\log(k+1)\rceil^2}\right)\end{aligned}$$

where $E_1^{(k)}, E_2^{(k)}$ are the events in the kth block corresponding to E_1, E_2 defined in Section V-A. Thus, $E_1^{(k)} \cup E_2^{(k)}$ constitutes a necessary condition for an update decoding error in the kth block. At this stage in the finite-horizon proof, we have used the union bound over update decoding error events in each block to obtain an upper bound for the error probability in the finite-horizon scheme. However, in this case taking the union bound would result in an error probability that is dominated by the first block, hence not decaying to zero. From this point, assume the transmission scheme was terminated at time $n = n_k$ after *precisely k arithmetically growing blocks were sent*,[20] which means that $\sqrt{2n}-1 \le k \le \sqrt{2n}$. Let us divide the transmission period into two *batches*: The *first batch* includes the first k^{a_3} blocks for some $a_3 \in (0,1)$, while the *last batch* includes all the rest $k - k^{a_3}$ blocks. Let us also set the *recency threshold* to be $\rho_k \triangleq k^{a_3}$, which means that if the last accepted block resides in the first batch, no bits are decoded.

Define V_0 to be the following event:

$$V_0 \triangleq \bigcap_{j=\lceil \rho_k \rceil}^{k} \left\{ \left(E_1^{(j)}\right)^c \cap \left(E_2^{(j)}\right)^c \right\}.$$

Using the same ideas as in the fixed-horizon analysis, we can show V_0 implies that no update decoding error occurred in the last batch. Due to the recency threshold, this implies in turn that either the decoded message interval is correct, or that no bits are decoded. Therefore, V_0^c is a necessary condition for an error, and so

$$\begin{aligned}\sup_{\substack{\mathcal{W}\in M_{\mathcal{X}} \\ \theta_0\in[0,1)}} & p_e(n,\mathcal{W},\theta_0) \\ &\le \mathbb{P}(V_0^c) \le k\cdot\mathbb{P}\left(E_1^{(k^{a_3})} \cup E_2^{(k^{a_3})}\right) \\ &\le 5|\mathcal{X}|^3 k\lceil\log(k^{a_3}+1)\rceil \\ &\quad\times\exp\left(-\frac{k^{a_3(2(a_1-a_2)-1)}}{8|\mathcal{X}|^2\lceil\log(k^{a_3}+1)\rceil^2}\right) \\ &\le 5|\mathcal{X}|^3\sqrt{2n}\lceil\log((2n)^{\frac{a_3}{2}}+1)\rceil \\ &\quad\times\exp\left(-\frac{(\sqrt{2n}-1)^{a_3(2(a_1-a_2)-1)}}{8|\mathcal{X}|^4\lceil\log((2n)^{\frac{a_3}{2}}+1)\rceil^2}\right) \triangleq \varepsilon_1(n)\end{aligned} \tag{50}$$

where we have used the union bound over blocks in the last batch, and the fact that the update decoding error probability of the first block in that batch dominates the others. We get

$$-\log\varepsilon_1(n) = \Omega\left(\frac{n^{a_3(a_1-a_2-\frac{1}{2})}}{\log^2 n}\right) \tag{51}$$

so the error probability tends to zero uniformly for any selection $a_1 - a_2 > \frac{1}{2}$. This concludes the error probability part of the proof.

We now show that at any time point, the decoding rate attained by the scheme is close to the empirical capacity with probability approaching one. Let V_1 be the event where none of the blocks in the last batch were discarded due to an improper selection of M_t, M_u made by the receiver. Using Hoeffding's inequality as in (39), it is readily verified that

$$-\log\mathbb{P}(V_1^c) = \Omega\left(n^{a_3(a_1-\frac{1}{2})}\right) \tag{52}$$

and so for any selection $a_1 - a_2 > \frac{1}{2}$ both $\mathbb{P}(V_0), \mathbb{P}(V_1) \to 1$. Now, let $\boldsymbol{p}_a^{\mathrm{reg,f}}$ be the empirical distribution over accepted regular positions in the first batch, and $\boldsymbol{p}_a^{\mathrm{reg},\ell}$ the corresponding distribution in the last batch. Let us express $\boldsymbol{p}_a^{\mathrm{reg}}$ as

$$\boldsymbol{p}_a^{\mathrm{reg}} = \lambda\boldsymbol{p}_a^{\mathrm{reg,f}} + (1-\lambda)\boldsymbol{p}_a^{\mathrm{reg},\ell}.$$

Due to possible nonnegligible erroneous update decoding, the receiver might use noisy observations for its KT estimates. In the finite-horizon case, this problem was averted since the update error probability in *each block* was negligible, and so the event of noisy observations had a vanishing impact incorporated into the redundancy term $\varepsilon_3(n)$. However, in the horizon-free case there is a nonvanishing update error probability dominated by the first blocks. Nevertheless, under V_0, only the first batch may include erroneous blocks, and thus the *Hamming distance* between the actual noise sequence (over accepted regular positions) and the one used by the receiver when updating the KT estimates, is upper-bounded by $d = \lambda n$. Now, since $b_k \le \sqrt{2n}$, the receiver uses a $\mathrm{KT}(2\sqrt{2n})$ estimator and using Lemmas 2 and 3 with noisy observations we have

$$R^{\mathrm{reg}} \ge R_\beta - K_7\left(\frac{\log n}{\sqrt{n}} + \lambda\log n\right) \quad (\text{given } V_0\cap V_1) \tag{53}$$

for some $K_7 > 0$ large enough, where

$$R_\beta \triangleq \beta(\log|\mathcal{X}| - H(\boldsymbol{p}_a^{\mathrm{reg}})).$$

Let ξ_j be the maximal possible number of non-regular positions in blocks j to k (where k is the last block), i.e.,

$$\xi_j \triangleq \sum_{i=j}^{k} 4m_j(1+2\log b_j) \le K_3\sum_{i=j}^{k} m_j\log b_j \tag{54}$$

[20]In general, the scheme may be terminated arbitrarily in the middle of some block, and decoding will be carried out w.r.t. the end of the previous block. Since the maximal block size is $\mathrm{O}(\sqrt{n})$, this can cause a maximal rate fluctuation of $\mathrm{O}(n^{-\frac{1}{2}})$, which should be added to the redundancy term $\varepsilon_2(n)$, but turns out to be asymptotically negligible.

where K_2 was defined in (32). Simple algebraic manipulations yield the following bound for β:

$$\beta \geq \frac{\beta_\ell}{1-\lambda}\left(\frac{n - n_{\rho_k} - \xi_{\rho_k+1}}{n}\right) = \frac{\beta_\ell}{1-\lambda}\left(1 - 2n^{a_3-1} - K_8 n^{\frac{a_1-1}{2}}\log n\right) \quad (55)$$

for some $K_8 > 0$, where β_ℓ is the fraction of regular positions in the last batch that were accepted. To continue, we need the following lemma.

Lemma 8: Let $\boldsymbol{p}$ and $\boldsymbol{q}$ be any two probability distributions over a finite alphabet $\mathcal{X}$. Then for any $\lambda \in [0,1]$

$$H(\lambda\boldsymbol{q} + (1-\lambda)\boldsymbol{p}) \leq H(\boldsymbol{p}) + 3|\mathcal{X}|\lambda\log\frac{2}{\lambda}.$$

Proof: Let $\boldsymbol{p} = (p_1,\ldots,p_{|\mathcal{X}|})$ be a probability distribution over $\mathcal{X}$ with nonzero elements. Let $\boldsymbol{v}$ be the representation of $\boldsymbol{p}$ over the $|\mathcal{X}|-1$-dimensional probability simplex $\mathbb{S}^{|\mathcal{X}|-1}$, i.e., a vector of the first $|\mathcal{X}|-1$ elements of $\boldsymbol{p}$. With some abuse of notations, denote by $H(\boldsymbol{v})$ the entropy function of $\boldsymbol{p}$, calculated over $\mathbb{S}^{|\mathcal{X}|-1}$. We take the partial derivatives of H and get

$$\frac{\partial H(\boldsymbol{v})}{\partial v_i} = \log\frac{1-\sum_j v_j}{v_i}, \quad i = 1,\ldots,|\mathcal{X}|-1.$$

Since $H(\cdot)$ is concave over $\mathbb{S}^{|\mathcal{X}|-1}$, its tangents at any point are always above it. Therefore, for any $\boldsymbol{r} \in \mathbb{R}^{|\mathcal{X}|-1}$ that satisfies $(\boldsymbol{v}+\boldsymbol{r}) \in \mathbb{S}^{|\mathcal{X}|-1}$, we have that

$$H(\boldsymbol{v}+\boldsymbol{r}) \leq H(\boldsymbol{v}) + \sum_{i=1}^{|\mathcal{X}|-1} r_i \log\frac{1-\sum_j v_j}{v_i}. \quad (56)$$

Now let $\boldsymbol{v}$ and $\boldsymbol{u}$ be vectors over the $(|\mathcal{X}|-1)$-dimensional simplex that correspond to $\boldsymbol{p}$ and $\boldsymbol{q}$, respectively, and $0 \leq \lambda \leq 1$ some constant. With the same abuse of notations, we use (56)

$$\begin{aligned}H(\lambda\boldsymbol{q} + (1-\lambda)\boldsymbol{p}) &= H(\boldsymbol{v} + \lambda(\boldsymbol{u}-\boldsymbol{v}))\\ &\leq H(\boldsymbol{v}) + \lambda\sum_{i=1}^{|\mathcal{X}|-1}(u_i - v_i)\log\frac{1-\sum_j v_j}{v_i}\\ &= H(\boldsymbol{p}) + \lambda\sum_{i=1}^{|\mathcal{X}|-1}(q_i-p_i)\log\frac{p_{|\mathcal{X}|}}{p_i}\\ &\leq H(\boldsymbol{p}) + \lambda\sum_{i=1}^{|\mathcal{X}|-1}\left|\log\frac{p_{|\mathcal{X}|}}{p_i}\right|.\end{aligned}$$

Assume for the moment that $\lambda < \frac{1}{4}$. If it so happens and all the symbol probabilities satisfy $p_i > \lambda$, then from the above we have

$$\begin{aligned}H(\lambda\boldsymbol{q}+(1-\lambda)\boldsymbol{p}) &\leq H(\boldsymbol{p}) + \lambda\sum_{i=1}^{|\mathcal{X}|-1}\left|\log\frac{p_{|\mathcal{X}|}}{p_i}\right|\\ &\leq H(\boldsymbol{p}) + (|\mathcal{X}|-1)\lambda\log\frac{1}{\lambda} \quad (57)\end{aligned}$$

which satisfies the statement in the lemma. Otherwise, assume that there are precisely t symbols that do not satisfy that requirement. Without loss of generality we assume that $p_1 \leq p_2 \leq \cdots \leq p_{|\mathcal{X}|}$, and therefore $p_t \leq \lambda$. Define

$$\boldsymbol{\psi} \triangleq \lambda\boldsymbol{q} + (1-\lambda)\boldsymbol{p} = (\psi_1,\ldots,\psi_{|\mathcal{X}|}).$$

The first t elements of $\boldsymbol{\psi}$ are all smaller than 2λ. Without loss of generality we assume that al least one of those t elements is nonzero, as otherwise we can reduce the dimension of the problem. We have the following:

$$\begin{aligned}H(\boldsymbol{\psi}) &\overset{\text{(a)}}{=} H(\psi_1+\psi_{|\mathcal{X}|},\psi_2,\ldots,\psi_{|\mathcal{X}|-1})\\ &\quad + (\psi_1+\psi_{|\mathcal{X}|})h_B\left(\frac{\psi_1}{\psi_1+\psi_{|\mathcal{X}|}}\right)\\ &\overset{\text{(b)}}{\leq} H(\psi_1+\psi_{|\mathcal{X}|},\psi_2,\ldots,\psi_{|\mathcal{X}|-1}) + h_B(2\lambda)\\ &\overset{\text{(c)}}{\leq} H\left(\sum_{i=1}^t\psi_i+\psi_{|\mathcal{X}|},\psi_{t+1},\ldots,\psi_{|\mathcal{X}|-1}\right) + t\cdot h_B(2\lambda)\\ &\overset{\text{(d)}}{\leq} H\left(\sum_{i=1}^t p_i + p_{|\mathcal{X}|},p_{t+1},\ldots,p_{|\mathcal{X}|-1}\right)\\ &\quad + (|\mathcal{X}|-t-1)\lambda\log\frac{1}{\lambda} + t\cdot h_B(2\lambda)\\ &\overset{\text{(e)}}{\leq} H(\boldsymbol{p}) + |\mathcal{X}|\lambda\log\frac{1}{\lambda}\\ &\quad + |\mathcal{X}|\left(2\lambda\log\frac{1}{2\lambda} + 2\lambda\log e\right)\\ &\leq H(\boldsymbol{p}) + 3|\mathcal{X}|\lambda\log\frac{2}{\lambda}.\end{aligned}$$

In (a) we applied the entropy's grouping property [28], in (b) we used the assumption that $\lambda < \frac{1}{4}$, in (c) we repeated the two previous steps $t-1$ more times, in (d) we used (57) since the probability vector argument of the entropy function has a minimal symbol probability exceeding λ, and in (e) we used $0 \leq t \leq |\mathcal{X}|$, the entropy's grouping property, and the inequality $h_B(p) \leq p\log\frac{1}{p} + p\log e$. This proves the result for $\lambda < \frac{1}{4}$. The proof is now concluded by noticing that for $\lambda \geq \frac{1}{4}$ the excess term satisfies $3|\mathcal{X}|\lambda\log\frac{2}{\lambda} > \log|\mathcal{X}|$. □

Applying Lemma 8 to R_β and using inequality (55), we have

$$\begin{aligned}R_\beta &= \beta\left(\log|\mathcal{X}| - H\left(\lambda\boldsymbol{p}_a^{\text{reg,f}} + (1-\lambda)\boldsymbol{p}_a^{\text{reg},\ell}\right)\right)\\ &\geq \frac{\beta_\ell}{1-\lambda}\left(1-2n^{a_3-1} - K_8 n^{\frac{a_1-1}{2}}\log n\right)\\ &\quad\times\left(\log|\mathcal{X}| - H\left(\boldsymbol{p}_a^{\text{reg},\ell}\right) - 3|\mathcal{X}|\lambda\log\frac{2}{\lambda}\right)\\ &\geq \left(\frac{1-2n^{a_3-1}-K_8 n^{\frac{a_1-1}{2}}\log n}{1-\lambda}\right)R_\beta^\ell - 3|\mathcal{X}|\frac{\lambda}{1-\lambda}\log\frac{2}{\lambda}\end{aligned} \quad (58)$$

where

$$R_\beta^\ell \triangleq \beta_\ell\left(\log|\mathcal{X}| - H(\boldsymbol{p}_a^{\text{reg},\ell})\right).$$

R_β^ℓ is a quantity similar to R_β, but only for the last batch. Define $\boldsymbol{p}^\ell, \boldsymbol{p}^{\text{reg},\ell}, \boldsymbol{p}_d^{\text{reg},\ell}$ to be the empirical distribution of the noise sequence in the entire last batch, over regular positions, and over discarded regular positions, respectively. We can now repeat the finite-horizon analysis over the last batch only, using the parameters of the ρ_kth block which is the smallest in the batch. Namely, one can set the an auxiliary parameter $\gamma(n) = o(1)$ to satisfy (the equivalent of (41))

$$\begin{aligned}\gamma^2 &= \Omega\left(m_{\rho_k} b_{\rho_k}^{-1}\log b_{\rho_k}\right) + \Omega(\tau_{\rho_k})\\ &= \Omega\left(n^{\frac{a_1-1}{2}}\log n\right) + \Omega\left(n^{-\frac{a_2}{2}}\right)\end{aligned}$$

and define the events

$$V_2 \triangleq \left\{ 2|\mathcal{X}| \left\| \boldsymbol{p}_d^{\mathrm{reg},\ell} - \boldsymbol{p}_u \right\|_\infty^{\frac{1}{2}} \leq \frac{\gamma}{2} \right\}$$
$$V_3 \triangleq \{ \|\boldsymbol{p}^\ell - \boldsymbol{p}_u\|_\infty \geq \gamma \} \tag{59}$$

to obtain

$$R_\beta^\ell \geq R_{\beta=1}^\ell \geq \log|\mathcal{X}| - H(\boldsymbol{p}^{\mathrm{reg},\ell}) \quad \left(\text{given } \bigcap_{i=0}^{3} V_i\right) \tag{60}$$

and with

$$\begin{aligned} -\log \mathbb{P}(V_2^c) &= \Omega\left(b_{\rho_k}^{-1} m_{\rho_k}^2 \gamma^4\right) \\ &= \Omega\left(\log^2 n \cdot n^{a_3(a_1-\frac{1}{2})-(1-a_1)}\right) \\ &\quad + \Omega\left(n^{a_3(a_1-\frac{1}{2})-a_2}\right) \end{aligned}$$

and so setting $a_3(a_1 - \frac{1}{2}) > \max(1-a_1, a_2)$ we have $\mathbb{P}(V_2) \to 1$.

Let us now define V_4 as the event where $\lambda < n^{-a_4}$ for some $a_4 \in (0,1)$. Using (53), (58), and (60) we get[21]

$$\begin{aligned} R^{\mathrm{reg}} &\geq \log|\mathcal{X}| - H(\boldsymbol{p}^{\mathrm{reg},\ell}) \\ &\quad - \frac{\left|2n^{a_3-1} + K_8 n^{\frac{a_1-1}{2}} \log n - n^{-a_4}\right| \log|\mathcal{X}|}{1-n^{-a_4}} \\ &\quad - 3|\mathcal{X}| \frac{n^{-a_4}\log(2n^{a_4})}{1-n^{-a_4}} - K_7\left(\frac{\log n}{\sqrt{n}} + n^{-a_4}\log n\right) \\ &= \log|\mathcal{X}| - H(\boldsymbol{p}^{\mathrm{reg},\ell}) \\ &\quad + \mathrm{O}(n^{-a_5}\log n) \quad \left(\text{given } \bigcap_{i=0}^{4} V_i\right) \end{aligned} \tag{61}$$

where $a_5 \triangleq \min(1-a_3, \frac{1-a_1}{2}, a_4, \frac{1}{2})$. To express (61) in terms of the rate R_n and the empirical entropy $H(\boldsymbol{p}_{\mathrm{emp}}(Z^n))$ we use Jensen's inequality and standard manipulations, yielding (given V_1)

$$\begin{aligned} R_n &\geq \left(1 - \frac{\xi_1}{n}\right) R^{\mathrm{reg}} \\ H(\boldsymbol{p}^\ell) &\geq \left(1 - \frac{\xi_{\rho_k}}{n}\right) H(\boldsymbol{p}^{\mathrm{reg},\ell}) \\ H(\boldsymbol{p}_{\mathrm{emp}}(Z^n)) &\geq \left(1 - \frac{\zeta}{n}\right) H(\boldsymbol{p}^{\mathrm{reg},\ell}) \end{aligned} \tag{62}$$

where the term ξ_j was defined in (54) and ζ is given by

$$\zeta = \xi_{\rho_k} + \sum_{j=1}^{\rho_k} b_j = \mathrm{O}\left(n^{\frac{1+a_1}{2}} \log n\right) + \mathrm{O}(n^{a_3})$$

and corresponds to the maximal number of channel uses wasted on the first batch and on non-regular transmission in the second batch together. Thus, since a_5 already involves all the relevant terms, we get

$$R_n(\mathcal{W},\theta_0) = C_n^{\mathrm{emp}}(\mathcal{W},\theta_0) + \mathrm{O}(n^{-a_5}\log n) \quad \left(\text{given } \bigcap_{i=0}^{4} V_i\right). \tag{63}$$

[21]Notice that the minimal positive value for λ is always greater than $\frac{1}{n}$ (single accepted block in the first batch), and for $\lambda = 0$ (first batch fully discarded) the penalty term $\frac{\lambda}{1-\lambda}\log\frac{2}{\lambda}$ is zero.

As before, under V_3^c the $\mathcal{L}_\infty$ bound for the entropy (Lemma 1) yields

$$\begin{aligned} \log|\mathcal{X}| - H(\boldsymbol{p}^\ell) &\leq |\mathcal{X}| \log|\mathcal{X}| \gamma(n) \\ &= \mathrm{O}(n^{\frac{a_1-1}{4}}\sqrt{\log n}) + \mathrm{O}(n^{-\frac{a_2}{4}}) \end{aligned}$$

and using (62) yields in turn

$$\begin{aligned} C_n^{\mathrm{emp}}(\mathcal{W},\theta_0) &= \mathrm{O}\left(n^{\frac{a_1-1}{4}}\sqrt{\log n}\right) \\ &\quad + \mathrm{O}\left(n^{-\frac{a_2}{4}}\right) + \mathrm{O}\left(n^{\frac{a_1-1}{2}}\log n\right) \\ &\quad + \mathrm{O}(n^{a_3-1}) \quad (\text{given } V_3^c). \end{aligned} \tag{64}$$

This penalty can be incorporated into the redundancy term to remove the dependence on the event V_3.

We would also like to remove the dependence on V_4, and to that end consider the event $V_4^c \cap V_0$. Under this event and assuming $a_3 + a_4 < \frac{1}{2}$, there can be no accepted block of size $\Omega(n^{a_3+a_4})$, as otherwise V_4^c is contradicted. Due to V_0, the empirical distribution (over passive positions) of each of these larger blocks is τ_j close to the training estimate, which in turn is τ_j close to being uniform, where j is the index/length of the corresponding block. Using the $\mathcal{L}_\infty$ bound for the entropy, the empirical capacity of each of these blocks (over passive positions) is therefore $\mathrm{O}(\tau_j) = \mathrm{O}(n^{-a_2(a_3+a_4)})$, where we have used the fact that τ_j of the smallest such block dominates the others. Moreover, the empirical capacity of all the blocks of size $\mathrm{O}(n^{a_3+a_4})$ is $\mathrm{O}(1)$ (which is true of course for any block). By convexity, the empirical capacity over the entire transmission period is no larger than the average of the empirical capacities over some segmentation. Hence

$$\begin{aligned} &C_n^{\mathrm{emp}}(\mathcal{W},\theta_0) \\ &\quad = \mathrm{O}\left(n^{-a_2(a_3+a_4)}\right) + \mathrm{O}\left(n^{-2(a_3+a_4)}\right) \\ &\qquad + \mathrm{O}\left(n^{\frac{1-a_1}{2}}\log n\right) + \mathrm{O}(n^{a_3-1}) \quad (\text{given } V_4^c \cap V_0) \end{aligned} \tag{65}$$

where the first term is the contribution of blocks of size $\Omega(n^{a_3+a_4})$, the second term is the contribution of the smaller blocks (after averaging w.r.t. the fraction of time they occupy), and the last two terms correspond to the deviation possibly incurred by considering only passive positions.

Let us now combine all the above results. The following inclusion is easily verified:

$$\left\{\bigcap_{i=0}^{4} V_i\right\} \cup \{V_3^c\} \cup \{V_4^c \cap V_0\} \supseteq \bigcap_{i=0}^{2} V_i. \tag{66}$$

Under the event on the left above (and thus also under the event on the right) at least one of (63), (64), or (65) must hold. We therefore conclude that

$$\begin{aligned} \mathbb{P}\left(R_n(\mathcal{W},\theta_0) \geq C_n^{\mathrm{emp}}(\mathcal{W},\theta_0) - \varepsilon_2(n)\right) \\ \geq \mathbb{P}\left(\bigcap_{i=0}^{2} V_i\right) \geq 1 - \varepsilon_3(n) \end{aligned} \tag{67}$$

where all the redundancy terms are now incorporated into $\varepsilon_2(n)$ using all the constraints set thus far (with some further relaxations)

$$\varepsilon_2(n) = \mathrm{O}(n^{-a_6}\log n),$$
$$a_6 \triangleq \min\left(\frac{1-a_1}{4}, \frac{a_2}{4}, 1-a_3, a_2(a_3+a_4), a_4, \frac{1}{2}\right)$$

and where $\varepsilon_3(n) \leq \sum_{i=0}^{2}\mathbb{P}(V_i^c)$, hence

$$-\log\varepsilon_3(n) = \Omega\left(n^{a_3(a_1-\frac{1}{2})-\max(1-a_1,a_2)}\right).$$

To conclude, we summarize the constraints on the constants $a_i \in (0,1)$ which guarantee that $\varepsilon_i(n)\to 0$, so that the empirical capacity is achieved:

$$\max(1-a_1,a_2) < a_3\left(a_1-\frac{1}{2}\right), \quad a_3+a_4<\frac{1}{2}.$$

There are many parameter selections that satisfy the conditions above, e.g., $(a_1,a_2,a_3,a_4) = (\frac{7}{8},\frac{1}{8},\frac{3}{8},\frac{1}{16})$.

Appendix C
A Universal Scheme for $\mathcal{C}_{\mathcal{X}}$ Utilizing Common Randomness

In this appendix, we show how the horizon-free universal scheme developed in the previous appendix can be modified, using common randomness, to achieve the empirical capacity over the larger family $\mathcal{C}_{\mathcal{X}}$ of all causal channels. To that end, we first describe a more general communication setting using common randomness, and later show how our scheme is adapted into this setting. Note that only passive feedback of the received sequence is assumed, due to the availability of common randomness. A *feedback transmission scheme with common randomness* is a triplet $(G,\mathcal{P},\Delta)$ and can operate either *with or without dithering* (the definitions and details appear below). Using the scheme over a channel $\mathcal{W}\in\mathcal{C}_{\mathcal{X}}$ with a message point $\theta_0\in[0,1)$, is described by the following construction.

- Common randomness resources are assumed to be available in the following form.
  - A *control sequence* A^∞ of independent r.v.'s taking values over some countable alphabet $\mathcal{A}$, with a given sequence of marginal distributions $\mathcal{P} \triangleq \{P_k(\cdot) = \mathbb{P}_{A_k}(\cdot)\}_{k=1}^\infty$.
  - An i.i.d. *dithering sequence* Φ^∞ taking values over $\mathcal{X}$. When the scheme operates with dithering then $\Phi_k\sim\mathrm{Uniform}(\mathcal{X})$, and when it operates without dithering we set $\Phi_k=0$ for any $k\in\mathbb{N}$.
- $(\tilde{X}^\infty,\tilde{Y}^\infty)$ constitute an input/output pair for the channel $\mathcal{W}\in\mathcal{C}_{\mathcal{X}}$.
- (X^∞,Y^∞) are related to the input/output pair via

$$\tilde{X}_k = X_k+\Phi_k \quad \tilde{Y}_k = Y_k+\Phi_k,\quad k\in\mathbb{N}. \tag{68}$$

- For any message point $\theta_0\in[0,1)$ and any $k\in\mathbb{N}$

$$X_k = g_k(\theta_0,Y^{k-1},A^k) \tag{69}$$

  where $G=\{g_k : [0,1)\times\mathcal{X}^{k-1}\times\mathcal{A}^k\mapsto\mathcal{X}\}_{k=1}^\infty$ is a sequence of *transmission functions*.
- A_k is statistically independent of $(X^{k-1},Y^{k-1},\Phi^{k-1},A^{k-1})$ for any $k\in\mathbb{N}$.
- Φ_k is statistically independent of $(X^k,Y^{k-1},\Phi^{k-1},A^k)$ for any $k\in\mathbb{N}$.
- The following Markov relation holds for any $k\in\mathbb{N}$:

$$\tilde{Y}_k \leftrightarrow \tilde{X}^k\tilde{Y}^{k-1}\leftrightarrow A^k\Phi^k. \tag{70}$$

  Loosely speaking, this relation guarantees privacy of randomness resources, namely, that the adversary/channel cannot utilize common randomness shared by the terminals. This is the common randomness counterpart of (9).
- $\Delta=\{\Delta_k : \mathcal{X}^k\times\mathcal{A}^{k-1}\mapsto\mathfrak{J}\}_{k=1}^\infty$ is a sequence of *decoding rules*, such that $\Delta_k(Y^k,A^{k-1})$ is the decoded interval at time k.

For any given channel $\mathcal{W}\in\mathcal{C}_{\mathcal{X}}$, feedback transmission scheme $(G,\mathcal{P},\Delta)$ with/without dithering and message point $\theta_0\in[0,1)$, the above construction uniquely determines the joint statistics of $(X^\infty,Y^\infty,\tilde{X}^\infty,\tilde{Y}^\infty,\Phi^\infty,A^\infty)$. The error probability $p_e(n,\mathcal{W},\theta_0)$ and instantaneous rate $R_n(\mathcal{W},\theta_0)$ are defined similarly to (10), with $\Delta_n(Y^n,A^{n-1})$ replacing $\Delta_n(Y^n,U^{n-1})$. The empirical capacity $C_n^{\mathrm{emp}}(\mathcal{W},\theta_0)$ is defined as in (11), using the realized noise sequence Z^∞ pertaining to the input/output pair $(\tilde{X}^\infty,\tilde{Y}^\infty)$. A scheme is said to *locally achieve* the empirical capacity for a specific channel/message point pair $(\mathcal{W},\theta_0)$, if

$$p_e(n,\mathcal{W},\theta_0)<\varepsilon_1(n)$$
$$\mathbb{P}\left(R_n(\mathcal{W},\theta_0)>C_n^{\mathrm{emp}}(\mathcal{W},\theta_0)-\varepsilon_2(n)\right)>1-\varepsilon_3(n)$$

for some $\varepsilon_1(n),\varepsilon_1(n),\varepsilon_1(n)\to 0$. As in (12), the scheme is said to (uniformly) achieve the empirical capacity over a family of channels $\mathcal{F}$, if the above is satisfied uniformly over $\mathcal{W}\in\mathcal{F}$ and $\theta_0\in[0,1)$.

Assuming the scheme operates with dithering, let $\mathcal{W}^\dagger\in\mathcal{C}_{\mathcal{X}}$ denote the causal channel induced by the input/output pair (X^∞,Y^∞), i.e., the channel defined by

$$\mathcal{W}^\dagger(G,\mathcal{P},\Delta,\mathcal{W},\theta_0)\triangleq\Big\{W_k^\dagger(y_k|x^k,y^{k-1}) = \mathbb{P}_{Y_k|X^kY^{k-1}}(y_k|x^k,y^{k-1})\Big\}_{k=1}^\infty. \tag{71}$$

The *induced channel* $\mathcal{W}^\dagger$ depends in general on the transmission scheme and the message point, and is therefore not a "true channel" in the regular operational sense. Moreover, generally $\mathcal{W}^\dagger\notin\mathcal{M}_{\mathcal{X}}$ despite the modulo-additive dithering, due to the statistical coupling generated by feedback.[22] Nevertheless, the following observation provides an operational meaning to the induced channel.

Lemma 9: Fix a channel $\mathcal{W}\in\mathcal{C}_{\mathcal{X}}$ and a message point $\theta_0\in[0,1)$. Let $\mathcal{W}^\dagger=\mathcal{W}^\dagger(G,\mathcal{P},\Delta,\mathcal{W},\theta_0)$ be the corresponding induced channel. The following two statements are equivalent.

(i) The scheme $(G,\mathcal{P},\Delta)$ operating *with dithering* locally achieves the empirical capacity for $(\mathcal{W},\theta_0)$, with the convergence parameters $\varepsilon_1(n),\varepsilon_2(n),\varepsilon_3(n)$.

(ii) The scheme $(G,\mathcal{P},\Delta)$ operating *without dithering* locally achieves the empirical capacity for $(\mathcal{W}^\dagger,\theta_0)$, with the convergence parameters $\varepsilon_1(n),\varepsilon_2(n),\varepsilon_3(n)$.

[22]Note however that in the special case where $\mathcal{W}$ is memoryless, the induced channel $\mathcal{W}^\dagger$ is independent of the transmission scheme and the message point, is memoryless and modulo-additive, and is obtained by averaging "cyclicly shifted" versions of $\mathcal{W}$ (see the discussion in the end of Section III for the binary case).

Proof: For any feedback transmission scheme operating with or without dithering, the decoded interval Δ_n is a function of (Y^n, A^{n-1}), and the rate and error probability are in turn functions of Δ_n. Moreover, the realized noise sequence corresponding to $(\tilde{X}^\infty, \tilde{Y}^\infty)$ and to (X^∞, Y^∞) is exactly the same sequence due to (68), hence, the empirical capacity $C_n^{\rm emp}(\mathcal{W}, \theta_0)$ is a function of (X^n, Y^n). Therefore, in general, $(\Delta_n, p_e, R_n, C_n^{\rm emp})$ are functions of (X^n, Y^n, A^{n-1}). Now for case (ii) above, the induced channel $\mathcal{W}^\dagger$ together with θ_0 and $(G, \mathcal{P}, \Delta)$ uniquely defines the joint distribution of $(X^\infty, Y^\infty, A^\infty)$. But by definition, this distribution must coincide with the joint distribution of $(X^\infty, Y^\infty, A^\infty)$ obtained in case (i), concluding the proof. □

It should be emphasized that the two statements in the lemma above correspond to two separate constructions. The following important observation is due.

Lemma 10: Let $\mathcal{W} \in \mathcal{C}_\mathcal{X}$, and suppose the scheme $(G, \mathcal{P}, \Delta)$ operates without dithering over the channel $\mathcal{W}^\dagger(G, \mathcal{P}, \Delta, \mathcal{W}, \theta_0)$ with the message point θ_0. Then for any $a \in \mathcal{A}$, the indicator sequence $\{\mathbb{1}_a(A_k)\}_{k=1}^\infty$ is a (not necessarily identically distributed) causal sampling sequence for the noise sequence Z^∞.

Proof: This is an analogue to the statement made in Lemma 6, and the proof is of the same spirit. We will prove for the case where $(G, \mathcal{P}, \Delta)$ operates with dithering over $\mathcal{W}$ with θ_0, and the result will follow as in Lemma 9, since the distribution of (Z^∞, A^∞) under both settings is the same. Clearly, $\mathbb{1}_a(A_k) \sim \mathrm{Ber}(P_k(a))$ and the indicator sequence is not necessarily identically distributed. However as we now show, A_k is statistically independent of (Z^k, A^{k-1}) for any $k \in \mathbb{N}$, from which the result follows immediately. Since A^∞ is a sequence of independent r.v.'s it is sufficient to show that $Z^k \leftrightarrow A^{k-1} \leftrightarrow A_k$. To this end, we repeat the derivation in (48) to the letter, replacing transition justifications $\mathrm{(a)}$ and $\mathrm{(b)}$ with the following (a^*) and (c^*), respectively:

(a*) $Z_k \leftrightarrow X^{k-1}Y^{k-1}A^{k-1} \leftrightarrow A_k$.

Proof: Again we omit the r.v. subscripts where there is no confusion, vector additions over $\mathcal{X}^k$ are taken to be element by element modulo-addition.

$$
\begin{aligned}
&\mathbb{P}(z_k|x^{k-1}, y^{k-1}, a^k)\\
&= \sum_{\phi^k} \mathbb{P}(\phi_k|\phi^{k-1}, x^{k-1}, y^{k-1}, a^k)\\
&\quad \times \mathbb{P}(\phi^{k-1}|x^{k-1}, y^{k-1}, a^k)\\
&\quad \times \mathbb{P}(z_k|x^{k-1}, y^{k-1}, a^k, \phi^k)\\
&\overset{\rm (a1)}{=} \sum_{\phi^k} \mathbb{P}(\phi_k)\mathbb{P}(\phi^{k-1}|x^{k-1}, y^{k-1}, a^k)\\
&\quad \times \mathbb{P}_{\tilde{Y}_k|\tilde{X}^k, \tilde{Y}^{k-1}A^k\Phi^k}(g_k(\theta_0, y^{k-1}, a^k)\\
&\quad + \phi_k + z_k | g_k(\theta_0, y^{k-1}, a^k) + \phi_k, x^{k-1}\\
&\quad + \phi^{k-1}, y^{k-1} + \phi^{k-1}, a^k, \phi^k)\\
&\overset{\rm (a2)}{=} |\mathcal{X}|^{-1} \sum_{\phi^{k-1}} \Bigg[\mathbb{P}(\phi^{k-1}|x^{k-1}, y^{k-1}, a^{k-1})\\
&\quad \times \sum_{\phi_k \in \mathcal{X}} W_k(g_k(\theta_0, y^{k-1}, a^k)\\
&\quad + \phi_k + z_k | g_k(\theta_0, y^{k-1}, a^k)\\
&\quad + \phi_k, x^{k-1} + \phi^{k-1}, y^{k-1} + \phi^{k-1}) \Bigg]\\
&\overset{\rm (a3)}{=} |\mathcal{X}|^{-1} \sum_{\phi^{k-1}} \mathbb{P}(\phi^{k-1}|x^{k-1}, y^{k-1}, a^{k-1})\\
&\quad \sum_{\phi'_k \in \mathcal{X}} W_k(\phi'_k + z_k|\phi'_k, x^{k-1} + \phi^{k-1}, y^{k-1} + \phi^{k-1})\\
&\overset{\rm (a4)}{=} \mathbb{P}(z_k|x^{k-1}, y^{k-1}, a^{k-1})
\end{aligned}
$$

where transitions are justified as follows:

(a1) Φ_k is statistically independent of $(X^{k-1}, Y^{k-1}, \Phi^{k-1}, A^k)$, together with (68) and (69).

(a2) Φ_k is uniformly distributed, A_k is statistically independent of $(X^{k-1}, Y^{k-1}, \Phi^{k-1})$, the Markov relation (70), and the definition of the channel $\mathcal{W}$.

(a3) A change of variables $\phi'_k = g_k(\theta_0, y^{k-1}, a^k) + \phi_k$ reveals that the inner sum does not depend on the value of $g_k(\theta_0, y^{k-1}, a^k)$.

$(a4)$ The dependence of the expression on a_k has been removed.

(c*) By construction, A_{k+1}^∞ is independent of (X^k, Y^k, A^k). □

Our horizon-free finite-alphabet universal scheme is now easily adapted to use common randomness within the framework of this section, as follows. First, active positions are removed (i.e., $b_a = 0$). Instead, the type of each position and the repetition position information for the update bits are directly provided by the control sequence A^∞. This is achieved (say) by using an alphabet $\mathcal{A} = \{\text{training, regular}\} \cup \mathbb{N} \cup \{0\}$, where numerical values correspond to update positions and determine which update bit is to be transmitted using which input pair, taking the place of the Γ^{M_u} described in Section IV-E. Thus, X_k is now generated from (θ_0, Y^{k-1}, A^k) instead of from (θ_0, U^{k-1}). Finally, the sequence of marginal distributions $\mathcal{P}$ is suitably defined taking into account the removal of active positions (which can only improve the redundancy term). Namely, for any position j within the kth block we have $P_j(\text{training}) = m_k b_k^{-1}$, $P_j(\text{regular}) = 1 - 2m_k b_k^{-1}$, and a uniform distribution over the numerical values $\langle s_k \rangle$ which constitute the rest of the support of $P_j(\cdot)$ (where $s_k = 2|\mathcal{X}|\lceil \log(k+1) \rceil$ corresponds to the number of update bits, see (49)). Given the modifications described above, the adapted universal transmission scheme under the new construction, either operating with or without dithering, is well defined.

We are now ready to show that the adapted scheme with dithering achieves the empirical capacity over $\mathcal{C}_\mathcal{X}$. Clearly, the adapted scheme without dithering is essentially equivalent to the scheme without common randomness discussed in previous sections (up to the minor issue of active feedback replaced by common randomness), and by repeating the same proof it is readily verified that it achieves the empirical capacity over $M_\mathcal{X}$ as well. Moreover, note that the fact that $\mathcal{W} \in M_\mathcal{X}$ was used in that proof solely for the sake of Lemma 6, namely, to show that the training pattern sequence and each of the update pattern sequences constitute causal sampling sequences for the noise sequence within each block. Now, for a given channel $\mathcal{W} \in \mathcal{C}_{\mathcal{X}}$ and a message point $\theta_0 \in [0,1)$, suppose the adapted scheme operates without dithering over the corresponding induced channel $\mathcal{W}^{\dagger}(G, \mathcal{P}, \Delta, \mathcal{W}, \theta_0)$ with the same message point θ_0. In this case, Lemma 10 verifies that under the construction considered in this section, it still holds that the training and update pattern sequences constitute causal sampling sequences for the noise sequence within each block. Therefore, we conclude that for any $\mathcal{W} \in \mathcal{C}_{\mathcal{X}}$ and $\theta_0 \in [0,1)$, the adapted scheme operating without dithering locally achieves the empirical capacity for $(\mathcal{W}^{\dagger}, \theta_0)$. Furthermore, note that although the induced channel $\mathcal{W}^{\dagger}$ depends both on the message point and on the channel $\mathcal{W}$, the convergence parameters $\varepsilon_1(n), \varepsilon_2(n), \varepsilon_3(n)$ do not. Finally, according to Lemma 9, the above implies that the adapted scheme with dithering locally achieves the empirical capacity for any pair of channel $\mathcal{W} \in \mathcal{C}_{\mathcal{X}}$ and message point $\theta_0 \in [0,1)$, with convergence parameters $\varepsilon_1(n), \varepsilon_2(n), \varepsilon_3(n)$ independent of $(\mathcal{W}, \theta_0)$. Hence, by definition this scheme uniformly achieves the empirical capacity over the family $\mathcal{C}_{\mathcal{X}}$, and the proof is concluded.

As discussed in Section III, when operating over $\mathcal{C}_{\mathcal{X}}$, a uniform input distribution is essential in order for the defined modulo-additive empirical capacity to be meaningful, and in turn achievable. The discussion in this section reveals the operational significance of this requirement within the framework of our universal scheme. The entire scheme hinges on the ability of a training sample to estimate the empirical capacity of a block, and on the capability to reliably transmit update information over any block whose empirical capacity is not too small. Roughly speaking, a uniform input distribution (obtained here by dithering) guarantees that with high probability, the empirical distribution of the realized noise sequence over an i.i.d. sample (i.e., training or update positions) is close to that of the entire realized noise sequence.

REFERENCES

[1] C. E. Shannon, "The zero-error capacity of a noisy channel," *IRE. Trans. Inf. Theory*, vol. IT-2, no. 3, pp. 8–19, Sep. 1956.

[2] F. Alajaji, "Feedback does not increase the capacity of discrete channels with additive noise," *IEEE Trans. Inf. Theory*, vol. 41, no. 2, pp. 546–549, Mar. 1995.

[3] T. T. Kadota, M. Zakai, and J. Ziv, "Capacity of a continuous memoryless channel with feedback," *IEEE Trans. Inf. Theory*, vol. IT-17, no. 4, pp. 372–378, Jul. 1971.

[4] M. Horstein, "Sequential transmission using noiseless feedback," *IEEE Trans. Inf. Theory*, vol. IT–9, no. 3, pp. 136–143, Ju. 1963.

[5] J. P. M. Schalkwijk and K. A. Post, "On the error probability for a class of binary recursive feedback strategies," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 4, pp. 498–511, Jul. 1973.

[6] J. P. M. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback part I: No bandwidth constraint," *IEEE Trans. Inf. Theory*, vol. IT-12, no. 2, pp. 172–182, Apr. 1966.

[7] H. Yamamoto and K. Itoh, "Asymptotic performance of a modified Schalkwijk-Barron scheme for channels with noiseless feedback," *IEEE Trans. Inf. Theory*, vol. IT-25, no. 6, pp. 729–733, Nov. 1979.

[8] M. Horstein, "Sequential Transmission of Digital Information With Feedback," Res. Lab. Electronics, MIT, Cambridge, MA, Tech. Rep. 375, 1960.

[9] J. M. Ooi and G. W. Wornell, "Fast iterative coding techniques for feedback channels," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2960–2976, Nov. 1998.

[10] N. Shulman, "Communication over Unknown Channels via Common Broadcasting," doctoral dissertation, Tel-Aviv University, Tel-Aviv, Israel, 2003.

[11] N. Shulman and M. Feder, "Static broadcasting," in *Proc. IEEE Int. Symp. Information Theory*, Sorrento, Italy, Jun. 2000, p. 23.

[12] A. Tchamkerten and İ. E. Telatar, "Variable length coding over unknown channels," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 2126–2145, May 2006.

[13] M. V. Burnashev, "Data transmission over a discrete channel with feedback: Random transmission time," *Probl. Inf. Transm.*, vol. 12, no. 4, pp. 250–265, 1976.

[14] J. M. Ooi, "A Framework for Low-Complexity Communication Over Channels With Feedback," Ph.D. dissertation, MIT, Cambridge, MA, 1998.

[15] E. R. Berlekamp, "Block Coding with Noiseless Feedback," Ph.D. dissertation, MIT, Cambridge, MA, 1964.

[16] K. Sh. Zigangirov, "Number of correctable errors for transmission over a binary symmetrical channel with feedback," *Probl. Inf. Transm.*, vol. 12, pp. 85–97, 1976.

[17] D. Blackwell, L. Breiman, and A. J. Thomasian, "The capacities of certain channel classes under random coding," *Ann. Math. Statist.*, vol. 31, pp. 558–567, 1960.

[18] R. Ahlswede, "Elimination of correlation in random codes for arbitrarily varying channels," *Z. Wahrscheinlichkeitstheorie Verw. Gehiete*, vol. 44, pp. 159–175, 1978.

[19] I. Csiszár and P. Narayan, "The capacity of the arbitrarily varying channel revisited: Positivity, constraints," *IEEE Trans. Inf. Theory*, vol. 43, no. 2, pp. 181–193, Mar. 1988.

[20] M. Langberg, "Private codes or succinct random codes that are (almost) perfect," in *Proc. Conf. Foundation of Computer Science (FOCS)*, Oct. 2004, pp. 325–334.

[21] R. Ahlswede and I. Csiszár, "Common randomness in information theory and cryptography—Part II: CR capacity," *IEEE Trans. Inf. Theory*, vol. 44, no. 1, pp. 225–240, Jan. 1998.

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

[23] Y. M. Shtrakov, "Universal sequential coding of single messages," *Probl. Inf. Transm.*, vol. 23, pp. 175–186, Jul. 1987.

[24] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, vol. 58, no. 301, pp. 13–30, Mar. 1963.

[25] F. Chung and L. Lu, "Concentration inequalities and martingale inequalities: A survey," *Internet Math.*, vol. 3, no. 1, pp. 79–127, 2006.

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

[27] O. Shayevitz and M. Feder, "The posterior matching feedback scheme: Capacity achieving and error analysis," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Toronto, ON, Canada, Jul. 2008, pp. 900–904.

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

Ofer Shayevitz (S'04–M'08) received the B.Sc. degree (*summa cum laude*) and the M.Sc. degree (*magna cum laude*) both in electrical engineering, from the Technion–Israel Institute of Technology, Haifa, Israel, and the Tel-Aviv University, Tel-Aviv, Israel, in 1997 and 2004, respectively. He has concluded his Ph.D. studies in electrical engineering in the Tel-Aviv University in 2008.

He is currently a Postdoctoral Fellow in the Information Theory and Applications (ITA) Center at the University of California, San Diego (UCSD). He has served as an Engineer in the Israeli Defense Forces (1997–2003), and as an Algorithms Engineer at CellGuide (2003–2004), a high-tech company developing low-power GPS navigation solutions.

Dr. Shayevitz is the recipient of the Adams Fellowship (2006–2008) awarded by the Israel Academy of Sciences and Humanities, of the Weinstein award (2006–2008), and was a member of the Technion Excellence program (1993–1997).

Meir Feder (S'81–M'93–SM'98–F'08) received the B.Sc. and M.Sc. degrees from Tel-Aviv University, Tel-Aviv, Israel and the Sc.D degree from the Massachusetts Institute of Technology (MIT), Cambridge, and the Woods Hole Oceanographic Institution, Woods Hole, MA, all in electrical engineering in 1980, 1984, and 1987, respectively.

After being a Research Associate and Lecturer at 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 cofounded 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 ASICs for wireless high-definition A/V connectivity at the home.

Prof. Feder is a corecipient 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, U.K., awarded by Ben-Gurion University. Between June 1993 and June 1996 he served as an Associate Editor for Source Coding of the IEEE TRANSACTIONS ON INFORMATION THEORY.