REFERENCES

[1] R. A. Brualdi and V. S. Pless, "Weight enumerators of self-dual codes," *IEEE Trans. Inf. Theory*, vol. 37, no. 4, pp. 1222–1225, Jul. 1991.

[2] S. Buyuklieva and I. Boukliev, "Extremal self-dual codes with an automorphism of order 2," *IEEE Trans. Inf. Theory*, vol. 44, no. 1, pp. 323–328, Jan. 1998.

[3] J. H. Conway and N. J. A. Sloane, "A new upper bound on the minimal distance of self-dual codes," *IEEE Trans. Inf. Theory*, vol. 36, no. 6, pp. 1319–1333, Nov. 1990.

[4] S. T. Dougherty, T. A. Gulliver, and M. Harada, "Extremal binary self-dual codes," *IEEE Trans. Inf. Theory.*, vol. 43, no. 6, pp. 2036–2047, Nov. 1997.

[5] S. T. Dougherty and M. Harada, "New extremal self-dual codes of length 68," *IEEE Trans. Inf. Theory*, vol. 45, no. 6, pp. 2133–2136, Sep. 1999.

[6] T. A. Gulliver and M. Harada, "Classification of extremal double circulant self-dual codes of lengths 64 to 72," *Des., Codes Cryptogr.*, vol. 13, pp. 257–269, 1998.

[7] M. Harada, "Existence of new extremal doubly-even codes and extremal singly-even codes," *Des., Codes Cryptogr.*, vol. 8, pp. 273–283, 1996.

[8] M. Harada, T. Nishimura, and R. Yorgova, "New extremal self-dual codes of length 66," *Mathematica Balkanica*, vol. 21, pp. 113–121, 2007.

[9] W. C. Huffman, "On the classification and enumeration of self-dual codes," *Finite Fields Appl.*, vol. 11, pp. 451–490, 2005.

[10] V. Pless, V. Tonchev, and J. Leon, "On the existence of a certain $(64, 32, 12)$ extremal code," *IEEE Trans. Inf. Theory*, vol. 39, no. 1, pp. 214–215, Jan. 1993.

[11] R. Russeva and N. Yankov, "On binary self-dual codes of lengths 60, 62, 64 and 66 having an automorphism of order 9," *Des., Codes Cryptogr.*, vol. 45, pp. 335–346, 2007.

[12] H. P. Tsai, "Existence of some extremal self-dual codes," *IEEE Trans. Inf. Theory*, vol. 38, no. 6, pp. 1829–1833, Nov. 1992.

[13] H. P. Tsai, "Extremal self-dual codes of lengths 66 and 68," *IEEE Trans. Inf. Theory*, vol. 45, no. 6, pp. 2129–2133, Sep. 1999.

How to Filter an "Individual Sequence With Feedback"

Tsachy Weissman, *Senior Member, IEEE*

***Abstract*—We consider causally estimating (filtering) the components of a noise-corrupted sequence relative to a reference class of filters. The noiseless sequence to be filtered is designed by a "well-informed antagonist," meaning it may evolve according to an arbitrary law, unknown to the filter, based on past noiseless *and noisy* sequence components. We show that this setting is more challenging than that of an individual noiseless sequence (a.k.a. the "semi-stochastic" setting) in the sense that *any* deterministic filter, even one guaranteed to do well on every noiseless individual sequence, fails under some well-informed antagonist. On the other hand, we constructively establish the existence of a *randomized* filter which successfully competes with an arbitrary given finite reference class of filters, under *every* antagonist. Thus, unlike in the semi-stochastic setting, randomization is crucial in the antagonist framework. Our noise model allows for channels whose noisy output depends on the l past channel outputs (in addition to the noiseless channel input symbol). Memoryless channels are obtained as a special case of our model by taking $l = 0$. In this case, our scheme coincides with one that was recently shown to compete with an arbitrary reference class when the underlying noiseless sequence is an individual sequence. Hence, our results show that the latter scheme is universal not only for the semi-stochastic setting in which it was originally proposed, but also under the well-informed antagonist.**

***Index Terms*—Compound sequential decisions, individual sequence with feedback, universal filtering, well-informed antagonist.**

I. INTRODUCTION

Stated in modern terms, the "compound decision problem" is concerned with estimating the components of an individual sequence corrupted by memoryless noise. This problem was formulated by Robbins in the pioneering work [14], where it was shown that such estimation can be done essentially as well as the best time-invariant symbol-by-symbol scheme (that can be chosen by a genie with access to the noiseless sequence). In [16], Samuel showed that this goal is attainable also sequentially, i.e., when the estimate of each clean symbol can depend only on the present and past noisy observations. The schemes that Samuel constructed, which were based on ideas from Hannan's seminal work on prediction of individual sequences [11], were randomized, i.e., utilized external randomization variables which were assumed available. Subsequently, Van Ryzin showed in [15] that randomization is not necessary, and that deterministic versions of the schemes in [16] attain the same goal of doing as well as the best "symbol-by-symbol" scheme regardless of the underlying noiseless individual sequence.

While earlier work on the compound sequential decision problem concentrated on competing with the class of time-invariant "symbol-by-symbol" estimation rules, later developments extended the scope to reference classes of "Markov" estimators of a fixed and

Manuscript received October 18, 2006; revised January 31, 2008. This work was supported in part by the National Science Foundation (NSF) awards 0512140 and 0546535. The material in this correspondence was presented in part at the 44th Annual Allerton Conference on Communications, Control, and Computing, Monticello, IL, September 2006. Part of this work was performed while the author was visiting Hewlett-Packard Laboratories.

The author is with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Technion City, Haifa 32000, Israel, on leave from the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: tsachy@stanford.edu).

Communicated by P. L. Bartlett, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Color versions Figures 1, 3, and 4 in this correspondence are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2008.926457

known order [1], [2], [18], [19], culminating in the recent work [22] which considered competition with an arbitrary finite reference class of filtering schemes, as well as the set of finite-state filters of arbitrary order.

The compound decision framework, where no assumptions are made on a probabilistic or any other type of mechanism that may have generated the noiseless data, leads to performance guarantees that hold uniformly for all individual sequences. Results for the case where the sequence is assumed stochastically generated by an unknown source (though evolving independently of past noisy sequence components) follow as corollaries. For this reason, the individual sequence setting is generally regarded a rather strong one.

Indeed, in prediction of individual sequences [11], [7], [8], [20], [3], [4], [12], [5], a bound that holds for all individual sequences immediately holds also under any stochastic mechanism for generating the sequence, no matter how adversarial one tries to make it. Furthermore, regret bounds that hold under an "oblivious opponent" hold also under a "nonoblivious opponent" who is fully cognizant of the randomization and hence the predictions of the competing predictor (cf. [5, Sec. 4.1]). Tempting as it may seem to extend to the filtering setting, this kind of a conclusion was actually shown by Vardeman [17] to fail for the compound sequential decision problem. Vardeman showed, in the setting of a binary symmetric channel (BSC)-corrupted binary sequence, that "playing Bayes against the estimated past empirical distribution," shown by Van Ryzin in [15] to yield a filter that successfully competes with the best symbol-by-symbol scheme when the underlying signal is an individual sequence, can perform disastrously if the noiseless sequence is allowed to evolve based on past *noisy* sequence components (a.k.a. "channel outputs"). Vardeman referred to this as a situation where the sequence components are chosen by a "well-informed antagonist," a terminology we adopt herein, using "antagonist" for short. Alternatively, taking a communications-oriented viewpoint, as is alluded in the title of this correspondence , one may choose to think of this framework as one where the underlying noiseless sequence is an "individual sequence with feedback."

The framework of a well-informed antagonist is of interest not merely for its conceptual and theoretical significance, but also since it may better capture the reality of some filtering scenarios. Consider, for example, a case where the filter represents a device for tracking a target sequentially based on a noisy observation of its trajectory. The antagonist can then represent the trajectory of a target that has *feedback* on its past noisy trajectory (and hence knowledge of what the tracker is doing), and that can use this feedback to better evade the tracker. A filter that can perform well under the antagonist would correspond then to a device that can successfully track such a potentially more elusive target.

A natural question arising in the context of Vardeman's result is whether failure to perform under the well-informed antagonist is due to peculiarities of Van Ryzin's scheme or perhaps, on the other extreme, due to some more basic limitation shared by all deterministic filters. We answer this question in Section II by showing that *any deterministic* filter will fail to compete with the class of symbol-by-symbol filters under some antagonist. Evidently, competing under an antagonist is more challenging than under an individual noiseless sequence.

This negative result raises the question of whether it is at all possible for a filter to compete with a nontrivial reference class under the well-informed antagonist. We answer this question in the affirmative in Sections III–VI by constructing a *randomized* filter that can successfully compete in the following generality.

1. An arbitrary finite reference class of (possibly randomized) filters.
2. Noisy time-invariant channels satisfying a mild invertibility condition, possibly having memory, where the noisy output depends on the l past channel outputs (in addition to the noiseless input symbol). This is a large class of channels, that includes many channels arising naturally in signal processing and communications (cf. [6] and references therein).
3. A possibly randomized antagonist, allowed to base its choice for the next noiseless sequence component on knowledge of the past *noisy* sequence, as well as all past randomization variables. Thus, in particular, when choosing the current noiseless sequence component, the antagonist knows all history of the noiseless sequence, noisy sequence, actions of competing filters, and actions of all filters in the reference class (and, in particular, their respective cumulative losses).

Memoryless channels are obtained as a special case of our setting by taking $l = 0$. In this case, our filter coincides with the one in [22], shown therein to compete with an arbitrary finite reference class in the individual sequence (semi-stochastic[1]) setting. Thus, for memoryless channels, our main contribution is in establishing that the filter of [22] is universal not only for the semi-stochastic setting in which it was originally proposed, but also under the more challenging setting of a well-informed antagonist. Further, in this more challenging setting, our results imply that randomization is an absolute necessity.

The main problem we treat, as well as the approach we take to its solution via the prediction-filtering correspondence, form a natural continuation and extension to the framework of [22]. Thus, though we give a self-contained account of our framework, results, and analysis, familiarity with [22] will make for an easier read of the present work.

The remainder of the correspondence is organized as follows. As mentioned, we show in Section II that any deterministic filter will fail under some antagonist, thereby both establishing the antagonist framework as more challenging than the semi-stochastic one, and making the case for the use of randomized filters. In Section III, we describe the antagonist framework in its full generality, and state our main result, Theorem 1, on the existence of a randomized filter that successfully competes with an arbitrary finite reference class regardless of the underlying antagonist. Section IV is a short detour into prediction of individual sequences, to introduce notation and recap a known result (with a slight twist) on prediction of individual sequences relative to a set of experts that will serve us well in later sections. In Section V, we describe the correspondence between filters and predictors which allows us to present Theorem 3 and its corollary, Corollary 1, the main technical results underlying Theorem 1. In Section VI, we detail the construction of a filter which, equipped with the results of the previous sections, we show is competitive in the senses asserted in Theorem 1, thereby proving the theorem. It will be clear that our main technical increment in proving Theorem 1, beyond ingredients that were already used in [22], is Theorem 3, which extends the martingale lemma [22, Lemma 2] to the setting of an arbitrary antagonist and channels with memory. We conclude in Section VII with a summary of our findings and related questions for future research.

II. Individual Sequence Versus the Well-informed Antagonist

The main point of this section is to show that the well-informed antagonist is a more challenging adversary than the individual sequence for the filtering problem. To this end, it suffices to restrict attention, as we do throughout this section, to the case of a memoryless channel, and only a certain type of antagonist (namely, a nonrandomized one). The full generality in which our main result holds will be specified in Section III.

[1]The term "semi-stochastic" is due to the fact that the individual noiseless sequence is deterministic, while the noise in the channel and, hence, the noisy sequence are stochastic.

A. Filtering a Corrupted Individual Sequence Relative to a Set of Experts

Let $x_1, x_2, \ldots, x_i \in \mathcal{X}$, be an individual (deterministic) sequence and $Z_1, Z_2, \ldots, Z_i \in \mathcal{Z}$, denote its noisy observation when corrupted by some known memoryless channel, where $\mathcal{X}$, $\mathcal{Z}$ are finite alphabets (without loss of generality, we will identify the elements of any finite alphabet $\mathcal{V}$ with $\{0, 1, \ldots, |\mathcal{V}| - 1\}$).[2] Concretely, $Z_t = g(x_t, W_t)$, where $W_1, W_2, \ldots$ are independent and identically distributed (i.i.d.) $\sim U[0,1]$ (with $U[0,1]$ denoting the uniform distribution on the $[0,1]$ interval) and the function[3] g is known. Note that g induces the *channel matrix* Π defined by

$$\Pi(x, z) = \lambda(\{w : g(x, w) = z\}) \tag{1}$$

where λ denotes Lebesgue measure. In words, $\Pi(x, z)$ is the probability of a channel output z when the channel input is x. A *deterministic filter* is a sequential (causal) estimator for the components of the individual sequence based on its noisy observations. Specifically, a deterministic filter $\hat{\boldsymbol{X}}$ is given by a sequence of mappings $\{\hat{X}_t\}_{t \geq 1}$, where $\hat{X}_t : \mathcal{Z}^t \to \hat{\mathcal{X}}$, and $\hat{\mathcal{X}}$ is a finite reconstruction alphabet, so that the filter's estimate of x_t is given by $\hat{X}_t = \hat{X}_t(Z^t)$, where Z^t is shorthand notation for the sequence $(Z_1, Z_2, \ldots, Z_t)$. We will also use, e.g., $Z_{t_1}^{t_2}$ to denote the sequence $(Z_{t_1}, \ldots, Z_{t_2})$ (which is to be understood as the empty string when $t_1 > t_2$).

A *randomized filter*, or just *filter* for short, is one that accesses, in addition to the noisy sequence, a randomization sequence $U_1, U_2, \ldots$ of i.i.d. $\sim U[0,1]$ components, independent of the channel noise $W_1, W_2, \ldots$, on which it can base its estimates. More specifically, the randomized filter is given as $\hat{\boldsymbol{X}} = \{\hat{X}_t\}_{t \geq 1}$, where $\hat{X}_t : \mathcal{Z}^t \times [0,1] \to \hat{\mathcal{X}}$, the filter's estimate of x_t being $\hat{X}_t = \hat{X}_t\left(Z^t, U_t\right)$. We further assume, for concreteness that will enable a simplified presentation of later results, that for each z^t there exists a partition of the unit interval into subintervals $\alpha_{-1} = 0 \leq \alpha_0 \leq \alpha_1 \leq \cdots \leq \alpha_{|\hat{\mathcal{X}}|-1} = 1$ such that

$$\hat{X}_t\left(z^t, u_t\right) = \hat{x}, \quad \text{iff } u_t \in [\alpha_{\hat{x}-1}, \alpha_{\hat{x}}). \tag{2}$$

In other words, the randomization variable is used by "slicing" the unit interval into subintervals of lengths corresponding to the sought probabilities of the possible reconstruction symbols. This entails no essential loss of generality since clearly, for any filter $\hat{\boldsymbol{X}}$, not necessarily of the form in (2), there exists another filter $\tilde{\boldsymbol{X}}$ such that

$$\hat{X}_t\left(z^t, u_t\right) \stackrel{d}{=} \tilde{X}_t\left(z^t, u_t\right), \quad \forall\, t \text{ and } z^t \in \mathcal{Z}^t \tag{3}$$

where $\stackrel{d}{=}$ denotes equality in distribution.[4] The per-symbol loss of the filter (deterministic or randomized) is denoted by

$$L_{\hat{\boldsymbol{X}}}(x^n, Z^n) = \frac{1}{n} \sum_{t=1}^{n} \Lambda(x_t, \hat{X}_t) \tag{4}$$

[2]We use $|\cdot|$ to denote cardinality or absolute value according to whether the argument is set- or real-valued.

[3]Here and throughout, all functions considered are assumed to be measurable.

[4]Two filters satisfying (3) were said in [22] to be "equivalent." Every equivalence class of filters contains one (and only one) filter of the form in (2). The restriction to filters of this form will allow us in the sequel to make statements about filters which would otherwise have to involve qualifications about equivalence classes.

where $\Lambda : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}$ is a given loss function (when $\hat{\boldsymbol{X}}$ is a randomized filter, $L_{\hat{\boldsymbol{X}}}(x^n, Z^n)$ may depend on the randomization sequence as well, though we suppress this dependence in the notation for simplicity). Given a set of filters $\mathcal{G}$, the goal is to find a filter (not necessarily a member of $\mathcal{G}$) that does essentially as well as the best in the set, regardless of the underlying noiseless individual sequence. More concretely, one seeks a filter $\hat{\boldsymbol{X}}$ satisfying

$$\max_{x^n \in \mathcal{X}^n} \left[EL_{\hat{\boldsymbol{X}}}(x^n, Z^n) - \min_{\hat{\boldsymbol{X}}' \in \mathcal{G}} EL_{\hat{\boldsymbol{X}}'}(x^n, Z^n) \right] \leq \varepsilon_n \tag{5}$$

where $\varepsilon_n \to 0$. In [15], it was shown, under benign assumptions on the channel, that for the case where $\mathcal{G} = \mathcal{G}_s$, the class of constant "symbol-by-symbol" filters, there exists a deterministic filter $\hat{\boldsymbol{X}}$ satisfying

$$\max_{x^n \in \mathcal{X}^n} \left[EL_{\hat{\boldsymbol{X}}}(x^n, Z^n) - \min_{\hat{\boldsymbol{X}}' \in \mathcal{G}_s} EL_{\hat{\boldsymbol{X}}'}(x^n, Z^n) \right] \leq C/\sqrt{n} \tag{6}$$

for some constant C (depending on $|\mathcal{X}|$, $|\mathcal{Z}|$, Π). This result was extended in [2] to the case of reference classes consisting of "Markov" (finite-memory, time-invariant, sliding-window) filters of a known order.

B. Any Deterministic Filter Fails Under Some Antagonist

Consider now the case where the sequence to be filtered, rather than a predetermined individual sequence, is allowed to evolve based on the past noisy (channel output) sequence. Specifically, $\{X_t\}_{t \geq 1}$ is now a sequence of mappings where $X_t : \mathcal{Z}^{t-1} \to \mathcal{X}$, so that the tth noiseless sequence component is $X_t = X_t(Z^{t-1})$. Let A_n ("A" standing for "antagonists") denote the set of all such length-n sequences $X^n = \{X_t\}_{t=1}^n$. While clearly

$$\begin{aligned} \max_{X^n \in A_n} &\left[EL_{\hat{\boldsymbol{X}}}(X^n, Z^n) - \min_{\hat{\boldsymbol{X}}' \in \mathcal{G}_s} EL_{\hat{\boldsymbol{X}}'}(X^n, Z^n) \right] \\ &\geq \max_{x^n \in \mathcal{X}^n} \left[EL_{\hat{\boldsymbol{X}}}(x^n, Z^n) - \min_{\hat{\boldsymbol{X}}' \in \mathcal{G}_s} EL_{\hat{\boldsymbol{X}}'}(x^n, Z^n) \right] \end{aligned} \tag{7}$$

it may be tempting to assume that the inequality in (7) is satisfied with equality. Vardeman disproved this supposition in [17] by establishing the existence of a filter $\hat{\boldsymbol{X}}$ satisfying (6) but for which there exists a constant $\eta > 0$ such that for all n

$$\max_{X^n \in A_n} \left[EL_{\hat{\boldsymbol{X}}}(X^n, Z^n) - \min_{\hat{\boldsymbol{X}}' \in \mathcal{G}_s} EL_{\hat{\boldsymbol{X}}'}(X^n, Z^n) \right] \geq \eta. \tag{8}$$

To do this, Vardeman considered the setting of filtering a binary sequence corrupted by a BSC of some crossover probability δ, under Hamming loss. In this case, $\mathcal{X} = \mathcal{Z} = \hat{\mathcal{X}} = \{0,1\}$, g is given by equation (9) at the bottom of the page, and Λ is Hamming loss, i.e., $\Lambda(x, \hat{x}) = 1_{\{x \neq \hat{x}\}}$. Van Ryzin's "play Bayes against the estimated past empiric distribution" filter [15], for this setting, assuming $0 \leq \delta < 1/2$, is given as

$$\begin{aligned} \hat{X}_t(z^t) &= B\left(\frac{\overline{n}_1(z^{t-1}) - \delta}{1 - 2\delta}, z_t\right) \\ &= \begin{cases} 0, & \text{if } \overline{n}_1(z^{t-1}) < 2\delta(1-\delta) \\ 1, & \text{if } \overline{n}_1(z^{t-1}) > \delta^2 + (1-\delta)^2 \\ z_t, & \text{otherwise} \end{cases} \end{aligned} \tag{10}$$

$$g(x, w) = \begin{cases} 0, & \text{if } \{x = 0 \text{ and } w \in [0, 1-\delta)\} \text{ or } \{x = 1 \text{ and } w \in [0, \delta)\} \\ 1, & \text{otherwise} \end{cases} \tag{9}$$

where $\overline{n}_1(z^t)$ denotes the fraction of 1's in z^t (and $\overline{n}_1(z^0) \equiv 1/2$) and $B(p,z)$ is defined by

$$B(p,z) = \begin{cases} 0, & \text{if } p < \delta \\ 1, & \text{if } 1-p < \delta \\ z, & \text{otherwise} \end{cases} \tag{11}$$

i.e., the optimal estimate of X given $Z = z$, when $X \sim \text{Bernoulli}(p)$ and Z is the output of the BSC (δ) whose input is X. The bottom line of [17] was the construction of an antagonist, on which the filter in (10), for $\delta = 1/3$, has $\lim_n EL_{\hat{\boldsymbol{X}}}(X^n, Z^n) = 5/9$, which is nowhere near the expected per-symbol loss of the best symbol-by-symbol filter (which is $\leq \delta = 1/3$, since δ is achievable by the "say what you see" scheme).

Thus, Vardeman's result shows that Van Ryzin's deterministic filter, shown in [15], that competes with the filter set $\mathcal{G}_s$ under any noiseless individual sequence, fails under some antagonist. It does not, however, exclude the possibility of the existence of some other deterministic filter that successfully competes with that set under all antagonists. This possibility is excluded by the result that follows (stated assuming the BSC setting).

Proposition 1: For any deterministic filter $\hat{\boldsymbol{X}}$ there exists an antagonist such that, for all n

$$EL_{\hat{\boldsymbol{X}}}(X^n, Z^n) - \min_{\hat{\boldsymbol{X}}' \in \mathcal{G}_s} EL_{\hat{\boldsymbol{X}}'}(X^n, Z^n) \geq \delta(1-\delta). \tag{12}$$

Proof: Given the deterministic filter $\hat{\boldsymbol{X}} = \{\hat{X}_t\}$, we construct the adversarial sequence (antagonist) by

$$X_t(z^{t-1}) = \begin{cases} 1, & \text{if } \hat{X}_t(z^{t-1}, \cdot) \equiv 0 \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

where $\hat{X}_t(z^{t-1}, \cdot)$ denotes the mapping $s : \mathcal{Z} \to \hat{\mathcal{X}}$ satisfying $s(z) = \hat{X}_t(z^{t-1}, z)$ for all $z \in \mathcal{Z}$. This gives

$$E\left[\Lambda(X_t, \hat{X}_t) | z^{t-1}\right] = \begin{cases} 1, & \text{if } \hat{X}_t(z^{t-1}, \cdot) \equiv 0 \\ 1, & \text{if } \hat{X}_t(z^{t-1}, \cdot) \equiv 1 \\ \delta, & \text{if } \hat{X}_t(z^{t-1}, \cdot) = \text{SWYS} \\ 1-\delta, & \text{if } \hat{X}_t(z^{t-1}, \cdot) = \overline{\text{SWYS}}, \end{cases} \tag{14}$$

where $\text{SWYS}(z) = z$ and $\overline{\text{SWYS}}(z) = \overline{z}$ with $\overline{z}$ denoting the binary complement of z. Evidently[5]

$$\begin{aligned} E\left[\Lambda(X_t, \hat{X}_t) | Z^{t-1}\right] &\geq 1_{\{X_t = 1\}} + 1_{\{X_t = 0\}} \cdot \delta \\ &= \delta + 1_{\{X_t = 1\}} \cdot (1-\delta) \end{aligned} \tag{15}$$

which, taking expectations over both sides of (15) and summing from $t = 1$ to $t = n$, implies

$$\begin{aligned} EL_{\hat{\boldsymbol{X}}}(X^n, Z^n) &= \frac{1}{n} \sum_{t=1}^n E\Lambda(X_t, \hat{X}_t) \\ &\geq \delta + (1-\delta) \frac{1}{n} \sum_{t=1}^n P(X_t = 1). \end{aligned} \tag{16}$$

On the other hand

$$\begin{aligned} &\min_{\hat{\boldsymbol{X}}' \in \mathcal{G}_s} EL_{\hat{\boldsymbol{X}}'}(X^n, Z^n) \\ &= \min\left\{ \frac{1}{n} \sum_{t=1}^n P(X_t = 1), 1 - \frac{1}{n} \sum_{t=1}^n P(X_t = 1), \delta, 1-\delta \right\} \end{aligned} \tag{17}$$

[5]Throughout, equalities and inequalities between random variables should be understood in the almost-sure sense.

where the four terms inside the brackets on the right-hand side correspond, respectively, to the expected per-symbol loss of the schemes "say all zeros," "say all ones," "say what you see," and "say the binary complement of what you see." Consequently

$$\begin{aligned} &EL_{\hat{\boldsymbol{X}}}(X^n, Z^n) - \min_{\hat{\boldsymbol{X}}' \in \mathcal{G}_s} EL_{\hat{\boldsymbol{X}}'}(X^n, Z^n) \\ &\quad \geq f\left(\delta, \frac{1}{n} \sum_{t=1}^n P(X_t = 1)\right) \\ &\quad \geq \delta(1-\delta) \end{aligned} \tag{18}$$

where the left inequality follows upon defining

$$f(\delta, p) = p(1-\delta) + \delta - \min\{p, 1-p, \delta, 1-\delta\} \tag{19}$$

and the right inequality by the fact that $f(\delta, p) \geq \delta(1-\delta)$ for $0 \leq p \leq 1$. □

It is interesting to examine the form that the antagonist constructed in the above proof assumes when applied to a familiar filter. Specifically, the form of this antagonist (given in (13)), when applied to Van Ryzin's scheme (10), is

$$X_t(z^{t-1}) = \begin{cases} 1, & \text{if } \overline{n}_1(z^{t-1}) < 2\delta(1-\delta) \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

Thus, under this antagonist, a typical sample path of $\overline{n}_1(Z^t)$ will fluctuate around the value $2\delta(1-\delta)$, which is a boundary point in the decision region of the filter it was designed to impede. Indeed, each time that $\overline{n}_1(Z^t)$ "down-crosses" the level $2\delta(1-\delta)$, the antagonist produces the channel input symbol "1" while the Van Ryzin filter, as can be seen in (10), outputs a "0," hence incurring a loss of "1." On the other hand, at each up-crossing of that level, the antagonist produces an input symbol of "0" while the Van Ryzin filter "says what it sees," thus incurring an expected loss of δ. Thus, overall, the expected per-symbol loss of the Van Ryzin filter under the antagonist in (20) exceeds that of the "say what you see" filter (and, *a fortiori*, of the best filter in the reference class of symbol-by-symbol schemes) by a nondiminishing amount.

Fig. 1 plots the performance (cumulative *non*-normalized loss) of the Van Ryzin filter, employed on a BSC($\delta = 1/3$)-corrupted individual sequence. Also plotted are the respective cumulative losses of the four symbol-by-symbol filters in the reference class with which the Van Ryzin filter was designed to compete. The filter is indeed seen to be doing as well as the best in the reference class, as may be expected from the fact that this filter satisfies (6). The individual sequence is obtained via a raster scan of the binary image in Fig. 2. Fig. 3 plots the performance of the same filter, again for a BSC($\delta = 1/3$), but this time under the antagonist in (20). As can be expected from Proposition 1, and as discussed above, this filter performs poorly relative to the best in the reference class. Indeed, inspection of the graph shows that essentially for all n plotted

$$L_{\hat{\boldsymbol{X}}}(X^n, Z^n) - \min_{\hat{\boldsymbol{X}}' \in \mathcal{G}_s} L_{\hat{\boldsymbol{X}}'}(X^n, Z^n) \approx \frac{5}{9} - \frac{1}{3} = \frac{1}{3} \cdot \frac{2}{3} = \delta(1-\delta)$$

which is in line with the lower bound (12).

Clearly, Proposition 1, which is stated in the setting of the BSC and the (small) reference class of symbol-by-symbol schemes, implies *a fortiori* that in the more general case of larger alphabets, larger reference classes, and more general loss functions, deterministic filters that successfully compete under all antagonists do not exist. Evidently, deterministic filters, even those that can successfully compete with a reference class on all individual noiseless sequences, fail under the well-informed antagonist. Viewed positively, feedback about the past

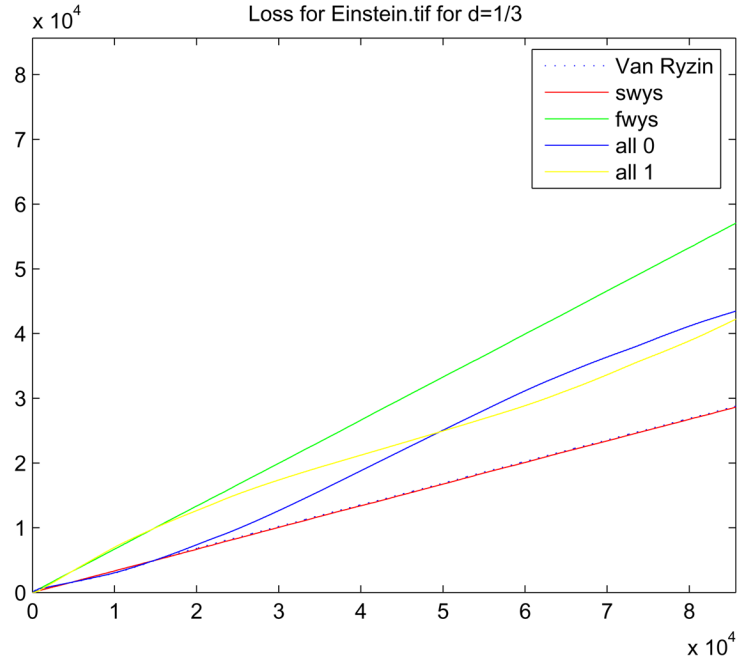

Fig. 1. Performance of Van Ryzin's filter on an individual sequence.

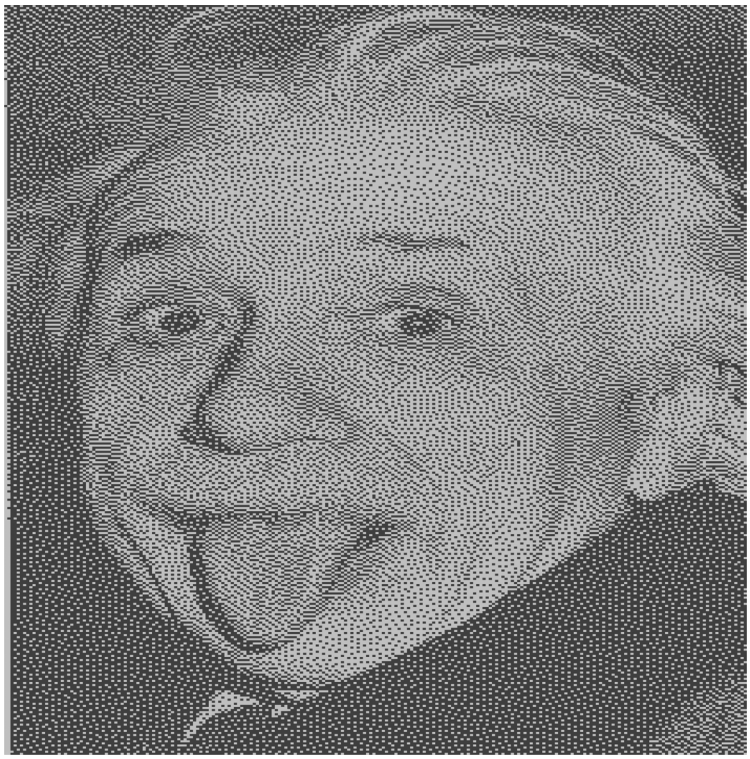

Fig. 2. The individual sequence.

noisy channel outputs can be of substantial benefit to an antagonist trying to avoid being tracked by a deterministic filter. The question now arises as to whether allowing randomization could enable a filter to successfully compete under this more adversarial antagonist. We answer this question in the affirmative in what follows, for a problem setting more general than considered above.

III. Filtering Against the Well-Informed Antagonist

In what follows $\boldsymbol{U} = (U_1, U_2, \ldots)$, $\boldsymbol{V} = (V_1, V_2, \ldots)$, $\boldsymbol{W} = (W_1, W_2, \ldots)$ are independent sequences of i.i.d. $\sim U[0,1]$ random variables, constituting the sources of randomness, respectively, for the filters, the antagonist, and the channel, in ways that will be detailed below. Given a finite reference class of filters $\mathcal{G}$, our goal is to choose a filter $\hat{\boldsymbol{X}}$, after which the antagonist, based on knowledge of the reference class and our choice of filter, chooses the sequence of mappings $\{X_t\}$ to be used in the following game: For each $t \geq 1$

[Nature:] Generates randomization variables U_t, V_t, W_t

[Antagonist:] Generates noiseless sequence component $X_t = X_t(W^{t-1}, U^{t-1}, V^t)$

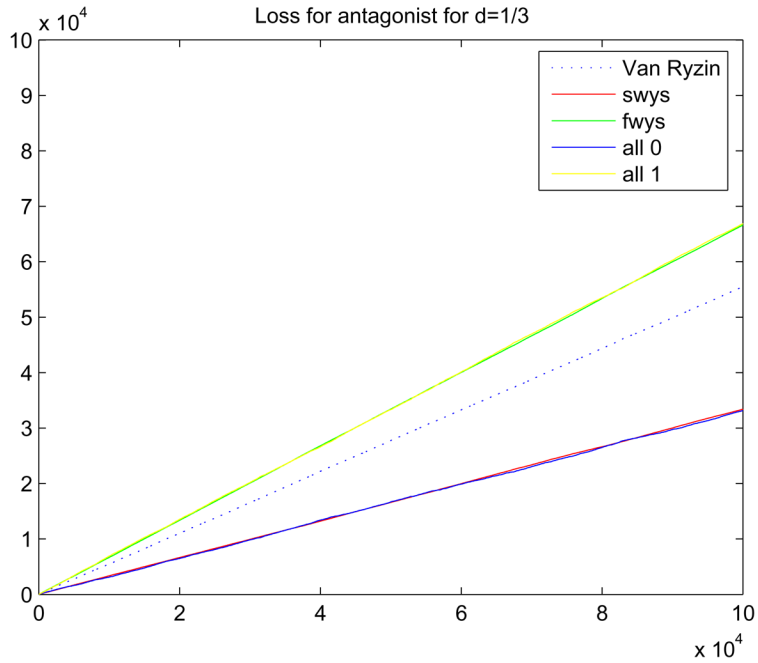

Fig. 3. Performance of Van Ryzin's filter under the antagonist of Proposition 1.

[Channel:] Generates noisy sequence component $Z_t = g\left(Z_{t-l}^{t-1}, X_t, W_t\right)$
[Filter:] Generates estimate of noiseless sequence component $\hat{X}_t = \hat{X}_t\left(Z^t, U_t\right)$
[Reference Class:] Each filter in reference class $\hat{\boldsymbol{X}}' \in \mathcal{G}$ generates estimate $\hat{X}_t' = \hat{X}_t'\left(Z^t, U_t\right)$.

Assumptions:

1. Z_{-l+1}^0 is an arbitrary deterministic sequence known to all sides (the particular values are inconsequential). The sole goal of Z_{-l+1}^0 is to initiate the noisy context for the channel.
2. The function g is known to all sides. Note that, since g characterizes the channel, this assumption is equivalent to that of a known channel (cf. [21] for a discussion of why this assumption is realistic in many practical scenarios). Note also that l corresponds to the effective channel memory (the case $l = 0$ reducing to a memoryless channel). The function g induces the channel matrices

$$\Pi_{z^l}(x, z) = \lambda\left(\{w : g(z^l, x, w) = z\}\right). \tag{21}$$

$\Pi_{z^l}(x, z)$ is thus the probability of a channel output symbol z when the input symbol is x and the previous l noisy symbols are z^l. Our assumption on the channel is that the matrix Π_{z^l} is of full row rank for every z^l. This assumption guarantees that, for every z^l, there exists a function $h_{z^l} : \mathcal{Z} \to \mathbb{R}^{\mathcal{X}}$ such that

$$\sum_z h_{z^l}(z)[x']\Pi_{z^l}(x, z) = \delta(x, x') \tag{22}$$

where $h_{z^l}(z)[x']$ denotes the x'th component of $h_{z^l}(z)$ and $\delta(x, x')$ denotes the Kronecker delta. In vector notation, viewing $h_{z^l}(z)$ as a column vector, (22) is equivalent to

$$\sum_z h_{z^l}(z)\Pi_{z^l}(x, z) = \delta_x \tag{23}$$

where $\delta_x \in \mathbb{R}^{\mathcal{X}}$ denotes the column vector all of whose components are 0 except for the xth one which is 1. Throughout the remainder of the paper we assume a fixed set of functions $\{h_{z^l}\}_{z^l}$ satisfying (23).

Remarks:

- Let H_{z^l} denote the $|\mathcal{Z}| \times |\mathcal{X}|$ matrix whose zth row is $h_{z^l}^T(z)$, i.e., $H_{z^l}(z, b) = h_{z^l}(z)[b]$. Equality (23), in matrix form, becomes

$$\Pi_{z^l} \cdot H_{z^l} = I \tag{24}$$

where I is the $|\mathcal{X}| \times |\mathcal{X}|$ identity matrix. Thus, as stated, Π_{z^l} being of full row rank guarantees the existence of a matrix H_{z^l} satisfying (24). Note further that for the case where $|\mathcal{X}| = |\mathcal{Z}|$, the matrix H_{z^l} is uniquely determined and given by the inverse of Π_{z^l}. Thus, H_{z^l} can be thought of an operation for "inverting" the effect of the channel.
- The requirement that Π_{z^l} be of full row rank for every z^l is rather benign, and satisfied by most channels of interest. For example, in the binary case where $\mathcal{X} = \mathcal{Z} = \{0, 1\}$, this requirement boils down to the condition that $\Pi_{z^l}(0, 0)\Pi_{z^l}(1, 1) \neq (1 - \Pi_{z^l}(0, 0))(1 - \Pi_{z^l}(1, 1))$ for every z^l. On the other hand, it is easy to show that when the full-rankness assumption does not hold, competition with any nontrivial reference class in the setting of an unknown individual sequence (and, *a fortiori*, the setting of an antagonist) is infeasible (cf. [15], [21], [10]).
- Let for each $t \geq 1$

$$\mathcal{F}_t = \sigma(U^t, V^t, W^t) \tag{25}$$

where the right-hand side denotes the sigma-field generated by U^t, V^t, W^t. Note that since Z^t is completely determined by W^t and X^t, and since both $W^t \in \mathcal{F}_t$ and $X^t \in \mathcal{F}_t$,[6] it follows that $Z^t \in \mathcal{F}_t$. Also, for any filter $\hat{\boldsymbol{X}}$, $\hat{X}^t$ is completely determined by Z^t and U^t, so, since both $Z^t \in \mathcal{F}_t$ and $U^t \in \mathcal{F}_t$, it follows that $\hat{X}^t \in \mathcal{F}_t$. Note further that when the antagonist chooses the tth noiseless component X_t, it has access to $\mathcal{F}_{t-1}$ so, in particular, is fully cognizant not only of the past noisy sequence Z^{t-1} (which

[6]When X is a random variable and σ a sigma field, we shall write $X \in \sigma$ to denote the fact that X is σ-measurable. Whether $\in$ denotes set membership or measurability will be clear from the context.

was the case with the antagonist constructed in the previous section), but also of the actual past reconstruction symbols generated by the filter $\hat{\boldsymbol{X}}$, as well as all those generated by the reference filters in $\mathcal{G}$. The additional variable available to the antagonist at time t, V_t, is to allow it to also randomize if it so chooses.

Toward stating our main result, we now let $L^n_{\hat{\boldsymbol{X}}}$ denote the normalized cumulative loss of the filter $\hat{\boldsymbol{X}}$ in the first n rounds of the game, i.e.,

$$L^n_{\hat{\boldsymbol{X}}} = \frac{1}{n}\sum_{t=1}^{n} \Lambda(X_t, \hat{X}_t). \tag{26}$$

Though we suppress this dependence in the notation for simplicity, $L^n_{\hat{\boldsymbol{X}}}$ depends on the underlying antagonist (i.e., the mappings $\{X_t\}_{t=1}^n$) and, consequently, also on the variables U^n, V^n, W^n. Similarly, for each filter in the reference class $\hat{\boldsymbol{X}}' \in \mathcal{G}$, we let $L^n_{\hat{\boldsymbol{X}}'}$ denote its normalized cumulative loss. Our ultimate goal in the game is to construct a filter $\hat{\boldsymbol{X}}$ competing with the reference class $\mathcal{G}$ in the sense of ensuring that the difference

$$L^n_{\hat{\boldsymbol{X}}} - \min_{\hat{\boldsymbol{X}}' \in \mathcal{G}} L^n_{\hat{\boldsymbol{X}}'} \tag{27}$$

is small, regardless of the underlying antagonist who may be trying to make this difference as large as possible. This goal turns out to be achievable in senses we make precise in Theorem 1 below. To state the theorem, we define for every z^l and mapping $s : \mathcal{Z} \to \hat{\mathcal{X}}$ the column vector $\rho_{z^l}(s)$, whose xth component is given by

$$\rho_{z^l}(s)[x] = \sum_{z} \Lambda(x, s(z)) \Pi_{z^l}(x, z). \tag{28}$$

One can think of $\rho_{z^l}(s)[x]$ as the expected loss when employing the estimation rule s on the channel output, when the channel input is x and the noisy context (past noisy l-tuple) is z^l. Defining now

$$\ell_{\max} = \max_{z^l, z, s} h_{z^l}(z)^T \cdot \rho_{z^l}(s) - \min_{z^l, z, s} h_{z^l}(z)^T \cdot \rho_{z^l}(s) \tag{29}$$

and

$$C_{\max} = \max_{x, \hat{x}, z^l, z, s} \left| \Lambda(x, \hat{x}) - h_{z^l}(z)^T \cdot \rho_{z^l}(s) \right| \tag{30}$$

our main result can be stated as follows.

Theorem 1: For any finite set of filters $\mathcal{G}$, and any $n \geq 1$, there exists a filter $\hat{\boldsymbol{X}}$ such that for all $\varepsilon > 0$, and under all antagonists

1.
$$EL^n_{\hat{\boldsymbol{X}}} - \min_{\hat{\boldsymbol{X}}' \in \mathcal{G}} EL^n_{\hat{\boldsymbol{X}}'} \leq \ell_{\max}\sqrt{\frac{\ln|\mathcal{G}|}{2n}} \tag{31}$$

2.
$$P\left(L^n_{\hat{\boldsymbol{X}}} - \min_{\hat{\boldsymbol{X}}' \in \mathcal{G}} L^n_{\hat{\boldsymbol{X}}'} \geq \varepsilon + \ell_{\max}\sqrt{\frac{\ln|\mathcal{G}|}{2n}}\right) \leq 2(|\mathcal{G}|+1)\exp\left[-n\frac{\varepsilon^2}{2C^2_{\max}}\right]. \tag{32}$$

Remarks:

- Note that inequality (32) implies

$$\begin{aligned} &E\left[L^n_{\hat{\boldsymbol{X}}} - \min_{\hat{\boldsymbol{X}}' \in \mathcal{G}} L^n_{\hat{\boldsymbol{X}}'}\right] \\ &= E\left[L^n_{\hat{\boldsymbol{X}}} - \min_{\hat{\boldsymbol{X}}' \in \mathcal{G}} L^n_{\hat{\boldsymbol{X}}'} - \ell_{\max}\sqrt{\frac{\ln|\mathcal{G}|}{2n}}\right] + \ell_{\max}\sqrt{\frac{\ln|\mathcal{G}|}{2n}} \\ &\leq \int_0^\infty P\left(L^n_{\hat{\boldsymbol{X}}} - \min_{\hat{\boldsymbol{X}}' \in \mathcal{G}} L^n_{\hat{\boldsymbol{X}}'} - \ell_{\max}\sqrt{\frac{\ln|\mathcal{G}|}{2n}} \geq \varepsilon\right) d\varepsilon \\ &\quad + \ell_{\max}\sqrt{\frac{\ln|\mathcal{G}|}{2n}} \\ &\overset{(a)}{\leq} 2(|\mathcal{G}|+1)\int_0^\infty \exp\left[-n\frac{\varepsilon^2}{2C^2_{\max}}\right] d\varepsilon + \ell_{\max}\sqrt{\frac{\ln|\mathcal{G}|}{2n}} \\ &= \left[(|\mathcal{G}|+1)\sqrt{2C^2_{\max}\pi} + \ell_{\max}\sqrt{\frac{\ln|\mathcal{G}|}{2}}\right]\frac{1}{\sqrt{n}} \end{aligned} \tag{33}$$

  where (a) follows from (32). Thus, (32) implies also a bound on the expected regret, in addition to the bound in (31) which is on the regret under the expected losses. Of course, $E\left[L^n_{\hat{\boldsymbol{X}}} - \min_{\hat{\boldsymbol{X}}' \in \mathcal{G}} L^n_{\hat{\boldsymbol{X}}'}\right] \geq EL^n_{\hat{\boldsymbol{X}}} - \min_{\hat{\boldsymbol{X}}' \in \mathcal{G}} EL^n_{\hat{\boldsymbol{X}}'}$, so (32), which implies (33), is enough to imply a bound of the form $const./\sqrt{n}$ on the left-hand side of (31). The bound in (31), however, is much better in terms of the constant and, in particular, the dependence of the constant on $|\mathcal{G}|$.
- The proof of Theorem 1 will constructively establish existence of a filter $\hat{\boldsymbol{X}}$ satisfying (31) and (32).
- For the case of a memoryless channel, i.e., $l = 0$, the filtering scheme we will construct coincides with the one in Theorem 1 of [22]. Thus, for this case, our main contribution is in showing that the scheme of [22] competes under the well-informed antagonist just as efficiently as it does under an individual noiseless sequence. In turn, this can be shown to imply that all other performance guarantees that are given in [22] for the semi-stochastic setting in fact hold under the well-informed antagonist. In particular, the "incremental parsing filter" of [22, Sec. 5] is guaranteed to attain the "finite-state filterability" that can be associated with any antagonist (analogously, as is associated with a noise-corrupted individual sequence).
- As previously discussed, our (benign) assumption that the channel matrices Π_{z^l} are of full row rank guarantees the existence of matrices $\{H_{z^l}\}_{z^l}$ satisfying (24). However, when $|\mathcal{X}| < |\mathcal{Z}|$, the requirement to satisfy (24) does not uniquely determine $\{H_{z^l}\}$. In this case, the upper bound in (31) suggests that the minimization of $\ell_{\max}$ (recall (29) for the dependence of $\ell_{\max}$ on $\{H_{z^l}\}$) may be a reasonable guideline for the choice of matrices $\{H_{z^l}\}$, among all those that satisfy (24).
- The filter of Theorem 1 is "horizon-dependent" (i.e., dependent on the sequence length n). However, as was the case for the filter of [22, Theorem 1], a similar result holds for a "strongly sequential" (horizon- independent) filter at the price of slightly larger constants multiplying the square-root term in (31) and in (32). Such a strongly sequential filter can be constructed from the one in Theorem 1 via a standard "doubling trick," as described, e.g., in [3].

Theorem 1 will be proved in Section VI by combining a bound from the realm of prediction of individual sequences (presented in Section IV) with an association between the problems of prediction and filtering that we extend in Section V from the setting of [22] to our present setting of the well-informed antagonist.

IV. Prediction Under a Loss-Function With Memory

In this section, we digress briefly to the setting of prediction of individual sequences relative to an arbitrary expert set, under a loss function with memory. Let the finite sets $\mathcal{Y}, \mathcal{A}$ be, respectively, a source alphabet and a prediction alphabet (also referred to as the "action space"). A predictor, $F = \{F_t\}_{t\geq 1}$, is a sequence of functions $F_t : \mathcal{Y}^{t-1} \to \mathcal{M}(\mathcal{A})$, where $\mathcal{M}(\mathcal{A})$ denotes the simplex of probability distributions on the alphabet $\mathcal{A}$. The interpretation is that the prediction for time t is given by $a \in \mathcal{A}$ with probability $F_t(y^{t-1})[a]$, where $F_t(y^{t-1})[a]$ denotes the ath component of $F_t(y^{t-1})$. Note that, unlike for the filtering setting of the previous section, where the filter output was a reconstruction symbol (rather than a distribution on the reconstruction alphabet) with access to a randomization variable, here we view the prediction as a distribution on the prediction alphabet, with no access to external randomization. This view simplifies the statement of Theorem 2 below, and will suffice for our later goal of transferring results from prediction to filtering. Assuming a given loss function $\Gamma : \mathcal{Y}^{l+1} \times \mathcal{A} \to \mathbb{R}$, for any n and $y^n \in \mathcal{Y}^n$ we define the *normalized cumulative loss* of the predictor F by

$$L_F(y^n) = \frac{1}{n}\sum_{t=1}^{n}\sum_{a\in\mathcal{A}} \Gamma(y_{t-l}^t, a) F_t(y^{t-1})[a], \tag{34}$$

which can be interpreted as the expected prediction loss on the individual sequence y^n,[7] when averaging over the randomization. Note that the loss function may depend on sequence components from some fixed portion of the past of length l, but not on past predictions. The latter case was treated in [13] and is more challenging than our situation, which is easily reduced to the standard prediction problem by considering the sequence $\tilde{y}^n$, where $\tilde{y}_i = y_{i-l}^i$. The following theorem is then a direct consequence of [22, Theorem 2] (which, in turn, follows from [4, Theorem 1]).

Theorem 2: For every finite predictor set $\mathcal{F}$ there exists a predictor F such that for all $y^n \in \mathcal{Y}^n$

$$L_F(y^n) - \min_{F'\in\mathcal{F}} L_{F'}(y^n) \leq \Gamma_{\max}\sqrt{\frac{\ln|\mathcal{F}|}{2n}} \tag{35}$$

where

$$\Gamma_{\max} = \max_{y_{-l}^0, a} \Gamma\left(y_{-l}^0, a\right) - \min_{y_{-l}^0, a} \Gamma\left(y_{-l}^0, a\right). \tag{36}$$

The proof of [22, Theorem 2] implies that a predictor F satisfying (35) is

$$F_t(y^{t-1}) = \frac{\sum_{F'\in\mathcal{F}} e^{-\eta L_{F'}(y^{t-1})} F'_t(y^{t-1})}{\sum_{F'\in\mathcal{F}} e^{-\eta L_{F'}(y^{t-1})}} \tag{37}$$

where $\eta = \frac{\sqrt{8n\ln|\mathcal{F}|}}{\Gamma_{\max}}$.

V. From Prediction to Filtering and Back

In this section, we establish a correspondence between prediction and filtering under an arbitrary antagonist that will be key for proving Theorem 1. This correspondence is the natural generalization of that established in [22] to the case of channels with memory.

[7]The right-hand side of (34) actually depends also on y_{-l+1}^0 which we assume is set to some arbitrary value.

Let F be a predictor (from the setting of the previous section), where the source alphabet is taken to be the alphabet of the noisy sequence from the filtering problem $\mathcal{Y} = \mathcal{Z}$. As the prediction alphabet we take $\mathcal{A} = \mathcal{S}$, where $\mathcal{S}$ is the (finite) set of mappings that take $\mathcal{Z}$ into $\hat{\mathcal{X}}$, i.e., $\mathcal{S} = \{s : \mathcal{Z} \to \hat{\mathcal{X}}\}$. Thus, for each $z^{t-1} \in \mathcal{Z}^{t-1}$, $F_t(z^{t-1})$ is a distribution on the set of mappings $\mathcal{S}$, i.e., $F_t(z^{t-1}) \in \mathcal{M}(\mathcal{S})$. With any such predictor we associate a filter $\hat{\boldsymbol{X}}^F$ as follows:

$$\hat{X}_t^F\left(z^t, u_t\right) = \hat{x} \text{ if } \sum_{a=0}^{\hat{x}-1}\sum_{s:s(z_t)=a} F_t(z^{t-1})[s] \leq u_t < \sum_{a=0}^{\hat{x}}\sum_{s:s(z_t)=a} F_t(z^{t-1})[s] \tag{38}$$

where, without loss of generality, we assume the reconstruction alphabet to be $\{0, 1, \ldots, |\hat{\mathcal{X}}| - 1\}$. In words, $\hat{X}^F$ is defined so that the probability that $\hat{X}_t^F\left(Z^t, U_t\right) = \hat{x}$ is the probability that the mapping S, generated according to the distribution $F_t(z^{t-1})$, maps z_t to $\hat{x}$. Conversely, for any filter $\hat{\boldsymbol{X}}$, we define the associated predictor $F^{\hat{\boldsymbol{X}}}$ by

$$\begin{aligned} F_t^{\hat{\boldsymbol{X}}}(z^{t-1})[s] &= \lambda\left(\left\{u \in [0,1] : \hat{X}_t(z^{t-1}z, u) = s(z)\ \forall\, z\right\}\right) \\ &= \lambda\left(\left\{u \in [0,1] : \hat{X}_t(z^{t-1}\cdot, u) = s\right\}\right) \end{aligned} \tag{39}$$

where $z^{t-1}z$ denotes the sequence of length t formed by concatenation of the symbol z to the right of z^{t-1}. The associations in (38) and (39), taken with the convention on the structure of filters stated in (2), are readily verified (cf. [22, Sec. 4] for a similar derivation) to be consistent in the sense that, for any filter $\hat{\boldsymbol{X}}$

$$\hat{\boldsymbol{X}}^{(F^{\hat{\boldsymbol{X}}})} = \hat{\boldsymbol{X}}. \tag{40}$$

We are now ready to present the main technical result underlying the proof of Theorem 1, where the setting of Section III is assumed (recall, in particular, the σ-field $\mathcal{F}_t$ defined in (25)).

Theorem 3: Let $L_F(z^n)$ denote the normalized cumulative loss of a predictor F from the prediction setting of Section IV when employed on the sequence z^n, for source alphabet $\mathcal{Y} = \mathcal{Z}$, prediction alphabet $\mathcal{A} = \mathcal{S}$, and under the loss function with memory

$$\ell(z^l z, s) = h_{z^l}(z)^T \cdot \rho_{z^l}(s). \tag{41}$$

For any predictor F, and under *any* antagonist, $\{n[L_{\hat{\boldsymbol{X}}^F}^n - L_F(Z^n)]\}_{n\geq 1}$ is an $\{\mathcal{F}_n\}$-martingale.

Theorem 3 extends [22, Lemma 2] to accommodate our current setting of channels with memory and an underlying antagonist (rather than an individual sequence). For the proof of Theorem 3, it will be convenient to introduce the following notation: For each t, z^t, define $P_{\hat{\boldsymbol{X}}}(z^t) \in \mathcal{M}(\hat{\mathcal{X}})$ by

$$P_{\hat{\boldsymbol{X}}}(z^t)[\hat{x}] = \int_{u\in[0,1]:\hat{X}_t(z^t,u)=\hat{x}} du \tag{42}$$

namely, the probability that $\hat{X}_t\left(z^t, U_t\right) = \hat{x}$.

Proof of Theorem 3: We fix a predictor F, $t \geq 1$, an antagonist, and consider (43)–(44) at the bottom of the next page, where

- (a) follows since $\left(X_t, Z^t\right) \in \sigma(V_t, W_t, \mathcal{F}_{t-1})$ and U_t is independent of $\sigma(V_t, W_t, \mathcal{F}_{t-1})$ so

$$E\left[\Lambda\left(X_t, \hat{X}_t^F\left(z^t, u_t\right)\right)\middle| V_t, W_t, \mathcal{F}_{t-1}\right] = \sum_{\hat{x}} \Lambda(X_t, \hat{x}) P_{\hat{\boldsymbol{X}}^F}(Z^t)[\hat{x}]$$

- (b) follows since $(X_t, Z^{t-1}) \in \sigma(V_t, \mathcal{F}_{t-1})$ and, conditioned on $\sigma(V_t, \mathcal{F}_{t-1})$, Z_t is distributed according to $\Pi_{Z_{t-l}^{t-1}}(X_t, \cdot)$ so

$$E\left[\sum_{\hat{x}} \Lambda(X_t, \hat{x}) P_{\hat{\boldsymbol{X}}^F}(Z^t)[\hat{x}] \middle| V_t, \mathcal{F}_{t-1}\right] = \sum_z \Pi_{Z_{t-l}^{t-1}}(X_t, z) \left(\sum_{\hat{x}} \Lambda(X_t, \hat{x}) P_{\hat{\boldsymbol{X}}^F}(Z^{t-1}z)[\hat{x}]\right)$$

- (c) follows since $Z^{t-1} \in \mathcal{F}_{t-1}$
- (d) follows from

$$\begin{aligned}
&\sum_z \Pi_{Z_{t-l}^{t-1}}(x, z) \sum_{\hat{x}} \Lambda(x, \hat{x}) P_{\hat{\boldsymbol{X}}^F}(Z^{t-1}z)[\hat{x}] \\
&= \sum_z \Pi_{Z_{t-l}^{t-1}}(x, z) \sum_{\hat{x}} \Lambda(x, \hat{x}) \left[\sum_{s: s(z)=\hat{x}} F_t(Z^{t-1})[s]\right] \\
&= \sum_z \Pi_{Z_{t-l}^{t-1}}(x, z) \sum_s \Lambda(x, s(z)) F_t(Z^{t-1})[s] \\
&= \sum_s \rho_{Z_{t-l}^{t-1}}(s)[x] F_t(Z^{t-1})[s] \\
&= \sum_s \left[\delta_x^T \cdot \rho_{Z_{t-l}^{t-1}}(s)\right] F_t(Z^{t-1})[s]
\end{aligned}$$

where the equality before last follows from the definition of $\rho_{zl}(s)$ (recall (28))

- (e) follows from

$$\begin{aligned}
&E\left[h_{Z_{t-l}^{t-1}}(Z_t)^T \middle| \mathcal{F}_{t-1}\right] \\
&= \sum_z P(Z_t = z|\mathcal{F}_{t-1}) h_{Z_{t-l}^{t-1}}(z)^T \\
&= \sum_z \left[\sum_x P(X_t = x|\mathcal{F}_{t-1}) \Pi_{Z_{t-l}^{t-1}}(x, z)\right] h_{Z_{t-l}^{t-1}}(z)^T \\
&= \sum_x P(X_t = x|\mathcal{F}_{t-1}) \left[\sum_z \Pi_{Z_{t-l}^{t-1}}(x, z) h_{Z_{t-l}^{t-1}}(z)^T\right] \\
&= \sum_x P(X_t = x|\mathcal{F}_{t-1}) \delta_x^T
\end{aligned}$$

where the last equality follows from (23).

The proof is concluded by noting that

$$\begin{aligned}
&n\left[L_{\hat{\boldsymbol{X}}^F}^n - L_F(Z^n)\right] - (n-1)\left[L_{\hat{\boldsymbol{X}}^F}^{n-1} - L_F(Z^{n-1})\right] \\
&\quad = \Lambda(X_n, \hat{X}_n^F(Z^n, U_n)) - \sum_s \ell(Z_{n-l}^n, s) F_n(Z^{n-1})[s] \qquad (45)
\end{aligned}$$

that, by (44), the right-hand side of (45) equals 0 a.s. when conditioned on $\mathcal{F}_{n-1}$, and that $L_{\hat{\boldsymbol{X}}^F}^{n-1} - L_F(Z^{n-1}) \in \mathcal{F}_{n-1}$. □

Corollary 1: For all n, any predictor F, and under any antagonist, [Unbiasedness:]

$$EL_{\hat{\boldsymbol{X}}^F}^n = EL_F(Z^n). \qquad (46)$$

[Concentration:]

$$P\left(\left|L_{\hat{\boldsymbol{X}}^F}^n - L_F(Z^n)\right| \geq \varepsilon\right) \leq 2\exp\left(-n\frac{2\varepsilon^2}{C_{\max}^2}\right). \qquad (47)$$

Proof: The equality in (46) is immediate from Theorem 3. To establish (47) note first that since $\ell(z^l z, s) = h_{zl}(z)^T \cdot \rho_{zl}(s)$ (recall (41)), $C_{\max}$, as defined in (30), is equivalently expressed as

$$C_{\max} = \max_{x, \hat{x}, z^l, z, s} \left|\Lambda(x, \hat{x}) - \ell(z^l z, s)\right|. \qquad (48)$$

From (45), (48), and Theorem 3 it then follows that $L_{\hat{\boldsymbol{X}}^F}^n - L_F(Z^n)$ is a martingale with differences bounded by $C_{\max}$. Inequality (47) now

---

$$\begin{aligned}
&E\left[\Lambda(X_t, \hat{X}_t^F(z^t, u_t)) \middle| \mathcal{F}_{t-1}\right] \qquad (43) \\
&= E\left[E\left[\Lambda(X_t, \hat{X}_t^F(z^t, u_t)) \middle| V_t, W_t, \mathcal{F}_{t-1}\right] \middle| \mathcal{F}_{t-1}\right] \\
&\overset{(a)}{=} E\left[E\left[\sum_{\hat{x}} \Lambda(X_t, \hat{x}) P_{\hat{\boldsymbol{X}}^F}(Z^t)[\hat{x}] \middle| V_t, \mathcal{F}_{t-1}\right] \middle| \mathcal{F}_{t-1}\right] \\
&\overset{(b)}{=} E\left[\sum_z \Pi_{Z_{t-l}^{t-1}}(X_t, z) \left(\sum_{\hat{x}} \Lambda(X_t, \hat{x}) P_{\hat{\boldsymbol{X}}^F}(Z^{t-1}z)[\hat{x}]\right) \middle| \mathcal{F}_{t-1}\right] \\
&\overset{(c)}{=} \sum_x P(X_t = x|\mathcal{F}_{t-1}) \left[\sum_z \Pi_{Z_{t-l}^{t-1}}(x, z) \sum_{\hat{x}} \Lambda(x, \hat{x}) P_{\hat{\boldsymbol{X}}^F}(Z^{t-1}z)[\hat{x}]\right] \\
&\overset{(d)}{=} \sum_x P(X_t = x|\mathcal{F}_{t-1}) \left[\sum_s [\delta_x^T \cdot \rho_{Z_{t-l}^{t-1}}(s)] F_t(Z^{t-1})[s]\right] \\
&= \sum_s \left[\left(\sum_x P(X_t = x|\mathcal{F}_{t-1}) \delta_x^T\right) \cdot \rho_{Z_{t-l}^{t-1}}(s)\right] F_t(Z^{t-1})[s] \\
&\overset{(e)}{=} \sum_s \left[E\left[h_{Z_{t-l}^{t-1}}(Z_t)^T \middle| \mathcal{F}_{t-1}\right] \cdot \rho_{Z_{t-l}^{t-1}}(s)\right] F_t(Z^{t-1})[s] \\
&= E\left[\sum_s \left[h_{Z_{t-l}^{t-1}}(Z_t)^T \cdot \rho_{Z_{t-l}^{t-1}}(s)\right] F_t(Z^{t-1})[s] \middle| \mathcal{F}_{t-1}\right] \\
&= E\left[\sum_s \ell(Z_{t-l}^t, s) F_t(Z^{t-1})[s] \middle| \mathcal{F}_{t-1}\right] \qquad (44)
\end{aligned}$$

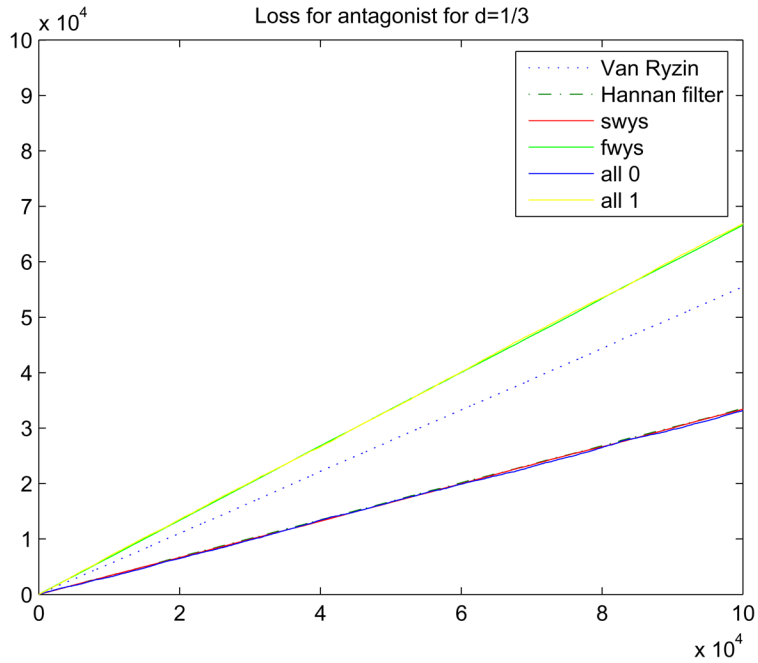

Fig. 4. Randomized filter of Theorem 1 (the "Hannan filter") and Van Ryzin's deterministic filter under the antagonist.

follows from an application of the Hoeffding–Azuma inequality [9, Theorem 9.1]. □

VI. The Competing Filter

The important conclusion of the previous section, encapsulated in Corollary 1, is that the $L_F(Z^n)$ (which can be computed based on observing only the noisy channel output sequence Z^n) is an unbiased—and quite efficient—estimate of $L_{\hat{\boldsymbol{X}}^F}^n$ (the unobserved loss of the filter $\hat{\boldsymbol{X}}^F$), regardless of what the (unknown) underlying antagonist may be. This conclusion suggests the following recipe for construction of a filter competing with the reference class of filters $\mathcal{G}$.

- Transform each of the filters in $\mathcal{G}$ into its associated predictor to obtain the predictor set

$$\mathcal{F} = \left\{ F^{\hat{\boldsymbol{X}}'} : \hat{\boldsymbol{X}}' \in \mathcal{G} \right\}. \tag{49}$$

- Construct a predictor F that competes with $\mathcal{F}$ in the sense of Theorem 2.
- Let the competing filter $\hat{\boldsymbol{X}}$ be given by $\hat{\boldsymbol{X}}^F$.

With this recipe, Theorem 1 is proved using Corollary 1 similarly to the way in which [22, Theorem 5] was proved using [22, Theorem 4]. We give the full proof in Appendix A for completeness.

Fig. 4 shows the performance of the filter constructed in the above proof for the binary setting of a BSC($\delta = 1/3$), Hamming loss, and the reference class $\mathcal{G}_s$ of symbol-by-symbol filters, when the noiseless sequence evolves according to the antagonist in (20). We refer to it as the "Hannan filter," as it is induced by the "Hannan predictor" competing with the set of constant predictors in the individual sequence setting of Section IV (under the loss function, source alphabet, and prediction space detailed in Theorem 3). In line with Theorem 1, this filter is seen to be doing as well as the best in the reference class (its curve in the graph is obscured by those of the best schemes). Also plotted for comparison is the performance of Van Ryzin's deterministic filter [15] which was shown in Section III to fail to compete under this antagonist.

VII. Conclusion

We have considered the problem of filtering relative to an arbitrary finite set of filtering experts when the noiseless sequence to be filtered is allowed to evolve based on knowledge of the past noisy sequence components (that is, chosen by the "well-informed antagonist" or, alternatively, the noiseless sequence can be thought of as an "individual sequence with feedback").

We have shown that this framework is more challenging than its "semi-stochastic" origin of an underlying noiseless individual sequence. Specifically, that while there exist deterministic filters that can compete with certain reference classes of filters in the semi-stochastic setting, *all* such filters fail under an appropriately chosen antagonist.

On the positive side, we have constructively established the existence of a *randomized* filter that can compete with an arbitrary given finite reference class of filters regardless of what the underlying antagonist may be. This was done in the generality of channels that may have memory of a finite number of past noisy symbols.

Our findings suggest a fundamental difference between the problems of prediction and of filtering: while in the former, randomization is needed whether the sequence outcomes are formed by an individual sequence or a "nonoblivious opponent" [5], in the latter we have shown that, at least for some reference classes, randomization is not needed when the underlying sequence components form an individual sequence but is crucial when they are generated by the well-informed antagonist. These findings reiterate and further motivate consideration of the open question mentioned in [22]: is randomization needed to successfully compete in the semi-stochastic setting with an arbitrary (finite) reference class of filters? We conjecture that the answer is negative, for reasons elaborated on in [22, Sec. VI].

Another question arising naturally in the context of our findings is: What happens when the "feedback" is noisy, i.e., if the antagonist driving the noiseless sequence gets to see only a noisy version of the past channel outputs? Is it still the case that any deterministic filter will fail under some antagonist of this type? As motivation, we note that for the application discussed in the Introduction, of tracking a target with access to its past noisy trajectory, it is more realistic to assume the presence of noisy rather than noise-free feedback.

APPENDIX

A. Proof of Theorem 1

Let $\mathcal{F}$ be the predictor set in (49) and F be the predictor that competes with this set in the sense of Theorem 2. Letting $\hat{\boldsymbol{X}} = \hat{\boldsymbol{X}}^F$

$$
\begin{aligned}
EL^n_{\hat{\boldsymbol{X}}} - \min_{\hat{\boldsymbol{X}}' \in \mathcal{G}} EL^n_{\hat{\boldsymbol{X}}'} &= EL_F(Z^n) - \min_{\hat{\boldsymbol{X}}' \in \mathcal{G}} EL_{F\hat{\boldsymbol{X}}'}(Z^n) && (50)\\
&= EL_F(Z^n) - \min_{F' \in \mathcal{F}} EL_{F'}(Z^n) && (51)\\
&\le E\left[L_F(Z^n) - \min_{F' \in \mathcal{F}} L_{F'}(Z^n)\right] && (52)\\
&\le \ell_{\max}\sqrt{\frac{\ln|\mathcal{F}|}{2n}} && (53)\\
&= \ell_{\max}\sqrt{\frac{\ln|\mathcal{G}|}{2n}} && (54)
\end{aligned}
$$

where (50) follows from the combination of (40) and (46), (51) follows from the definition of $\mathcal{F}$ in (49), (53) follows from Theorem 2 (and the association $\ell_{\max} \longleftrightarrow \Gamma_{\max}$), and the equality follows since $|\mathcal{F}| = |\mathcal{G}|$ (which is implied by (40)). This proves (31). For the second item note that, for all sample paths

$$
\begin{aligned}
&\left|L^n_{\hat{\boldsymbol{X}}} - \min_{\hat{\boldsymbol{X}}' \in \mathcal{G}} L^n_{\hat{\boldsymbol{X}}'} - \left(L_F(z^n) - \min_{F' \in \mathcal{F}} L_{F'}(z^n)\right)\right|\\
&\quad = \left|L^n_{\hat{\boldsymbol{X}}} - \min_{F' \in \mathcal{F}} L^n_{\hat{\boldsymbol{X}}^{F'}} - \left(L_F(z^n) - \min_{F' \in \mathcal{F}} L_{F'}(z^n)\right)\right| && (55)\\
&\quad \le \left|L^n_{\hat{\boldsymbol{X}}^F} - L_F(z^n)\right| + \max_{F' \in \mathcal{F}}\left|L^n_{\hat{\boldsymbol{X}}^{F'}} - L_{F'}(z^n)\right| && (56)
\end{aligned}
$$

where equality (55) follows from the fact that $\mathcal{G} = \left\{\hat{\boldsymbol{X}}^{F'} : F' \in \mathcal{F}\right\}$ (implied by (40)). It follows from (35), (47), (56), and a union bound that

$$
\begin{aligned}
&P\left(L^n_{\hat{\boldsymbol{X}}} - \min_{\hat{\boldsymbol{X}}' \in \mathcal{G}} L^n_{\hat{\boldsymbol{X}}'} \ge \varepsilon + \ell_{\max}\sqrt{\frac{\ln|\mathcal{F}|}{2n}}\right)\\
&\quad \le P\left(L^n_{\hat{\boldsymbol{X}}} - \min_{\hat{\boldsymbol{X}}' \in \mathcal{G}} L^n_{\hat{\boldsymbol{X}}'} \ge \varepsilon + L_F(Z^n) - \min_{F' \in \mathcal{F}} L_{F'}(Z^n)\right)\\
&\quad \le P\left(\left|L^n_{\hat{\boldsymbol{X}}^F} - L_F(Z^n)\right| \ge \varepsilon/2\right)\\
&\qquad + P\left(\max_{F' \in \mathcal{F}}\left|L^n_{\hat{\boldsymbol{X}}^{F'}} - L_{F'}(Z^n)\right| \ge \varepsilon/2\right)\\
&\quad \le 2(|\mathcal{F}|+1)\exp\left(-n\frac{\varepsilon^2}{2C^2_{\max}}\right)\\
&\quad = 2(|\mathcal{G}|+1)\exp\left(-n\frac{\varepsilon^2}{2C^2_{\max}}\right). \qquad \square
\end{aligned}
$$

ACKNOWLEDGMENT

The author is indebted to Erik Ordentlich for bringing Vardeman's paper [17] to his attention, to Tom Cover and Erik Ordentlich for stimulating discussions, and to Taesup Moon for his help in the experiments.

REFERENCES

[1] R. J. Ballard, "Extended Rules for the Sequence Compound Decision Problem With $m \times n$ Component," Ph.D. dissertation, Michicgan State Univ., East Lansing, MI, 1974.

[2] R. J. Ballard, D. C. Gilliland, and J. Hannan, "$O(N^{-1/2})$ convergence to k-extended Bayes risk in the sequence compound decision problem with $m \times n$ component," Statistics and Probability RM-333, Michigan State Univ., East Laning, MI, 1974.

[3] N. Cesa-Bianchi, Y. Freund, D. P. Helmbold, D. Haussler, R. Schapire, and M. K. Warmuth, "How to use expert advice," *J. ACM*, vol. 44, no. 3, pp. 427–485, 1997.

[4] N. Cesa-Bianchi and G. Lugosi, "On prediction of individual sequences," *Ann. Statist.*, vol. 27, no. 6, pp. 1865–1895, 1999.

[5] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. New York: Cambridge Univ. Press, 2006.

[6] J. Chen and T. Berger, "The capacity of finite-state Markov channels with feedback," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 780–789, Mar. 2005.

[7] T. M. Cover, "Behavior of sequential predictors of binary sequences," in *Trans. 4th Prague Conf. Information Theory*, Prague, Czechoslovakia, Sep. 1966.

[8] T. M. Cover and A. Shenhar, "Compound Bayes predictors for sequences with apparent Markov structure," *IEEE Trans. Syst., Man Cybern.*, vol. SMC-7, no. 6, pp. 421–424, Jun. 1977.

[9] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.

[10] G. Gemelos, S. Sigurjònsson, and T. Weissman, "Universal minimax discrete denoising under channel uncertainty," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3476–3497, Aug. 2006.

[11] J. Hannan, "Approximation to Bayes risk in repeated play," in *Contributions to the Theory of Games*. Princeton, NJ: Princeton Univ. Press, 1957, vol. III, pp. 97–139.

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

[13] N. Merhav, E. Ordentlich, G. Seroussi, and M. J. Weinberger, "On sequential strategies for loss functions with memory," *IEEE Trans. Inf. Theory*, vol. 48, no. 7, pp. 1947–1958, Jul. 2002.

[14] H. Robbins, "Asymptotically subminimax solutions of compound statistical decision problems," in *Proc. 2nd Berkeley Symp. Mathematical Statistics and Probability*, Berkeley, CA, 1951, pp. 131–148.

[15] J. Van Ryzin, "The sequential compound decision problem with $m \times n$ finite loss matrix," *Ann. Math. Statist.*, vol. 37, pp. 954–975, 1966.

[16] E. Samuel, "Asymptotic solutions of the sequential compound decision problem," *Ann. Math. Statist.*, pp. 1079–1095, 1963.

[17] S. B. Vardeman, "A note on the applicability of sequence compound decision schemes," *Scand. J. Statist.*, vol. 6, pp. 86–88, 1979.

[18] S. B. Vardeman, "Admissible solutions of k-extended finite state set and the sequence compound decision problems," *J. Multiv. Anal.*, vol. 10, pp. 426–441, 1980.

[19] S. B. Vardeman, "Approximation to minimum k-extended Bayes risk in sequences of finite state decision problems and games," *Bull. Inst. Math. Acad. Sinica*, vol. 10, no. 1, pp. 35–52, Mar. 1982.

[20] V. Vovk, "Aggregating strategies," in *Proc. 3rd Annu. Workshop on Computational Learning Theory*, San Mateo, CA, 1990, pp. 371–383.

[21] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger, "Universal discrete denoising: Known channel," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 5–28, Jan. 2005.

[22] T. Weissman, E. Ordentlich, M. J. Weinberger, A. Somekh-Baruch, and N. Merhav, "Universal filtering via prediction," *IEEE Trans. Inf. Theory*, vol. 53, no. 4, pp. 1253–1264, Apr. 2007.