Prediction of Priors for Communication over Arbitrarily Varying Channels

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

***Abstract*—We consider the problem of communicating over an unknown and arbitrarily varying channel, using feedback. This paper focuses on the problem of determining the input behavior, or more specifically, a prior which is used to randomly generate a codebook. We pose the problem of setting the prior as a universal sequential prediction problem using information theoretic abstractions of the communication channel. For the case where the channel is block-wise constant, we show it is possible to asymptotically approach the best rate that can be attained by any system using a fixed prior. For the case where the channel may change on each symbol, we combine a rateless coding scheme with a prior predictor and asymptotically approach the capacity of the average channel universally for every sequence of channels.**

I. Introduction

We consider the problem of communicating over an unknown and arbitrarily varying channel, with the help of feedback. We would like 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 those of a compound channel, in which the channel law is selected arbitrarily out of a family of known channels, and an 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 and 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, since for example it may happen that all W_i have zero capacity, and therefore the AVC capacity is zero (and may be non-zero only if a constraint on W_i is defined).

Other communication models, which allow positive communication rates over such AVC-s were proposed by us and other authors [2], [3], [4], [5]. Although the channel models these papers consider are different, the common features distinguishing them from the traditional AVC setting are that the communication rate is adaptively modified by using feedback, and that 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. Clearly, with this limitation these systems are incapable of universally attaining the channel capacity in many cases of interest. In a crude way we may say that the aforementioned 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 and rate adaptation. However to obtain more than the "mutual information", i.e. the "capacity", one would need to select the prior in a universal way.

Prior adaptation using feedback is well known for static or semi-static channels. Two familiar examples are bit and power loading performed in Digital Subscriber Lines (DSL-s) [6], and precoding for in multi-antenna systems [7] which is performed in practice in wireless standards such as WiFi, WiMAX and LTE. The idea is that 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. All known models for prior adaptation 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 that we deal with in this paper is: assuming a channel which is *arbitrarily* changing over time, is there any merit in using feedback based prior adaptation, and what rates can be guaranteed?

To answer the question we adopt an abstract model of the communication system. In addition to assuming perfect feedback, we make two assumptions which are only approximately true:

1) Given a prior Q, the mean mutual information between the channel input and output $\frac{1}{n}I(\mathbf{X}^n;\mathbf{Y}^n)$ is an achievable rate.
2) Given a large enough number of channel inputs and outputs, the average channel $W(y|x)=\frac{1}{n}\sum_{i=1}^{n}\Pr(y_i|x_i)$ can be perfectly known at the receiver (and fed back to the transmitter).

Under this abstract model we consider two scenarios, one in which the channel is changing in a block-wise manner, and one in which the channel is changing on each symbol individually. For each model we define attainable rates which have a competitive interpretation, and using tools from the theory of universal prediction, present prediction systems that attain these rates universally. The first model, which is rather artificial, is used mainly as a tool to gain insight into the problem. The attainable rate in this model is the maximum over the prior of the block-averaged mutual information (i.e. the best rate that can be attained by any system using a fixed prior). For the second model, the attainable rate is the capacity of the time-averaged channel (which is a bound on the rate achievable with per-symbol operation). Although we do not present and analyze the full communication system, it is reasonable to assume that by applying these methods, such a system can be devised, and provide improved results over the previous ones [5]. Due to space limits, the details and proofs will appear in a full paper on the subject [8], whose draft is currently available on the web.

The paper is organized as follows: Section II deals with the problem of prior prediction for a block-wise arbitrarily varying channel, Section III deals with the problem of prior prediction for a symbol-wise varying channel, and Section IV is devoted to discussion and extensions.

II. Block-wise arbitrary channel variation

A. Problem statement

Let $\{W_i\}$ be a sequence of memoryless channels, defined through conditional distributions $W_i(y|x)$ where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ represent an input and output symbol respectively. Except when specifically noted, we assume $\mathcal{X}$ is a finite alphabet, and $\mathcal{Y}$ may be discrete or finite.

We assume the channel is changing arbitrarily over blocks $i = 1, \ldots, n$. Each block contains a large number of channel uses $j = 1, \ldots, N$, in each of which the same memoryless channel law W_i applies.

Let $\{Q_i\}$ be a sequence of priors. We assume that the receiver knows W_i during block i, and that this information can be fed bad to the transmitter at the end of the i-th block. So the sequence of past channels is known at the transmitter and receiver, and can be used to determine the prior for the next block. A predictor $\hat{Q}_i$ is a function $\hat{Q}_i(W_1^{i-1})$ which determines Q_i as a function of the past channels.

We assume that the following rate is be achievable:

$$R = \frac{1}{n}\sum_{i=1}^{n} I(Q_i, W_i) \tag{1}$$

where $I(Q, W)$ denotes the mutual information with a prior Q and a channel W. We would like to characterize rates R as a function of the sequence of channels $\{W_i\}$ which have an operational or competitive meaning, and which can be achieved universally for every $\{W_i\}$, by a scheme sequentially determining Q_i as a function of W_1^{i-1}.

B. Potential target rates

With respect to the sequence $\{W_i\}$ we can define various meaningful information theoretic measures which result from optimizing rate (1) with respect to the priors. The maximum rate is the capacity when the sequence is known a-priori:

$$C_1(W_1^n) = \max_{\{Q_i\}:\left(\frac{1}{n}\sum Q_i\right)\in\mathcal{Q}} \frac{1}{n}\sum_{i=1}^{n} I(Q_i, W_i) \tag{2}$$

In fading channels this value is termed the "water pouring" capacity (water pouring in time [9]), where it is required to meet the constraint only on average. A lower target is the mean of the individual capacities:

$$C_2(W_1^n) = \frac{1}{n}\sum_{i=1}^{n} C(W_i) = \frac{1}{n}\sum_{i=1}^{n} \max_{Q\in\mathcal{Q}} I(Q, W_i) \tag{3}$$

In a fading channel this would mean constraining to an equal power in time. The maximum rate that can be obtained with a single *fixed* prior when the sequence is known, or alternatively the maximum rate that can be attained when only the sequence distribution is known (i.e. it is known up to order) is:

$$C_3(W_1^n) = \max_{Q\in\mathcal{Q}} \frac{1}{n}\sum_{i=1}^{n} I\left(P, W_i\right) \tag{4}$$

Lastly, the capacity of the averaged channel is:

$$C_4(W_1^n) = \max_{Q\in\mathcal{Q}} I\left(P, \frac{1}{n}\sum_{i=1}^{n} W_i\right) \tag{5}$$

C_4 is an upper bound on the achievable rate of a system operating symbol-by-symbol (since this system effectively sees the averaged channel, see the definition of the collapsed channel [5]).

Clearly, $C_1 \geq C_2 \geq C_3 \geq C_4$ where the first three inequalities result from the order of maximization and the last one results from the convexity of the mutual information with respect to the channel. If there are no constraints then $C_1 = C_2$.

The question we can ask regarding each of these targets $C(W_1^n)$, is: does there exist a predictor $\hat{Q}_i(W_1^{i-1})$ such that for every sequence we have:

$$\forall\{W_i\} : R = \frac{1}{n}\sum_{i=1}^{n} I\left(\hat{Q}_i(W_1^{i-1}), W_i\right) \geq C(W_1^n) - \delta_n$$

With $\delta_n \to 0$? If such a predictor exists, we say that the target is universally attainable.

C. Categorization of the problem

The target rate C_3 is special in being an additive function for each value of Q. Universally attaining C_3 falls into a widely studied category of universal prediction problems [10], [11], [12]. In the full paper [8] we show that C_2 is not in general universally attainable, and that the regret δ_n with respect to C_3 is at least $O(\sqrt{n})$. Therefore C_3 is a reasonable target for universal prediction. Furthermore, in the case of C_3, the simple predictor defined by following the prior which optimizes the a-posteriori rate $\hat{Q}_i = \operatorname*{argmax}_Q \sum_{t=1}^{i-1} I\left(Q, W_t\right)$ does yield a vanishing regret.

D. A prediction algorithm

The prediction algorithm we propose is based on a well known technique of a weighted average predictor, using exponential weighting [11, Secion 2.1]. The main novelty is the extension to a continuous set of references.

We assume the input alphabet is discrete. Let $\Delta_{|\mathcal{X}|}$ be the unit simplex $\Delta_{|\mathcal{X}|} \triangleq \{\mathbf{q} : \sum_{i=1}^{|\mathcal{X}|} q_i = 1\}$. The constraint set $\mathcal{Q}$ is a subset of $\Delta_{|\mathcal{X}|}$ and we assume that this subset is convex. A weight function $w(Q)$ is any non-negative function $w : \mathcal{Q} \to \mathbb{R}^+$ with $\int_{\mathcal{Q}} w(Q) dQ = 1$. All integrals in the sequel are by default over $\mathcal{Q}$.

Define the following weight function:

$$w_i(Q) = \frac{e^{\eta \sum_{t=1}^{i-1} I(Q,W_t)}}{\int_{\mathcal{Q}} e^{\eta \sum_{t=1}^{i-1} I(Q,W_t)} dQ} \tag{6}$$

and the predictor:

$$\hat{Q}_i = \int_{\mathcal{Q}} Q \cdot w_i(Q) \cdot dQ \tag{7}$$

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. We define the regret as

$$\mathcal{R}_n(Q) = \sum_{i=1}^{n} I(Q, W_i) - \sum_{i=1}^{n} I(\hat{Q}_i, W_i) \tag{8}$$

The following theorem gives a bound on the regret of this predictor, which is proven in the full paper [8].

Theorem 1. *Let $\mathcal{Q} \subset \Delta_{|\mathcal{X}|}$ be a convex subset of the unit simplex defined over the input alphabet size $|\mathcal{X}|$, and $I(Q,W), Q \in \mathcal{Q}$ be any bounded function $0 \leq I(Q,W) \leq I_{\max}$ which is concave in its first argument. Then for any $n \geq 3$, the predictor defined by* (6) *and* (7) *with $\eta = \sqrt{\frac{|\mathcal{X}| \ln n}{n}} \cdot I_{\max}$ satisfies the constraint $\hat{Q}_i \in \mathcal{Q}$ and yields a regret* (8) *bounded by*

$$\forall Q : \mathcal{R}_n(Q) \leq 3 I_{\max} \cdot \sqrt{\dim(\mathcal{Q}) \cdot n \ln n} \tag{9}$$

Where $\dim(\mathcal{Q}) \leq |\mathcal{X}| - 1$ is the dimension of the set $\mathcal{Q}$.[1]

Note that the theorem applies to more general gain functions than the mutual information, since it uses only the properties of concavity and boundness. In the case of mutual information we have $I_{\max} = \log \max(|\mathcal{X}|, |\mathcal{Y}|)$.

Dividing (9) by n we obtain a convergence rate of $O\left(\sqrt{\frac{\ln n}{n}}\right)$ of the normalized regret, which is slightly worse than the asymptotic bound of $O\left(\sqrt{\frac{1}{n}}\right)$ from Section II-C. The additional $\sqrt{\ln n}$ may be attributed to the fact the space of reference predictors is continuous, but we did not prove that this is the best convergence rate.

[1]We define a dimension of a convex set S to be the dimension of the smallest affine set containing S. Loosely speaking, this is the number of parameters required to specify a point in S.

E. Performance analysis

In this section we introduce the exponential weighting concept and present the proof outline of Theorem 1.

Define the instantaneous regret $r_i(Q)$ and the cumulative regret $\mathcal{R}_i(Q)$ as functions of Q:

$$r_i(Q) = I(Q, W_i) - I(\hat{Q}_i, W_i) \tag{10}$$

$$\mathcal{R}_i(Q) = \sum_{t=1}^{i} r_i(Q) = \sum_{i=1}^{n} I(Q, W_i) - \sum_{i=1}^{n} I(\hat{Q}_i, W_i) \tag{11}$$

These functions express the regret with respect to a competing fixed prior Q. We sometimes omit the dependence on Q for brevity. For $\eta > 0$ of our choice, we define the following potential function:

$$\Phi(u) = \int_{\mathcal{Q}} e^{\eta u(Q)} dQ \tag{12}$$

where $u : \mathcal{Q} \to \mathbb{R}$ is an arbitrary function defined over the unit simplex. Note that for large values of $\eta \cdot u$, $\Phi(u)$ approximates $\max_Q(u)$. Following the ideas of weighted average predictors using potential functions, 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, based on the fact that the growth occurs in a direction orthogonal to the gradient of this function with respect to $\mathcal{R}_i(Q)$.
2) Relating $\max_Q \mathcal{R}_n(Q)$ to $\Phi(\mathcal{R}_n(Q))$

The techniques we use are based on Cesa-Bianchi and Lugosi's book [11] (see Theorem 2.1, Corollary 2.2, Theorem 3.3). The proof is given in the full paper [8].

III. Symbol-wise arbitrary channel variation

In this section we define a more realistic problem, where the channel may change arbitrarily every symbol. We show that under this scenario we can only obtain the target rate C_4, and present an iterative-rateless coding scheme, which under the abstractions used in this paper, achieves the target rate with an asymptotically vanishing regret.

A. Problem setting

We assume that there are n channel uses $i = 1, 2, \ldots, n$ (not blocks, as in the previous case), and the channel in symbol i is $W_i(y|x)$ The sequence of channels W_i arbitrary and unknown to the predictor. Let $\overline{W}_{[i,i+N-1]}$ be the averaged channel over the segment $\{i, i+1, \ldots, i+N-1\}$, i.e.

$$\overline{W}_{[i,i+N-1]} = \frac{1}{N} \sum_{t=i}^{i+N-1} W_i(y|x) \tag{13}$$

We assume that, if all input symbols $x \in \mathcal{X}$ are transmitted with non zero probability, and N is large enough, then assuming the receiver knows the transmitted signal $\mathbf{x}$ (e.g. after decoding, or by using known symbols), the averaged channel could be perfectly known by the receiver. However, clearly, it is not possible to measure the channel over a single use. When the channel is known at the receiver it can be fed back to the transmitter. We also assume that it is possible to transmit with an i.i.d. input distribution $Q(x)$ over a large enough segment $\{i, i+1, \ldots, i+N-1\}$, and achieve a rate of $R = I(Q, \overline{W}_{[i,i+N-1]})$. As opposed to Section II-A, we make the scenario more realistic by not assuming the transmitter knows R in advance. For the sake of simplicity we assume that there is no constraint on the input, i.e. $\mathcal{Q} = \Delta_{|\mathcal{X}|}$. See comments about the validity of these assumptions in the full paper [8].

Since in this scenario, we are not constrained to use specific encoding blocks, we need to determine the coding blocks and the times that the transmitted signal is known, and feedback is conveyed to the transmitter. Under these assumptions, we would like to construct a coding scheme (in the sense of priors and code blocks) and a prediction scheme that will universally approach one of the target rates defined in Section II-B. In the full paper [8] we present an example showing why C_3 cannot be universally attained in this scenarios, and therefore our target rate is C_4.

B. A rateless coding scheme

In this section we propose an outline of a coding scheme, and pose the resulting prediction problem. One of the problems is the determination of the rate R before knowing the channel. To solve this problem we suggest using rateless codes [13]. We send K bits on each block. A codebook of $\exp(K)$ infinite sequences is generated, and the sequence representing the message is transmitted symbol by symbol, until the receiver decides to decode, and informs the transmitter that the block ended. This means that when the channel is good, the block will become shorter, and vice versa. We divide the time into multiple such blocks as done in [3], [4].

We choose to use an i.i.d. prior during each block, and update the prior only at the end of the block. This choice is motivated by the following considerations:

- Varying the prior throughout the block creates complex relations between the past channel input and output values x, y and the future values of x, and inserts memory which complicates the analysis.
- Assuming that no constant symbols (pilots) 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).

The high level scheme we propose is as follows:

1) The transmitter sends blocks of K bits to the receiver
2) Each block i is transmitted using the i.i.d. prior $\hat{Q}_i$, which is computed by a prior predictor that will be defined later on.
3) The receiver decides when the block terminates, by estimating when there is enough information from the channel output to reliably decode the bits.
4) At the end of the block, the receiver estimates the averaged channel over the block, and informs the transmitter through the feedback link that the block has ended, as well as the estimated averaged channel.
5) Both sides compute, based on the sequence of previously measured averaged channels, a prior $\hat{Q}_i$ to be used for the next block.

We 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 we denote by $\overline{W}_i \triangleq \overline{W}_{[k_i, k_{i+1}-1]}$ the average channel over the block, and by $\hat{Q}_i$ the (i.i.d.) prior used. Under the abstraction, the length of the i-th block is:

$$m_i = \frac{K}{I(\hat{Q}_i, \overline{W}_i)} \tag{14}$$

where K is the number of bits.

C. A prediction algorithm

One of the issues in the rateless scheme is 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 we must avoid a situation where the channel has non-zero capacity (which our competition enjoys), while a badly chosen prior yields $I(\hat{Q}_i, \overline{W}_i) = 0$. 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 we construct the predictor as a mixture between an exponentially weighted predictor and a uniform prior. We use a result by Shulman and Feder [14], which is a bound on the loss that the uniform prior experiences with respect to the optimal prior. 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.

Denote by i the block index, and by m_i the block length. We define $t_i = \sum_{i=1}^{j} m_j$ as the time at the end of the i-th block. $\overline{W}_i$ is the averaged channel over block i, and $\overline{W}^i$ is the averaged channel from the beginning of transmission until the end of block i. Suppose that at time n, B blocks have been sent (and the $B+1$-th block is under transmission), then the regret at time n is:

$$\mathcal{R}_n(Q) = n \cdot I(Q, \overline{W}_{[1,n]}) - K \cdot B \tag{15}$$

The regret includes the loss from the not decoding the last block which started before time n. We use an exponentially weighted predictor mixed with a uniform prior. The weight function is defined as

$$w_i(Q) = c \cdot e^{\eta t_i \cdot I(Q, \overline{W}^i)} \tag{16}$$

where c is a constant normalizing to $\int w_i(Q) dQ = 1$. Let $U = \frac{1}{|\mathcal{X}|}\mathbf{1}$ be the uniform prior over $\mathcal{X}$. Then the predictor is defined as:

$$\hat{Q}_i = (1-\lambda) \int_{\Delta_{|\mathcal{X}|}} w_i(Q) Q dQ + \lambda U \tag{17}$$

The parameters λ, η and K will be chosen later on.

The following theorem states a bound on the regret of this predictor, which is proven in the full paper [8].

Theorem 2. *Let $I(Q, W)$ denote the mutual information with prior Q and channel W, where the input alphabet $\mathcal{X}$ and the output alphabet $\mathcal{Y}$ are finite. Consider the predictor defined by*

(17)*, in conjunction with the rateless communication scheme defined in Section III-B. Then for a suitable choice of the parameters* K, η, λ *as functions of* n *the regret* (15) *satisfies:*

$$\forall Q : \mathcal{R}_n(Q) \le r_0 \cdot n \cdot \left(\frac{\ln n}{n}\right)^{\frac{1}{4}} \tag{18}$$

for any $n \ge 3$*, where* r_0 *is constant in* n*.*

The values of K, η, λ and r_0 are specified in the full paper [8]. In other words, the normalized regret is bounded by $O\left(\frac{\ln n}{n}\right)^{\frac{1}{4}}$, which converges to zero but at a worse rate, by a square root, than we had in Section II-D.

IV. Discussion

The scheme we have proposed in Section III is based on an abstraction of the communication channel. To make it an actual communication scheme one may use a rateless scheme similar to the one proposed in [4]. However the predictor needs to be adapted to deal with overheads of the rateless scheme in achieving the mutual information (i.e. excess block length compared to (14)), as well as estimation errors of the averaged channel.

In a previous paper [5] we have obtained a result that over the modulo-additive channel with an unknown noise sequence, it is possible to attain universally rates comparable to those obtained by any fixed length encoder and decoder operating over sub-blocks. This result relies on the fact that the capacity achieving prior is fixed for any noise sequence. The current work is a step toward removing this assumption, since the capacity of the averaged channel C_4 is a bound on the rate that can be obtained reliably by an encoder and decoder operating on a single symbol (see the "collapsed channel" [5]). By combining k symbols into a single super-symbol, we can extend the result to obtaining a rate which is better than the rate obtained by block encoder and decoder operating over chunks of k symbols. Therefore the current result suggests that it is possible to compete with finite block systems universally for all vector channels that are memoryless in the input, i.e. that have the form $\Pr(\mathbf{Y}_1^n|\mathbf{X}_1^n) = \prod_{i=1}^n W_i(Y_i|X_i)$, for an arbitrary sequence of channels W_i (compared to an arbitrary noise sequence, in the previous result).

It is interesting to compare the current results with the AVC capacity. The discrete memoryless AVC capacity without constraints [1, Theorem 2] 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 $y = 1 - x$). Then the randomized code capacity of the AVC is:

$$\begin{aligned} C_{AVC} &= \max_Q \min_{W \in \mathrm{conv}(\mathcal{W})} I(Q, W) \\ &= \min_{W \in \mathrm{conv}(\mathcal{W})} \max_Q I(Q, W) = \min_{W \in \mathrm{conv}(\mathcal{W})} C(W) \end{aligned} \tag{19}$$

where $\mathrm{conv}(\mathcal{W})$ is the convex hull of $\mathcal{W}$, which represents all channels which are realizable by a random selection of the channel state (in the example, $\mathrm{conv}(\mathcal{W})$ is the set of all binary symmetric channels). In the current work, the target rate is the capacity of the averaged channel $C(\overline{W})$. Since by definition $\overline{W} \in \mathrm{conv}(\mathcal{W})$, we have $C(\overline{W}) \ge C_{AVC}$. What we possibly gain is that the rate depends on the actual occurrence of $\overline{W}$, rather than on the worst case. This is especially important when $C_{AVC} = 0$, i.e. we cannot a-priori preclude the possibility of having zero capacity. In this case, by adaptation we may have $C(\overline{W}) > 0$, depending on the actual channel occurrence.

Acknowledgment

We would like to thank Prof. Yishay Mansour for helpful discussions on this problem.

References

[1] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Transactions on 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 Transactions on 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 Transactions on Information Theory*, vol. 58, no. 1, Jan. 2010.

[4] Y. Lomnitz and M. Feder, "Communication over individual channels," *Submitted to IEEE Transactions on Information Theory, arXiv:0901.1473v2 [cs.IT]*, Oct. 2009.

[5] ——, "Universal communication over modulo-additive channels with an individual noise sequence," *arXiv:1012.2751v1 [cs.IT]*, Dec. 2010.

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

[7] 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.

[8] Y. Lomnitz and M. Feder, "Universal prediction of priors for communication," *uploaded to arXiv*, Jan. 2011.

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

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

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

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

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

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