Universal Communication over Arbitrarily Varying Channels

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

***Abstract*—Consider the problem of universally communicating over an arbitrarily varying channel, i.e., a channel comprised of an unknown, arbitrary sequence of memoryless channels. Interestingly, it is shown that there is a communication system using feedback and common randomness that asymptotically attains, with high probability, the capacity of the time-averaged channel, universally for every sequence of channels. This attainable rate is optimal under certain conditions. While no prior knowledge of the channel sequence is assumed, the capacity of the time-averaged channel meets or exceeds the traditional arbitrarily varying channel (AVC) capacity for every memoryless AVC defined over the same alphabets, and therefore the system universally attains the random code AVC capacity, without knowledge of the AVC parameters. The presented system combines rateless coding with a universal prediction scheme for the input "prior" distribution. The determination of the input behavior by universally predicting the prior used to randomly generate the codebook, plays a major role in the presented result.**

I. Introduction

Let us consider the problem of communicating over an unknown and arbitrarily varying channel, with the help of feedback. The target is to minimize the assumptions on the communication channel as much as possible, while using the feedback link to learn the channel. The main questions with respect to such channels are how to define the expected communication rates, and how to attain them universally, without channel knowledge.

The traditional models for unknown channels [1] are compound channels, in which a fixed channel law is selected arbitrarily out of a family of known channels, and arbitrarily varying channels (AVC's), in which a sequence of channel states is selected arbitrarily. The well known results for these models [1] do not assume adaptation. Therefore, the AVC capacity, which is the supremum of the communication rates that can be obtained with vanishing error probability over any possible occurrence of the channel state sequence, is in essence a worst-case result. For example, if one assumes that y_i, the channel output at time i, is determined by the probability law $W_i(y_i|x_i)$ where x_i is the channel input, and W_i is an arbitrary sequence of conditional distributions, clearly no positive rate can be guaranteed a-priori, as it may happen that all W_i have zero capacity. Therefore, the AVC capacity is zero. This capacity may be non-zero only if a constraint on W_i is defined. In this paper the term "arbitrarily varying channel" is used in a loose manner, to describe any kind of unknown and arbitrary change of the channel over time, while the acronym "AVC" refers to the traditional model [1].

Other communication models, which allow positive communication rates over such AVC's were proposed by the authors and others [2], [3], [4], [5]. Although the channel models considered in these papers are different, the common feature distinguishing them from the traditional AVC setting is that the communication rate is adaptively modified using feedback. The target rate is known only a-posteriori, and is gradually learned throughout the communication process. By adapting the rate, one avoids worst case assumptions on the channel, and can achieve positive communication rates when the channel is good. However, in the aforementioned communication models, the distribution of the transmitted signal is fixed and independent of the feedback, and only the rate is adapted. Specifically in the "individual channel" model [4] for reasons explained therein, the distribution of the channel input is fixed to a predefined prior. Likewise, Eswaran *et al* [3] show that for a fixed prior, the mutual information of the averaged channel can be attained. Clearly, with this limitation, these systems are incapable of universally attaining the channel capacity in many cases of interest. Even in the simple case where the channel is a compound memoryless channel, i.e. the conditional distributions $W_i = W$ are all constant but unknown, capacity cannot be attained this way.

In a more recent paper [5], the problem of universal communication was formulated as that of a competition against a reference system, comprised of an encoder and a decoder with limited capabilities. For the case where the channel is modulo-additive with an individual, arbitrary noise sequence, it was shown possible to asymptotically perform at least as well as any finite-block system (which may be designed knowing the noise sequence), without prior knowledge of the noise sequence. However, this result crucially relies on the property of the modulo-additive channel, that the capacity achieving prior is the uniform i.i.d. prior for any noise distribution. To extend the result to more general models, the input behavior needs to be adapted. The key parameter to be adapted is the "prior", i.e. the distribution of the codebook (or equivalently the channel input), since it plays a vital role in the converse as well as the attainability proof of channel capacity and is the main factor in adapting the message to the channel [6].

Loosely speaking, previous works achieve various kinds of "mutual information" for a fixed prior and any channel from a wide class, by mainly solving problems of universal decoding

Parts of this paper were presented at ISIT-2011, St.Petersburg, and CISS-2012, Princeton [expected]

and rate adaptation. However to obtain more than the "mutual information", i.e. the "capacity", the prior would need to be selected in a universal way.

Prior adaptation using feedback is a well known practice for static or semi-static channels. Two familiar examples are bit and power loading performed in Digital Subscriber Lines (DSL-s) [7], and precoding for in multi-antenna systems [8] which is performed in practice in wireless standards such as WiFi, WiMAX and LTE. If the channel can be assumed to be static for a period of time sufficient to close a loop of channel measurement, feedback and coding, then an input prior close to the optimal one can be chosen. In the theoretical setting of the compound memoryless channel where $\Pr(Y_i|X_i) = W(Y_i|X_i)$, where W is unknown but fixed, a system with feedback can asymptotically attain the channel capacity of W, without prior knowledge of it, by using an asymptotically small portion of the transmission time to estimate the channel, and using an estimate of the optimal prior and the suitable rate during the rest of the time [9]. All models for prior adaptation that we are aware of, use the assumption that the knowledge of the channel at a given time yields non trivial statistical information about future channel states, but do not deal with arbitrary variation.

The question dealt with in this paper is: assuming a channel which is *arbitrarily* changing over time, is there any merit in using feedback to adapt the input distribution, and what rates can be guaranteed? Although the goal is to cope with the most general variation of the channel (as in the unknown vector channel model [5]), to start this exploration, let us focus on channel models which are memoryless in the input, i.e. whose behavior at a certain time does not depend on any previous channel *inputs*. Specifically, the model assumed here is of an unknown sequence of memoryless channels (which is in essence an AVC without constraints). The motivation for avoiding memory of the input can be appreciated by considering the negative examples in [5].

Following is a brief overview of the structure and the results of this paper. In Section II the problem is stated, and several communication rates of interest are defined (as a function of the channel sequence). In order to focus thoughts on questions related to the problem of determining the *prior*, an abstract model of the communication system is initially adopted, stripping off the details of communication, such as decoding, channel estimation, overheads, error probability, etc. An easier synthetic problem is first presented, in which all previous channels are known (Section III). This problem may represent a "realistic" case where the channel changes its behavior in a block-wise manner and remains i.i.d. memoryless during each block (a subset of the original problem). This problem is related to standard prediction problems (Section III-B), and used as a tool to gain insight into the prediction problem involved, present bounds on what can be achieved universally, and develop the techniques that will be used later on. Even for this easier problem, it is shown that there is no hope to attain the channel capacity universally and one would have to settle for lower rates (Section III-C). The attained rate is the maximum over the prior, of the averaged mutual information (Theorem 1). In Section IV, returning to the main problem, it is shown that the previously attained rate is no longer attainable. On the other hand, the capacity of the time-averaged channel is the best achievable rate that does not depend on the order of the channel sequence (Theorem 2), this rate is indeed achievable (Theorem 3). Furthermore, this rate meets or exceeds the AVC capacity, and essentially equals the "empirical capacity" defined by Eswaran *et al* [3]. The scheme that attains this rate is based on rateless coding and combines a prior predictor. In Section IV-C, the communication scheme and the prior predictor are presented, and in Section V the main result (Theorem 3) is proven. Section VI is devoted to discussion and comments.

II. Notation and problem statement

A. Notation

Random variables are denoted by capital letters and vectors by boldface. However, for probability distributions, which are sometimes treated as vectors, regular capital letters are used. Superscript and subscript indices are applied to vectors to define subsequences in the standard way, i.e. $\mathbf{x}_i^j \triangleq (x_i, x_{i+1}, ..., x_j)$, $\mathbf{x}^i \triangleq \mathbf{x}_1^i$

$I(Q,W)$ denotes the mutual information obtained when using a prior Q over a channel W, i.e. it is the mutual information $I(Q,W) = I(X;Y)$ between two random variables with the joint probability $\Pr(X,Y) = Q(X) \cdot W(Y|X)$. $C(W)$ denotes the channel capacity $C(W) = \max_Q I(Q,W)$. For discrete channels, the channel $W(y|x)$ is sometimes presented as a matrix where $W(y|x)$ is in the x-th column and the y-th row. Logarithms and all information quantities are base 2 unless specified otherwise.

The unit simplex, i.e. the set of all probability measures on $\mathcal{X}$, is denoted by $\Delta_{\mathcal{X}} \triangleq \{Q : \sum_{x\in\mathcal{X}} Q(x) = 1\}$.

$\mathrm{Ber}(p)$ denotes a Bernoulli random variable with probability p to be 1. $\mathrm{Ind}(\cdot)$ denotes an indicator function of an event or a condition, and equals 1 if the event occurs and 0 otherwise. The notation "..." is used to denote simple mathematical inductions, where the same rule is repeatedly applied, for example $a_n \le n \cdot a_{n-1} \le \ldots \le n! \cdot a_0$.

A hat $\hat{\square}$ denotes an estimated value, and a line $\overline{\square}$ denotes an average value. The empirical distribution of a vector $\mathbf{x}$ of length n is a function representing the relative frequency of each letter,

$$\hat{P}_{\mathbf{x}}(x) = \frac{\sum_{i=1}^{n} \mathrm{Ind}(x_i = x)}{n}, \tag{1}$$

where the subscript identifies the vector. The conditional empirical distribution of two equal length vectors $\mathbf{x}, \mathbf{y}$ is defined as

$$\hat{P}_{\mathbf{y}|\mathbf{x}}(y|x) = \frac{\hat{P}_{\mathbf{x},\mathbf{y}}(x,y)}{\hat{P}_{\mathbf{x}}(x)}. \tag{2}$$

B. Problem setting

Let $\mathcal{X}, \mathcal{Y}$ be sets defining the input and output alphabets, respectively. Both $\mathcal{X}, \mathcal{Y}$ are assumed to be finite, unless stated otherwise.[1] Let $\{W_i\}_{i=1}^n$ be a sequence of memoryless channels over n channel uses. Each W_i is a conditional distribution

[1]Note that the results in Section III,IV do not require $\mathcal{Y}$ to be finite

$W_i(y|x)$ where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ represent an input and output symbol respectively. The conditional distribution of the output vector $\mathbf{Y}$ given the input vector $\mathbf{X}$ is given by:

$$\Pr(\mathbf{Y}|\mathbf{X}) = \prod_{i=1}^{n} W_i(Y_i|X_i). \tag{3}$$

The sequence of channels W_i is arbitrary and unknown to the transmitter and the receiver. The existence of common randomness (i.e. that the transmitter and the receiver both have access to some random variable of choice) is assumed. There exists a feedback link between the receiver and the transmitter. To simplify, let us assume the feedback is completely reliable, has unlimited bandwidth and is instantaneous, i.e. arrives to the encoder before the next symbol.[2] The system is rate adaptive, which means that the message is represented by an infinite bit sequence $\mathbf{m}_0^\infty$, and the system may choose how many bits to send. The error probability is measured only over the bits which were actually sent (i.e. over the first $\lceil nR \rceil$ bits, where R is the rate reported by the receiver). The system setup is presented in Figure 1.

To simplify, it is assumed that there are no constraints on the channel input (such as power constraints). If such constraints exist they can be accommodated by changing the set of potential priors.

Since the channel sequence is arbitrary there is no positive rate which can be guaranteed a-priori. Instead, a target rate $R(W_1^n)$ can be defined as a function of the channel sequence W_1^n.

Definition 1. A sequence of rate functions $R(W_1^n)$ is said to be asymptotically attainable, if for every $\epsilon, \delta, \Delta > 0$ there is n large enough such that there is a system with feedback and common randomness over n channel uses, in which, for *every* sequence $\{W_i\}_{i=1}^n$, the rate is $R(W_1^n) - \Delta$ or more, with probability of at least $1 - \delta$, while the probability of error is at most ϵ.

In the next section several potential target rates are proposed, and in what follows, we would ask which of these are attainable.

C. Potential target rates

With respect to the sequence $\{W_i\}$ various meaningful information theoretic measures can be defined. The maximum possible rate of reliable communication is the capacity when the sequence is known a-priori (in other words, the capacity with full, non causal, channel state information at the transmitter and the receiver) and is given by:

$$\begin{aligned} C_1(W_1^n) &= \max_{\{Q_i\}} \frac{1}{n} \sum_{i=1}^{n} I(Q_i, W_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} \max_{Q} I(Q, W_i) = \frac{1}{n} \sum_{i=1}^{n} C(W_i). \end{aligned} \tag{4}$$

Note that if constraints on the sequence $\{Q_i\}$ existed, (4) would be an inequality (see [10]). The maximum rate that can be obtained with a single *fixed* prior when the sequence is known is:

$$C_2(W_1^n) = \max_{Q} \frac{1}{n} \sum_{i=1}^{n} I\left(Q, W_i\right). \tag{5}$$

Lastly, the capacity of the time-averaged channel is:

$$C_3(W_1^n) = \max_{Q} I\left(Q, \frac{1}{n} \sum_{i=1}^{n} W_i\right) = C(\overline{W}), \tag{6}$$

where the time-averaged channel is defined as

$$\overline{W}(y|x) = \frac{1}{n} \sum_{i=1}^{n} W_i(y|x). \tag{7}$$

Clearly, $C_1 \geq C_2 \geq C_3$ where the first inequality results from the order of maximization and the other results from the convexity of the mutual information with respect to the channel. For each of the above target rates we would like to find out whether it is achievable under the definitions above. As shall be seen, C_1 is not achievable, C_3 is achievable, and C_2 is achievable only under further constraints imposed on the problem.

A rigorous proof that C_1 is the capacity of the channel sequence is left out of the scope of this paper. For our purpose, it is sufficient to observe that C_1 is an upper bound on the achievable rate, because the mutual information between channel input and output is maximized by a memoryless (not i.i.d.) input distribution $\prod_{i=1}^{n} Q_i(x_i)$. To see intuitively how C_1 can be achieved, consider that since n can be arbitrarily large while the input and output alphabets, and thus the set of channels, remain constant, one may sort the channels into groups of similar channels, and apply block coding to each group. A close result pertaining to stationary ergodic channels appears in [11, (3.3.5)].

III. A SYNTHETIC "TOY" PROBLEM

In this section a synthetic problem is presented. This problem will help examine the achievability of the target rates defined above in a simplified scenario, draw the links to universal prediction, and introduce the techniques that will be used in the sequel.

A. Problem description

Let us focus on the problem of setting a prior $\hat{Q}_i$ at time i. Assume that at each time instance i, the system has full knowledge of the sequence of past channels W_1^{i-1}. The prior prediction mechanism sets $\hat{Q}_i$ based on the knowledge of W_1^{i-1}. Then, $I(\hat{Q}_i, W_i)$ bits are conveyed during time instance i. A predictor $\hat{Q}_i(W_1^{i-1})$ attains a given target rate $R(W_1^n)$ if $\frac{1}{n}\sum_{i=1}^{n} I(\hat{Q}_i, W_i) \geq R(W_1^n) - \delta_n$ for all sequences W_1^n, and $\delta_n \underset{n\to\infty}{\longrightarrow} 0$.

This abstract problem can apply to a situation where the channel sequence is constant during long blocks, and changes its value only from block to block, or from one transmission to another. In this case i denotes the block index, and denoting by m the constant block length, at most $m \cdot I(\hat{Q}_i, W_i)$ bits can be sent in block i. If the channel is constant over long

[2] The asymptotical results hold also when feedback is band limited and delayed.

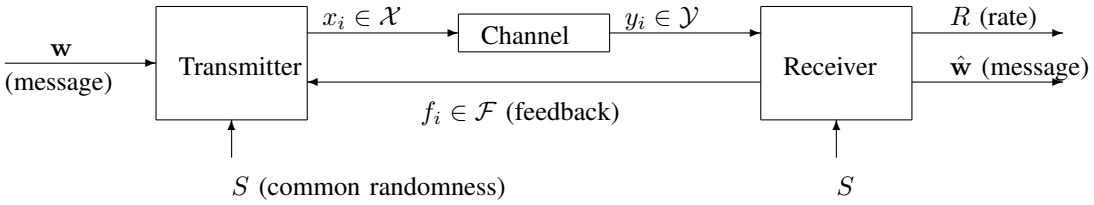

Fig. 1. A rate adaptive system with feedback

blocks it is reasonable to assume that past channels can be estimated. The assumption that $I(\hat{Q}_i, W_i)$ is achievable was made, although this communication rate is unknown to the transmitter in advance, i.e., the problem of rate adaptation is ignored. Therefore the synthetic problem is a subset of the original problem and upper bounds shown here apply also to the original problem.

B. Classification as a universal prediction problem

Let us begin by discussing the achievability of C_2 for the synthetic problem. The target rate C_2 is special in being an additive function for each value of Q. Universally attaining C_2 under the conditions specified above, falls into a widely studied category of universal prediction problems [12], [13], [14], [15]. Below, this class of problems and some relevant known results are reviewed.

These prediction problems have the following form: let $b \in \mathcal{B}$ be a strategy in a set of possible strategies $\mathcal{B}$, and $x \in \mathcal{X}$ be a state of nature. A loss function $l(b, x)$ associates a loss with each combination of a strategy and a state of nature. The total loss over n occurrences is defined as $L = \sum_{i=1}^{n} l(b_i, x_i)$. The universal predictor $\hat{b}_i(\mathbf{x}_1^{i-1})$ assigns the next strategy given the past values of the sequence, and before seeing the current value. There is a set of reference strategies $\{b_i^{(k)}\}_{k=1}^{N}$ (sometimes called experts), which are visible to the universal predictor. The target of universal prediction is to provide a predictor $\hat{b}_i$ which is asymptotically and universally better than any of the reference strategies, in the sense defined below.

For a given sequence $\mathbf{x}_1^n$, denote the losses of the universal predictor and the reference strategies as $\hat{L} \triangleq \sum_{i=1}^{n} l(\hat{b}_i, x_i)$ and $L_k \triangleq \sum_{i=1}^{n} l(b_i^{(k)}, x_i)$, respectively. Denote the regret of the universal predictor with respect a specific reference strategy as the excessive loss:

$$\mathcal{R}(k) \triangleq \hat{L} - L_k. \tag{8}$$

$\mathcal{R}_k$ is a function of the sequence $\mathbf{x}_1^n$ and the predictor. The target of the universal predictor is to minimize the worst case regret, i.e. attain

$$\mathcal{R}_{\text{minimax}} \triangleq \min_{\{\hat{b}_i(\cdot)\}} \max_{k} \max_{\mathbf{x}_1^n} \mathcal{R}(k). \tag{9}$$

The reference strategies may be defined in several different ways. In the simplest form of the problem the competition is against the set of fixed strategies $b_i^{(k)} = b(k)$. The exact minimax solution is known only for very specific loss functions [13, §8], and a solution guaranteeing $\max_{\mathbf{x}_1^n, k} \mathcal{R}(k) \underset{n\to\infty}{\longrightarrow} 0$ is not known for general loss functions. However there are many prediction schemes which perform well for a wide range of loss functions (see references above).

In the information theoretic framework, the log-loss $l(b, x) = \log\left(\frac{1}{b(x)}\right)$, where $b(x)$ is a probability distribution over $\mathcal{X}$ is the most familiar loss function, and used in analyzing universal source encoding schemes [12], since $l(b, x)$ represents the optimal encoding length of the symbol x when assigned a probability $b(x)$. It exhibits an asymptotical minimax regret of $\frac{1}{n}\mathcal{R}_{\text{minimax}} = O\left(\frac{\log n}{n}\right)$. However in the more general setting the asymptotical minimax regret decreases in a slower rate of $\frac{1}{n}\mathcal{R}_{\text{minimax}} = O\left(\frac{1}{\sqrt{n}}\right)$. There are several loss functions which are characterized by a "smoother" behavior for which better minimax regret is obtained [13, Theorem 3.1, Proposition 3.1]. For some of these loss functions, a simple forecasting algorithm termed "Follow the leader" (FL) can be used [13, §3.2] [16, Theorem 1]. In FL, the universal forecaster picks at every iteration i the strategy that performed best in the past, i.e. minimizes the cumulative loss over the instances from 1 to $i-1$.

The archetype of loss functions for which it is not possible to obtain a better convergence rate than $O\left(\frac{1}{\sqrt{n}}\right)$ is the absolute loss $l(b, x) = |b - x|$, where $x \in \mathcal{X} = \{0, 1\}$ and $b \in \mathcal{B} = [0, 1]$. The proof for the lower bound on the minimax regret [13, Theorem 3.7] is based on generating the sequence $\mathbf{x}_1^n$ randomly, and calculating the minimum *expected* regret (over $\mathbf{x}$). This value is a lower bound for the minimum-maximum regret (9). To show that the regret is $\omega(\sqrt{n})$ it is enough to consider only two competitors – one forecasting a constant zero, and one a constant one, and observe that since the cumulative losses of the two competitors always sum up to n, the minimum loss of the two competitors is a random variable with a standard deviation of $O(\sqrt{n})$ which is upper bounded by $\frac{n}{2}$, and therefore its expected value is $\frac{n}{2} - O(\sqrt{n})$, whereas the expected loss of the best single strategy over the random sequence cannot be better than $\frac{n}{2}$. A similar idea is used in the current paper, to prove lower bounds on the regret in the current problems. For general loss functions, and specifically for the absolute loss, the simple FL strategy does not converge.

The problem of asymptotically attaining $C_2(W_1^n)$ is analogous to the standard prediction problem, where the prior Q_i represents a strategy, and the channel W_i represents a state of nature. The current problem is given in terms of gains rather than losses, i.e. the loss is $l(Q, W) = -I(Q, W)$. The regret

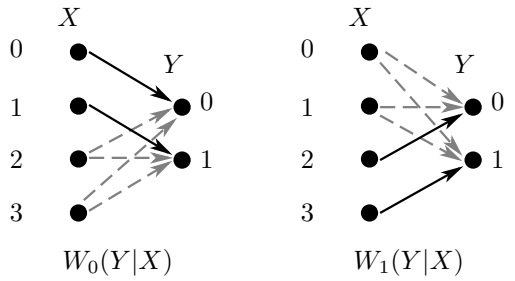

Fig. 2. Example channels W_0, W_1

is therefore:

$$\mathcal{R}_n(Q) = \sum_{i=1}^{n} I(Q, W_i) - \sum_{i=1}^{n} I(\hat{Q}_i, W_i). \tag{10}$$

Note that the regret is defined in terms of bits rather than rates (i.e. it is not normalized), from technical reasons.

C. A lower bound on the regret

A natural question to ask is, then: what is the asymptotical form of the minimax regret expected in the current problem? As will be shown, the prior prediction problem posed above, includes as a special case the prediction problem with the absolute loss function. Therefore, the asymptotical behavior cannot be better than $O(\sqrt{n})$, and it is not possible to apply the simple FL strategy.

The following example shows why the problem of attaining C_2 includes as a particular case the absolute loss function:

Example 1. Consider the quaternary to binary channel ($|\mathcal{X}| = 4, |\mathcal{Y}| = 2$), which may be in one of two states $s \in \{0, 1\}$, which define two conditional probability functions (shown as $|\mathcal{Y}| \times |\mathcal{X}|$ matrices below):

$$\begin{aligned} W_0(Y|X) &= \begin{bmatrix} 1 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \\ W_1(Y|X) &= \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 1 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 1 \end{bmatrix}. \end{aligned} \tag{11}$$

By writing the input as two binary digits $X = [X_1, X_2]$, the channel can be defined as follows: if $X_2 = s$ then $Y = X_1$, otherwise, $Y = \mathrm{Ber}\left(\frac{1}{2}\right)$. These channels are depicted in Figure 2, where transitions are denoted by solid lines for probability 1, and dashed lines for probability $\frac{1}{2}$. Now consider the same prediction problem, under the simplifying assumption that the channel $W_i = W_{s_i}$ is chosen only between the two channels above, and the forecaster knows this limitation, i.e. only the sequence of states $s_i \in \{0, 1\}$ is unknown.

It is clear from convexity of the mutual information, and the symmetry with respect to X_1 (interchanging the values of X_1 leads to the same mutual information), that any solution can only be improved by taking a uniform distribution over X_1. Therefore, without loss of generality, the input distribution Q can be defined by a single value $q = \Pr(X_2 = 1) \in [0, 1]$, and be written $Q = [\frac{1}{2}(1-q), \frac{1}{2}(1-q), \frac{1}{2}q, \frac{1}{2}q]$. For this choice the output will always be uniformly distributed $\mathrm{Ber}\left(\frac{1}{2}\right)$. Now,

$$\begin{aligned} I(Q, W_0) &= H(Y) - H(Y|X) \\ &= 1 - \sum_x Q(x) H(Y|X = x) = 1 - q, \end{aligned} \tag{12}$$

and similarly $I(Q, W_1) = q$, therefore:

$$I(Q, W_s) = 1 - |s - q|. \tag{13}$$

Hence, even under this limited scenario, the loss function $1 - I(Q, W)$ behaves like the absolute loss function, and therefore the normalized minimax regret (and the redundancy in attaining C_2) is at least $O\left(\sqrt{\frac{1}{n}}\right)$.

Note that the relation to the absolute loss implies that the simple FL predictor $\hat{Q}_i = \operatorname*{argmax}_Q \sum_{t=1}^{i-1} I(Q, W_t)$, cannot be applied to the current problem. An example to illustrate this and some further details are given in Appendix L.

Since the rest of the paper focuses on the rate function C_3, it is interesting to note that, although this rate is smaller, in general, than C_2, the minimum redundancy in obtaining it cannot be better than $O\left(\sqrt{\frac{1}{n}}\right)$. To show this, it only need to be shown that in the context of the counter-example shown above, $C_2 = C_3$. For a specific sequence of channels, denote by p the relative frequency with which channel W_1 appears. The averaged channel is $(1-p)W_0 + pW_1$. It is easy to see that the capacity of this channel is obtained by placing the entire input probability on the two useful inputs of the channel that appears most of the time. That is, if $p \geq \frac{1}{2}$ the input probability is placed on the useful inputs of W_1 and the rate $p \cdot C(W_1) = p$ is obtained, and otherwise $(1-p) \cdot C(W_0) = 1 - p$ is obtained. Hence the capacity of the averaged channel is $C_3 = \max(p, 1-p)$. On the other hand,

$$\begin{aligned} C_2 &= \max_Q \left((1-p) \cdot I(Q, W_0) + p \cdot I(Q, W_1)\right) \\ &= \max_{q \in [0,1]} \left((1-p) \cdot (1-q) + pq\right) = \max(p, 1-p). \end{aligned} \tag{14}$$

The example above also shows that C_1 is not universally achievable. In the example, the capacities of the two channels are $C(W_s) = 1$. Suppose the sequence of channel states $\mathbf{s}_1^n \in \{0, 1\}^n$ is generated randomly i.i.d. $\mathrm{Ber}\left(\frac{1}{2}\right)$. Then for any sequential predictor of q, the expected loss in each time instance is $\mathbb{E}[I(Q, W_s)] = \frac{1}{2}(1-q) + \frac{1}{2}q = \frac{1}{2}$, while the target rate is $C_1 = 1$. Therefore the expected normalized regret with respect to C_1 is $\frac{1}{2}$, and the maximum regret (maximum over the sequence $\{W_i\}$) is lower bounded by the expected regret.

To summarize, C_1 is not universally achievable, and therefore C_2 constitutes a reasonable target. Furthermore, the minimax regret with respect to C_2 is at least $O\left(\sqrt{\frac{1}{n}}\right)$, and the simple FL predictor following the best a-posteriori strategy does yield a vanishing regret.

D. A prediction algorithm

The prediction algorithm proposed below is based on a well known technique of a weighted average predictor, using exponential weighting [13, §2.1]. A minor difference with

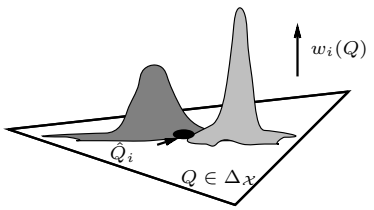

Fig. 3. An illustration of exponential weighting. The triangle represents the unit simplex. The two peaks represent two priors Q which have a relatively large gain $\sum_{t=1}^{i-1} I(Q, W_i)$. The weight function $w_i(Q)$ combines them exponentialy, and the predictor $\hat{Q}_i$ (represented as a black spot) is the weighted average.

respect to known results is the extension to a continuous set of reference strategies.

A weight function $w(Q)$ is any non-negative function $w : \Delta_{\mathcal{X}} \to \mathbb{R}^+$ with $\int_{\Delta_{\mathcal{X}}} w(Q)dQ = 1$. All integrals in the sequel are by default over $\Delta_{\mathcal{X}}$.

Define the following weight function:

$$w_i(Q) = \frac{e^{\eta \sum_{t=1}^{i-1} I(Q,W_t)}}{\int_{\Delta_{\mathcal{X}}} e^{\eta \sum_{t=1}^{i-1} I(\tilde{Q},W_t)} d\tilde{Q}}, \tag{15}$$

and the predictor:

$$\hat{Q}_i = \int_{\Delta_{\mathcal{X}}} Q \cdot w_i(Q) \cdot dQ. \tag{16}$$

The weighting function gives a higher weight to priors that succeeded in the past and the predictor averages the potential priors with respect to the weight. This is illustrated in Fig. 3. The following theorem gives a bound on the regret of this predictor, which is proven in the next section.

Theorem 1. *Let $I(Q, W), Q \in \Delta_{\mathcal{X}}$ be bounded function $0 \leq I(Q, W) \leq I_{\max}$ which is concave in its first argument. Then for n large enough so that $\frac{\ln(n)}{n} \leq e^{-2}$, the predictor defined by* (15) *and* (16) *with $\eta = \sqrt{\frac{|\mathcal{X}| \ln n}{n}} \cdot I_{\max}^{-1}$ yields*

$$R = \frac{1}{n} \sum_{i=1}^{n} I(\hat{Q}_i, W_i) \geq C_2(W_1^n) - \delta, \tag{17}$$

with

$$\delta = 2I_{\max} \cdot \sqrt{\frac{(|X| - 1) \ln n}{n}}. \tag{18}$$

Note that the theorem applies to gain functions more general than the mutual information, since it uses only the properties of concavity and boundness. In the case of mutual information $I_{\max}$ equals

$$I_{\max} = \log \min(|\mathcal{X}|, |\mathcal{Y}|). \tag{19}$$

The convergence rate is $O\left(\sqrt{\frac{\ln n}{n}}\right)$ and is slightly worse than the asymptotic bound of $O\left(\sqrt{\frac{1}{n}}\right)$ from Section III-C. The additional $\sqrt{\ln n}$ may be attributed to the fact the space of reference predictors is continuous (it results from Lemma 2 stated below), but we do not know if this is the best convergence rate.

E. Proof of Theorem 1

In this section the performance of the predictor (16) is analyzed Theorem 1 is proven. Define the instantaneous regret $r_i(Q)$ and the cumulative regret $\mathcal{R}_i(Q)$ as functions of Q:

$$r_i(Q) = I(Q, W_i) - I(\hat{Q}_i, W_i), \tag{20}$$

$$\mathcal{R}_i(Q) = \sum_{t=1}^{i} r_t(Q) = \sum_{t=1}^{i} I(Q, W_t) - \sum_{t=1}^{i} I(\hat{Q}_i, W_t). \tag{21}$$

These functions express the regret with respect to a fixed competing prior Q. The claim of the theorem is equivalent to the claim that for all Q, $\mathcal{R}_n(Q) \leq n\delta$. The dependence on Q is sometimes omitted for brevity.

For $\eta > 0$ of choice, define the following potential function:

$$\Phi(u) = \int_{\Delta_{\mathcal{X}}} e^{\eta u(Q)} dQ, \tag{22}$$

where $u : \Delta_{\mathcal{X}} \to \mathbb{R}$ is an arbitrary function defined over the unit simplex. Note that for large values of $\eta \cdot u$, $\Phi(u)$ approximates $\max_Q(u)$. As customary in this prediction technique, the proof consists of two parts:

1) Bounding the growth rate of $\Phi(\mathcal{R}_i(Q))$ over $i = 1, 2, \ldots, n$ for any Q.
2) Relating $\max_Q\{\mathcal{R}_n(Q)\}$ to $\Phi(\mathcal{R}_n(Q))$.

The techniques used below are based on Cesa-Bianchi and Lugosi's [13] (see Theorem 2.1, Corollary 2.2, Theorem 3.3).

Since $I(Q, W)$ is concave with respect to Q, then for any weight function $w(Q)$ and any W_i:

$$\begin{aligned} \int w(Q) r_i(Q) dQ &= \int w(Q) I(Q, W_i) dQ - I(\hat{Q}_i, W_i) \\ &\leq I\left(\underbrace{\int w(Q) Q dQ}_{\hat{Q}_i}, W_i \right) - I(\hat{Q}_i, W_i) \\ &= 0. \end{aligned} \tag{23}$$

Following [13] this inequality may be termed the "Blackwell condition". The meaning of this condition is that by choice of $w(Q)$ one can prevent an increase of $\mathcal{R}_i(Q)$ in a chosen direction ($w(Q)$ can be thought of as a unit vector in the Hilbert space of functions over $\Delta_{\mathcal{X}}$). For the specific choice of the weight function (15), this direction is proportional to the gradient of $\Phi(R)$ with respect to R, thus preventing any growth in this direction and leaving only second order terms that contribute to the increase of $\Phi(\mathcal{R}_n(Q))$. Since the factor $\sum_{t=1}^{i} I(\hat{Q}_i, W_t)$ in (21) does not depend on Q, the weight function (15) can be alternatively written as:

$$w_i(Q) = \frac{e^{\eta \mathcal{R}_{i-1}(Q)}}{\int e^{\eta \mathcal{R}_{i-1}(Q)} dQ}. \tag{24}$$

$w_i(Q)$ is indifferent to any constant addition to $\mathcal{R}_{i-1}(Q)$ due to the normalization. The growth of the potential can be

bounded as follows:

$$\begin{aligned}\Phi(\mathcal{R}_i) &= \Phi(\mathcal{R}_{i-1}+r_i) = \int e^{\eta\mathcal{R}_{i-1}+\eta r_i}dQ \\ &= \int e^{\eta\mathcal{R}_{i-1}}\cdot e^{\eta r_i}dQ \\ &\overset{(24)}{=} \int e^{\eta\mathcal{R}_{i-1}}dQ\cdot\int w_i(Q)e^{\eta r_i}dQ \\ &= \Phi(\mathcal{R}_{i-1})\cdot\int w_i(Q)e^{\eta r_i}dQ,\end{aligned} \tag{25}$$

Notice that $r_i \le I_{\max}$. Take η small enough that $\eta r_i \le \eta I_{\max} \le 1$ and use the following inequality (proven in Appendix E):

Lemma 1. *For $x\in[-1,1]$:*

$$1+x\le e^x\le 1+x+x^2. \tag{26}$$

Returning to (25):

$$\begin{aligned}&\int w_i(Q)e^{\eta r_i}dQ \overset{(26)}{\le} \int w_i(Q)\left(1+\eta r_i+(\eta r_i)^2\right)dQ \\ &= \int w(Q)dQ+\eta\underbrace{\int w(Q)r_i dQ}_{\le 0,(23)}+\eta^2\int w(Q)r_i^2dQ \\ &\overset{(23)}{\le} 1+\eta^2 I_{\max}^2 \overset{(26)}{\le} e^{\eta^2 I_{\max}^2}.\end{aligned} \tag{27}$$

Therefore recursively applying (25):

$$\Phi(\mathcal{R}_n)\overset{(25),(27)}{\le} e^{\eta^2 I_{\max}^2}\Phi(\mathcal{R}_{n-1})\le\ldots\le e^{n\eta^2 I_{\max}^2}\cdot\Phi(0). \tag{28}$$

Notice that $\Phi(0)=\int 1dQ=\mathrm{vol}(\Delta_{\mathcal{X}})$. This completes the first part of showing that the increase in $\Phi(\mathcal{R}_n)$ is bounded. For the second part, the exponential weighting of a function is related to its maximum, using the following lemma, which proven in Appendix A:

Lemma 2. *Let $F(\mathbf{x})$ be a real non-negative bounded function $F: S\to[a,b]$ concave in S, where S is a closed convex vector region of dimension d, and let η satisfy $\eta(b-a)\ge d$, then*

$$\begin{aligned}\max_{\mathbf{x}\in S}F(\mathbf{x}) &\le \frac{1}{\eta}\ln\left[\frac{\int_S e^{\eta F(\mathbf{x})}d\mathbf{x}}{\mathrm{vol}(S)}\right]+\frac{d}{\eta}\ln\left(\frac{\eta e(b-a)}{d}\right) \\ &= \frac{1}{\eta}\ln\left[\frac{\Phi(F)}{\Phi(0)}\right]+\frac{d}{\eta}\ln\left(\frac{\eta e(b-a)}{d}\right).\end{aligned} \tag{29}$$

Let $F(Q)=\mathcal{R}_n(Q)$. In this case the convex region is $\Delta_{\mathcal{X}}$ and therefore $d=\dim(\Delta_{\mathcal{X}})=|X|-1$. By (21) F can be bounded by:

$$\underbrace{-\sum_{i=1}^n I(\hat{Q}_i,W_i)}_{\triangleq a}\le F(Q)\le\underbrace{nI_{\max}-\sum_{i=1}^n I(\hat{Q}_i,W_i)}_{\triangleq b}, \tag{30}$$

where the factor $\sum_{i=1}^n I(\hat{Q}_i,W_i)$ is constant in Q, and $b-a=nI_{\max}$. Assuming $\eta n I_{\max}\ge d$ to satisfy the conditions of Lemma 2, by (29):

$$\begin{aligned}\mathcal{R}_n(Q) &\le \frac{1}{\eta}\ln\frac{\Phi(\mathcal{R}_n(Q))}{\Phi(0)}+\frac{d}{\eta}\ln\left(\frac{\eta enI_{\max}}{d}\right) \\ &\overset{(28)}{\le} n\eta I_{\max}^2+\frac{d}{\eta}\ln\left(\frac{\eta enI_{\max}}{d}\right). \\ &\le n\eta I_{\max}^2+\frac{d}{\eta}\ln(n)\triangleq\Delta,\end{aligned} \tag{31}$$

where in the last inequality it was assumed that $\frac{\eta eI_{\max}}{d}\le 1$ (this would hold for η small enough). The following lemma is used to optimize the RHS of (31) with respect to η:

Lemma 3. *The unique minimum over $t\in\mathbb{R}^+$ of $f(t)=a\cdot t^\alpha+b\cdot t^{-\beta}$ ($a,b,\alpha,\beta>0$) is obtained at $t^*=\left(\frac{b\beta}{a\alpha}\right)^{\frac{1}{\alpha+\beta}}$ and equals*

$$f(t^*)=\left(\frac{\beta}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta}}\left[1+\frac{\alpha}{\beta}\right]\cdot a^{\frac{\beta}{\alpha+\beta}}\cdot b^{\frac{\alpha}{\alpha+\beta}}. \tag{32}$$

Particularly, for $\alpha=\beta=1$, i.e. $f(t)=a\cdot t+\frac{b}{t}$ the above results in $t^*=\sqrt{\frac{b}{a}}$ and $f(t^*)=2\sqrt{ab}$.

The proof of the lemma is simple by a direct derivation (see Appendix E). Applying the lemma to the optimization of η in (31) yields:

$$\eta^*=\sqrt{\frac{d\ln(n)}{nI_{\max}^2}}, \tag{33}$$

and

$$\Delta^*=\Delta\Big|_{\eta=\eta^*}=2I_{\max}\sqrt{dn\ln(n)}. \tag{34}$$

Let us now verify the assumptions that have been made along the way. In (27) it was assumed that $\eta I_{\max}\le 1$. If the contrary holds $\eta I_{\max}>1$ then considering the first term in the RHS of (31), yields $\Delta>nI_{\max}$, and therefore the theorem holds in a void way. To apply Lemma 2 it was require that $\eta n I_{\max}\ge d$. If the opposite is true, i.e. $\eta n I_{\max}<d$ then the second term the RHS of (31) becomes $\frac{d}{\eta}\ln(n)>nI_{\max}\ln(n)$, and so for $n\ge e$, $\Delta>nI_{\max}$ and the theorem will hold in a void way. Thus for the two last conditions, it is enough that $n\ge 3$, since in this case if either of the conditions does not hold, the theorem becomes true automatically (in a void way). Lastly, in (31) it was assumed that $\frac{\eta eI_{\max}}{d}\le 1$. Substituting η^* yields $\frac{\eta eI_{\max}}{d}=e\cdot\sqrt{\frac{e^2\ln(n)}{dn}}\le e\cdot\sqrt{\frac{\ln(n)}{n}}$, which becomes smaller than 1 for n large enough. The last condition supersedes $n\ge e$, and is specified as a requirement in the theorem. □

IV. Arbitrary channel variation

In this section, the main results of this paper are presented, with respect to the problem defined in Section II-B: the achievability of the capacity of the averaged channel, and a converse showing that this is the best rate, under some conditions. The communication system attaining this rate is described, while leaving out some of the technical details, such as decoding and channel estimation (these will be completed in the next section). It is shown that under abstract assumptions, the system achieves the desired rate.

A. Target rate

The synthetic problem differs from the problem defined in Section II-B, in two main aspects:

1) It assumes that the sequence of past channels is fully known. Since the receiver observes only one output sample from each channel, this assumption is not realistic. On the other hand, the time-averaged channel over "large" chunks of symbols can be measured.
2) It assumes that a rate corresponding to a sum of the per-symbol mutual information can be attained, whereas with an arbitrarily varying channel, the amount of mutual information between the input and output vectors is potentially lower.

Therefore, as shall be seen, C_2 is no longer achievable in the context of the arbitrarily varying channel defined in Section II-B. In Appendix H it is shown that, even imposing on the synthetic problem only the limitation that the past channels are not given, but need to be estimated, leads to the conclusion that C_2 is not attainable. The compromise is the alternative target of obtaining $C_3 = C(\overline{W})$, i.e. the capacity of the averaged channel. This rate is optimal in a sense described below, and is indeed asymptotically achievable.

The rate $C(\overline{W})$ is certainly not the maximum achievable target rate. As an example, if $C(\overline{W})$ is achievable for large n then by operating the same scheme on two halves of the transmission time one could attain $R = \frac{1}{2}C\left(\overline{W_1^{n/2}}\right) + \frac{1}{2}C\left(\overline{W_{n/2+1}^{n}}\right)$, where $\overline{W_1^{n/2}}, \overline{W_{n/2+1}^{n}}$ denote the averaged channels on the two halves. This rate is in general higher, because due to the convexity of the mutual information with respect to the channel $C(\overline{W}) = \max_Q I(Q,\overline{W}) \leq \max_Q \left[\frac{1}{2}I\left(Q,\overline{W_1^{n/2}}\right) + \frac{1}{2}I\left(Q,\overline{W_{n/2+1}^{n}}\right)\right] \leq R$.

On the other hand, $C(\overline{W})$ is the maximum achievable rate which is independent of the order of the sequence $\{W_i\}$, or, in other words, which is fixed under permutation of the sequence. This observation is formalized in the following theorem:

Theorem 2. *Let $R(W_1^n)$ (for $n = 1, 2, ..$) be a sequence of rate functions, which are oblivious to the order of W_1^n. If the sequence is asymptotically attainable according to Definition 1, then there exists a sequence $\delta_n \underset{n\to\infty}{\longrightarrow} 0$ such that $R(W_1^n) \leq C(\overline{W}) + \delta_n$.*

Note that $C(\overline{W})$ depends on n through the average over n channels $\{W_i\}_1^n$. Since both C_1 and C_2 are oblivious to the order of W_1^n, Theorem 2 implies they are not achievable.

Following is a rough outline of the proof. Consider the channel generated by uniformly drawing a random permutation π of the indices $i = 1, \ldots, n$, using the channels W_i in a permuted order. If a system guarantees a rate $R(W_1^n)$, which is fixed under permutation, then this rate would be fixed for all drawing of π, and therefore for the channel described, the system can guarantee the rate $R(W_1^n)$ a-priori. Hence, the capacity of this channel must be at least $R(W_1^n)$. The next stage is to show that the feedback capacity of this channel is at most $C(\overline{W})$. Due to the fact the channels are selected from the set $\{W_i\}_{i=1}^n$ without replacement, the proof is a little technical and will be deferred to Appendix F. However to give an intuitive argument, replace the channel described above, by a similar channel, obtained by randomly drawing at each time instance one of $\{W_i\}_{i=1}^n$, this time *with* replacement. This new channel is simply the DMC with channel law $\overline{W}$. Therefore feedback does not increase the capacity and its feedback capacity is simply $C(\overline{W})$. The main point in the proof is to show there is no difference in feedback-capacity between the two channels, and the main tool is Hoeffding's bounds on sampling without replacement [17].

Another interesting property of the rate $C(\overline{W})$ is that it meets or exceeds the random-code capacity of any memoryless AVC defined over the same alphabet, and thus attaining $C(\overline{W})$ yields universality over all AVC's (see Section VI-A). Through the relation to AVC capacity one can see that common randomness is essential to obtain $C(\overline{W})$, as it is essential for obtaining the random-code capacity [1].

After settling for $C(\overline{W})$, the next question that naturally arises is: what is the best convergence rate of the regret, with respect to this target? In Section III-C it was shown, that even in the context of the synthetic problem of Section III (with full knowledge of past channels), the regret with respect to C_3 is at least $O(n^{-\frac{1}{2}})$, and this lower bound naturally holds in the current problem, where only partial knowledge of past channels is available.

The following theorem formalizes claim that $C(\overline{W})$ is achievable according to Definition 1:

Theorem 3. *For every $\epsilon, \delta > 0$ there exists N and a constant c_Δ, such that for any $n \geq N$ there is an adaptive rate system with feedback and common randomness, where for the problem of Section II-B, over any sequence of channels $\{W_i(y|x)\}_{i=1}^n$:*

1) *The probability of error is at most ϵ*
2) *The rate satisfies $R \geq C(\overline{W}) - \Delta_C$ with probability at least $1 - \delta$*
3) $\Delta_C = c_\Delta \cdot \left(\frac{\ln^2(n)}{n}\right)^{\frac{1}{4}}$

where the probabilities are with respect to the channel and the common randomness, and hold for any transmitted message.

Corollary 1. *Specific values for $\epsilon, \delta, \Delta_C$ can be obtained as follows. Let $d_\epsilon, \delta_0, c_\lambda > 0$ be parameters of choice. Then the constants $n_{\min}$ and c_Δ are given in the proof, by* (114)*,* (117)*, where constants used in these equations are defined in* (19)*,* (42)*,* (54)*,* (105)-(107)*,* (109)*. For any $n \geq n_{\min}$, $\epsilon = n^{-d_\epsilon}$ and $\delta = \epsilon + \delta_0$.*

Corollary 2. *The same holds if W_i is determined (e.g. by an adversary) as a function of the message and all previous channel inputs and outputs $\mathbf{X}^{i-1}, \mathbf{Y}^{i-1}$.*

The proof of the theorem is given in Section V. A numerical example is given after the proof (Example 2). To easily see how the asymptotical promise of Theorem 3 can be achieved (without the specific convergence rate), consider the following crude scheme, which combines the results of Eswaran [3] or our previous paper [4], i.e. the fact that the empirical mutual information is achievable, with the prior prediction scheme of Theorem 1. The transmission time n may be divided

into multiple fixed-size blocks $i = 1, \ldots, N$, and in each block, one of these schemes is operated, with an i.i.d. prior chosen by a predictor. Using Eswaran's result, for example, and ignoring some details such as finite-state assumptions, one would obtain the rate $I(\hat{Q}_i, \overline{W}_i)$ over each block, where $\overline{W}_i$ is the averaged channel over the block. The channel $\overline{W}_i$ can be well estimated (e.g. using training symbols or using the communication scheme itself). Assuming it is known, if the prediction scheme of Theorem 1 is operated over $\overline{W}_i$ it will guarantee the average rate over the N blocks will be asymptotically at least $\frac{1}{N}\sum_{i=1}^{N} I(Q, \overline{W}_i)$ for any Q, and using convexity, $\frac{1}{N}\sum_{i=1}^{N} I(Q, \overline{W}_i) \geq I\left(Q, \frac{1}{N}\sum_{i=1}^{N} \overline{W}_i\right) = I(Q, \overline{W})$. Since this holds for any Q this achieves the capacity of the average channel.

The scheme used for the proof of Theorem 3 combines the rate-less scheme with the prior prediction in a cleaner way. The communication and prediction scheme are described in the remainder of this section.

B. The communication scheme

In this section give the communication scheme, up to some details which will be completed later on (Section V-B). One of the issues ignored in the synthetic problem is the determination of the rate R before knowing the channel. To solve this problem, rateless codes [18] are applied. The available time is divided into multiple blocks as done by Eswaran *et al* [3] and in [4].

Fix a number K of bits per block. In each block, K bits from the message string are sent. At each block $i = 1, 2, \ldots$, a codebook of $\exp(K)$ codewords is generated randomly and i.i.d. (in time and message index) according to the prior $\hat{Q}_i(x)$. $\hat{Q}_i(x)$ is determined by a prediction scheme which is specified below. The random drawing of the codewords is carried out by using the common randomness, and the codebook is known to both sides. The relevant codeword matching the message substring is sent to the receiver symbol by symbol. At each symbol of the block and for each codeword $\mathbf{x}_l, l = 1, \ldots, \exp(K)$ in the codebook, the receiver evaluates a decoding condition (59) that will be specified later on. Roughly speaking, the condition measures whether there is enough information from the channel output to reliably decode the message.

The receiver decides to terminate the block if the condition (59) holds, and informs the transmitter. When this happens, the receiver determines the decoded codeword as one of the codewords that satisfied (59). Then, using the known channel output $\mathbf{y}$, and the decoded input $\mathbf{x}$ over the block which was decoded, the receiver computes an estimate of the averaged channel over the block. The specific estimation scheme will be specified in Section V-B.

The receiver calculates a new prior for the next block according to the prediction scheme that will be specified below. The receiver sends the new prior to the transmitter. Alternatively, the receiver may send the estimated channel, and the new prior can be calculated at each side separately. The new block $i + 1$ starts at the next symbol, and the process continues, until symbol n is reached. The last block may terminate before decoding.

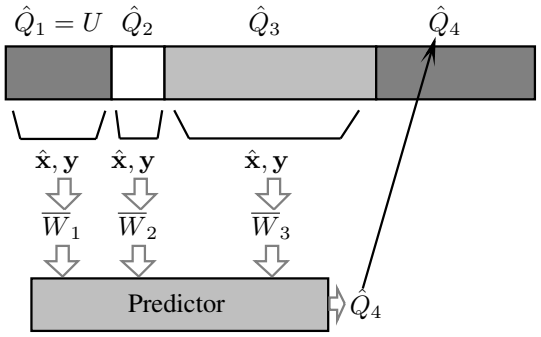

Fig. 4. An illustration of the combination of a rateless scheme with prior prediction. Each box represents a rateless block in which K bits are transmitted.

C. The prediction algorithm

In this section the prediction algorithm is presented. Denote by i the index of the block, and by $\overline{W}_i$ the averaged channel over the block, i.e. if the block i starts at symbol k_i and ends at $k_{i+1} - 1$, then $\overline{W}_i(y|x) \triangleq \frac{1}{k_{i+1}-k_i}\sum_{t=k_i}^{k_{i+1}-1} W_t(y|x)$. The length of the i-th block is denoted $m_i = k_{i+1} - k_i$. An exponentially weighted predictor mixed with a uniform prior is used. The motivation for using the uniform prior is explained in the next section. Let $U = \frac{1}{|\mathcal{X}|}\mathbf{1}$ be the uniform prior over $\mathcal{X}$. Define the predictor as:

$$\hat{Q}_i = (1 - \lambda)\int_{\Delta_{\mathcal{X}}} w_i(Q) Q dQ + \lambda U. \tag{35}$$

where

$$w_i(Q) = \frac{1}{\Phi\left(\sum_{j=1}^{i-1} m_j \cdot F_j(\tilde{Q})\right)} \cdot e^{\eta \sum_{j=1}^{i-1} m_j \cdot F_j(Q)}, \tag{36}$$

where $F_i(Q)$ is an estimate of the mutual information of the averaged channel over block i, $I(Q, \overline{W}_j)$, and is interpreted as an estimate of the number of bits that would have been sent with the alternative prior Q. This estimate is defined later on in Section V-E. The parameters λ, η and K will be chosen later on. Φ is the potential function defined in (22). The term $\frac{1}{\Phi(\ldots)}$ normalizes $w_i(Q)$ to $\int_{\Delta_{\mathcal{X}}} w_i(Q) dQ = 1$.

The following Lemma formalizes the claim that the predictor resulting of (35)-(36), asymptotically achieves a rate $R \geq \sum_{i=1}^{B+1} \frac{m_i}{n} F_i(Q)$:

Lemma 4. *Let $F_i(Q)$, $i = 1, \ldots, B+1$ be a set of $B+1$ non-negative concave functions of the prior $Q \in \Delta_{\mathcal{X}}$, let $\{m_i\}_{i=1}^{B+1}$ denote a set of non-negative numbers, and $K, n, I_{\max}$ be arbitrary positive constants satisfying $n > e$ and $K \geq 2I_{\max}$.*

Define the target rate

$$R_T = \max_{Q \in \Delta_{\mathcal{X}}} \sum_{i=1}^{B+1} \frac{m_i}{n} F_i(Q). \tag{37}$$

Define the actual rate R over n channel uses as:

$$R = \frac{KB}{n}. \tag{38}$$

Define the sequential predictor $\hat{Q}_i$ as the result of (35) *and* (36)*. Let $\{m_i\}_{i=1}^{B+1}$ satisfy:*

$$m_i F_i(\hat{Q}_i) \leq K. \tag{39}$$

Then for the value of η specified below (43) *it is guaranteed that:*

$$R \geq \min(R_T, I_{\max}) - \Delta_{\text{pred}}, \tag{40}$$

where

$$\Delta_{\text{pred}} = \frac{K}{n} + I_{\max} \cdot \lambda + c_1 \sqrt{\frac{\ln(n)}{n}} \lambda^{-\frac{1}{2}}, \tag{41}$$

and

$$c_1 = 2\sqrt{K \cdot |\mathcal{X}|(|\mathcal{X}| - 1) \cdot I_{\max}}. \tag{42}$$

The value of η attaining the result above is:

$$\eta = \sqrt{\frac{|\mathcal{X}| - 1}{K \cdot |\mathcal{X}| \cdot I_{\max}} \cdot \frac{\ln(n) \cdot \lambda}{n}}. \tag{43}$$

The lemma is proven in Appendix B. The proof uses similar techniques to those introduced in Section III-E, however, different from the previous analysis, due to mixing with the uniform prior, the "Blackwell" condition ((23) in the previous case) only approximately holds. On the other hand, the use of the uniform prior enables relating $F_i(\hat{Q}_i)$ to $F_i(Q)$ for any other Q, and thus obtain from (39) an upper bound on the gain $m_i F_i(Q)$ related to an alternative prior Q. The trade-off between the two is expressed in the two last factors in (41), one of which is increasing with λ and the other decreasing.

Since by (39), $R \geq \sum_{i=1}^{B+1} \frac{m_i}{n} F_i(\hat{Q}_i) - \frac{K}{n}$, the claim of the lemma appears similar to Theorem 1, with $m_i F_i(Q)$ taking the place of the function $I(Q, W_i)$. However two important properties of the lemma, distinguishing it from the rather standard claim of Theorem 1 are that the bound does not depend on the number of blocks (i.e. the number of prediction steps), and that no upper bound on $F_i(Q)$ is assumed.

The rate $I_{\max}$ represents a bound on mutual information, but in the context of the lemma it enough to consider it as an arbitrary rate that caps R_T. It affects the setting of η and the resulting loss. Also, n does not have to correspond to the actual number of symbols and serves here merely as a scaling parameter for the communication rate. The lemma sets a value of η but not for λ, since λ will have additional roles in the next section.

D. Motivation for the prediction algorithm

In this section a motivation for the prediction algorithm, and especially for the use of the uniform prior is given. Under abstract assumptions it is shown to achieve the capacity of the averaged channel. This section is intended merely to give motivation and is not formally necessary for the proof of Theorem 3.

To simplify the discussion, let us make abstract assumptions regarding the decoding condition and the channel estimation:

1) The decoding condition yields block lengths satisfying:

$$m_i \leq \frac{K}{I(\hat{Q}_i, \overline{W}_i)}, \tag{44}$$

with an equality for all blocks except the last one which is not decoded. This implies the rate $\frac{K}{m_i}$ equals the mutual information of the averaged channel.

2) The averaged channels over all previous blocks are known and available for the predictor

With these assumptions, the prediction problem can be considered separately from decoding and channel estimation issues. Supposing that B blocks were transmitted, the achieved rate is $R = \frac{KB}{n}$. Since $n \approx \sum_i m_i$, using (44) this can be written as $R \approx \left(\frac{1}{B} \sum_{i=1}^{B} \frac{1}{I(\hat{Q}_i, \overline{W}_i)}\right)^{-1}$. The target is to find a prediction scheme for $\hat{Q}_i$, such that for any sequence W_i, one will have $R \geq C(\overline{W}) - \delta_n$ with $\delta_n \to 0$. There are two main difficulties compared to the prediction problem discussed in Section III:

1) The problem is not directly posed as a prediction problem with an additive loss.
2) The loss is not bounded: if for some i, $I(\hat{Q}_i, \overline{W}_i) = 0$ then the rate becomes zero regardless of other blocks.

The first issue is resolved by posing an alternative problem which has an additive loss, and using the convexity of the mutual information with respect to the channel (as will be exemplified below in the abstract case). Regarding the second issue, notice that if the channel has zero capacity (always, or from some point in time onward), it is possible that one of the blocks will extend forever and will never be decoded. However one must avoid a situation where the channel has non-zero capacity (which the competition enjoys), while a badly chosen prior yields $I(\hat{Q}_i, \overline{W}_i) = 0$. This may happen for example in the channels of Example 1, if the predictor selects to use the pair of inputs that yield zero capacity. If this happens then the scheme will get stuck since the block will never be decoded, and hence there will be no chance to update the prior. In addition, notice that selecting some inputs with zero probability makes the predictor blind to the channel values over these inputs. To resolve these difficulties the predictor is constructed as a mixture between an exponentially weighted predictor and a uniform prior. A result by Shulman and Feder [19] bounds the loss from capacity by using the uniform prior U:

$$I(U; W) \overset{[19,(3)]}{\geq} C \cdot \beta(C) \overset{[19,(17)]}{\geq} \frac{C}{|\mathcal{X}| \cdot (1 - e^{-1})}, \tag{45}$$

where C is the channel capacity and $\beta(C)$ is defined therein. This guarantees that if the capacity is non-zero, then the uniform prior will yield a non-zero rate, and hence the block will not last indefinitely.

Under the abstract assumptions made here, the following F_i is known and can be substituted in Lemma 4:

$$F_i(Q) = I(Q, \overline{W}_i). \tag{46}$$

This yields the following result:

Lemma 5. *For the scheme of Section IV-B under the abstraction specified above, with $n \geq 3$ and $K \geq 2I_{\max}$ and properly chosen η, λ, the following holds: for any sequence of channels, the rate satisfies:*

$$R = \frac{K \cdot B}{n} \geq C(\overline{W}) - \Delta_{\text{pred}}, \tag{47}$$

where $C(\overline{W})$ is the capacity of the averaged channel and

$$\Delta_{\text{pred}} = 4 \cdot I_{\max}^{\frac{2}{3}} \cdot |\mathcal{X}|^{\frac{2}{3}} \cdot K^{\frac{1}{3}} \cdot \left(\frac{\ln(n)}{n}\right)^{\frac{1}{3}} \underset{n \to \infty}{\longrightarrow} 0, \tag{48}$$

where $I_{\max} = \log\min(|\mathcal{X}|, |\mathcal{Y}|)$. *The parameters of the scheme* η, λ *required to attain the result are specified in* (43) *and* (191) *respectively.*

Note that the bound (48) is increasing with K, so it appears that that it can be improved by taking the minimal value of K. However in the actual system, there are be fixed overheads related to the communication scheme, and a large block size would be needed to overcome them. Taking any fixed and large enough K, the normalized regret is bounded by $O\left(\frac{\ln n}{n}\right)^{\frac{1}{3}}$, which converges to zero, but at a worse rate than that of Theorem 1.

Note that the claims of Lemma 5 are stronger than the claims that appeared in the conference paper on the subject [10], for the same problem, mainly in terms of the improved convergence rate with n. Also, the scheme used here is slightly different than the one in the conference paper (in Equation (36)). The proof corresponding to the scheme presented in the conference paper can be found in an early version uploaded to arXiv [20].

To prove Lemma 5, Lemma 4 is used with F_i defined in (46). The rate guaranteed by Lemma 4 is approximately $R_T \geq \sum_{i=1}^{B+1} \frac{m_i}{n} I(Q, \overline{W}_i)$. Using convexity of the mutual information with respect to the channel this is at least $I\left(Q, \sum_{i=1}^{B+1} \frac{m_i}{n} \overline{W}_i\right) = I\left(Q, \overline{W}\right)$, and since this is true for any Q, the rate is at least $C\left(\overline{W}\right)$. The detailed proof appears in Appendix G.

Notice that in the alternative scheme described after Theorem 3, it appears that there is no need for the uniform prior, however this is somewhat hidden in the assumption that the channel is known. Furthermore in that scheme there is no need to worry about rateless blocks extending "forever" since the commnication scheme is re-started on each of the N blocks.

V. Proof of the main result

In this section Theorem 3, regarding the attainability of $C(\overline{W})$ is proven.

A. Preliminaries

Suppose that during a certain block of length m the scheme applied the i.i.d. prior $Q(x)$. In order to estimate the channel after the block has ended and $\mathbf{x}$ was decoded, the following estimate is used:

$$\breve{W}(y|x) = \frac{\hat{P}_{\mathbf{x},\mathbf{y}}(x,y)}{Q(x)}, \tag{49}$$

where here and throughout the current section, $\mathbf{x}, \mathbf{y}$ denote the m-length input and output vectors over the block, and $\hat{P}_{\mathbf{x},\mathbf{y}}(x,y)$ is the empirical distribution of the pair (x_i, y_i) (for $i = 1, \ldots, m$). The estimator is the joint empirical distribution divided by the (known) marginal distribution of the input X. Since a uniform prior is mixed into $Q(x)$ (35), all $Q(x)$ are bounded away from zero, which makes the estimator (49) statistically stable, in comparison with the more natural estimator given by the empirical conditional distribution:

$$\hat{W}(y|x) = \hat{P}_{\mathbf{y}|\mathbf{x}}(x,y) = \frac{\hat{P}_{\mathbf{x},\mathbf{y}}(x,y)}{\hat{P}_{\mathbf{x}}(x)}, \tag{50}$$

in which the denominator may turn out to be zero. A drawback of the proposed estimator (49) is, that it does not generally yield a legitimate probability distribution, i.e. $\sum_y \breve{W}(y|x) \neq 1$. The result of using this estimator is that the calculations below include values that formally appear like probabilities but are not. To distinguish them from legitimate probabilities these values are termed "false" probabilities, and are marked with a $\breve{\square}$. These functions usually approximate or estimate a legitimate probability. Formally, a false probability $\breve{p}(y)$ or $\breve{p}(y|x)$ can be any non-negative function of y or x, y (respectively). Note that until this point, the assumption that the output alphabet $\mathcal{Y}$ is finite was not needed, since the channel was given to the predictor rather than being estimated, and it is the first time this assumption is used.

The function that used as an optimization target for selecting the prior for the next block is, as before, the mutual information. The reason is that since the aim is to achieve the capacity of the averaged channels, the "competing" schemes, for each prior Q, achieve the mutual information of the averaged channel. Since the estimate of the channel is a false probability, the mutual information function is extended to receive a false-probability in its second argument, by simply plugging-in into the standard formula of $I(Q,W)$. This substitution results in what is defined as the *false mutual information* $\breve{I}(Q, \breve{W})$:

$$\breve{I}(Q, \breve{W}) \triangleq \sum_{x,y} Q(x) \breve{W}(y|x) \log\left(\frac{\breve{W}(y|x)}{\sum_{x'} Q(x') \breve{W}(y|x')}\right), \tag{51}$$

where cases of $Q(x) = 0$ or $\breve{W}(y|x)$ are resolved using the convention $0 \cdot \log 0 = 0$. The following lemma shows that most of the properties of the mutual information function $I(P,W)$ needed for the previous analysis in Section IV-C are maintained.

Lemma 6 (Properties of false mutual information)**.** *The function* $\breve{I}(Q, \breve{W})$ *defined in* (51) *is*

1) *Non negative*
2) *Concave with respect to* Q
3) *Convex with respect to* $\breve{W}$
4) *Upper bounded by* $\sigma \cdot \log|\mathcal{X}|$, *where* $\sigma = \max_x \left[\sum_y \breve{W}(y|x)\right]$.

The proof is technical and appears in Appendix C. In addition to the properties above, the proof relies on the next property which is more surprising. When the prior Q used for estimating the channel in (49) is the same prior Q used as input in (51), the false mutual information attains a form which is familiar from [21] as a prototype of the zero order rate function. As in [21], this form can be used to obtain a bound on the probability of $\breve{I}(Q, \breve{W})$ to exceed a threshold for a random drawing of $\mathbf{x}$. This bound, in turn, allows constructing the rate-adaptive system attaining a block length m_i that depends on $\breve{I}(Q, \breve{W})$.

Following [21], let us define conditional empirical *probability* of the discrete sequence $\mathbf{x}$ given the sequence $\mathbf{y}$ as $\hat{p}(\mathbf{x}|\mathbf{y}) \triangleq \prod_{i=1}^{m} \hat{P}_{\mathbf{x}|\mathbf{y}}(x_i|y_i)$, i.e. the probability of the sequence $\mathbf{x}$ under the conditionally i.i.d. distribution $P(y|x) = \hat{P}_{\mathbf{x}|\mathbf{y}}(y|x)$. Also, when vectors are substituted into Q, then

Q is implicitly extended in an i.i.d. fashion, i.e. $Q(\mathbf{x}) \triangleq \prod_{i=1}^{m} Q(x_i)$. The following lemma will be used to bound the error probability:

Lemma 7 (False mutual information as a decoding metric)**.** *The false MI with prior $Q(x)$ and $\breve{W}(y|x) = \frac{\hat{P}_{\mathbf{x},\mathbf{y}}(x,y)}{Q(x)}$ where $\mathbf{x},\mathbf{y}$ are m-length vectors can be written as:*

$$\breve{I}(Q,\breve{W}) = \breve{I}\left(Q(x), \frac{\hat{P}_{\mathbf{xy}}(x,y)}{Q(x)}\right) = \frac{1}{m}\log\frac{\hat{p}(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}. \quad (52)$$

Furthermore, for any Q and any $\mathbf{y}$, when $\mathbf{X}$ is distributed i.i.d. $\mathbf{X} \sim Q^n$,

$$\begin{aligned}\Pr\left(\breve{I}(Q,\breve{W}) \geq T|\mathbf{y}\right) &= \Pr\left(\frac{\hat{p}(\mathbf{X}|\mathbf{y})}{Q(\mathbf{X})} \geq \exp(mT)\Big|\mathbf{y}\right) \\ &\leq \exp(-(mT - k_0\log m - k_1)),\end{aligned} \quad (53)$$

where

$$k_0 = k_1 = (|\mathcal{X}|-1)\cdot|\mathcal{Y}|. \quad (54)$$

Note that from the results in [21, Theorem 9?][3] (by using the result of the Theorem and the definition of intrinsic redundancy therein) a tighter upper bound can be obtained, with $k_0 = \frac{|\mathcal{Y}|\cdot(|\mathcal{X}|-1)}{2}\log m$ ($k_0\log m + k_1 = r_m$ where r_m is explicitly stated in [21, Theorem 9?]). For the sake of simplicity, a looser result is presented here, as this does not change the asymptotical results significantly.

Proof of Lemma 7: The first part is shown by direct substitution. When $\breve{W} = \frac{\hat{P}_{\mathbf{xy}}(x,y)}{Q(x)}$:

$$\begin{aligned}\sum_{x'} Q(x')\breve{W}(y|x') &= \sum_{x'} Q(x')\frac{\hat{P}_{\mathbf{xy}}(x',y)}{Q(x')} \\ &= \sum_{x'} \hat{P}_{\mathbf{xy}}(x',y) = \hat{P}_{\mathbf{y}}(y).\end{aligned} \quad (55)$$

Therefore

$$\begin{aligned}\breve{I}(Q,\breve{W}) &= \breve{I}\left(Q(x), \frac{\hat{P}_{\mathbf{xy}}(x,y)}{Q(x)}\right) \\ &\overset{(51),(55)}{=} \sum_{x,y} Q(x)\frac{\hat{P}_{\mathbf{xy}}(x,y)}{Q(x)}\log\left(\frac{\hat{P}_{\mathbf{xy}}(x,y)}{Q(x)\hat{P}_{\mathbf{y}}(y)}\right) \\ &= \sum_{x,y}\hat{P}_{\mathbf{xy}}(x,y)\log\left(\frac{\hat{P}_{\mathbf{x}|\mathbf{y}}(x|y)}{Q(x)}\right) \\ &= \frac{1}{m}\sum_{i=1}^{m}\log\left(\frac{\hat{P}_{\mathbf{x}|\mathbf{y}}(x_i|y_i)}{Q(x_i)}\right) \\ &= \frac{1}{m}\log\frac{\hat{p}(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}.\end{aligned} \quad (56)$$

[3]Reference is to be updated in the final revision.

As for the second claim, by Markov bound :

$$\begin{aligned}&\Pr\left(\frac{\hat{p}(\mathbf{X}|\mathbf{y})}{Q(\mathbf{X})} \geq \exp(mT)\Big|\mathbf{y}\right) \\ &\leq \frac{1}{\exp(mT)}\mathbb{E}\left[\frac{\hat{p}(\mathbf{X}|\mathbf{y})}{Q(\mathbf{X})}\Big|\mathbf{y}\right] \\ &\overset{(a)}{=} \exp(-mT)\sum_{\mathbf{x}\in\mathcal{X}^m}\frac{\hat{p}(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}Q(\mathbf{x}) \\ &= \exp(-mT)\sum_{\mathbf{x}\in\mathcal{X}^m}\hat{p}(\mathbf{x}|\mathbf{y}),\end{aligned} \quad (57)$$

where (a) is because $\mathbf{X}$ is distributed Q independently of $\mathbf{y}$. To bound the sum above, let us split the set of sequences $\mathbf{x}$ to sub-sets having the same conditional empirical probability $\hat{P}_{\mathbf{x}|\mathbf{y}}(x|y)$ (i.e. same conditional type [22][23, §11]). In a subset having $\hat{P}_{\mathbf{x}|\mathbf{y}}(x|y) = p(x|y)$, the empirical probability $\hat{p}(\mathbf{x}|\mathbf{y}) = \prod_i p(x_i|y_i)$ equals the (legitimate) probability of the sequence under the i.i.d. distribution p, and as a result $\sum_{\mathbf{x}:\hat{P}_{\mathbf{x}|\mathbf{y}}(x|y)=p(x|y)}\hat{p}(\mathbf{x}|\mathbf{y}) \leq 1$. The number of subsets is upper bounded (similarly to bounds on the number types [23, Theorem 11.1.1]) by which is upper bounded by $(m+1)^{(|\mathcal{X}|-1)\cdot|\mathcal{Y}|}$, since $p(x|y) \in \left\{0, \frac{1}{m}, \frac{2}{m}, \ldots, 1\right\}$ is completely defined by $(|\mathcal{X}|-1)\cdot|\mathcal{Y}|$ integers in $\{0,\ldots,m\}$.

$$\begin{aligned}\sum_{\mathbf{x}\in\mathcal{X}^m}\hat{p}(\mathbf{x}|\mathbf{y}) &= \sum_{p}\sum_{\mathbf{x}:\hat{P}_{\mathbf{x}|\mathbf{y}}(x|y)=p(x|y)}\hat{p}(\mathbf{x}|\mathbf{y}) \\ &\leq (m+1)^{(|\mathcal{X}|-1)\cdot|\mathcal{Y}|}.\end{aligned} \quad (58)$$

Substituting in (57) and using $m+1 \leq 2m$ yields the desired result.

B. Decoding condition and estimated channel for the scheme

When the communication scheme was described in Section IV-B, the details of the decoding condition and channel estimation were omitted. These are specified below. At each symbol of the block and for each codeword $\mathbf{x}_l, l = 1,\ldots,\exp(K)$ in the codebook, the receiver evaluates the following decoding condition:

$$\log\frac{\hat{p}(\mathbf{x}_l|\mathbf{y})}{\hat{Q}_i(\mathbf{x}_l)} > \beta K, \quad (59)$$

where β is a parameter to be specified later on, and the vectors $\mathbf{x}_l$ and $\mathbf{y}$ are taken over the symbols of the block.

Equivalently, by Lemma 7, the decoding condition can be written as:

$$m\cdot\breve{I}(\hat{Q}_i,\breve{W}) > \beta K, \quad (60)$$

where m is the number of the symbol in the block and $\breve{W}(y|x)$ is a channel estimate according to (49), where $\mathbf{x}$ is substituted with the hypothesized input $\mathbf{x}_l$ and $\mathbf{y}$ is known output vector over the block.

After decoding, the receiver sets the estimated channel $\breve{W}_i$ as the false channel $\breve{W}(y|x)$ measured according to (49), where $\mathbf{x},\mathbf{y}$ are the m length vectors denoting the (hypothesized) input and output vectors over the duration of the block.

To produce the next prior, this false channel is fed into the prediction scheme of Lemma 4, with $F_i(Q) = \breve{I}(Q,\breve{W}_i)$ and

where m_i denotes the length of block i. The parameters β, η, λ (the latter are required for the prediction scheme of Lemma 4) will be determined in the course of the proof.

C. Proof outline

The following proof outline conveys the main ideas in the proof, while some details were intentionally dropped out, for simplicity.

1) Using the results of Lemma 7 it is shown that the block lengths can satisfy the inequality (39) required by Lemma 4, up to a small overhead term in K, while still attaining a small probability of error.
2) Operating the prior prediction scheme of Lemma 4, with $F_i(Q) = \breve{I}(Q, \breve{W}_i)$ as the metric with $\breve{W}_i$ the *measured* channels, guarantees that if no errors were made, the rate achieved by the system exceeds $\max_Q \sum_{i=1}^{B+1} \frac{m_i}{n} \breve{I}(Q, \breve{W}_i)$ up to vanishing factors, where B is the number of blocks that were sent.
3) Due to the convexity of the false mutual information with respect to the channel, the rate above exceeds $\max_Q \breve{I}(Q, \breve{W}_A)$ where $\breve{W}_A = \sum_{i=1}^{B+1} \frac{m_i}{n} \breve{W}_i$.
4) Since the rate above exceeds $\breve{I}(Q, \breve{W}_A)$ for any Q, it exceeds $\breve{C}(\breve{W}_A) = \max_Q \breve{I}(Q, \breve{W}_A)$.
5) All is left is to show the convergence in probability of $\breve{W}_A$ to the true average channel $\overline{W}$, and by using the continuity of the capacity this proves the convergence in probability of $\breve{C}(\breve{W}_A)$ to the capacity of the averaged channel $C(\overline{W})$.
6) In order to attain explicit bounds on the convergence rate, bounds relating the difference in capacity to the difference in the channels are used, and the system parameters are optimized.

Note that there are several delicate issues caused by the relations between $\breve{W}_i$, m_i and $\hat{Q}_i$. For example, the correct operation of the prior predictor relies on the assumption of correct decoding which is required to obtain the correct channel estimators (i.e. that $\mathbf{x}$ used in (49) is the true channel input). However, conditioning on the event of correct decoding changes the distribution of the average estimated channel $\breve{W}_A$. Another example is that, although the convergence of $\sum_{i=0}^{B+1} \frac{m_i}{n} \breve{W}_i$ to $\overline{W}$ appears to be trivial at first sight, the proof is complicated by the fact that the block lengths m_i are random variables, which themselves depend on the estimated channels $\breve{W}_i$. One embodiment of this dependence is that the block would never end with an estimated channel which has zero capacity. Another dependence is between $m_i, \breve{W}_i$ of different blocks, created through the prior prediction $\hat{Q}_i$.

The proof starts with a set of propositions formalizing the claims made in the proof outline above. k to denotes the symbol index and i denotes the block index. The block index of a certain symbol is denoted $i = b_k$ (i.e. $i = b_k$ if symbol k belongs to block i). The length of each block is denoted m_i ($i = 1, \ldots, B+1$, including the last block). The last block is not accounted for in the rate, even if it is decoded.

D. Error probability

Proposition 1 (Error probability and decoding thresholds)**.** For the value of β given below (63), the probability of any decoding error occurring in any of the blocks is at most ϵ.

Proof: Consider a specific block and denote by m the number of the symbol inside the block. Since codewords other than the one which is actually transmitted are independent of $\mathbf{x}, \mathbf{y}$, the probability to decide in favor of a specific erroneous codeword $\mathbf{X}_l$, at any specific symbol k (i.e. that (59) will hold with respect to it), is upper bounded using (53) by:

$$\begin{aligned} P_{err}(l,k) &= \Pr\left(\log\frac{\hat{p}(\mathbf{X}_l|\mathbf{y})}{Q(\mathbf{X}_l)} > \beta K \middle| \mathbf{y}\right) \\ &= \Pr\left(\frac{\hat{p}(\mathbf{X}_l|\mathbf{y})}{Q(\mathbf{X}_l)} > \exp(mT) \middle| \mathbf{y}\right)\Big|_{T=\beta K/m} \\ &\leq \exp(-(\beta K - k_0 \log m - k_1)), \end{aligned} \tag{61}$$

where k_0, k_1 are defined in Lemma 7. And by taking expected value over $\mathbf{Y}$, the same bound holds when not conditioning on $\mathbf{y}$. Since there are $\exp(K) - 1$ competing codewords, and n symbols, the probability to decide in favor of any erroneous codeword at any symbol (i.e. to make any decoding error), is upper bounded using the union bound, by:

$$\begin{aligned} P_{err} &\leq \exp(K) \cdot n \cdot \exp(-(\beta K - k_0 \log n - k_1)) \\ &= \exp(-((\beta-1)K - (k_0+1)\log n - k_1)), \end{aligned} \tag{62}$$

where $\log m$ was replaced by $\log n \geq \log m$. β is determined so as to make the RHS equal ϵ, and thereby guarantee the error probability is at most ϵ:

$$\beta = 1 + \frac{\log(\epsilon^{-1}) + (k_0+1)\log n + k_1}{K}. \tag{63}$$

A suitable choice of K would yield $\beta \underset{n\to\infty}{\longrightarrow} 1^+$. □

E. Attained rate

The following lemma relates the rate to the averaged estimated channel $\breve{W}_A$:

Proposition 2 (Rate as a function of average estimated channel)**.** If there are no decoding errors, the rate of the scheme satisfies:

$$R = \frac{KB}{n} \geq (1-\delta_1) \cdot \min\left(\breve{C}\left(\breve{W}_A\right), I_{\max}\right) - \Delta_{\text{pred}}, \tag{64}$$

where $\breve{C}\left(\breve{W}\right) \triangleq \max_{Q\in\Delta_{\mathcal{X}}} \breve{I}(Q, \breve{W})$ is the false capacity, $\breve{W}_A$ is the averaged estimated channel

$$\breve{W}_A(y|x) = \frac{1}{n}\sum_{i=1}^{B+1} m_i \breve{W}_i(y|x), \tag{65}$$

Δ_{pred} is defined in Lemma 4 (for the relevant parameters n, K, λ), and

$$\delta_1 = \frac{1}{K}\left[\log(\epsilon^{-1}) + (k_0+1)\log n + k_1 + \log\left(\frac{|\mathcal{X}|}{\lambda}\right)\right]. \tag{66}$$

Proof: Denote by $\breve{W}_i^{(l)}(y|x)$ the channel estimate according to (49), taken over the symbols of the i-th block, with respect

to the hypothesized input sequence $\mathbf{x}_l$. By definition of $\breve{W}_i$ (Section V-B), $\breve{W}_i = \breve{W}^{(l)}(y|x)$ when l is the index of the correct codeword. Denote by $\breve{W}_i^*$ the value of $\breve{W}_i^{(l)}(y|x)$ when l is the index of the hypothesized codeword. When there are no errors, $\breve{W}_i^* = \breve{W}_i$.

The prediction scheme of Lemma 4 is applied with $F_i(Q) = \breve{I}(Q, \breve{W}_i^*)$. By Lemma 6, this choice satisfies the conditions of the lemma with respect to $F_i(Q)$. Assuming there are no errors, then $F_i(Q) = \breve{I}(Q, \breve{W}_i)$.

In the following, the decoding condition is used to show, that the requirements of Lemma 4 with respect to the block length (39) hold.

Denote by $\breve{W}_i^{(B)}$ and $\breve{W}_i^{(E)}$, the channel estimates taken with respect to the true $\mathbf{x}$ over the first $m_i - 1$ symbols of the block i, and over the last symbol of the block, respectively. In other words, if block i spans symbols $[k_i, l_i]$ where $l_i - k_i + 1 = m_i$ then

$$\breve{W}_i(y|x) = \frac{\sum_{k=k_i}^{l_i} \mathrm{Ind}(X_k = x, Y_k = y)}{m_i \cdot \hat{Q}_i(x)} \quad (67)$$

$$\breve{W}_i^{(B)}(y|x) = \frac{\sum_{k=k_i}^{l_i-1} \mathrm{Ind}(X_k = x, Y_k = y)}{(m_i - 1)\hat{Q}_i(x)} \quad (68)$$

$$\breve{W}_i^{(E)}(y|x) = \frac{\mathrm{Ind}(X_{l_i} = x, Y_{l_i} = y)}{\hat{Q}_i(x)}, \quad (69)$$

where in the equations above the empirical distribution in (49) is written explicitly as a normalized sum of indicator functions. Let us assume $m_i > 1$ and return to the case of $m_i = 1$ at the end. From the above:

$$\breve{W}_i(y|x) = \frac{m_i - 1}{m_i}\breve{W}_i^{(B)}(y|x) + \frac{1}{m_i}\breve{W}_i^{(E)}(y|x). \quad (70)$$

Since at symbol $m_i - 1$ in the block, which is one symbol before decoding, none of the codewords satisfies the decoding condition (60), including the correct codeword (which corresponds to the true channel input $\mathbf{X}$):

$$(m_i - 1) \cdot \breve{I}\left(\hat{Q}_i, \breve{W}_i^{(B)}\right) \le \beta K. \quad (71)$$

The same holds for the last block $i = B + 1$. As for $\breve{W}_i^{(E)}$, (35) yields:

$$\hat{Q}_i(x) \ge \frac{\lambda}{|\mathcal{X}|}, \quad (72)$$

and because $\breve{W}_i^{(E)}$ is measured on a single symbol, the following bound holds:

$$\breve{I}\left(\hat{Q}_i, \breve{W}_i^{(E)}\right) = \log\left(\frac{1}{\hat{Q}_i(X_{l_i})}\right) \le \log\left(\frac{|\mathcal{X}|}{\lambda}\right). \quad (73)$$

The equality above can be obtained using Lemma 7, or by definition (51), using the fact that only for a single pair (x, y), $\breve{W}_i^{(E)}(y|x) > 0$. Combining (71) and (73) using (70):

$$\begin{aligned} m_i \cdot \breve{I}\left(\hat{Q}_i, \breve{W}_i\right) &= m_i \cdot \breve{I}\left(\hat{Q}_i, \frac{m_i - 1}{m_i}\breve{W}_i^{(B)} + \frac{1}{m_i}\breve{W}_i^{(E)}\right) \\ &\le (m_i - 1) \cdot \breve{I}\left(\hat{Q}_i, \breve{W}_i^{(B)}\right) + 1 \cdot \breve{I}\left(\hat{Q}_i, \breve{W}_i^{(E)}\right) \\ &\le \beta K + \log\left(\frac{|\mathcal{X}|}{\lambda}\right) \triangleq \tilde{K}. \end{aligned} \quad (74)$$

In the case of $m_i = 1$, $\breve{W}_i = \breve{W}_i^{(E)}$ and (74) holds due to (73). The last inequality means the conditions of Lemma 4 with respect to m_i are satisfied, with K replaced by $\tilde{K}$. Under the conditions of the lemma, it guarantees that:

$$\tilde{R} \triangleq \frac{\tilde{K}B}{n} \ge \min\left(\max_Q \sum_{i=1}^{B+1} \frac{m_i}{n} \cdot \breve{I}(Q, \breve{W}_i), I_{\max}\right) - \tilde{\Delta_{\mathrm{pred}}}, \quad (75)$$

where $\tilde{\Delta_{\mathrm{pred}}} = \Delta_{\mathrm{pred}}(\tilde{K})$ is the offset defined in the lemma, with K replaced by $\tilde{K}$. The convexity of $\breve{I}$ with respect to the channel (Lemma 6) is now used to relate the sum above to the capacity of the estimated averaged channel $\breve{W}_A$:

$$\sum_{i=1}^{B+1} \frac{m_i}{n} \cdot \breve{I}(Q, \breve{W}_i) \ge \breve{I}\left(Q, \sum_{i=1}^{B+1} \frac{m_i}{n} \cdot \breve{W}_i\right) = \breve{I}\left(Q, \breve{W}_A\right). \quad (76)$$

Substituting in (75) yields:

$$\begin{aligned} \tilde{R} &\ge \min\left(\max_Q \breve{I}\left(Q, \breve{W}_A\right), I_{\max}\right) - \tilde{\Delta_{\mathrm{pred}}} \\ &= \min\left(\breve{C}\left(\breve{W}_A\right), I_{\max}\right) - \tilde{\Delta_{\mathrm{pred}}}. \end{aligned} \quad (77)$$

Because the actual rate that the scheme achieves is not $\tilde{R}$ but $R = \frac{K \cdot B}{n}$, the rate is at least:

$$R = \tilde{R} \cdot \frac{K}{\tilde{K}} \ge \frac{K}{\tilde{K}} \cdot \min\left(\breve{C}\left(\breve{W}_A\right), I_{\max}\right) - \frac{K}{\tilde{K}}\tilde{\Delta_{\mathrm{pred}}}. \quad (78)$$

Considering the second term, notice that the expression for $\Delta_{\mathrm{pred}}(K)$ in Lemma 4, is sublinear in K, i.e. $\frac{1}{K}\Delta_{\mathrm{pred}}(K)$ is decreasing with K, and therefore $\frac{K}{\tilde{K}}\tilde{\Delta_{\mathrm{pred}}}(\tilde{K}) \le \frac{K}{K}\tilde{\Delta_{\mathrm{pred}}}(K)$, and the offset term in (78) can be replaced by $\Delta_{\mathrm{pred}}(K)$.

As for the factor $\frac{K}{\tilde{K}}$:

$$\begin{aligned} \frac{\tilde{K}}{K} &= \beta + \frac{1}{K}\log\left(\frac{|\mathcal{X}|}{\lambda}\right) \\ &= 1 + \underbrace{\frac{1}{K}\left[\log(\epsilon^{-1}) + (k_0 + 1)\log n + k_1 + \log\left(\frac{|\mathcal{X}|}{\lambda}\right)\right]}_{\delta_1}. \end{aligned} \quad (79)$$

and using $\frac{K}{\tilde{K}} = \frac{1}{1+\delta_1} \ge 1 - \delta_1$ yields the desired result. □

F. Channel convergence

The following discusses the convergence of $\breve{W}_A$ to $\overline{W}$. As mentioned above, m_i and $\breve{W}_i$ are statistically dependent. To avoid conditioning on m_i, $\breve{W}_A$ can be written in an alternative form. Plugging the explicit form of $\breve{W}_i$ from (67) into the definition of $\breve{W}_A$ (65):

$$\begin{aligned} \breve{W}_A &= \frac{1}{n}\sum_{i=1}^{B+1} m_i \breve{W}_i(y|x) \\ &= \frac{1}{n}\sum_{i=1}^{B+1} m_i \frac{\sum_{k=k_i}^{l_i} \mathrm{Ind}(X_k = x, Y_k = y)}{m_i \cdot \hat{Q}_i(x)} \\ &= \frac{1}{n}\sum_{k=1}^{n} \frac{\mathrm{Ind}(X_k = x, Y_k = y)}{\hat{Q}_{b_k}(x)}. \end{aligned} \quad (80)$$

Recall that the averaged channel is

$$\overline{W} = \frac{1}{n}\sum_{k=1}^{n} W_k(y|x). \tag{81}$$

To show that $\breve{W}_A - \overline{W} \overset{Prob.}{\underset{n\to\infty}{\longrightarrow}} 0$, define

$$\gamma_k(x,y) \triangleq \frac{1}{n}\left[\frac{\mathrm{Ind}(X_k = x, Y_k = y)}{\hat{Q}_{b_k}(x)} - W_k(y|x)\right], \tag{82}$$

then

$$\breve{W}_A - \overline{W} = \sum_{k=1}^{n}\gamma_k(x,y). \tag{83}$$

Although $\gamma_k(x,y)$ are not i.i.d., they constitute a bounded martingale difference sequence, where the martingale is $\sum_{j=1}^{k}\gamma_j$. First, by (72), each component $\gamma_k(x,y)$ is bounded $-\frac{1}{n} \le \gamma_k(x,y) \le \frac{1}{n}|\mathcal{X}|\lambda^{-1} \triangleq \gamma_{\max}$, so they be bounded in absolute value by $\gamma_{\max}$. On average over the common randomness, each symbol X_k is generated $X_k \sim \hat{Q}_{b_k}(x)$ independent of the past (given $\hat{Q}_{b_k}(x)$). In other words, for someone not knowing the specific codebook, the knowledge of past values of $\mathbf{X}_1^{k-1}, \mathbf{Y}_1^{k-1}$ does not yield any information about X_k when $\hat{Q}_{b_k}(x)$ is given. Define the state variable $S_{k-1} = \left(\mathbf{X}_1^{k-1}, \mathbf{Y}_1^{k-1}, \{\hat{Q}_{b_j}\}_{j=1}^{k}\right)$. Note that $\hat{Q}_{b_k}$ is only generated as a function of past symbols and therefore can be considered as part of the state at time k. The conditional expectation of γ_k is:

$$\begin{aligned}\mathbb{E}\left[\gamma_k(x,y)\middle|S_{k-1}\right] &= \frac{\Pr(X_k = x, Y_k = y|S_{k-1})}{n\cdot\hat{Q}_{b_k}(x)} - \frac{W_k(y|x)}{n} \\ &= \frac{\hat{Q}_{b_k}(x)\cdot W_k(y|x)}{n\cdot\hat{Q}_{b_k}(x)} - \frac{W_k(y|x)}{n} = 0.\end{aligned} \tag{84}$$

Now, since the previous value of the sum $\sum_{j=1}^{k-1}\gamma_j$ is only a function of S_{k-1}, by applying the iterated expectations law:

$$\begin{aligned}&\mathbb{E}\left[\gamma_k(x,y)\middle|\sum_{j=1}^{k-1}\gamma_j\right] \\ &= \mathbb{E}\left\{\mathbb{E}\left[\gamma_k(x,y)\middle|S_{k-1}, \sum_{j=1}^{k-1}\gamma_j\right]\middle|\sum_{j=1}^{k-1}\gamma_j\right\} = 0,\end{aligned} \tag{85}$$

which shows $\sum_{j=1}^{k}\gamma_j$ is a martingale. Applying Hoeffding-Azuma Inequality [13, A.1.3][24][17] yields:

$$\begin{aligned}&\Pr\left\{\left|\breve{W}_A(y|x) - \overline{W}(y|x)\right| > t\right\} = \Pr\left\{\left|\sum_{k=1}^{n}\gamma_k(x,y)\right| > t\right\} \\ &\le 2e^{-\frac{2t^2}{n\gamma_{\max}^2}} = 2e^{-\frac{2n\lambda^2 t^2}{|\mathcal{X}|^2}}.\end{aligned} \tag{86}$$

The above holds for each value of (x,y) separately. To bound the L_∞ norm the union bound is used:

$$\begin{aligned}&\Pr\left\{\|\breve{W}_A - \overline{W}\|_\infty > t\right\} \\ &= \Pr\left\{\bigcup_{x,y}\left[\left|\breve{W}_A(y|x) - \overline{W}(y|x)\right| > t\right]\right\} \\ &\le \sum_{x,y}\Pr\left\{\left|\breve{W}_A(y|x) - \overline{W}(y|x)\right| > t\right\} \\ &\overset{(86)}{\le} 2|\mathcal{X}|\cdot|\mathcal{Y}|\cdot e^{-\frac{2n\lambda^2 t^2}{|\mathcal{X}|^2}}.\end{aligned} \tag{87}$$

To guarantee the above holds with probability at most δ_0, choose t to make the RHS equal δ_0:

$$t = \delta_W = \frac{|\mathcal{X}|}{\lambda}\sqrt{\frac{1}{2n}\ln\left(\frac{2|\mathcal{X}|\cdot|\mathcal{Y}|}{\delta_0}\right)}. \tag{88}$$

This is summarized in the following proposition:

Proposition 3 (Average estimated channel convergence)**.** For any $\delta_0 > 0$, and for δ_W defined above,

$$\Pr\left\{\|\breve{W}_A - \overline{W}\|_\infty > \delta_W\right\} \le \delta_0. \tag{89}$$

Observe that a large λ improves the channel estimate convergence (reduces δ_W), since it increases the minimum rate at which each input symbol is sampled. This is an additional role of λ which is not considered in Lemma 5.

G. Convergence of capacity

The final step is to link the difference in the channels $\|\breve{W}_A - \overline{W}\|$ to the difference in capacities. The following lemma is used:

Lemma 8 (L_p bound on difference of false mutual information and capacity)**.** *Let $Q(x)$ be an input distribution on the discrete alphabet $\mathcal{X}$, $W(y|x), y\in\mathcal{Y}$ a conditional distribution, and $\breve{W}(y|x)$ a false conditional distribution. Define*

$$\Delta_p = \|\breve{W}(y|x) - W(y|x)\|_p, \tag{90}$$

where

$$\|f(x,y)\|_p \triangleq \begin{cases}\left(\sum_{x,y}|f(x,y)|^p\right)^{1/p} & p<\infty \\ \max_{x,y}|f(x,y)| & p=\infty\end{cases}. \tag{91}$$

Assuming $\Delta_p \le \frac{1}{4}$, then:

$$\forall Q: \left|\breve{I}(Q,\breve{W}) - I(Q,W)\right| \le 2f_p(\Delta_p), \tag{92}$$

and

$$\left|\breve{C}(\breve{W}) - C(W)\right| \le 2f_p(\Delta_p), \tag{93}$$

where

$$f_p(t) = -t\cdot|\mathcal{Y}|^{1-1/p}\log\left(\frac{t}{|\mathcal{Y}|^{1/p}}\right). \tag{94}$$

For $p=\infty$, by convention $1/p = 0$. Furthermore $f_p(t)$ is concave and monotonically non-decreasing for $t \le \frac{1}{4}$.

Note that the lemma is also true with respect to legitimate distributions. The proof of the lemma is based on Cover and Thomas' L_1 bound on entropy [23], and Hölder's inequality, and appears in Appendix D.

H. Main argument of the proof

The results above are combined as follows: Choose a value of δ_0. Denote by E the event of any decoding error occurring in any of the blocks, and by D the event $\|\breve{W}_A - \overline{W}\|_\infty > \delta_W$. Below, an over-line $\overline{\square}$ denotes complementary events.

Consider the event $\overline{D} \cap \overline{E}$. In this case, $\|\breve{W}_A - \overline{W}\|_\infty \leq \delta_W$ and from Lemma 8 this implies $|\breve{C}(\breve{W}_A) - C(\overline{W})| \leq \delta_C$ where $\delta_C = 2f_\infty(\delta_W) = -2\delta_W \cdot |\mathcal{Y}| \log(\delta_W)$. From Proposition 2:

$$\begin{aligned} R &\geq (1-\delta_1) \cdot \min\left(\breve{C}\left(\breve{W}_A\right), I_{\max}\right) - \Delta_{\text{pred}} \\ &\geq (1-\delta_1) \cdot \min\left(C(\overline{W}) - \delta_C, I_{\max}\right) - \Delta_{\text{pred}} \\ &\geq (1-\delta_1) \cdot \left(\min\left(C(\overline{W}), I_{\max}\right) - \delta_C\right) - \Delta_{\text{pred}} \\ &= (1-\delta_1) \cdot \left(C(\overline{W}) - \delta_C\right) - \Delta_{\text{pred}} \\ &= C(\overline{W}) - \delta_1 \cdot C(\overline{W}) - \delta_C \cdot (1-\delta_1) - \Delta_{\text{pred}} \\ &\geq C(\overline{W}) - \underbrace{\left(\delta_1 \cdot I_{\max} + \delta_C + \Delta_{\text{pred}}\right)}_{\triangleq \Delta_C}. \end{aligned} \tag{95}$$

To summarize, if $\overline{D} \cap \overline{E}$ then $R \geq C(\overline{W}) - \Delta_C$. By the union bound and Propositions 3,1:

$$\begin{aligned} \Pr\{R < C(\overline{W}) - \Delta_C\} &\leq \Pr\{D \cup E\} \leq \Pr\{D\} + \Pr\{E\} \\ &\leq \delta_0 + \epsilon. \end{aligned} \tag{96}$$

Note that although Lemma 8 is stated for general L_p norms, it was used here only with respect to the L_∞ norm, since it is relatively simple to obtain bounds on the convergence of $\breve{W}_A - \overline{W}$ by using the well known Hoeffding-Azuma inequality per channel element (x, y) and the union bound. However as the distribution of $\breve{W}_A$ tends to a mutlivariate Gaussian distribution, using L_2 norm seems to be more suited. Indeed, applying Lemma 8 with L_2 norm, together with the (yet unpublished) bound on the L_2 convergence of vector martingales due to Hayes [25] yields tighter bounds on the probability of having a small difference $\breve{C}(\breve{W}_A) - C(\overline{W})$ for large alphabet sizes.

I. Choice of the parameters

Finally, the numerical expressions for the various overheads are substituted, and the parameters of the scheme are chosen to approximately optimize the convergence rate. δ_0, ϵ are parameters of choice, and together with λ, K they determine Δ_C. The purpose is to choose λ, K that will approximately minimize Δ_C. This part is rather tedious. The relations leading to Δ_C are collected below:

$$\Delta_C = \delta_1 \cdot I_{\max} + \delta_C + \Delta_{\text{pred}} \tag{97}$$

$$\delta_1 = \frac{1}{K}\Big[\log(\epsilon^{-1}) + (k_0+1)\log n + k_1 + \log\left(\frac{|\mathcal{X}|}{\lambda}\right)\Big] \tag{98}$$

$$\delta_C = -2\delta_W \cdot |\mathcal{Y}| \log(\delta_W) \tag{99}$$

$$\delta_W = \frac{|\mathcal{X}|}{\lambda}\sqrt{\frac{1}{2n}\ln\left(\frac{2|\mathcal{X}|\cdot|\mathcal{Y}|}{\delta_0}\right)} \tag{100}$$

$$\Delta_{\text{pred}} = \frac{K}{n} + I_{\max}\cdot\lambda + c_1\sqrt{\frac{\ln(n)}{n}}\lambda^{-\frac{1}{2}}. \tag{101}$$

$$c_1 = 2\sqrt{K\cdot|\mathcal{X}|(|\mathcal{X}|-1)\cdot I_{\max}} \tag{102}$$

Since $\delta_W \geq \sqrt{\frac{1}{n}}$, $-2\log(\delta_W) \leq \log n$, therefore $\delta_C \leq \delta_W \cdot |\mathcal{Y}|\log(n)$. To make $\delta_W \underset{n\to\infty}{\longrightarrow} 0$ it is required that $\frac{|\mathcal{X}|}{\lambda} \leq \sqrt{n}$, and making this assumption, the last element in δ_1 is bounded by $\log\left(\frac{|\mathcal{X}|}{\lambda}\right) \leq \frac{1}{2}\log n$. Further assuming that $k_1 \leq \frac{1}{4}k_0 \log n$ (this holds trivially for the values of k_0, k_1 of Lemma 7 when $n > 2^4$), and $\epsilon \geq \frac{1}{n^{d_\epsilon}}$ (for some arbitrary polynomial decay rate d_ϵ) yields

$$\begin{aligned} \delta_1 &\leq \frac{1}{K}\left[d_\epsilon\log(n) + (k_0+1)\log n + \tfrac{1}{4}k_0\log n + \tfrac{1}{2}\log n\right] \\ &= \frac{\log n}{K}(d_\epsilon + \tfrac{5}{4}k_0 + \tfrac{3}{2}). \end{aligned} \tag{103}$$

Using these bounds and extracting the constants Δ_C is upper bounded by:

$$\Delta_C \leq \underbrace{c_2\frac{\ln n}{K}}_{(1)} + \underbrace{\frac{c_3}{\lambda}\frac{\ln(n)}{\sqrt{n}}}_{(2)} + \underbrace{I_{\max}\cdot\lambda}_{(3)} + \underbrace{c_4\sqrt{\frac{\ln(n)}{n}\cdot\frac{K}{\lambda}}}_{(4)} + \underbrace{\frac{K}{n}}_{(5)}, \tag{104}$$

where element (1) stems from δ_1, (2) from δ_C and $(3)-(5)$ from Δ_{pred}, and the constants are:

$$c_2 = \left(d_\epsilon + \tfrac{5}{4}k_0 + \tfrac{3}{2}\right)\cdot I_{\max}\cdot\log e \tag{105}$$

$$c_3 = |\mathcal{X}|\cdot|\mathcal{Y}|\cdot\log(e)\cdot\sqrt{\frac{1}{2}\ln\left(\frac{2|\mathcal{X}|\cdot|\mathcal{Y}|}{\delta_0}\right)} \tag{106}$$

$$c_4 = \frac{c_1}{\sqrt{K}} = 2\sqrt{|\mathcal{X}|(|\mathcal{X}|-1)\cdot I_{\max}}. \tag{107}$$

As shall be seen, element (5) is negligible. Therefore let us first optimize the sum of (1) and (4) with respect to K, using Lemma 3. The sum can be written as $aK^\alpha + bK^{-\beta}$ with $\alpha = \frac{1}{2}, \beta = 1, a = c_4\sqrt{\frac{\ln(n)}{n}\cdot\frac{1}{\lambda}}, b = c_2\ln n$. Since K is required to be integer the Lemma 3 applies to real numbers t, first write K as a function of a real valued t: $K = \lfloor t \rfloor$, and assume $t \geq 5$. Then $\frac{1}{K} \leq \frac{1}{t-1} = \frac{1}{t}\frac{t}{t-1} \leq \frac{5}{4}\frac{1}{t}$, and therefore $aK^\alpha + bK^{-\beta} \leq at^\alpha + \underbrace{b\left(\frac{5}{4}\right)^\beta}_{b'}\cdot t^{-\beta}$. Optimizing the bound with respect to t using Lemma 3, yields

$$t^* = \left(\frac{b'\beta}{a\alpha}\right)^{\frac{1}{\alpha+\beta}} = \underbrace{\left(\tfrac{5}{2}c_2c_4^{-1}\right)^{\frac{2}{3}}}_{c_5}\cdot(\lambda\cdot n\ln n)^{\frac{1}{3}}, \tag{108}$$

where

$$c_5 \triangleq \left(\tfrac{5}{2}c_2c_4^{-1}\right)^{\frac{2}{3}}. \tag{109}$$

$$\begin{aligned} aK^\alpha + bK^{-\beta} &\overset{(32)}{\leq} 2^{\frac{1}{3}}\tfrac{3}{2}\cdot a^{\frac{2}{3}}\cdot(b')^{\frac{1}{3}} \\ &= \underbrace{\tfrac{3}{2}\cdot\left(\tfrac{5}{2}c_2c_4^2\right)^{\frac{1}{3}}}_{c_6}\cdot\left(\frac{\ln^2(n)}{n}\cdot\frac{1}{\lambda}\right)^{\frac{1}{3}}. \end{aligned} \tag{110}$$

Substituting in (104) (and upper bounding element (5) by t^*/n) yields:

$$\Delta_C \le \underbrace{c_6 \cdot \left(\frac{\ln^2(n)}{n} \cdot \frac{1}{\lambda}\right)^{\frac{1}{3}}}_{(1)+(4)} + \underbrace{\frac{c_3}{\lambda}\frac{\ln(n)}{\sqrt{n}}}_{(2)} + \underbrace{I_{\max} \cdot \lambda}_{(3)} + \underbrace{c_5 \cdot \left(\lambda \cdot \frac{\ln n}{n^2}\right)^{\frac{1}{3}}}_{(5)}, \tag{111}$$

To determine λ, notice that it is a trade-off between element (3) which is increasing in λ and either $(1)+(4)$ or (2) which are decreasing. Minimizing any combination separately (i.e. $((1)+(4))+(3)$ or $(2)+(3)$) using Lemma 3, yields the same decay rate $O\left(\left(\frac{\ln^2(n)}{n}\right)^{\frac{1}{4}}\right)$, and λ of the form

$$\lambda = c_\lambda \cdot \left(\frac{\ln^2(n)}{n}\right)^{\frac{1}{4}}. \tag{112}$$

Therefore this determines the best decay rate possible for (111). Note that one does not have to worry about the case $\lambda > 1$, since in this case the term $\lambda I_{\max}$ in (104) will exceed $I_{\max}$ and Theorem 3 will be true in a void way. Substituting λ:

$$\begin{aligned}
\Delta_C &\le \underbrace{\frac{c_6}{c_\lambda^{\frac{1}{3}}} \cdot \left(\left(\frac{\ln^2(n)}{n}\right)^{1-\frac{1}{4}}\right)^{\frac{1}{3}}}_{(1)+(4)} + \underbrace{\frac{c_3}{c_\lambda}\left(\frac{\ln^2(n)}{n}\right)^{\frac{1}{2}-\frac{1}{4}}}_{(2)} \\
&\quad + \underbrace{I_{\max} \cdot c_\lambda \cdot \left(\frac{\ln^2(n)}{n}\right)^{\frac{1}{4}}}_{(3)} + \underbrace{c_5 \cdot \left(\lambda \frac{\ln n}{n^2}\right)^{\frac{1}{3}}}_{(5)} \\
&\le \left[\frac{c_6}{c_\lambda^{\frac{1}{3}}} + \frac{c_3}{c_\lambda} + I_{\max} \cdot c_\lambda\right] \cdot \left(\frac{\ln^2(n)}{n}\right)^{\frac{1}{4}} + c_5 \cdot \left(\frac{\ln n}{n^2}\right)^{\frac{1}{3}} \\
&\le \left[\frac{3}{2} \cdot \left(\frac{5}{2}\frac{c_2 c_4^2}{c_\lambda}\right)^{\frac{1}{3}} + \frac{c_3}{c_\lambda} + I_{\max} \cdot c_\lambda + 1\right] \cdot \left(\frac{\ln^2(n)}{n}\right)^{\frac{1}{4}} \\
&= c_\Delta \cdot \left(\frac{\ln^2(n)}{n}\right)^{\frac{1}{4}},
\end{aligned} \tag{113}$$

where in the last inequality the expression for c_6 was substituted and it was that assumed $c_5 \cdot \left(\frac{\ln n}{n^2}\right)^{\frac{1}{3}} \le \left(\frac{\ln^2(n)}{n}\right)^{\frac{1}{4}}$. In the last step c_Δ was defined as:

$$c_\Delta \triangleq \frac{3}{2} \cdot \left(\frac{5}{2}\frac{c_2 c_4^2}{c_\lambda}\right)^{\frac{1}{3}} + \frac{c_3}{c_\lambda} + I_{\max} \cdot c_\lambda + 1. \tag{114}$$

Now, let us revisit the assumptions made along the way.

- In (113), it was assumed that $c_5 \cdot \left(\frac{\ln n}{n^2}\right)^{\frac{1}{3}} \le \left(\frac{\ln^2(n)}{n}\right)^{\frac{1}{4}}$. This requires that $(\ln n)^{\frac{1}{6}} n^{\frac{5}{12}} \ge c_5 = \left(\frac{5}{2}\frac{c_2}{c_4}\right)^{\frac{2}{3}}$, and a sufficient condition is $n \ge \left(\frac{5}{2}\frac{c_2}{c_4}\right)^{\frac{8}{5}}$.
- For (103), it was assumed that $\frac{|\mathcal{X}|}{\lambda} \le \sqrt{n}$. Substituting λ leads to $n \ln^2(n) \ge \frac{|\mathcal{X}|^4}{c_\lambda^4}$, and a sufficient condition is
$$n \ge \frac{|\mathcal{X}|^4}{c_\lambda^4}. \tag{115}$$
- For (103), it was assumed that $\epsilon \ge \frac{1}{n^{d_\epsilon}}$. This holds by simply determining d_ϵ and setting $\epsilon = \frac{1}{n^{d_\epsilon}}$.
- For (103), it was assumed that $k_1 \le \frac{1}{4}k_0 \log n$, i.e. $n \ge \exp(4k_1/k_0)$
- The application of Lemma 4 to obtain Proposition 2 requires that $n \ge e$ and $\tilde{K} \ge 2 \cdot I_{\max}$. Since $\tilde{K} > K$ it is sufficient that $K \ge 2I_{\max}$, or $t^* \ge 2I_{\max} + 1$. Furthermore for (110) it was assumed that $t^* \ge 5$, so it is required that $t^* \ge \max(2I_{\max}+1, 5)$. Substituting $t^* = c_5 \cdot (\lambda \cdot n \ln n)^{\frac{1}{3}} = c_5 \cdot c_\lambda^{\frac{1}{3}} (n \ln^2 n)^{1/4} \ge c_5 \cdot c_\lambda^{\frac{1}{3}} n^{1/4}$ leads to the sufficient condition:
$$n \ge \left(\max(2I_{\max}+1, 5) \cdot c_5^{-1} \cdot c_\lambda^{-\frac{1}{3}}\right)^4. \tag{116}$$

  To summarize, the results holds for $n \ge n_{\min}$ where $n_{\min}$ is the maximum of the conditions of (115),(116), (103) and of $n \ge e$:

$$\begin{aligned}
n_{\min} = \max\Bigg[&e, \frac{|\mathcal{X}|^4}{c_\lambda^4}, \left(\max(2I_{\max}+1, 5) \cdot c_5^{-1} \cdot c_\lambda^{-\frac{1}{3}}\right)^4, \\
&\exp(4k_1/k_0)\Bigg].
\end{aligned} \tag{117}$$

This proves Corollary (1). □

The claims of the Theorem are milder and are easily deduced from this Corollary. Given ϵ, δ, let $\delta_0 = \frac{1}{2}\delta$, and choose any $d_\epsilon > 0$ and $c_\lambda > 0$. Choose N large enough so that the error probability given by the Corollary satisfies $\epsilon(N) = N^{-d_\epsilon} < \min(\epsilon, \frac{1}{2}\delta)$, and $N \ge n_{\min}$. This guarantees that for $n \ge N$, the requirements of the Corollary are met the error probability is $\epsilon(n) \le \epsilon$, and the probability to fall short of the rate is at most $\epsilon(n) + \delta_0 \le \delta$. This concludes the proof of Theorem 3. □

Following is a numerical example for the calculation of c_Δ and $n_{\min}$ in Theorem 3.

Example 2. The parameters $|\mathcal{X}| = 4, |\mathcal{Y}| = 6$, $d_\epsilon = 1$ and $\delta_0 = 10^{-10}$ result in $I_{\max} = 2$ and $c_2 = 72.1, c_3 = 127, c_4 = 9.8, c_5 = 6.97$. Choosing $c_\gamma = 10$ yields $c_\Delta = 51.7$ and $n_{\min} = \min(e, 0.0256, 0.0123, 16) = 16$. The convergence rate is rather slow and $\Delta_C \le 0.2$ only for $n > 3.98 \cdot 10^{12}$.

J. Proof of Corollary 2

During the proof of Theorem 3 it was assumed that the channel sequence is unknown but fixed. It is easy to see that the same proof holds even if the channel sequence is determined by an online adversary.

The error probability (Proposition 1) is maintained regardless of channel behavior, because the probabilistic assumptions made (61) refer to the distribution of codewords that were *not* transmitted. Proposition 2 does not make any assumptions on the channel as it connects the communication rate with the *measured* channel. The main difference is with respect to channel convergence. For the proof of Proposition 3 to hold it needs to be shown that γ_k remains a bounded martingale difference sequence, which boils down to verifying (120) still holds, i.e. that γ_k has zero mean conditioned on the past. Adding the message to the state variable S_{k-1} defined before (120), i.e. redefining $S_{k-1} = \left(\mathbf{X}_1^{k-1}, \mathbf{Y}_1^{k-1}, \{\hat{Q}_{b_j}\}_{j=1}^{k}, \mathbf{b}_1^{\infty}\right)$, where $\mathbf{b}_1^{\infty}$ is the message bit sequence, it can be seen that (120) holds even when the channel $W_k(y|x)$ is a function of S_{k-1}. □

K. A result for channels with memory of the input

Although channels with memory of the input are not considered in the current setting, the scheme presented above can be used over such channels as well. In this case, the performance of the scheme can be characterized as follows:

Lemma 9. *When the scheme of Theorem 3 is operated over a general channel* $\Pr(\mathbf{Y}_1^n|\mathbf{X}_1^n)$*, the results of the theorem hold if the averaged channel is redefined as follows:*

$$\overline{W} = \frac{1}{n}\sum_{k=1}^{n} \Pr(Y_k = y|X_k = x, \mathbf{X}^{k-1}, \mathbf{Y}^{k-1}) \tag{118}$$

Note that for each pair x, y, $\Pr(Y_k = y|X_k = x, \mathbf{X}^{k-1}, \mathbf{Y}^{k-1})$ is a random variable depending on the history $\mathbf{X}^{k-1}, \mathbf{Y}^{k-1}$, and therefore, different from the main setting considered in this paper, $\overline{W}$ is also a random variable. The definition above (118) coincides with the previous definition of $\overline{W}$ (7) when the channel is memoryless with respect to the history $\mathbf{X}^{k-1}, \mathbf{Y}^{k-1}$. This lemma is used in [32] to show competitive universality for channels with memory of the input.

Proof: As in the proof of Corollary 2 it is easy to see that assumptions on the channel apply only to Proposition 3 showing the convergence of the average estimated channel $\breve{W}_A$ to $\overline{W}$. To show Proposition 3 holds, it needs to be shown that γ_k remains a bounded martingale difference sequence, where now γ_k is defined as:

$$\gamma_k(x,y) \triangleq \frac{1}{n}\Big[\frac{\mathrm{Ind}(X_k = x, Y_k = y)}{\hat{Q}_{b_k}(x)} - \Pr(Y_k = y|X_k = x, \mathbf{X}^{k-1}, \mathbf{Y}^{k-1})\Big]. \tag{119}$$

As in (83), the relation $\breve{W}_A - \overline{W} = \sum_{k=1}^{n}\gamma_k(x,y)$ holds. Equation (120) now becomes

$$\begin{aligned}\mathbb{E}\left[\gamma_k\middle|S_{k-1}\right] &= \frac{\Pr(X_k = x, Y_k = y|S_{k-1})}{n\cdot\hat{Q}_{b_k}(x)} \\ &\quad - \frac{1}{n}\Pr(Y_k = y|X_k = x, \mathbf{X}^{k-1}, \mathbf{Y}^{k-1}) \\ &= \frac{\hat{Q}_{b_k}(x)\cdot\Pr(Y_k = y|X_k = x, \mathbf{X}^{k-1}, \mathbf{Y}^{k-1})}{n\cdot\hat{Q}_{b_k}(x)} \\ &\quad - \frac{1}{n}\Pr(Y_k = y|X_k = x, \mathbf{X}^{k-1}, \mathbf{Y}^{k-1}) \\ &= 0.\end{aligned} \tag{120}$$

The rest of the proof of Proposition 3 remains the same. □

VI. Discussion and comments

In this section, some comments are made on the schemes presented here, and the relation of the current results to existing results is discussed.

A. A comparison with AVC capacity

It is interesting to compare the target rate $C(\overline{W})$ with the AVC capacity. Let us start with a short background on the AVC and the relation to the current problem.

In the traditional AVC setting [1], the channel model is similar to the setting assumed here, but slightly more constrained. The channel in each time instance is assumed to be chosen arbitrarily out of a set of channels, each of which is determined by a state. Frequently, constrains on the state sequence (such as maximum power, number of errors) are defined. The AVC capacity is the maximum rate that can be transmitted reliably, for every sequence of states that obeys the constraints.

The AVC capacity may be different depending on whether the maximum or the average error probability over messages is required to tend to zero with block length, on the existence of feedback, and on whether common randomness is allowed, i.e. whether the transmitter and the receiver have access to a shared random variable. The last factor has a crucial effect on the achievable rate as well as on the complexity of the underlying mathematical problem: the characterization of AVC capacity with randomized codes is relatively simple and independent on whether maximum or average error probability is considered, while the characterization of AVC capacity for deterministic codes is, in general, still an open problem. Randomization has a crucial role, since the worst-case sequence of channels is considered. This sequence of channels is chosen after the deterministic code was selected (and therefore sometimes viewed as an adversary), enabling the worst-case sequence of channels to exploit vulnerabilities that exist in the specific code. As an example, for every symmetrizable AVC [27, Definition 2], the AVC capacity for deterministic codes is zero [27, Theorem 1]. When randomization does exist, the random seed is selected "after" the channel sequence was selected (mathematically, the probability over random seeds is taken after the maximum error probability over all possible sequences), and therefore prevents tuning the channel to the worst-case code. When randomization exists, the channel

inputs may be made to appear independent from the point of view of the adversary, thus limiting effective adversary strategies. Therefore the results in the current paper assume common randomness exists.

Let us compare the target rate $C(\overline{W})$ with the randomized AVC capacity. The discrete memoryless AVC capacity without constraints may be characterized as follows: let $\mathcal{W}$ be the set of possible channels that are realized by different channel states (for example in a binary modulo-additive channel with an unknown noise sequence, there are two channels in the set – one in which $y = x$ and another in which $y = 1 - x$). This set is traditionally assumed to be finite, i.e. there is a finite number of "states", however this constraint is immaterial for the comparison. The randomized code capacity of the AVC is [1, Theorem 2]:

$$\begin{aligned} C_{\mathrm{AVC}} &= \max_{Q} \min_{W \in \mathrm{conv}(\mathcal{W})} I(Q, W) \\ &= \min_{W \in \mathrm{conv}(\mathcal{W})} \max_{Q} I(Q, W) = \min_{W \in \mathrm{conv}(\mathcal{W})} C(W), \end{aligned} \tag{121}$$

where $\mathrm{conv}(\mathcal{W})$ is the convex hull of $\mathcal{W}$, which represents all channels which are realizable by a random drawing of channels from $\mathcal{W}$.[4] In the example, $\mathrm{conv}(\mathcal{W})$ would be the set of all binary symmetric channels. When input or state constraints exist, they affect (121) simply by including in the set of Q-s and in $\mathrm{conv}(\mathcal{W})$ only those priors, or channels, that satisfy the constraints (respectively). The converse of (121) is obtained by choosing the worst-case channel $W^* = \operatorname*{argmin}_{W \in \mathrm{conv}(\mathcal{W})} C(W)$ and implementing a discrete memoryless channel (DMC) where the channel law is W^*, by a random selection of channels from $\mathcal{W}$. Hence it is clear that the randomized code capacity cannot be improved by feedback. In contrast, the deterministic code AVC capacity can be improved by feedback, and in some cases made to equal to the randomized code capacity [28][29][30]. Therefore, most existing works on feedback in AVC deal with the deterministic case.

Since by definition $\overline{W} \in \mathrm{conv}(\mathcal{W})$, by (121), $C(\overline{W}) \geq C_{\mathrm{AVC}}$, i.e. the target rate meets or exceeds the AVC capacity. While in the traditional setting, a-priori knowledge of $\mathcal{W}$ or state constraints on the channel is necessary in order to obtain a positive rate, here a rate possibly higher than the AVC capacity is attained, without prior knowledge of $\mathcal{W}$. This is important since without such constraints, i.e. when the channel sequence is completely arbitrary, the AVC capacity is zero. This property makes the system presented here universal, with respect to the AVC parameters, a universality which also holds in an online-adversary setting (Corollary 2).

The difference between C_{AVC} (121) and $C(\overline{W})$ can be regarded as the difference between the capacities of the worst realizable channel $W^* \in \mathrm{conv}(\mathcal{W})$, and the specific channel $\overline{W} \in \mathrm{conv}(\mathcal{W})$ representing the average of the sequence of channels that actually occurred. This difference is obtained by adapting the communication rate to the capacity of the average channel, and adapting the input prior to the prior that achieves this capacity, whereas in the AVC setting, the rate and the prior are determined a-priori, based on the worst-case realizable channel.

As noted above, feedback cannot improve the randomized AVC capacity. Therefore the improvement is attained not merely by the use of feedback, but by allowing the communication rate to vary, whereas in the traditional AVC setting, one looks for a fixed rate of communication which can be guaranteed a-priori (note that the improvement is not in the worst case). In allowing the rate to vary, the formal notion of capacity (as the supremum of achievable rates) is lost, thereby making the question of setting the target rate more ambiguous, but nevertheless the achieved rates are improved.

B. Relation to empirical capacity and mutual information

The capacity of the averaged channel $C(\overline{W})$ is a slight generalization of the notion of *empirical capacity* defined by Eswaran *et al* [3, §D]. The only difference is releasing the assumption made there, that the set of channel states is finite. The empirical capacity of [3] is in itself a generalization of the empirical capacity for modulo additive channels defined by Shayevitz and Feder [2]. Eswaran *et al* [3] assume the prior Q is given a-priori and attain the empirical mutual information $I(Q, \overline{W})$. The scheme used here is similar to the scheme they presented in its high level structure. The current result (Theorem 3) can be regarded as an improvement over the previous work, i.e. attaining the capacity $C(\overline{W}) \geq I(Q, \overline{W})$, rather than the mutual information, by the addition of the universal predictor. This answers the question raised there [3, §D], whether the empirical capacity is attainable.

Another small extension is in Corollary 2, showing that the result holds in an adversarial setting. This extension is outside the main focus of communicating over unknown channels, and is only used to strengthen the claim on universality with respect to the AVC parameters.

C. Competitive universality

In a related paper [5], the concept of the iterated finite block capacity C_{IFB} of an infinite vector channel was presented. This concept is similar in spirit to the finite state compressibility defined by Lempel and Ziv [31]. Roughly speaking, it is the maximum rate that can be reliably attained by any block encoder and decoder, constrained to apply the same encoding and decoding rules over sub-blocks of finite length. The positive result is that C_{IFB} is universally attainable for all modulo-additive channels (i.e. over all noise sequences). The result is obtained by a system similar to the one described in Section IV-B, while the input prior is fixed to the uniform prior. The result uses two key properties of the modulo additive channel:

1) The channel is memoryless with respect to the input x_i (i.e. current behavior is not affected by previous values of the input).
2) The capacity achieving prior is fixed for any noise sequence.

The current work is a step toward removing the second assumption. The capacity of the averaged channel is a bound

[4] The convex hull replaces the distribution $\zeta(s)$ over channel states in [1].

on the rate that can be obtained reliably by a transmitter and a receiver operating on a single symbol, since the channel that this system "sees" can be modeled as a random uniform selection of a channel out of $\{W_i\}_{i=1}^n$, which is termed the "collapsed channel" [5]. By combining k symbols into a single super-symbol, the result can be extended to obtain a rate which is equal to or better from the rate obtained by block encoder and decoder operating over chunks of k symbols. Therefore the current result suggests that it is possible to attain C_{IFB} for all vector channels that are memoryless in the input, i.e. that have the form defined in (3), for an arbitrary sequence of channels W_i (compared to only an arbitrary noise sequence, in the previous result). A stronger result, which applies also to channel with memory, is shown in [32], based on the current scheme, and Lemma 9.

D. Notes on the converse

It is interesting to consider the converse (Theorem 2) from the following point of view: Suppose a competitor is given the entire sequence of channels W_1^n, but is allowed to take from this sequence only the "histogram" (a list of channels and how many times they occurred), and devise a communication system based on this information. The rate that can be guaranteed in this case is limited by $C(\overline{W})$. On the other hand, assuming common randomness exists, it is enough to know $\overline{W}$ in order to attain $C(\overline{W})$ without feedback. To see this intuitively, apply a random interleaver and use the fact the interleaved channel is similar to a DMC with the channel law $\overline{W}$. Therefore even if one knows the entire histogram of the sequence, the average channel $\overline{W}$, which contains less information, contains all information necessary for communication.

To illustrate this, consider the deterministic setting, where instead of a sequence of channel laws $W_i(y|x)$ there is a sequence of deterministic functions $f_i : \mathcal{X} \to \mathcal{Y}$. This is a particular case of the current problem, with $W_i(y|x) = \text{Ind}(y = f_i(x))$. Even in this case, according to Theorem 2, a competitor knowing the list of functions up to order, will not be able to guarantee a rate better than $C(\overline{W})$, where $\overline{W} = \frac{1}{n}\sum_{i=1}^n \text{Ind}(y = f_i(x))$, i.e. a channel created by counting for each x, the normalized number of times a certain y would appear as output.

Comparing the amount of information in the channel histogram and the averaged channel in this case, there are $|\mathcal{Y}|^{|\mathcal{X}|}$ functions, and therefore the distribution is given by $|\mathcal{Y}|^{|\mathcal{X}|} - 1$ real numbers. On the other hand, the average channel is a probability distribution from $|\mathcal{X}|$ to $|\mathcal{Y}|$ and is specified by $(|\mathcal{Y}| - 1) \cdot |\mathcal{X}|$ real numbers.

An interesting property revealed through the example, is that although the setting is deterministic, the result is given in terms of probability functions. These "probabilities" are only averages related to the deterministic function sequence, but this shows that the formulation via probabilities (or frequencies) is more natural than by specifying the function f_i between the input and output.

E. The required feedback rate

The feedback channel was so far assumed to be of unlimited rate and free of delays and errors. This was done mainly to focus the discussion and simplify the results. It is clear from the scheme presented, that because the amount of information required to be fed back to the transmitter can be made small, the capacity of the average channel could be attained even if the feedback link has any small positive rate and a fixed delay. If the feedback channel is such that errors can be mitigated by coding with finite delay, then errors can be accommodated as well. Specifically, as shown in Appendix J, when the feedback rate is limited, or there is a fixed delay, the penalty is a gap of at most $O(\log n)$ symbols between the blocks, and that the normalized loss from this effect tends to zero. Therefore $\Delta_C \underset{n\to\infty}{\longrightarrow} 0$ (with the notation of Theorem 3), with any positive feedback rate and any fixed delay.

F. Convergence rate

Throughout the course of this paper, as the assumptions have been made more realistic, a deterioration of the rate of convergence of the achieved rate to the target rate is seen. Denote by δ_n the gap between the guaranteed rate and the target rate, and focus on the dominant polynomial power $p = -\lim_{n\to\infty} \frac{\ln \delta_n}{n}$, while ignoring the $\ln n$ terms. In the synthetic problem of Section III (assuming "block-wise" variation) §III $p = \frac{1}{2}$, when using the rateless scheme under assumptions of perfect average channel knowledge §IV-D, $p = \frac{1}{3}$, and when releasing the abstract assumptions §V, $p = \frac{1}{4}$. The first deterioration (between $\frac{1}{2}$ and $\frac{1}{3}$) is mainly attributed to the rateless coding scheme. More specifically, it stems from mixing with the uniform prior, which is necessary to bound the regret per block when the blocks have variable lengths. The second deterioration (between $\frac{1}{3}$ and $\frac{1}{4}$) can be attributed mainly to the fact that the number of bits per block K has to increase in a certain rate in order to balance overheads created by the universal decoding procedure (and reduces the rate of adaptation). While the rate of convergence which was achieved deteriorates, the only upper bound presented on the convergence rate is $p \leq \frac{1}{2}$ (§III-C), which is tight only for the first case. We do not know whether better convergence rates can be attained in Theorems 5,3.

G. Comments on the prediction scheme

The results in this paper were obtained by exponential weighting. This scheme was selected mainly due to its simplicity and elegance. Unfortunately, the exponential weighting is performed over a continuous domain (of probabilities), and therefore it is not immediately implementable. Of course, the simplest practical solution could be discrete sampling of the unit simplex and replacement of the integrals by sums. Since the mutual information is continuous, it is possible to bound the error resulting from this discretization. An alternative way is to quantize the set of priors. Instead of competing against a continuum of reference schemes, the number of reference schemes may be reduced to a finite one, by creating a "codebook" of priors $\{Q_m\}$. This codebook is designed

	Synthetic problem ("Block-wise variation")	Arbitrarily varying channel, with side information on average channel and without communication overheads	Arbitrarily varying channel	Notes
Reference	§III, Theorem 1	§IV-D, Lemma 5	§II-B,§V, Theorem 3	
C_1 Attainability	No	No	No	C_1 = Capacity of $\{W_i\}_1^n$ = Mean capacity $\frac{1}{n}\sum_i C(W_i)$
C_2 Attainability	Yes	No	No	C_2 = Mean mutual information with fixed prior $\max_Q \frac{1}{n}\sum_i I(Q,W_i)$
C_3 Attainability	Yes	Yes	Yes	C_3 = Capacity of the time-averaged channel $C(\overline{W}) = C\left(\frac{1}{n}\sum_i W_i\right)$ 1) Best attainable rate not using time structure (Theorem 2). 2) $C_3 \geq C_{\mathrm{AVC}}$ (Section VI-A)
Normalized regret lower bound	$O\left(\frac{1}{n}\right)^{\frac{1}{2}}$	$O\left(\frac{1}{n}\right)^{\frac{1}{2}}$	$O\left(\frac{1}{n}\right)^{\frac{1}{2}}$	
Normalized regret attained	$O\left(\frac{\ln n}{n}\right)^{\frac{1}{2}}$	$O\left(\frac{\ln n}{n}\right)^{\frac{1}{3}}$	$O\left(\frac{\ln^2 n}{n}\right)^{\frac{1}{4}}$	

TABLE I
SUMMARY OF THE RESULTS

so that the penalty in the mutual information resulting from rounding to the nearest codeword, is small. This quantization is useful in terms of the feedback link, which now only has to convey the index m. Having quantized the priors, one may replace the predictors shown here by standard schemes used for competition against a finite set of references [13, §2],[15]. See a rough analysis of this approach in Appendix I. An alternative approach is to bypass the explicit calculation of the predictor $\hat{Q}_i$ and use a rejection-sampling based algorithm to generate a random variable $X \sim \hat{Q}_i$. This approach is demonstrated in Appendix K.

Zinkevich [33] proposed a computationally efficient online algorithm, based on gradient descent, to solve a problem of minimizing the sum of convex functions, each revealed to the forecaster after the decision was made (a similar setting to that of Lemma 4). To apply Zinkevich's results to the current problem, some modifications are required. The mutual information does not have a bounded gradient (which is required by [33]), but this could be bypassed by keeping away from the boundary of $\Delta_{\mathcal{X}}$, i.e. from these points for which one of the elements of Q is 0 or 1. One way to accomplish this is by mixing with the uniform prior when defining the target rate, and use $\max_Q \sum_i I((1-\lambda)Q + \lambda U, W_i)$ as a target, and then bounding the loss induced by this mixture. In the rateless scheme, a bound on the maximum value of $m_i F_i(Q)$ (of Lemma 4) is required and can be obtained using the same methods presented here.

Another application of sequential algorithms to solve problems related to AVC's was proposed by Buchbinder *et al* [34] who used a sequential algorithm to solve a problem of dynamic transmit power allocation, where the current channel state is known but future states are arbitrary.

H. The combination of the communication scheme with the predictor

In the communication scheme proposed in Section IV-B an i.i.d. prior is selected during each block, and is updated only at the end of the block. This choice is motivated by the following considerations:

- Assuming no explicit training symbols are transmitted, the estimation of the channel $\overline{W}$ is done based on the encoded sequence, which is known to the receiver only after decoding (at the end of the block).
- Varying the prior throughout the block inserts memory into the channel input, which complicates the analysis.

The result of this is a relatively slow update of the prior, essentially limited by the block size, which is determined based on communication related considerations (overheads and error probabilities). An alterative would be learning the channel through random training symbols (see for example [2]), and updating the prior from time to time, without relation to the rateless blocks.

I. The behavior of the regret for binary channels

In Section III-C a lower bound on the redundancy in attaining C_2 was shown by a counter example with $|\mathcal{X}| = 4, |\mathcal{Y}| = 2$. It is worth mentioning that for the set of binary channels $|\mathcal{X}| = |\mathcal{Y}| = 2$, the normalized regret is not necessarily $O\left(\sqrt{\frac{1}{n}}\right)$. For this set of channels, the optimal prior does not reach the boundaries of $[0,1]$: the two input probabilities $\Pr(X=x)$ are always in $[e^{-1}, 1-e^{-1}]$ [19]. It is possible to show that the loss function $l(Q,W) = 1 - \frac{I(Q,W)}{I_{\max}}$ satisfies conditions 1,2,4 in Cesa-Bianchi and Lugosi's book [13] Theorem 3.1 (but not condition 3). This fact together with experimental results showing convergence of the FL predictor, suggests that the normalized minimax regret in this case may converge like $O\left(\frac{\log n}{n}\right)$.

J. The uniform component in prior predictor

In the prediction scheme of Theorems 5,3, a uniform prior is mixed with an exponentially weighted predictor (35). This mixing has two advantages:

1) Enabling to bound the instantaneous regret caused by a large block due to a low mutual information
2) Enabling channel estimation by making sure all input symbols have a non zero probability.

Note that alternative solutions are use of training symbols at random locations and termination and re-transmission of blocks whose length exceeds a threshold.

Mixing the exponentially weighted predictor with a uniform distribution is a technique used in prediction problems with partial monitoring, where the predictor only has access to its own loss (or a function of it) and not to the loss of the competitors [13, §6], and effectively assigns some time instances for sampling the range of strategies. In the scheme presented here, the uniform prior plays two roles. One is related to the rateless communication scheme, which required to relate the gains of the predictor to the gain of any alternative prior Q (134) in order to have an upper bound on the latter (135). The second role is in the convergence of the estimated channel (Proposition 3). The second role is similar to the role of uniform distribution in partial monitoring problems: the channel $W(y|x)$ cannot be estimated for input values x that occur with zero probability.

Note that even without the explicit uniform component λU, the exponential weighting element $\int w_i(Q)QdQ$ in (35) includes a small uniform component. Particularly, since referring to (36), $1 \leq e^{\eta \sum_{j=1}^{i-1} m_j \cdot I(Q,\overline{W}_j)} \leq e^{\eta n I_{\max}}$, $w_i(Q) \geq \frac{1}{\mathrm{vol}(\Delta_{\mathcal{X}})} e^{-\eta n I_{\max}}$ and

$$\int_{\Delta_{\mathcal{X}}} w_i(Q)QdQ \geq e^{-\eta n I_{\max}} \underbrace{\frac{1}{\mathrm{vol}(\Delta_{\mathcal{X}})} \int_{\Delta_{\mathcal{X}}} QdQ}_{U} \tag{122}$$
$$= e^{-\eta n I_{\max}} \cdot U.$$

However this value is too small for both purposes.

K. Continuous channels

In the current paper it is assumed the input and output alphabets are finite. In general it is not possible to universally attain C_2 or C_3, even in the context of the synthetic problem of Section III, when the alphabet size $\mathcal{X}$ is infinite. This is since in the continuous case one is trying to assign a probability Q to an infinite set of values, where the values producing the capacity may be a small subgroup. Consider the following example:

Example 3. Let the channel W_a, with input x and output y ($x, y \in \mathbb{R}$) be defined by the arbitrary sequence $\{a_k\}_1^\infty$, $a_k \in \mathbb{R}$, with all $a_i \neq a_k (i \neq k)$. The channel rule is defined by:

$$y = \begin{cases} k & x = a_k \\ 0 & o.w. \end{cases}. \tag{123}$$

For any sequential predictor (even randomized) there is a sequence of channels $\{W_a\}$ such that the values of the sequence $\{a_i\}$ at each step have total probability zero (since the input distribution may have at most a countable group of discrete values with non zero probability). Therefore there is always a sequence of channels where the rate obtained by the predictor is zero. On the other hand, each channel W_a has infinite capacity (since it can transmit noiselessly any integer number). Therefore the value of C_2 is infinite (it is enough to choose a prior suitable for one of the channels in the sum (5)).

It stands to reason that under suitable continuity conditions on $W(y|x)$ and input constraints on $Q(x)$, the problem may be converted to a discrete one, while bounding the loss in this conversion, by discretization of the input – i.e. by selecting the input from a finite grid, or alternatively assuming a parametrization of the channel.

VII. Conclusion

The problem of adapting an input prior for communication over an unknown and arbitrarily varying channel, using feedback from the receiver, was considered. The channel is comprised of an arbitrary sequence of memoryless channels. It was shown possible to asymptotically approach the capacity of the time-averaged channel universally for every sequence of channels. This rate equals or exceeds the randomized AVC capacity of any memoryless channel with the same inputs, and thus the system is universal with respect to the AVC model. The result holds also when the channel sequence is determined adversatively. Negative results showing which communication rates or minimax regret convergence rates cannot be attained universally (see a summary in Table I) were presented. A simplified synthetic problem relating to prediction of the communication prior, which may have applications for block-fading channels was considered as well.

When examining the role of feedback in combating unknown channel, previous works mainly focused on the gains of rate adaptation, while here an additional aspect in which feedback improves the communication rate is seen, namely, selection of the communication prior. The results have implications on competitive universality in communication, and suggest that with feedback, it would be possible for any memoryless AVC, to universally achieve a rate comparable to that of any finite block system, without knowing the channel sequence.

When comparing the results to the traditional AVC results, the former setting was prevailed by the notion of capacity, and thus, even when feedback was assumed, it was not used for adapting the communication rate. Here, for the first time, it was shown that rates equal to or better from the AVC capacity can be attained universally, when releasing the constraint of an a-priori guaranteed rate. This demonstrates the validity of the alternative "opportunistic" problem setting that has been considered in the last decade, for feedback communication over unknown channels, a setting which does not focus on capacity.

Acknowledgment

The authors thank Yishay Mansour for helpful discussions on the universal prediction problem.

APPENDIX

A. Proof of Lemma 2

Lemma 2 relates the exponential weighting of a bounded and concave real function $a \leq F(\mathbf{x}) \leq b$ over a convex vector region $\mathbf{x} \in S \subset \mathbb{R}^d$ to its maximum.

Proof: Let $\mathbf{x}^*$ denote a global maximum of $F(\mathbf{x})$ in S (which exists since F is concave and S is closed). Then from the concavity of F for any $\lambda \in [0,1]$:

$$F(\lambda\mathbf{x}+(1-\lambda)\mathbf{x}^*) \geq \lambda F(\mathbf{x})+(1-\lambda)F(\mathbf{x}^*) \geq \lambda a+(1-\lambda)F(\mathbf{x}^*). \tag{124}$$

Note that the RHS is a constant. Denote $S_\lambda \triangleq \{\lambda\mathbf{x} + (1-\lambda)\mathbf{x}^* : \mathbf{x} \in S\} = \lambda S+(1-\lambda)\mathbf{x}^*$. Then due to convexity $S_\lambda \subset S$ and due to the shrinkage $\mathrm{vol}(S_\lambda) = \lambda^d \mathrm{vol}(S)$. Furthermore by (124), $\forall \mathbf{x} \in S_\lambda : F(\mathbf{x}) \geq \lambda a + (1-\lambda)F(\mathbf{x}^*)$. Write:

$$\begin{aligned}\int_S e^{\eta F(\mathbf{x})} d\mathbf{x} &\geq \int_{S_\lambda} e^{\eta F(\mathbf{x})} d\mathbf{x} = \int_{S_\lambda} e^{\eta(\lambda a+(1-\lambda)F(\mathbf{x}^*))} d\mathbf{x} \\ &= e^{\eta(\lambda a+(1-\lambda)F(\mathbf{x}^*))}\mathrm{vol}(S_\lambda) \\ &= e^{\eta F(\mathbf{x}^*)} \cdot e^{-\eta\lambda(F(\mathbf{x}^*)-a)}\lambda^d \mathrm{vol}(S) \\ &\geq e^{\eta F(\mathbf{x}^*)} \cdot e^{-\eta\lambda(b-a)}\lambda^d \mathrm{vol}(S).\end{aligned} \tag{125}$$

Therefore,

$$\overline{F} \triangleq \frac{1}{\eta}\ln\left[\frac{1}{\mathrm{vol}(S)}\int_S e^{\eta F(\mathbf{x})} d\mathbf{x}\right] \geq F(\mathbf{x}^*)-\lambda(b-a)+\frac{d\ln\lambda}{\eta}. \tag{126}$$

Maximizing the RHS with respect to λ yields:

$$\lambda = \frac{d}{\eta(b-a)}, \tag{127}$$

where $\lambda \leq 1$ by the assumptions of the lemma, and substituting λ yields:

$$\overline{F} \geq F(\mathbf{x}^*)-\frac{d}{\eta}\left(1+\ln\frac{\eta(b-a)}{d}\right) = F(\mathbf{x}^*)-\frac{d}{\eta}\ln\frac{\eta e(b-a)}{d}. \tag{128}$$

Rearranging yields the desired result. □

B. Proof of Lemma 4

During the course of the derivation below, in order to optimize the asymptotical form of the loss (up to constant factors), some simplifying assumptions on the parameters, which hold asymptotically for large enough n are made. For finite n these assumptions might lead to suboptimal results. The assumptions are collected and discussed at the end. All integrals below are by default over the unit simplex $Q \in \Delta_{\mathcal{X}}$.

In the block-wise variation setting (Section III), the target was to control the growth rate of the regret. Here, at each block i, by (37) the gain of the competitor using prior Q is $m_i F_i(Q)$ (bits), while the universal scheme sends a fixed number of bits K. Therefore the gain of the competitor $m_i F_i(Q)$ and the instantaneous regret $m_i F_i(Q) - K$ are related by a constant, and it is more convenient to base the derivation on the gain rather than the regret. The potential function Φ will be used as an approximation of the $\max$ in (37).

Denote the cumulative gain of the competitor with prior Q as:

$$G_i(Q) \triangleq \sum_{j=1}^{i} m_j F_j(Q), \tag{129}$$

And the potential function of $G_i(Q)$ as:

$$\Phi_i \triangleq \Phi(G_i(Q)). \tag{130}$$

Note that Φ_i is not a function of Q due to the integration over Q performed by $\Phi(\cdot)$. Now $w_i(Q)$ can be written as:

$$w_i(Q) = \frac{e^{\eta G_{i-1}(Q)}}{\Phi_{i-1}}. \tag{131}$$

The growth of the potential is bounded by:

$$\begin{aligned}\Phi_i &= \int e^{\eta G_i(Q)} dQ \\ &= \int e^{\eta G_{i-1}(Q)} e^{\eta m_i F_i(Q)} dQ \\ &\overset{(131)}{=} \int \Phi_{i-1} w_i(Q) e^{\eta m_i F_i(Q)} dQ \\ &\overset{(26)}{\leq} \Phi_{i-1} \int w_i \left[1 + \eta m_i F_i(Q) + \eta^2 m_i^2 F_i(Q)^2\right] dQ \\ &= \Phi_{i-1}\left[1 + \eta\int w_i m_i F_i dQ + \eta^2 \int w_i m_i^2 F_i^2 dQ\right],\end{aligned} \tag{132}$$

where in the last inequality Lemma 1 was used, and it was assumed that $\eta m_i F_i(Q) \leq 1$. The dependence of F_i and w_i on Q is suppressed for brevity. In the following, the integrals $\int w_i m_i F_i dQ$ and $\int w_i m_i^2 F_i^2 dQ$ are bounded. The property that a badly chosen prior may cause the iterative system to get stuck (not transmitting any block) translates into the fact that without placing any limitations on $\hat{Q}_i$, the competitor's gain, $m_i F_i(Q)$ may be unbounded, since m_i might be indefinitely large while $F_i(Q)$ can be any positive value. This is prevented by mixing with the uniform prior, which enables us to link $F_i(Q)$ with $F_i(\hat{Q}_i)$. Since in the context of the lemma $F_i(Q)$ is not assumed to be the mutual information, the bound below is slightly looser than Shulman and Feder's (45), but is based on the same technique [19], and only assumes concavity.

Define $x + z$ as modulo-addition over the set $\mathcal{X}$, and write $U(x) = \frac{1}{|\mathcal{X}|}\sum_{z\in\mathcal{X}} Q(x+z)$ for any Q, i.e. express the uniform prior as the mean of all cyclic rotations of Q. Using concavity and non-negativity of F:

$$\begin{aligned}F_i(U) &= F_i\left(\frac{1}{|\mathcal{X}|}\sum_{z\in\mathcal{X}} Q(x+z)\right) \\ &\geq \frac{1}{|\mathcal{X}|}\sum_{z\in\mathcal{X}} F_i\left(Q(x+z)\right) \geq \frac{F_i(Q)}{|\mathcal{X}|}.\end{aligned} \tag{133}$$

Because the prior (35) has the structure $\hat{Q}_i = (1-\lambda)Q' + \lambda U$, by the concavity of F_i:

$$\begin{aligned}\forall Q, i : F_i(\hat{Q}_i) &\geq (1-\lambda)F_i(\hat{Q}') + \lambda F_i(U) \\ &\geq \lambda F_i(U) \overset{(133)}{\geq} \frac{\lambda}{|\mathcal{X}|}F_i(Q),\end{aligned} \tag{134}$$

Using (134) in conjunction with (39) yields:

$$\begin{aligned} m_i F_i(Q) &\overset{(134)}{\leq} \frac{|\mathcal{X}|}{\lambda} m_i F_i(\hat{Q}_i) \\ &\overset{(39)}{\leq} \frac{|\mathcal{X}|}{\lambda} K, \end{aligned} \tag{135}$$

which yields a bound on the competitor gain in each block. Let us now bound the two integrals appearing in (132). Starting with the first integral, using the concavity of F_i:

$$\begin{aligned} K &\overset{(39)}{\geq} m_i F_i(\hat{Q}_i) \overset{(35)}{=} m_i F_i\left((1-\lambda)\int w_i(Q) Q dQ + \lambda U\right) \\ &\geq m_i(1-\lambda)\int w_i(Q) F_i(Q) dQ + \lambda m_i F_i(U) \\ &\geq m_i(1-\lambda)\int w_i(Q) F_i(Q) dQ, \end{aligned} \tag{136}$$

which yields

$$\int w_i(Q) m_i F_i(Q) dQ \leq \frac{K}{1-\lambda}, \tag{137}$$

The second order term is bounded as follows:

$$\begin{aligned} \int w_i(Q) m_i^2 F_i^2(Q) dQ &= \int w_i(Q)(m_i F_i(Q))(m_i F_i(Q)) dQ \\ &\overset{(135)}{\leq} \frac{|\mathcal{X}|}{\lambda} K \cdot \int w_i(Q) m_i F_i(Q) dQ \\ &\overset{(137)}{\leq} \frac{|\mathcal{X}|}{\lambda} K \cdot \frac{K}{1-\lambda}. \end{aligned} \tag{138}$$

Recall that in the classical weighted average predictor [13], the vector product of the instantaneous regret and the weighting function is guaranteed to be non positive (Blackwell condition). Similarly in Section III this product satisfies $\int w(Q) r_i(Q) dQ \leq 0$ (see (23)). In the present case, defining $r_i(Q) = m_i F_i(Q) - K$, then by (137) one obtains $\int w(Q) r_i(Q) dQ \leq \frac{K}{1-\lambda} - K = \frac{K \cdot \lambda}{1-\lambda}$, i.e. due to the inclusion of the uniform prior (which is needed for $m_i F_i$ to be bounded), this integral may be positive, although arbitrarily small. Thus, a price in the first order term is payed, in order to be able to bound the second order term.

Plugging the bounds (137), (138) into (132) yields:

$$\begin{aligned} \Phi_i &\overset{(132)}{\leq} \Phi_{i-1}\left[1 + \eta\int w_i m_i F_i dQ + \eta^2 \int w_i m_i^2 F_i^2 dQ\right] \\ &\overset{(137),(138)}{\leq} \Phi_{i-1}\left[1 + \eta \cdot \frac{K}{1-\lambda} + \eta^2 \frac{|\mathcal{X}|}{\lambda} K \cdot \frac{K}{1-\lambda}\right] \\ &\overset{(26)}{\leq} \Phi_{i-1} e^{\eta \cdot \frac{K}{1-\lambda}\left(1 + \frac{\eta \cdot K \cdot |\mathcal{X}|}{\lambda}\right)} \\ &\leq \ldots \leq \Phi_0 e^{\eta \cdot \frac{K \cdot i}{1-\lambda}\left(1 + \frac{\eta \cdot K \cdot |\mathcal{X}|}{\lambda}\right)}. \end{aligned} \tag{139}$$

In the last step the same relation was inductively applied. Using (139) this yields a bound on $\Phi(G_{B+1}(Q))$, and Lemma 2 is used to relate this bound to $G_{B+1}(Q)$ and to the target rate. $R_T = \frac{1}{n}\max_Q G_{B+1}(Q)$. The dimension is $d = \dim(\Delta_{\mathcal{X}}) = |\mathcal{X}| - 1$. From (37):

$$\underbrace{0}_{\triangleq a} \leq G_{B+1}(Q) \leq \underbrace{n \cdot \max(R_T, I_{\max})}_{\triangleq b}. \tag{140}$$

The reason for setting the upper bound as $b = n \cdot \max(R_T, I_{\max})$ rather than just $n \cdot R_T$, is technical, as this simplifies the conditions required to meet the requirements of the lemma. Satisfying $\eta(b-a) \geq d$ only requires $\eta \geq \frac{|\mathcal{X}|-1}{n I_{\max}}$. By Lemma 2 and (139):

$$\begin{aligned} G_{B+1}(Q) &\overset{(29)}{\leq} \frac{1}{\eta} \ln \frac{\Phi(G_{B+1}(\tilde{Q}))}{\Phi(0)} + n\delta_1(R_T) \\ &= \frac{1}{\eta} \ln \frac{\Phi_{B+1}}{\Phi_0} + n\delta_1(R_T) \\ &\overset{(139)}{\leq} \frac{K \cdot (B+1)}{1-\lambda}\left(1 + \frac{\eta \cdot K \cdot |\mathcal{X}|}{\lambda}\right) + n\delta_1(R_T), \end{aligned} \tag{141}$$

where

$$\delta_1(R_T) \triangleq \frac{|\mathcal{X}| - 1}{n\eta} \cdot \ln\left(\frac{\eta e n \max(R_T, I_{\max})}{|\mathcal{X}| - 1}\right), \tag{142}$$

is the redundancy term introduced by Lemma 2. Bounding R_T using (141), while substituting $K(B+1) = KB + K = nR + K$, yields:

$$\begin{aligned} R_T &= \frac{1}{n}\max_Q G_{B+1}(Q) \\ &\leq \left(R + \frac{K}{n}\right)\frac{1}{1-\lambda}\left(1 + \frac{\eta \cdot K \cdot |\mathcal{X}|}{\lambda}\right) + \delta_1(R_T). \end{aligned} \tag{143}$$

After rearrangement the following bound on R is obtained:

$$R \geq (R_T - \delta_1(R_T)) \cdot (1 - \delta_2) - \delta_3, \tag{144}$$

where

$$1 - \delta_2 \triangleq \left(1 + \frac{\eta \cdot K \cdot |\mathcal{X}|}{\lambda}\right)^{-1} \cdot (1 - \lambda) \tag{145}$$

$$\delta_3 \triangleq \frac{K}{n}. \tag{146}$$

The rest of the proof of Lemma 4 is an algebraic derivation focused on simplifying and optimizing the bound above. The lower bound on R in the RHS of (144) is increasing with respect to R_T. This is since $\frac{\partial}{\partial R_T}\delta_1$ is zero for $R_T \leq I_{\max}$ and for $R_T \geq I_{\max}$ the derivative $\frac{\partial}{\partial R_T}\delta_1$ is $\frac{|\mathcal{X}|-1}{n\eta R_T}$, which by the assumption $\eta \geq \frac{|\mathcal{X}|-1}{n I_{\max}}$ is smaller than 1. Therefore $\frac{\partial}{\partial R_T}(R_T - \delta_1(R_T)) \geq 0$. In order to optimize the parameters, it is assumed for now that $R_T \leq I_{\max}$ and bound the difference $R - R_T$. Using $\frac{1}{1+t} \geq 1 - t$:

$$1 - \delta_2 \geq \left(1 - \frac{\eta \cdot K \cdot |\mathcal{X}|}{\lambda}\right) \cdot (1 - \lambda) \geq 1 - \frac{\eta \cdot K \cdot |\mathcal{X}|}{\lambda} - \lambda. \tag{147}$$

Using (144), under the assumption $R_T \leq I_{\max}$ yields:

$$\begin{aligned} R &\geq (R_T - \delta_1(I_{\max})) \cdot (1 - \delta_2) - \delta_3 \\ &\geq R_T - \delta_1(I_{\max}) - \delta_2 \cdot I_{\max} - \delta_3 \\ &\geq R_T - \delta_1(I_{\max}) - \frac{\eta \cdot K \cdot |\mathcal{X}| \cdot I_{\max}}{\lambda} - \lambda \cdot I_{\max} - \delta_3, \end{aligned} \tag{148}$$

To further simplify $\delta_1(I_{\max})$, assume that $\eta \leq \frac{|\mathcal{X}|-1}{e I_{\max}}$ and therefore $\ln\left(\frac{\eta e \max(R_T, I_{\max})}{|\mathcal{X}|-1}\right) \leq 0$, and $\delta_1(I_{\max}) \leq \frac{|\mathcal{X}|-1}{n\eta} \cdot$

$\ln(n)$. Using these simplifications the RHS of (148) is further bounded by $R_T - \Delta_{\text{pred}}$ where

$$\Delta_{\text{pred}} = I_{\max}\cdot\lambda + \underbrace{\frac{c_0}{\lambda}}_{\triangleq a_1}\cdot\eta + \underbrace{(|\mathcal{X}|-1)\cdot\frac{\ln(n)}{n}}_{\triangleq b_1}\cdot\frac{1}{\eta} + \delta_3, \quad (149)$$

and $c_0 = K\cdot|\mathcal{X}|\cdot I_{\max}$.

Applying Lemma 3 to the optimization of the two terms depending on η in (149) (marked a_1, b_1, with powers $\alpha = 1, \beta = 1$) yields:

$$\eta^* = \sqrt{\frac{b_1}{a_1}} = \sqrt{\frac{|\mathcal{X}|-1}{c_0}\cdot\frac{\ln(n)\cdot\lambda}{n}}, \quad (150)$$

and

$$\begin{aligned}\Delta_{\text{pred}}\Big|_{\eta=\eta^*} &= I_{\max}\cdot\lambda + 2\sqrt{a_1 b_1} + \delta_3 \\ &= I_{\max}\cdot\lambda + 2\sqrt{\frac{c_0(|\mathcal{X}|-1)\cdot\ln(n)}{n\lambda}} + \delta_3.\end{aligned} \quad (151)$$

Substituting c_0 yields Δ_{pred} and η stated in the Lemma. Now, the derivation involving equations (148) – (151) assumes $R_T \le I_{\max}$. Since the lower bound (144) on R is increasing with respect to R_T, in the case that $R_T > I_{\max}$ the lower bound on R is guaranteed to be better than the lower bound $R \ge I_{\max} - \Delta_{\text{pred}}$ attained for $R_T = I_{\max}$ (in other words, the RHS of (144) for $R_T = I_{\max}$ is at least $I_{\max} - \Delta_{\text{pred}}$). Therefore the bound can be stated as $R \ge \min(R_T, I_{\max}) - \Delta_{\text{pred}}$.

The various assumptions made along the way are now considered. For most of these, the technique used in the proof of Theorem 1 applies, i.e. showing that if the assumptions do not hold, then (possibly under some simple conditions), $\Delta_{\text{pred}} \ge I_{\max}$ and therefore the lemma holds in a void way (since the RHS of (40) becomes non-positive).

In (132) it was assumed that $\eta m_i F_i(Q) \le 1$. Using the upper bound of (135) a sufficient condition is $\eta\frac{|\mathcal{X}|}{\lambda}K \le 1$. If this condition doesn't hold, i.e. $\eta\frac{|\mathcal{X}|}{\lambda}K > 1$, then the second term in (149) satisfies $\frac{c_0}{\lambda}\eta = \frac{K\cdot|\mathcal{X}|\eta}{\lambda}\cdot I_{\max} > I_{\max}$, so $\Delta_{\text{pred}} > I_{\max}$ and the lemma holds in a void way. Before (149) it was assumed that $\eta \le \frac{|\mathcal{X}|-1}{eI_{\max}}$. When the opposite is true, then second term in (149) satisfies the $\frac{c_0}{\lambda}\eta = \frac{K\cdot|\mathcal{X}|\eta}{\lambda}\cdot I_{\max} > \frac{K\cdot|\mathcal{X}|(|\mathcal{X}|-1)}{e\cdot\lambda} > \frac{2}{e}\cdot K > \frac{K}{2}$. By requiring $K \ge 2I_{\max}$ it follows that in this case the lemma is also true in a void way. To use Lemma 2 it was required that $\eta \ge \frac{d}{b-a} = \frac{|\mathcal{X}|-1}{nI_{\max}}$. If the opposite is true, then the third term in (149) satisfies $\frac{(|\mathcal{X}|-1)\cdot\ln(n)}{\eta n} > I_{\max}\ln(n)$, and thus if $n \ge e$, $\Delta_{\text{pred}} > I_{\max}$. Therefore, by requiring $n > e$ and $K \ge 2I_{\max}$, it follows that if any of the assumptions made does not hold, the lemma is true in a void way. This concludes the proof of Lemma 4. □

C. Proof of Lemma 6

In this proof log-s are natural base (information is measured in nats). This does not change the results since all values scale according to the base of the log-s. Also, all probabilities and false probabilities are assumed to be non-zero. It is easy to check that the results for zero probabilities follow by replacing zeros with small probabilities and taking the limit using $p\log p \underset{p\to 0}{\longrightarrow} 0$.

Non negativity Define $\breve{p}(y) = \sum_x Q(x)\breve{W}(y|x)$ and write:

$$\begin{aligned}-\breve{I}(Q,\breve{W}) &= \sum_{x,y} Q(x)\breve{W}(y|x)\log\left(\frac{\breve{p}(y)}{\breve{W}(y|x)}\right) \\ &\overset{\log t\le t-1}{\le} \sum_{x,y} Q(x)\breve{W}(y|x)\left(\frac{\breve{p}(y)}{\breve{W}(y|x)} - 1\right) \\ &= \sum_x Q(x)\cdot\sum_y \breve{p}(y) - \sum_{x,y} Q(x)\breve{W}(y|x) \\ &= 1 - 1 = 0.\end{aligned} \quad (152)$$

Concavity with respect to Q: Denote as above $\breve{p}(y) = \sum_x Q(x)\breve{W}(y|x)$ and write:

$$\begin{aligned}\breve{I}(Q,\breve{W}) &= \sum_{x,y} Q(x)\breve{W}(y|x)\log\frac{\breve{W}(y|x)}{\breve{p}(y)} \\ &= \sum_{x,y} Q(x)\breve{W}(y|x)\log\breve{W}(y|x) - \sum_y \breve{p}(y)\log\breve{p}(y).\end{aligned} \quad (153)$$

The left hand term is linear with respect to Q. The function $t\log t$ is convex in t (for all $t \ge 0$), and $\breve{p}(y)$ is linear in Q, therefore the right hand term is convex in Q, and so $\breve{I}$ is concave with respect to Q.

Convexity with respect to $\breve{W}$: Let $\lambda_i \ge 0, \sum\lambda_i = 1$, and $\breve{W}(y|x) = \sum_i \lambda_i\breve{W}_i(y|x)$. It needs to be shown that $\Delta \triangleq \breve{I}(Q,\breve{W}) - \sum_i \lambda_i\breve{I}(Q,\breve{W}_i) \le 0$. Define the respective output distributions as $\breve{p}_i(y) = \sum_x Q(x)\breve{W}_i(y|x)$ and $\breve{p}(y) = \sum_x Q(x)\breve{W}(y|x) = \sum_i \lambda_i\breve{p}_i(y)$, then

$$\begin{aligned}\Delta &= \breve{I}(Q,\breve{W}) - \sum_i \lambda_i\breve{I}(Q,\breve{W}_i) \\ &= \sum_{x,y} Q(x)\underbrace{\breve{W}(y|x)}_{\sum_i\lambda_i\breve{W}_i(y|x)}\log\left(\frac{\breve{W}(y|x)}{\breve{p}(y)}\right) \\ &\quad - \sum_{x,y,i}\lambda_i Q(x)\breve{W}_i(y|x)\log\left(\frac{\breve{W}_i(y|x)}{\breve{p}_i(y)}\right) \\ &= \sum_{x,y,i}\lambda_i Q(x)\breve{W}_i(y|x)\log\left(\frac{\breve{W}(y|x)\cdot\breve{p}_i(y)}{\breve{W}_i(y|x)\cdot\breve{p}(y)}\right) \\ &\le \sum_{x,y,i}\lambda_i Q(x)\breve{W}_i(y|x)\left(\frac{\breve{W}(y|x)\cdot\breve{p}_i(y)}{\breve{W}_i(y|x)\cdot\breve{p}(y)} - 1\right) \\ &= \sum_{x,y,i}\lambda_i Q(x)\cdot\frac{\breve{W}(y|x)\cdot\breve{p}_i(y)}{\breve{p}(y)} - \sum_{x,y,i}\lambda_i Q(x)\breve{W}_i(y|x) \\ &= \sum_{x,y} Q(x)\breve{W}(y|x) - \sum_{x,y} Q(x)\breve{W}(y|x) = 0.\end{aligned} \quad (154)$$

Boundness: From $\sum_{x'} Q(x')\breve{W}(y|x') \ge \breve{W}(y|x)Q(x)$ it follows that $\log\left(\frac{\breve{W}(y|x)}{\sum_{x'}Q(x')\breve{W}(y|x')}\right) \le \log\left(\frac{\breve{W}(y|x)}{Q(x)\breve{W}(y|x)}\right) =$

$\log\left(\frac{1}{Q(x)}\right)$. Now write:

$$\begin{aligned}\breve{I}(Q,\breve{W}) &\triangleq \sum_{x,y} Q(x)\breve{W}(y|x)\log\left(\frac{\breve{W}(y|x)}{\sum_{x'}Q(x')\breve{W}(y|x')}\right)\\ &\le \sum_{x,y} Q(x)\breve{W}(y|x)\log\left(\frac{1}{Q(x)}\right)\\ &\le \sum_{x} \sigma Q(x)\log\left(\frac{1}{Q(x)}\right)\\ &= \sigma\cdot H(Q) \le \sigma\cdot\log|\mathcal{X}|.\end{aligned} \tag{155}$$

D. Proof of Lemma 8 and L_p bounds on differences of entropies and capacities

Below is a proof of Lemma 8, relating the L_p norm difference of two channels (one of which may be a false distribution) to the difference in capacities. Two intermediate results that are captured in Lemmas 11,12 are an extension of the L_1 bound of Cover & Thomas to false distributions and a trivial extension of the same bound to L_p norms.

Let us begin with the following L_1 bound on entropy from Cover & Thomas [23]:

Lemma 10 (L_1 bound on entropy, Theorem 7.3.3 of [23])**.** *Let Q, P be two distributions on the finite alphabet $\mathcal{Y}$ with $\|Q-P\|_1 \le \frac{1}{2}$, then*

$$|H(Q)-H(P)| \le -\|Q-P\|_1\cdot\log\left(\frac{\|Q-P\|_1}{|\mathcal{Y}|}\right). \tag{156}$$

Also note that the function $-t\log\frac{t}{|\mathcal{Y}|}$ is monotonous non decreasing for $t \le e^{-1}|\mathcal{Y}|$, as can be verified by differentiation. The first step is to extend the lemma to a case where one of P, Q is a false distribution. In Cover and Thomas' proof, the first step is to write entropy as $H(P) = \sum_y f(P(y))$ where $f = -t\log t$ and to show that for all $0 \le v \le \frac{1}{2}$ and $0 \le t \le 1-v$, the difference in f is bounded by $|f(t+v)-f(t)| \le v\log v$. Here t represents the minimum of $P(y), Q(y)$ for a certain y, v the absolute difference, and $t+v$ the maximum of $P(y), Q(y)$. Then, the difference in entropy is bounded by the sum of the absolute values, this bound is substituted in the summation, and convexity arguments are use to bring it to the desired form. The only step that needs to be modified is showing that $|f(t+v)-f(t)| \le v\log v$, where now t is no longer bounded to $t \le 1-v$. It can be verified by differentiating the function $g(t) = f(t+v)-f(t)$ with respect to t that the derivative is always negative for $v > 0$. In addition, $g(0) > 0$, therefore the maximum absolute of this function, which is the absolute value of either the the maximum or the minimum, occurs at either end of the region to which t is limited. In the original proof this yields $|f(t+v)-f(t)| = |g(t)| \le \max(|g(0)|, |g(1-v)|) = \max(f(v), f(1-v)) = -v\log v$ (notice that $f(0) = f(1) = 0$). Here, since one of P, Q is a legitimate distribution, $t \le 1$ (as the minimum of the two) the following holds instead: $|f(t+v)-f(t)| = |g(t)| \le \max(|g(0)|, |g(1)|) = \max(f(v), -f(1+v))$. As shown below, limiting $v \le \frac{1}{4}$ leads to $f(v) \ge -f(1+v)$, and therefore the bound $|f(t+v)-f(t)| = |g(t)| \le f(v)$ applies as in the original proof and Cover & Thomas' result holds. To show this, consider the function $g(v) = -v\ln v - (v+1)\ln(v+1)$. This function is 0 for $v = 0$, and the derivative is $g'(v) = -\ln v - 1 - \ln(v+1) - 1 = -\ln(v(v+1)e^2)$, it is positive in a certain interval $(0, v_1)$ and negative for $v > v_1$, and therefore it crosses 0 only once. Calculating this function for $v = \frac{1}{4}$ yields a positive value, therefore it is positive for all $v \le \frac{1}{4}$. This variation of Cover & Thomas result is captured in the following lemma:

Lemma 11 (L_1 bound on false entropy difference)**.** *Let P be a distribution on the finite alphabet $\mathcal{Y}$ and $\breve{P}$ be a false distribution on the same alphabet, with $\|\breve{P}-P\|_1 \le \frac{1}{4}$, then*

$$|\breve{H}(\breve{P}) - H(P)| \le -\|\breve{P}-P\|_1\log\left(\frac{\|\breve{P}-P\|_1}{|\mathcal{Y}|}\right), \tag{157}$$

where the false entropy $\breve{H}$ is defined as

$$\breve{H}(\breve{P}) \triangleq -\sum_{y\in\mathcal{Y}}\breve{P}(y)\log\breve{P}(y). \tag{158}$$

Let us first convert the bound to the L_p norm ($p \ge 1$). To relate the norms, Hölder's inequality is used: for two vectors $\mathbf{a}, \mathbf{b}$, $\sum_i |a_i b_i| \le \|a\|_p\cdot\|a\|_{\overline{p}}$, where $\overline{p}^{-1} = 1-p^{-1}$ is the Hölder conjugate of p and by convention for $p = \infty$ the inverse is $1/p = 0$ (note that $\overline{p} \ge 1$ and the conjugate of $p = \infty$ is $\overline{p} = 1$). Then,

$$\begin{aligned}\|\breve{P}-P\|_1 &= \sum_y 1\cdot|\breve{P}(y)-P(y)| \le \|\breve{P}-P\|_p\cdot\|\mathbf{1}\|_{\overline{p}}\\ &= \|\breve{P}-P\|_p\cdot(\sum_{y\in\mathcal{Y}}1^{\overline{p}})^{1/\overline{p}} = \|\breve{P}-P\|_p\cdot|\mathcal{Y}|^{1/\overline{p}}\\ &= \|\breve{P}-P\|_p\cdot|\mathcal{Y}|^{1-1/p}.\end{aligned} \tag{159}$$

Assuming $\|\breve{P}-P\|_p\cdot|\mathcal{Y}|^{1-1/p} \le e^{-1}|\mathcal{Y}|$ and using the monotonicity of the bound of Lemma 11, write:

$$\begin{aligned}|\breve{H}(\breve{P}) - H(P)| &\le -\|\breve{P}-P\|_1\log\left(\frac{\|\breve{P}-P\|_1}{|\mathcal{Y}|}\right)\\ &\le -\|\breve{P}-P\|_p\cdot|\mathcal{Y}|^{1-1/p}\log\left(\frac{\|\breve{P}-P\|_p}{|\mathcal{Y}|^{1/p}}\right)\\ &\triangleq f_p\left(\|\breve{P}-P\|_p\right),\end{aligned} \tag{160}$$

where

$$f_p(t) = -t\cdot|\mathcal{Y}|^{1-1/p}\log\left(\frac{t}{|\mathcal{Y}|^{1/p}}\right). \tag{161}$$

$f_p(t)$ is concave with respect to t (because $-t\ln t$ is concave in $t \ge 0$), and is monotonically non decreasing for $t \le e^{-1}|\mathcal{Y}|^{1/p}$, as can be verified by differentiation. Furthermore, to meet the requirement $\|\breve{P}-P\|_1 \le \frac{1}{4}$ of Lemma 11, it is sufficient that $\|\breve{P}-P\|_p\cdot|\mathcal{Y}|^{1-1/p} \le \frac{1}{4}$ (by (159)), and in addition prior to (160) it was assumed that $\|\breve{P}-P\|_p \le e^{-1}|\mathcal{Y}|^{1/p}$, however it is easy to see that this condition is dominated by the previous one. Since $1-1/p \ge 1$, and $|\mathcal{Y}| > 1$, it is sufficient to require $\|\breve{P}-P\|_p \le \frac{1}{4}$. This result is summarized below:

Lemma 12 (L_p bound on false entropy difference)**.** *Let $p \geq 1$, P be a distribution on the finite alphabet $\mathcal{Y}$, and $\breve{P}$ be a false distribution on the same alphabet with $\|\breve{P} - P\|_p \leq \frac{1}{4}$, then:*

$$|\breve{H}(\breve{P}) - H(P)| \leq f_p\left(\|\breve{P} - P\|_p\right), \tag{162}$$

where f_p is defined in (161)*, and it is concave and monotonically non-decreasing for $t \leq \frac{1}{4}$.*

Now, write the false mutual information (51) as a difference of false entropies (158):

$$\breve{I}(Q, \breve{W}) = \breve{H}\left(\sum_x \breve{W}(y|x)Q(x)\right) - \sum_x Q(x)\breve{H}(\breve{W}(y|x)). \tag{163}$$

The above is analogous to the equality $I(X;Y) = H(Y) - H(Y|X)$. For the channels $W, \breve{W}$ define the difference as $\delta_{xy} = W(y|x) - \breve{W}(y|x)$ and define the output distributions as $P_Y(y) = \sum_x W(y|x)Q(x)$ and $\breve{P}_Y(y) = \sum_x \breve{W}(y|x)Q(x)$, then by the triangle inequality:

$$\begin{aligned} |\breve{I}(Q, \breve{W}) - I(Q, W)| &\leq |H(P_Y) - H(\breve{P}_Y)| \\ &+ \sum_x Q(x)\left|\breve{H}(\breve{W}(y|x)) - H(W(y|x))\right|. \end{aligned} \tag{164}$$

Let us begin with the difference $H(P_Y) - \breve{H}(\breve{P}_Y)$. The L_p bound of Lemma 12 yields:

$$H(P_Y) - H(\breve{P}_Y) \leq f_p(\|P_Y - \breve{P}_Y\|_p). \tag{165}$$

Using the triangle inequality:

$$\begin{aligned} \|P_Y - \breve{P}_Y\|_p &= \|\sum_x Q(x)(W(y|x) - \breve{W}(y|x))\|_p \\ &= \|\sum_x Q(x)\delta_{xy}\|_p \\ &\leq \sum_x \|Q(x)\delta_{xy}\|_{p,y} \\ &= \sum_x Q(x)\|\delta_{xy}\|_{p,y}, \end{aligned} \tag{166}$$

where the notation $\|\square\|_{p,y}$ is used to emphasize that the norm operation is with respect to y only. Using Hölder's inequality,

$$\begin{aligned} \sum_x Q(x)\|\delta_{xy}\|_{p,y} &\leq \|Q(x)\|_{\overline{p}} \cdot \left\|\|\delta_{xy}\|_{p,y}\right\|_p \\ &= \left(\sum_x Q(x)^{\overline{p}}\right)^{1/\overline{p}} \cdot \|\delta_{xy}\|_p \\ &\overset{\overline{p}\geq 1}{\leq} \left(\sum_x Q(x)\right)^{1/\overline{p}} \cdot \|\delta_{xy}\|_p \\ &= \|\delta_{xy}\|_p. \end{aligned} \tag{167}$$

Assuming $\|\delta_{xy}\|_p \leq \frac{1}{4}$, f_p is monotonously increasing, and combining the inequalities above:

$$H(P_Y) - \breve{H}(\breve{P}_Y) \leq f_p\left(\|\delta_{xy}\|_p\right). \tag{168}$$

For the second part of (164), by the L_p bound:

$$\left|\breve{H}(\breve{W}(y|x)) - H(W(y|x))\right| \leq f_p(\|\delta_{xy}\|_{p,y}). \tag{169}$$

Using the concavity and monotonicity of f_p:

$$\begin{aligned} &\sum_x Q(x)\left|\breve{H}(\breve{W}(y|x)) - H(W(y|x))\right| \\ &\leq \sum_x Q(x) f_p(\|\delta_{xy}\|_{p,y}) \\ &\leq f_p\left(\sum_x Q(x)\|\delta_{xy}\|_{p,y}\right) \\ &\overset{(167)}{\leq} f_p(\|\delta_{xy}\|_p), \end{aligned} \tag{170}$$

where the monotonicity of f_p is again guaranteed by the condition $\|\delta_{xy}\|_p \leq \frac{1}{4}$. Plugging (168) and (170) into (164) yields:

$$\left|\breve{I}(Q, \breve{W}) - I(Q, W)\right| \leq 2f_p\left(\|\delta_{xy}\|_p\right). \tag{171}$$

which proves the bound on mutual information. The bound on capacity is trivially obtained from (171) above by writing $\breve{I}(Q, \breve{W}) \geq I(Q, W) - 2f_p\left(\|\delta_{xy}\|_p\right)$ and maximizing both sides with respect to Q (and similarly for the other direction). □

E. Proofs of small Lemmas

Proof of Lemma 1: It need to be shown that $1 + x \leq e^x \leq 1 + x + x^2$. Using a finite tailor series yields:

$$e^x = 1 + e^0 \cdot x + \tfrac{1}{2}e^t x^2, \tag{172}$$

where $t \in [0, x] \cup [x, 0]$ is a point between 0 and x. This proves the lower bound. Also, for $x \leq 0$ since $e^t \leq 1$ this also proves the upper bound. For $0 < x \leq 1$, the right inequality can be made tighter, by writing the full Tailor expansion:

$$\begin{aligned} e^x &= \sum_{m=0}^{\infty} \frac{1}{m!}x^m = 1 + x + \sum_{m=2}^{\infty} \frac{1}{m!}x^m \\ &\leq 1 + x + x^2 \sum_{m=2}^{\infty} \frac{1}{m!} = 1 + x + x^2(e^1 - 1 - 1) \\ &= 1 + x + (e-2)x^2 \leq 1 + x + x^2. \end{aligned} \tag{173}$$

□

Proof of Lemma 3: $f(t)$ is continuous and differentiable therefore $f'(t) = 0$ at the maximum. Derivation yields $f'(t) = \alpha a \cdot t^{\alpha-1} - \beta b \cdot t^{-\beta-1}$, and $f'(t) = 0$ yields the single solution t^* stated in the Lemma. This is a single maximum since $f'(t)$ is positive for $t < t^*$ and negative for $t > t^*$. □

F. Proof of Theorem 2: the optimality of averaged channel capacity

Below is a proof of Theorem 2, presented in §IV-A (regarding the optimality of $C(\overline{W})$). For a given sequence W_1^n, consider the "permutation" channel generated by uniformly selecting a random permutation Π of the indices $i = 1, \ldots, n$, rearranging the sequence W_1^n to a permuted sequence $T_i = W_{\pi_i}$, and applying the channel $\Pr(\mathbf{Y}|\mathbf{X}, \pi) = \prod_i T_i(Y_i|X_i)$ to the input (i.e. using the channels W_i in permuted order). Suppose there is a system achieving the rate $R(W_1^n) - \Delta$ with probability $1 - \delta$ and error probability ϵ. Since this

rate is fixed for all drawing of Π, the system can guarantee the rate $R(W_1^n) - \Delta$ a-priori (with probability $1-\delta$), and the rate-adaptive system can be converted to a fixed-rate system, delivering a message $\mathbf{m}$ of $n(R(W_1^n)-\Delta)$ bits, with probability of error at most $\epsilon+\delta$. Once the discussion is constrained to the permutation channel induced by the deterministic sequence W_1^n, this sequence can be assumed to be known to the transmitter and the receiver.

A standard application of Fano's inequality [23, Theorem 2.10.1] yields:

$$\begin{aligned} I(\mathbf{m};\mathbf{Y}) &= H(\mathbf{m}) - H(\mathbf{m}|\mathbf{Y}) \\ &\geq n(R(W_1^n)-\Delta)(1-(\epsilon+\delta)) - h_b(\epsilon+\delta). \end{aligned} \tag{174}$$

Rearranging and using $h_b(p) \leq 1$:

$$R(W_1^n) \leq \frac{\frac{1}{n}I(\mathbf{m};\mathbf{Y}) + \frac{1}{n}}{1-\epsilon-\delta} + \Delta. \tag{175}$$

The main part of the proof shows that approximately, $\frac{1}{n}I(\mathbf{m};\mathbf{Y}) \leq C(\overline{W})$. Note that because of feedback, X_i may be a function of $\mathbf{m}$ and $\mathbf{Y}^{i-1}$, and therefore $I(\mathbf{X}^n;\mathbf{Y}^n)$ does not give a tight bound on the rate. As noted in the outline presented in Section IV-A, if the channels T_i were selected from W_1^n with replacement, this result would be obvious, since feedback would not be helpful. In the permuted channel, a system with feedback can use past channel outputs to gain some knowledge about the future behavior of the channel. The point of the proof is to show that there is no considerable gain from this knowledge, and even a knowledge of the actual list of channels that were already picked does not change the mutual information considerably.

Denote by Π the random permutation and by π a specific instance of the permutation. Let us bound the mutual information as follows:

$$\begin{aligned} I(\mathbf{Y}^n;\mathbf{m}) &= \sum_{i=1}^{n} I(Y_i;\mathbf{m}|\mathbf{Y}^{i-1}) \\ &= \sum_{i=1}^{n} \left(H(Y_i|\mathbf{Y}^{i-1}) - H(Y_i|\mathbf{Y}^{i-1},\mathbf{m})\right) \\ &\overset{(a)}{\leq} \sum_{i=1}^{n} \left(H(Y_i) - H(Y_i|\mathbf{Y}^{i-1},\mathbf{m},\Pi^{i-1},X_i)\right) \\ &\overset{(b)}{=} \sum_{i=1}^{n} \left(H(Y_i) - H(Y_i|\Pi^{i-1},X_i)\right), \end{aligned} \tag{176}$$

where (a) is because conditioning reduces entropy (used twice), and (b) is since $\mathbf{Y}^{i-1},\mathbf{m} \leftrightarrow T^{i-1},X_i \leftrightarrow Y_i$ (in other words, π^{i-1},X_i gives all relevant information on Y_i). This can be seen from the functional dependence graph in Fig.5. Let Z_i be a random variable generated by passing X_i through the channel $\overline{W}$ (i.e. $\Pr(Z_1^n|X_1^n) = \prod_{i=1}^n \overline{W}(Z_i|X_i)$). In the following it is shown that $H(Y_i) \approx H(Z_i)$ and $H(Y_i|\Pi^{i-1},X_i) \approx H(Z_i|X_i)$.

Given Π^{i-1}, the channel law between X_i and Y_i is a random pick from the group of $n-i+1$ channels that are not included in $\{\Pi_j\}_{j=1}^{i-1}$:

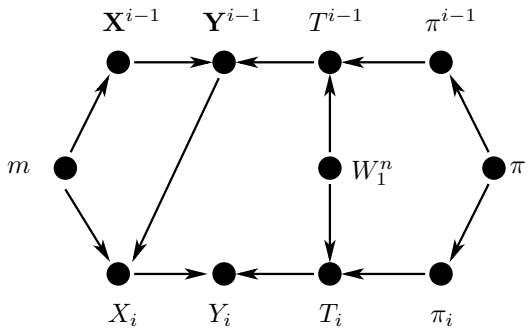

Fig. 5. A dependence graph for the variables of the permutation channel in Appendix F. Each node is a (potentially random) function of the nodes with arrows pointing toward it.

$$\begin{aligned} &\Pr(Y_i = y|\Pi^{i-1},X_i = x) \\ &= \sum_{k=1}^{n} \Pr(Y_i = y|\Pi^{i-1},\Pi_i = k, X_i = x) \\ &\quad \cdot \Pr(\Pi_i = k|\Pi^{i-1},X_i = x) \\ &= \sum_{k\notin\{\Pi^{i-1}\}} W_k(y|x)\cdot\frac{1}{n-i+1} \triangleq \overline{W}_{\Pi^{i-1}}(y|x). \end{aligned} \tag{177}$$

The average channel given the past indices $W_{\pi^{i-1}}(y|x)$ is an average of $n-i+1$ values $0 \leq W_k(y|x) \leq 1$. Note that the indices k belong to Π_i^n, so the notation may be confusing, but it is used to stress the causal dependence on Π^{i-1}.

Considering the random variable $\overline{W}_{\Pi^{i-1}}(y|x)$ generated by calculating this channel over all drawings of Π, the set $k \notin \{\Pi^{i-1}\}$ becomes a random set of $n-i+1$ distinct indices from $1,\ldots,n$, chosen uniformly from all such sets. $\overline{W}_{\Pi^{i-1}}(y|x)$ is an average of $n-i+1$ values $0 \leq W_k(y|x) \leq 1$, sampled uniformly without replacement from the set $\{W_k(y|x)\}_{k=1}^n$ (for any specific x,y). It was shown by Hoeffding [17, §6] that averages of variables sampled without replacement obey the same bounds [17, Theorem 1] with respect to the probability to deviate from their mean, as independent random variables. Specifically, applying Hoeffding's bounds (combining Theorem 1 with Section 6 in [17]), and using $\mathbb{E}\left[\overline{W}_{\Pi^{i-1}}(y|x)\right] = \overline{W}$, yields:

$$\Pr\{|\overline{W}_{\Pi^{i-1}}(y|x) - \overline{W}| \geq t\} \leq 2e^{-2(n-i+1)t^2}. \tag{178}$$

Using the union bound over all $|\mathcal{X}|\cdot|\mathcal{Y}|$ values of x,y (see the proof of Proposition 3), yields:

$$\Pr\{\|\overline{W}_{\Pi^{i-1}} - \overline{W}\|_\infty \geq t\} \leq 2|\mathcal{X}|\cdot|\mathcal{Y}|e^{-2(n-i+1)t^2}, \tag{179}$$

where the L_∞ norm is over x,y. To further simplify, pick a small value ϵ_0, and from now on, assume $i \leq (1-\epsilon_0)n$. Substituting in (179):

$$\Pr\{\|\overline{W}_{\Pi^{i-1}} - \overline{W}\|_\infty \geq t\} \leq 2|\mathcal{X}|\cdot|\mathcal{Y}|e^{-2\epsilon_0 n t^2} \triangleq p, \tag{180}$$

Since $H(\cdot)$ is uniformly continuous (see Lemma 12), for any ϵ_0 there is a t such that if $\|P_1(y) - P_2(y)\|_\infty \leq 2t$ then $|H(P_1) - H(P_2)| \leq \epsilon_0$. For a given ϵ_0 choose the value of

t such that this requirement is satisfied, so that together with (180) it holds that:

$$\forall x : \Pr\{|H(\overline{W}_{\pi^{i-1}}(\cdot|x)) - H(\overline{W})| \leq \epsilon_0\} \geq 1-p. \quad (181)$$

The following relation translates proximity in probability to proximity of the expected values: if $A, B \in [0, A_{\max}]$ are two random variables satisfying $\Pr\{|A-B| \leq \epsilon\} \geq 1-p$ (for some $\epsilon, p \in [0,1]$), then

$$\begin{aligned}\left|\mathbb{E}[A] - \mathbb{E}[B]\right| &= \Big|\mathbb{E}[(A-B)\cdot \mathrm{Ind}(|A-B| \leq \epsilon)] \\ &\qquad + \mathbb{E}[(A-B)\cdot \mathrm{Ind}(|A-B| > \epsilon)]\Big| \\ &\leq \mathbb{E}[|A-B|\cdot \mathrm{Ind}(|A-B| \leq \epsilon)] \\ &\qquad + \mathbb{E}[|A-B|\cdot \mathrm{Ind}(|A-B| > \epsilon)] \\ &\leq \epsilon + \mathbb{E}[A_{\max}\cdot \mathrm{Ind}(|A-B| > \epsilon)] \\ &\leq \epsilon + A_{\max}\cdot p. \end{aligned} \quad (182)$$

Applying this inequality to bound $H(Y_i|\Pi^{i-1}, X_i)$ yields:

$$\begin{aligned} H(Y_i|\Pi^{i-1}, X_i) &= \sum_{x,\pi} H(Y_i|\Pi^{i-1}=\pi^{i-1}, X_i = x) \\ &\qquad \cdot \Pr(\Pi = \pi, X_i = x) \\ &\overset{(177)}{=} \sum_{x,\pi} H(\overline{W}_{\pi^{i-1}}(\cdot|x))\cdot \Pr(\Pi=\pi, X_i = x) \\ &= \mathbb{E}\left[H(\overline{W}_{\Pi^{i-1}}(\cdot|X_i))\right] \\ &\overset{(181),(182)}{\geq} \mathbb{E}\left[H(\overline{W}(\cdot|X_i))\right] - \epsilon_0 - \log|\mathcal{Y}|\cdot p \\ &= H(Z_i|X_i) - \epsilon_0 - \log|\mathcal{Y}|\cdot p. \end{aligned} \quad (183)$$

In the following it is shown that the distributions of Y_i and Z_i are similar (note that they are not equal, due to the possible dependence of X_i on Π^{i-1}).

$$\begin{aligned} &|\Pr(Y_i = y) - \Pr(Z_i = y)| \\ &= \left|\mathbb{E}\left[\Pr(Y_i = y|\Pi^{i-1}, X_i)\right] - \mathbb{E}\left[\Pr(Z_i = y|X_i)\right]\right| \\ &\overset{(177)}{=} \left|\mathbb{E}\left[\overline{W}_{\Pi^{i-1}}(y|X_i)\right] - \mathbb{E}\left[\overline{W}(y|X_i)\right]\right| \overset{(180),(182)}{\leq} t + p. \end{aligned} \quad (184)$$

Since for any ϵ_0, t it holds that $p \underset{n\to\infty}{\longrightarrow} 0$ (180), n can be chosen large enough such that $p \leq t$ and yield $|\Pr(Y_i = y) - \Pr(Z_i = y)| \leq 2t$. Then, by the selection of t above (before (181)):

$$|H(Y_i) - H(Z_i)| \leq \epsilon_0. \quad (185)$$

Returning to (176), and treating the first $(1-\epsilon_0)n$ and the last $\epsilon_0 n$ symbols separately, yields:

$$\begin{aligned} I(\mathbf{Y}^n;\mathbf{m}) &\leq \sum_{i=1}^{n}\left(H(Y_i) - H(Y_i|\Pi^{i-1}, X_i)\right) \\ &\overset{(183),(185)}{\leq} \sum_{i=1}^{(1-\epsilon_0)n}\Big[(H(Z_i)+\epsilon_0) \\ &\qquad - (H(Z_i|X_i) - \epsilon_0 - \log|\mathcal{Y}|\cdot p)\Big] \\ &\qquad + \epsilon_0\cdot n\cdot\log|\mathcal{Y}| \\ &\leq \sum_{i=1}^{n} I(Z_i;X_i) + n\underbrace{(2\epsilon_0 + (\epsilon_0+p)\cdot\log|\mathcal{Y}|)}_{\delta_0} \\ &\leq n\cdot C(\overline{W}) + n\delta_0. \end{aligned} \quad (186)$$

Because ϵ_0 is a parameter of choice, and for any ϵ_0, t it holds that $p \underset{n\to\infty}{\longrightarrow} 0$ (180), δ_0 can be made as small as desired for n large enough. Returning to (175):

$$\begin{aligned} R(W_1^n) &\leq \frac{C(\overline{W}) + \delta_0 + 1/n}{(1-\epsilon-\delta)} + \Delta \\ &\leq (C(\overline{W}) + \delta_0 + 1/n)(1+\epsilon+\delta) + \Delta \\ &\leq C(\overline{W}) + \underbrace{(\delta_0 + 1/n)(1+\epsilon+\delta) + (\epsilon+\delta)I_{\max} + \Delta}_{\delta_1}, \end{aligned} \quad (187)$$

where $I_{\max}$ is defined in (19). Since by Definition 1, the above must hold for every ϵ, δ, Δ, for n large enough, and $\delta_0 \underset{n\to\infty}{\longrightarrow} 0$ (see (186) and the discussion following it), δ_1 can be made as small as desired by taking $n \to \infty$. This concludes the proof of Theorem 2. □

G. Proof of Lemma 5

Using the assumptions of Section IV-D, $F_i(Q) = I(Q, \overline{W}_i)$ satisfies the conditions of the lemma. The lemma assumes there are $B+1$ blocks and the rate is $\frac{KB}{n}$, which corresponds to a case where the last block was not decoded, however it holds as a lower bound even if the last block was decoded. Let us optimize the value of λ. Starting from (41):

$$\Delta^*_{\text{pred}}(\lambda) = \frac{K}{n} + I_{\max}\cdot\lambda + \underbrace{c_1\sqrt{\frac{\ln(n)}{n}}}_{b_2}\lambda^{-\frac{1}{2}}, \quad (188)$$

λ is determined using Lemma 3 (with $\alpha = 1, \beta = \frac{1}{2}$) which yields:

$$\begin{aligned}
\Delta^*_{\text{pred}}(\lambda^*) &= \left(\frac{\beta}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta}} \left[1 + \frac{\alpha}{\beta}\right] \cdot I_{\max}^{\frac{\beta}{\alpha+\beta}} \cdot b_2^{\frac{\alpha}{\alpha+\beta}} + \frac{K}{n} \\
&= 3 \cdot 2^{-\frac{2}{3}} \cdot I_{\max}^{\frac{1}{3}} \cdot \left(c_1 \sqrt{\frac{\ln(n)}{n}}\right)^{\frac{2}{3}} + \frac{K}{n} \\
&\overset{(42)}{=} 3 \left(K\mathcal{X}|(|\mathcal{X}|-1)\right)^{\frac{1}{3}} I_{\max}^{\frac{2}{3}} \cdot \left(\frac{\ln(n)}{n}\right)^{\frac{1}{3}} + \frac{K}{n} \\
&\leq \left(\frac{K}{n}\right)^{\frac{1}{3}} \cdot \left[3 \left(|\mathcal{X}| \cdot I_{\max}\right)^{\frac{2}{3}} \ln^{\frac{1}{3}}(n) + \left(\frac{K}{n}\right)^{\frac{2}{3}}\right] \\
&\overset{(a)}{\leq} \left(\frac{K}{n}\right)^{\frac{1}{3}} \cdot 4 \cdot \left(|\mathcal{X}| \cdot I_{\max}\right)^{\frac{2}{3}} \ln^{\frac{1}{3}}(n) \\
&= 4 \cdot K^{\frac{1}{3}} \cdot |\mathcal{X}|^{\frac{2}{3}} \cdot I_{\max}^{\frac{2}{3}} \cdot \left(\frac{\ln(n)}{n}\right)^{\frac{1}{3}} \triangleq \Delta_{\text{pred}},
\end{aligned} \tag{189}$$

where in (a) it was assumed that $K \leq |\mathcal{X}| \cdot n \cdot I_{\max}$. If the contrary is true, the first term in (188) yields $\Delta_{\text{pred}} > \frac{K}{n} > I_{\max}$ and the theorem is true in a void way. Similarly, one does not have to worry about the case $\lambda^* > 1$ since, also in this case, $\Delta_{\text{pred}} > I_{\max}$ due to the second term in (188).

If the conditions of Lemma 4 are satisfied, then for all Q (40):

$$\begin{aligned}
R &\geq \min\left(\sum_{i=1}^{B+1} \frac{m_i}{n} \cdot I(Q, \overline{W}_i) - \Delta_{\text{pred}}, I_{\max}\right) \\
&\geq I\left(Q, \sum_{i=1}^{B+1} \frac{m_i}{n} \overline{W}_i\right) - \Delta_{\text{pred}} \\
&= I\left(Q, \overline{W}\right) - \Delta_{\text{pred}},
\end{aligned} \tag{190}$$

where the convexity of $I(Q, W)$ with respect to the channel W was used. Maximizing both sides of (190) with respect to Q yields the desired result (47).

The conditions of Lemma 4 on n, K remain as conditions of the theorem. The application of Lemma 3 in (189) yields the following value of λ:

$$\begin{aligned}
\lambda^* &= \left(\frac{b_2 \beta}{I_{\max} \alpha}\right)^{\frac{1}{\alpha+\beta}} = \left(\frac{\frac{1}{2} c_1 \sqrt{\frac{\ln(n)}{n}}}{I_{\max}}\right)^{\frac{2}{3}} \\
&= \left(K \cdot |\mathcal{X}|(|\mathcal{X}|-1) \cdot I_{\max}^{-1} \cdot \frac{\ln(n)}{n}\right)^{\frac{1}{3}}.
\end{aligned} \tag{191}$$

This concludes the proof of Lemma 5. □

H. Channel knowledge compared to channel estimation

This section demonstrates the claim made in Section IV-A, that even imposing on the synthetic problem only the limitation that the past channels are not given, but need to be estimated, leads to the conclusion C_2 is not attainable.

This is shown by an example, based on randomization of the channel sequence. As in Section III, assume $I(\hat{Q}_i, W_i)$ bits are transmitted in time instance i (in other words, this is the gain obtained in retrospect for choosing $\hat{Q}_i$), however, instead of knowing the full channel sequence, the predictor is only allowed to base its decisions on measurements of the channel input and output, i.e. on the values of $(\mathbf{Y}_1^{i-1}, \mathbf{X}_1^{i-1})$ where Y_i is the result of W_i operated on X_i. It would make sense to also require that X_i be distributed $\hat{Q}_i(x)$ but this assumption is not required for the counter example.

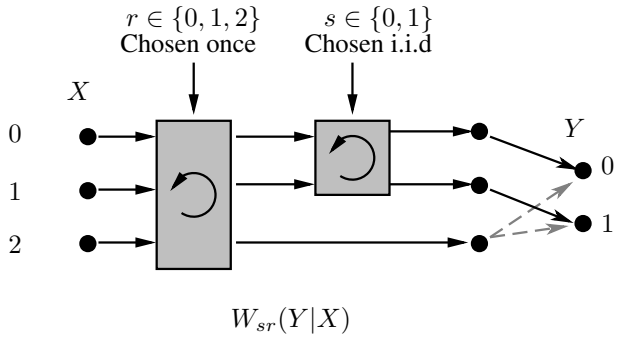

Fig. 6. An illustration of the generation of the channels W_{sr} in Example 4.

Example 4. Consider a ternary input binary output channel. The channel is chosen randomly, and the average gain of the predictor and the reference is considered (since the average regret is a lower bound for the maximum regret). The basic channels are $W_1 = \begin{bmatrix} \frac{1}{2} & 0 & 1 \\ \frac{1}{2} & 1 & 0 \end{bmatrix}$, $W_2 = \begin{bmatrix} \frac{1}{2} & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{bmatrix}$. Note that in the two channels, the first input is useless, and using only the two last inputs yields a rate of 1 bit/use. Now, add to this family of channels all 3 possible cyclic rotations of the inputs, and term the channel W_s^r $(s = 1, 2; r = 1, 2, 3)$. The resulting channels are depicted in Fig. 6. The sequence of channels is generated as follows: choose r randomly (one for the entire sequence), and choose a random (uniform, i.i.d.) sequence of s_i-s. The competitor, knowing r, easily selects a prior that optimizes $\sum_i I(Q, W_i)$, since W_1^r and W_2^r have the same optimizer for each r, and achieves a rate of 1. Because of the random generation of the sequence s_i, for any value of r, the channel output $\mathbf{Y}_1^{i-1}$ is uniform i.i.d. over $\{0, 1\}$ and independent of the input. Therefore the predictor cannot infer any information on r from the input-output distribution. Therefore the best the predictor can do (in terms of optimizing for the worst-case r), is place a uniform prior over all 3 inputs, and therefore obtain a rate of $\frac{2}{3}$, i.e. a regret of $\frac{1}{3}$ bit per channel use. By increasing the size of the channel input, this gap can be increased indefinitely.

The conclusion from the example is that C_2 cannot be attained universally when actual channel measurements are used.

I. An analysis of the prior quantization approach

In Section VI-G an alternative was mentioned, of using a "codebook" of priors, instead of the exponential weighting scheme over the continuum of priors used in this paper. Following is a rough analysis of this approach, for the block-wise variation setting. First, let us determine the accuracy

required of the codebook. Suppose there are two priors Q_1, Q_2 with $\|Q_1 - Q_2\|_\infty \le \Delta$, and for a certain channel W the resulting output distributions are P_1, P_2 respectively ($P_m = \sum_x Q_m(x)W(y|x), m = 1, 2$). Write $I(Q_m, W) = H(P_m) - \sum_x Q_m(x)H(W(\cdot|x))$ (output entropy minus output entropy given the input). Since by definition $\|P_1 - P_2\|_\infty \le |\mathcal{X}| \cdot \|Q_1 - Q_2\|_\infty$, by using Lemma 12, $|H(P_1) - H(P_2)| \le f_\infty(|\mathcal{X}| \cdot \Delta)$. Since the second term in $I(Q_m, W)$ may change by at most $\log|\mathcal{X}| \cdot \Delta$, then $|I(Q_1, W) - I(Q_2, W)| \le f_\infty(|\mathcal{X}| \cdot \Delta) + \log|\mathcal{X}| \cdot \Delta \triangleq \Delta_I$. Therefore, in order to bound the loss due to the codebook quantization to $\Delta_I = O\left(\frac{\ln n}{n}\right)^{\frac{1}{2}}$, it is required that $\Delta = O(n^{-\frac{1}{2}})$ (here, Q_1 represents the any prior, and Q_2 represents the closest point in the codebook). To have a density of $O(n^{-\frac{1}{2}})$ per dimension, $N = O\left(n^{\frac{1}{2}(|\mathcal{X}|-1)}\right)$ points are required. Now, since $\max_Q \frac{1}{n}\sum_{i=1}^n I(Q, W_i)$ differs from $\max_{m\in 1,\ldots,N} \frac{1}{n}\sum_{i=1}^n I(Q_m, W_i)$ by at most Δ_I, one can now consider the problem of competing against the N priors (considered as N experts). The best normalized redundancy than can be attained is $O\left(\sqrt{\frac{\ln N}{n}}\right) = O\left(\sqrt{\frac{\ln n}{n}}\right)$ (see the lower bound [13, Theorem 3.7] and the upper bound [13, Corollary 2.2] in Cesa-Bianchi and Lugosi's book). Note that since the predictor loss and the codebook loss are balanced, there is no potential gain from changing the codebook density. However, the bound on Δ_I is not necessarily tight.

J. Operation with any positive feedback rate

As mentioned, the scheme can be modified to operate with any positive feedback rate. Feedback is used in the scheme §IV-B for two purposes:

1) In order to report reception of a rateless block (using 1 bit per channel use)
2) In order to send the estimated averaged channel $\breve{W}_i$ after the end of each block (or alternatively, the next prior $\hat{Q}_{i+1}$).

Suppose feedback is limited to rate R_{FB}. Instead of reporting successful reception on each symbol, the scheme may report it each $N_1 = \lceil \frac{1}{R_{\mathrm{FB}}} \rceil$ symbols. The price would be wasting up to N_1 symbols per block, which essentially form an unused "gap" between successful decoding of block i and the start of block $i + 1$.

Here is a coarse bound on the number of bits required to represent the estimated averaged channel $\breve{W}_i$. $\breve{W}_i$ is completely specified to the transmitter by specifying the empirical distribution $\hat{P}_{\mathbf{x},\mathbf{y}}(x, y)$ which takes at most $(m + 1)^{|\mathcal{X}|\cdot|\mathcal{Y}|}$ values for a block of length m. Since $m \le n$, the number of bits is at most $N_2 = \log|\mathcal{X}| \cdot |\mathcal{Y}| \cdot \log(n + 1) = O(\ln n)$. These bits can be sent over $\frac{N_2}{R_{\mathrm{FB}}}$ channel uses at the end of each block, thus forming another unused "gap" between the blocks. Overall the gap between blocks is $N_1 + \frac{N_2}{R_{\mathrm{FB}}} = O\left(\frac{\log n}{R_{\mathrm{FB}}}\right)$. Since the maximum number of blocks grows sub-linearly in n, the overall loss can be made negligible.

Specifically, the effect of the additional gap on the rate can be analyzed using the same technique used to analyze the loss in the last symbol (the transition between (71) and (74)), and would effectively increase the term $\log\left(\frac{|\mathcal{X}|}{\lambda}\right)$ in δ_1 (66) by a factor of the gap $O(\log n)$. Since $K \in \omega(\log n)$ it is easy to see that under the same setting of of the parameters of the scheme, it still holds that $\delta_1 \underset{n\to\infty}{\longrightarrow} 0$ and $\Delta_C \underset{n\to\infty}{\longrightarrow} 0$, and nearly at the same convergence rate.

A delay in the feedback link would simply mean that an additional fixed gap will be added between the blocks, which also does not prevent asymptotical convergence.

K. Generation of the prior using rejection sampling

As mentioned, implementation of the prediction methods described in this paper, which are based on weighted average over the unit simplex, require the calculation of integrals. An alternative method is to generate the same results, using a method based on rejection sampling. Instead of explicitly calculating the predictor $\hat{Q}$, the algorithm generates a random variable $X \sim \hat{Q}$ (which can be used to generate a letter in the random codebook), based on multiple drawings of uniform random variables. The number of random drawings required in this algorithm is polynomial in n, but still prohibitively large, so unfortunately it is not practical.

First, any scalar random variable can be derived from a uniform $[0, 1]$ random variable by the inverse transform theorem. A generation of the mixture of an exponentially weighted and a uniform distribution such as in (35), only requires to toss a coin with probability λ, which determines whether X is generated using the exponentially weighted distribution or using a uniform distribution. Therefore the problem of generating the predictors described here (16), (35), boils down to the following problem: generate a random variable X distributed according to

$$\hat{Q} = \int w(Q)Q dQ, \tag{192}$$

where

$$w(Q) = \frac{e^{\eta g(Q)}}{\int_\Delta e^{\eta g(Q)} dQ}, \tag{193}$$

and where $g(Q)$ is a concave function and is bounded $0 \le g(Q) \le n \cdot g_0$. Δ is the unit simplex (which implicitly refers to the alphabet $\mathcal{X}$). This should be accomplished without computing any integrals.

All integrals below are over the unit simplex. The first observation is that instead of generating an X from $\hat{Q}$ it is enough to generate a the probability vector Q randomly with the probability distribution $w(Q)$ and then generate an X from the (specific) probability distribution Q. The last step can be accomplished using the inverse transform theorem. In this case:

$$\begin{aligned}\Pr(X = x) &= \mathop{\mathbb{E}}_{Q\sim w(Q)} [\Pr(X = x|Q)] \\ &= \mathop{\mathbb{E}}_{Q\sim w(Q)} [Q(x)] = \int Q(x)w(Q)dQ.\end{aligned} \tag{194}$$

There remains the problem of generating $Q \sim w(Q)$. This is accomplished by rejection sampling. I.e. by first generating a random variable with a different distribution, and if it does not satisfy a given condition, "rejecting" and re-generating it, until the condition is satisfied.

The first step is to generate a probability distribution P uniformly over the unit simplex Δ. There are several algorithms for uniform sampling over the unit simplex [35]. A simple algorithm, for example, is normalizing a vector of i.i.d. exponential random variables. Define $G(Q) = e^{\eta g(Q)}$, and $a(Q) = \alpha G(Q)$. α is to be determined later on, under the constraint $\forall Q : \alpha \cdot G(Q) \leq 1$. Having generated P, toss a coin with probability $a(P)$ for "accept". If P is accepted, this is the resulting random variable and $Q = P$. Otherwise, draw P again and repeat the process. Let A denote the event of acceptance, and f_P denote the distribution of P which is the uniform distribution over the simplex. The distribution of Q equals the distribution of P given that it was accepted. I.e.:

$$\begin{aligned} f_Q(q) = f_{P|A}(q) &= \frac{\Pr\{A|P=q\} \cdot f_P(q)}{\Pr\{A\}} \\ &= \frac{\Pr\{A|P=q\} \cdot f_P(q)}{\int \Pr\{A|P=q\} \cdot f_P(q) dq} = \frac{a(q) \cdot \frac{1}{\mathrm{vol}(\Delta)}}{\int a(q) \cdot \frac{1}{\mathrm{vol}(\Delta)} dq} \\ &= \frac{G(q)}{\int G(q) dq} = \frac{e^{\eta g(q)}}{\int e^{\eta g(q)} dq} = w(q), \end{aligned} \tag{195}$$

which is the desired distribution.

To determine α, suppose the maximum of $g(Q)$ is known. This is usually possible since it is a convex optimization problem. Even if this value is not known, a bound on this value is sufficient. Suppose that Q^* is the maximizer of $g(Q)$ and therefore also of $G(Q)$. Then it is enough to set $\alpha = \frac{1}{G(Q^*)} = e^{-\eta g(Q^*)}$.

An important question from implementation perspective is the average number of iterations required. Since the probability of acceptance $\Pr\{A\}$ in each iteration is fixed, the number of iterations is a geometrical random variable, with mean $\overline{N} = \frac{1}{\Pr\{A\}}$. By Lemma 2 one can relate $G(Q^*)$ to $\mathbb{E}G(Q)$ and bound the average number of iterations. Using the lemma:

$$\begin{aligned} g(Q^*) &\leq \frac{1}{\eta} \ln \left[\frac{\int e^{\eta g(Q)} dQ}{\mathrm{vol}(\Delta)} \right] + \frac{d}{\eta} \ln \left(\frac{\eta e n g_0}{d} \right) \\ &\leq \frac{1}{\eta} \ln \left(\mathbb{E}\left[G(P) \right] \right) + \frac{d}{\eta} \ln \left(\frac{\eta e n g_0}{d} \right), \end{aligned} \tag{196}$$

where $d = |\mathcal{X}| - 1$ is the dimension of the unit simplex. The following bound on α is obtained:

$$\alpha = e^{-\eta g(Q^*)} \geq \frac{1}{\mathbb{E}\left[G(P) \right]} \cdot \left(\frac{\eta e n g_0}{d} \right)^{-d}, \tag{197}$$

and the average number of iterations can be bounded:

$$\begin{aligned} \overline{N} &= \frac{1}{\Pr\{A\}} = \frac{1}{\mathbb{E}\left[\Pr\{A|P\} \right]} \\ &= \frac{1}{\mathbb{E}\left[a(P) \right]} = \frac{1}{\alpha \mathbb{E}\left[G(P) \right]} \leq \left(\frac{\eta e n g_0}{d} \right)^{d}. \end{aligned} \tag{198}$$

Since η is polynomial in n and tends to 0, $\overline{N}$ grows slower than n^d, however this number is still prohibitively large.

The algorithm described is summarized in Table II.

Generation of a random variable $X \sim \hat{Q}$**,** (192)**,** (193)
1) Compute the maximum of $g(Q)$ (a convex optimization problem), or a bound on it. 2) Set $\alpha \leq e^{-\eta \max_Q g(Q)}$. 3) Draw Q uniformly over the unit simplex [35]. 4) Toss a coin and with probability $1 - \alpha e^{\eta g(Q)}$ return to step 3. 5) Draw X randomly according to the distribution $Q(x)$.

TABLE II
AN ALGORITHM TO GENERATE $X \sim \hat{Q}$

L. Why "follow the leader" fails

As noted in Section III-B, the relation of the synthetic prediction problem to prediction under the absolute loss function, implies that the FL predictor cannot be applied to the "toy" problem presented here. Below is a specific example to show why FL fails, based on the channel defined in Section III-B. Construct the following sequence of channels: the channel at $i = 1$ is a mixture of W_0 with probability $\frac{1}{2}$ and a completely noisy channel $Y = \mathrm{Ber}\left(\frac{1}{2}\right)$. For this channel $I(Q, W) = \frac{1}{2} I(Q, W_0)$. At time $i = 2$, the best a-posteriori strategy is $q = 0$. The sequence of channels from time $i = 2$ onward is the alternating sequence $(W_1, W_0, W_1, W_0, \ldots)$. It is easy to see that the resulting cumulative rates are linear functions of q and thus the optimum is attained at the boundaries of $[0, 1]$ and $q_i = (0, 1, 0, 1, \ldots)$. At each time, since the channel that slightly dominates the past is opposite of the channel that is about to appear, the FL predictor chooses the prior that yields the *least* mutual information, and ends up having a zero rate in time instances $i = 2, \ldots, n$. On the other hand, by using a uniform fixed prior, a competitor may achieve an average rate of $\frac{1}{2}$ over these symbols. Therefore the normalized regret of FL would be at least $\frac{1}{2}$, and does not vanish asymptotically.

The problem with the FL predictor is that it takes a decision based on a slight inclination of the cumulative rate toward one of the extremes.

Note that for $|\mathcal{X}| = 4, |\mathcal{Y}| = 2$, $I(Q, W)$ does not satisfy the Lipschitz condition required in [36, Theorem 1] for this strategy to work.

REFERENCES

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

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

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

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

[5] ——. (2010, Dec.) Universal communication over modulo-additive channels with an individual noise sequence. arXiv:1012.2751v1 [cs.IT]. [Online]. Available: http://arxiv.org/abs/1012.2751

[6] C. E. Shannon, "A mathematical theory of communication," *The Bell System technical journal*, vol. 27, pp. 379–423, 1948.

[7] P. Chow, J. Cioffi, and J. Bingham, "A practical discrete multitone transceiver loading algorithm for data transmission over spectrally shaped channels," *IEEE Trans. Communications*, vol. 43, no. 234, pp. 773 –775, Apr. 1995.

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

[9] A. Mahajan and S. Tatikonda, "A training based scheme for communicating over unknown channels with feedback," in *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*, 30 2009-oct. 2 2009, pp. 1549 –1553.

[10] Y. Lomnitz and M. Feder, "Prediction of priors for communication over arbitrarily varying channels," in *IEEE Int. Symp. Information Theory (ISIT)*, Jul. 2011, pp. 219 –223.

[11] E. Biglieri, J. Proakis, and S. S. (shitz), "Fading channels: Information-theoretic and communications aspects," *IEEE Trans. Information Theory*, vol. 44, pp. 2619–2692, 1998.

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

[13] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning and games*. Cambridge University Press, 2006.

[14] D. Haussler, J. Kivinen, and M. K. Warmuth, "Sequential prediction of individual sequences under general loss functions," *IEEE Trans. Information Theory*, vol. 44, no. 5, Sep. 1998.

[15] V. Vovk, "A game of prediction with expert advice," *Journal of Computer and System Sciences*, vol. 56, pp. 153–173, 1997.

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

[17] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, Mar. 1963.

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

[19] N. Shulman and M. Feder, "The uniform distribution as a universal prior," *IEEE Trans. Information Theory*, vol. 50, no. 6, pp. 1356–1362, Jun. 2004.

[20] Y. Lomnitz and M. Feder. (2011, Jan.) Universal prior prediction for communication. arXiv:1102.0710v2 [cs.IT]. [Online]. Available: http://arxiv.org/abs/1102.0710v2

[21] Y. Lomnitz, "Universal communication over unknown channels with feedback," Ph.D. dissertation, Tel Aviv University, 2012, to be avaible online http://www.eng.tau.ac.il/~yuvall/publications/YuvalL_Phd_report.pdf.

[22] I. Csiszár, "The method of types [information theory]," *IEEE Trans. Information Theory*, vol. 44, no. 6, pp. 2505–2523, Oct. 1998.

[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & sons, 1991.

[24] W. Foundation. Wikipedia: The free encyclopedia. [Online]. Available: http://www.wikipedia.org

[25] T. Hayes. (2003, Feb.) A large-deviation inequality for vector-valued martingales. [Online]. Available: http://www.cs.unm.edu/~hayes/papers/VectorAzuma/

[26] Y. Lomnitz and M. Feder, "Universal communication over channels with memory," in preparation.

[27] I. Csiszár and P. Narayan, "The capacity of the arbitrarily varying channel revisited : Positivity, constraints," *IEEE Trans. Information Theory*, vol. 34, no. 2, Mar. 1988.

[28] E. R. Berlekamp, "Block coding for the binary symmetric channel with noiseless, delayless feedback," *Error-Correcting Codes, edited by H.B.*, 1968.

[29] R. Ahlswede, "Channels with arbitrarily varying channel probability functions in the presence of noiseless feedback," *Z. Wahrscheinlichkeitstheorie und verw. Geb.*, vol. 25, 1973.

[30] R. Ahleswede and N. Cai, "The AVC with noiseless feedback and maximal error probability: A capacity formula with a trichotomy," *Numbers, Information and Complexity*, pp. 151–176, 2000, special volume in honour of R. Ahlswede on occasion of his 60th birthday.

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

[32] Y. Lomnitz and M. Feder. (2012, Jan.) Universal communication over channels with memory. arXiv:1202.0417 [cs.IT]. [Online]. Available: http://arxiv.org/abs/1202.0417

[33] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *ICML*, 2003, pp. 928–936.

[34] N. Buchbinder, L. Lewin-Eytan, I. Menache, J. Naor, and A. Orda, "Dynamic power allocation under arbitrary varying channels - an online approach," in *INFOCOM 2009, IEEE*, Apr. 2009, pp. 145 –153.

[35] S. Onn and I. Weissman, "Generating uniform random vectors over a simplex with implications to the volume of a certain polytope and to multivariate extremes," *Annals of Operations Research*, vol. 189, pp. 331–342, 2011. [Online]. Available: http://dx.doi.org/10.1007/s10479-009-0567-7

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