CCIT Report #577 March 2006

Universal Filtering Via Prediction[*]

Tsachy Weissman[†] Erik Ordentlich[‡] Marcelo J. Weinberger[‡]
Anelia Somekh-Baruch[§] Neri Merhav[¶]

March 6, 2006

Abstract

We consider the filtering problem, where a finite-alphabet individual sequence is corrupted by a discrete memoryless channel, and the goal is to causally estimate each sequence component based on the past and present noisy observations. We establish a correspondence between the filtering problem and the problem of prediction of individual sequences which leads to the following result: Given an arbitrary finite set of filters, there exists a filter which performs, with high probability, essentially as well as the best in the set, regardless of the underlying noiseless individual sequence. We use this relationship between the problems to derive a filter guaranteed of attaining the "finite-state filterability" of any individual sequence by leveraging results from the prediction problem.

1 Introduction

The study of prediction of individual sequences with respect to a set of predictors (also known as experts) was pioneered by Hannan [18] and Blackwell [6, 5], who considered competition with the set of constant predictors. Their work prompted further research on and refinements of the problem throughout the late fifties, sixties, and seventies, with notable examples including [9, 10], and references therein. More recently, the problem has seen a resurgence of interest by both the information and learning theory communities, generalizing the original framework to accommodate competition with more general, and in fact arbitrary, predictors, cf. [32, 7, 8, 23] and references therein.

On a parallel thread, study of the problem of estimating the components of a noise-corrupted individual sequence was initiated by Robbins in the seminal [25] and dubbed 'the compound decision problem'. The problem has been the focus of much attention during the fifties and sixties, notable references including [19, 28, 26, 27] (cf. [39] for a comprehensive account of this literature). Much in

[*] Elements in this paper were presented at the 37th Annual Allerton Conference on Communication, Control, and Computing, September 1999, and at the 2004 Data Compression Conference (DCC'04), Snowbird, Utah, March 2004.

[†] Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA. Part of this work was done while the author was visiting Hewlett-Packard Laboratories.

[‡] Hewlett-Packard Laboratories, Palo Alto, CA 94304, USA

[§] Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA.

[¶] Department of Electrical Engineering, Technion, Haifa 32000, Israel.

this line of work was focused on the case where estimation of the components of the noise-corrupted individual sequence needs to be done causally, which was labelled 'the *sequential* compound decision problem'. Early work on the compound sequential decision problem concentrated on competing with the class of time-invariant "symbol by symbol" estimation rules. Later, references [1, 2, 30, 31] extended the scope to reference classes of "Markov" estimators of a fixed and known order. Unlike the prediction problem, however, this problem seems to have largely escaped the spotlight in the recent resurgence of interest in sequential decision problems. An exception is the work in [3, 4] on filtering a Discrete Memoryless Channel (DMC)-corrupted individual sequence with respect to filters implementable as finite-state machines. Another exception is the part of the work in [35] that deals with limited-delay coding of a noise-corrupted individual sequence.[1] The closely related problem of prediction for noise-corrupted individual sequences was considered in [34, 36].

In compliance with more modern terminology, used e.g. in the literature on hidden Markov models [16], we henceforth use the term 'filtering' in lieu of 'compound sequential decision problem' in referring to the problem of causally estimating the components of a noise-corrupted individual sequence. Our goal in this work is to establish a close relationship between the problem of predicting an individual sequence, and that of filtering a DMC-corrupted sequence. We show that with any filter one can associate a predictor for the noisy sequence, whose observable prediction loss (under the right prediction space and loss function) efficiently estimates that of the original filter (which depends also on the noiseless sequence and hence is not observable). This association allows us to transfer results on prediction relative to a set of experts to analogous results for the filtering problem: Given a set of filters, one constructs a predictor competing with the associated set of predictors, using existing theory on universal prediction. The filter associated with such a competing predictor can then be shown to successfully compete with the original set of filters. In other words, this approach yields a filter performing, with high probability, at least as well as the best in a given class of filters, regardless of the underlying noise-free individual sequence.

An approach similar in spirit to the one we follow here was taken in [34] for the problem of predicting a noise-corrupted individual sequence. There too, the idea was to transform the problem to one of prediction in the noiseless sense of the noisy sequence, under a modified loss function. The prediction space, however, remained that of the original problem. In contrast, in our filtering setting, the prediction space in the associated prediction problem will be a space of mappings from a noisy to a reconstruction symbol. Note that the idea of introducing a modified loss function (or distortion measure) to reduce a problem involving noise to a more familiar and basic noiseless one is used in other contexts as well. For example, rate distortion coding of noisy sources is readily reduced to the classical

[1] In particular, Theorem 5 of [35], when specialized to the case where the instantaneous encoding rate is as large as the cardinality of the alphabet of the noisy source, implies the existence of a filter competing with a reference class of finite-memory filters of a given order.

rate distortion problem via the introduction of a modified distortion measure [13, 38, 15]. The idea of reducing a problem to a more basic one by considering a richer alphabet consisting of mappings is also not new. Shannon, for example, used this idea in [29] to reduce the problem of channel coding with causal side information at the transmitter to the classical channel coding problem.

Perhaps the bottom line of the present work, taken with that of [34], is that problems involving sequential decision making in the presence of noise are not fundamentally different from the basic noiseless prediction problem. The former can be reduced to the latter via appropriate associations and modifications of the loss function and prediction space.

The remainder of this work is organized as follows. We shall start in Section 2 with a formal description of our filtering setting, and a statement of our main result, Theorem 1, on the existence of a filter that competes with any given finite set of "filtering experts". In Section 3, we then briefly state the problem of prediction of individual sequences, along with classical results on prediction relative to a set of experts which will be of later use. In Section 4, we establish a correspondence between prediction and filtering, which we then use to prove Theorem 1. In Section 5, we consider filtering relative to finite-state filters, and construct a filter based on incremental parsing [40] and attaining the "finite-state filterability" of any individual sequence. This filter builds on an incremental parsing-based predictor (similar to that of [17]) for the associated prediction problem, and we use results from previous sections for assessing the performance of the induced filter. We also use these results to show that, for any finite-state filter, there exists a finite-memory (also known as "Markov") filter which attains the same performance, a fact whose analogue for the prediction problem has been known since [17, 22]. In Section 6, we compare our results with those in the Markov-extended version of the sequential compound decision problem. We conclude in Section 7 with a summary of our results.

2 Problem Formulation and Main Result

Let $\mathcal{X}, \mathcal{Z}, \hat{\mathcal{X}}$ denote, respectively, the alphabets of the clean, noisy, and reconstructed source, which are assumed finite. As in [37], the noisy sequence is a DMC-corrupted version of the clean one, where the channel matrix Π, $\Pi(x,z)$ denoting the probability of a noisy symbol z when the clean symbol is x, is assumed to be known and of full row rank (implying $|\mathcal{X}| \leq |\mathcal{Z}|$).

Without loss of generality, we will identify the elements of any finite set $\mathcal{V}$ with $\{0, 1, \ldots, |\mathcal{V}| - 1\}$. $\mathbb{R}^{\mathcal{V}}$ will denote the space of $|\mathcal{V}|$-dimensional column vectors with real-valued components indexed by the elements of $\mathcal{V}$. $\mathcal{M}(\mathcal{V})$ will denote the simplex consisting of the elements of $\mathbb{R}^{\mathcal{V}}$ with non-negative components summing up to 1. The a-th component of $v \in \mathbb{R}^{\mathcal{V}}$ will be denoted either by $v[a]$ or by v_a (according to what will result in an overall simpler expression in each particular case). Subscripting a vector or a matrix by 'max' will stand for the difference between the maximum and the minimum of all its components. Thus, for example, if Γ is a $|\mathcal{Z}| \times |\mathcal{X}|$ matrix then Γ_{max} stands for $\max_{z \in \mathcal{Z}, x \in \mathcal{X}} \Gamma(z,x) -$

$\min_{z\in\mathcal{Z},x\in\mathcal{X}}\Gamma(z,x)$ (in particular, if the components of Γ are non-negative and $\Gamma(z,x)=0$ for some z and x, then $\Gamma_{max}=\max_{z\in\mathcal{Z},x\in\mathcal{X}}\Gamma(z,x)$).

Throughout, we shall be assuming a 'semi-stochastic' setting of a noiseless individual sequence $\mathbf{x}=(x_1,x_2,\ldots)$ corrupted by the DMC (of channel transition matrix Π), so that the noisy sequence $\mathbf{Z}=(Z_1,Z_2,\ldots)$ is stochastic. We will also be assuming a randomization sequence $\mathbf{U}=(U_1,U_2,\ldots)$ of i.i.d. components uniformly distributed on $[0,1]$, independent of $\mathbf{Z}$. Lower case letters will denote either individual deterministic quantities or specific realizations of random variables.

A *filter* is a sequence $\hat{\mathbf{X}}=\{\hat{X}_t\}_{t\geq 1}$, where $\hat{X}_t:\mathcal{Z}^t\times[0,1]\to\hat{\mathcal{X}}$ is a measurable mapping. The interpretation is that, upon observing $z^t=(z_1,\ldots,z_t)$, and accessing a randomization variable $u_t\in[0,1]$, the reconstruction for the unobserved x_t is given by $\hat{X}_t(z^t,u_t)$. The normalized cumulative loss of the filter on the individual triple (x^n,z^n,u^n) is denoted by

$$L_{\hat{\mathbf{X}}}(x^n,z^n,u^n)=\frac{1}{n}\sum_{t=1}^{n}\Lambda\left(x_t,\hat{X}_t(z^t,u_t)\right), \tag{1}$$

where $\Lambda:\mathcal{X}\times\hat{\mathcal{X}}\to[0,\infty)$ is the loss function[2]. We also let

$$L_{\hat{\mathbf{X}}}(x^n,z^n)=\frac{1}{n}\sum_{t=1}^{n}\int_0^1\Lambda\left(x_t,\hat{X}_t(z^t,u)\right)du, \tag{2}$$

where the integral here and throughout should be understood in the Lebesgue sense. Note that since $U_t\sim U[0,1]$ for each t,

$$L_{\hat{\mathbf{X}}}(x^n,z^n)=EL_{\hat{\mathbf{X}}}(x^n,z^n,U^n), \tag{3}$$

where the expectation on the right-hand side assumes that x^n,z^n are individual sequences. Furthermore, since the U_t are also independent, $n\left(L_{\hat{\mathbf{X}}}(x^n,z^n)-L_{\hat{\mathbf{X}}}(x^n,z^n,U^n)\right)$ is a sum of n independent random variables of magnitude bounded by Λ_{max}. Hoeffding's inequality [20] then implies:

Lemma 1 *For all individual sequences x^n,z^n*

$$P\left(\left|L_{\hat{\mathbf{X}}}(x^n,z^n)-L_{\hat{\mathbf{X}}}(x^n,z^n,U^n)\right|\geq\varepsilon\right)\leq 2\exp\left(-n\frac{2\varepsilon^2}{\Lambda_{max}^2}\right). \tag{4}$$

For each t,z^t, define now $P_{\hat{\mathbf{X}}}(z^t)\in\mathcal{M}(\hat{\mathcal{X}})$ by

$$P_{\hat{\mathbf{X}}}(z^t)[\hat{x}]=\int_{u\in[0,1]:\hat{X}_t(z^t,u)=\hat{x}}du, \tag{5}$$

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

$$\int_0^1\Lambda\left(x_t,\hat{X}_t(z^t,u)\right)du=\sum_{\hat{x}\in\hat{\mathcal{X}}}\Lambda(x_t,\hat{x})P_{\hat{\mathbf{X}}}(z^t)[\hat{x}]. \tag{6}$$

[2] Our assumption that Λ assumes non-negative values entails no loss of generality, since otherwise one can work with $\tilde{\Lambda}$ defined by $\tilde{\Lambda}(x,\hat{x})=\Lambda(x,\hat{x})-\min_{x',\hat{x}'}\Lambda(x',\hat{x}')$.

Thus, letting $\boldsymbol{\lambda}^x$ denote the xth row of the loss matrix Λ, substitution of (6) into (2) gives

$$L_{\hat{\mathbf{X}}}(x^n, z^n) = \frac{1}{n}\sum_{t=1}^{n}\sum_{\hat{x}} \Lambda(x_t, \hat{x}) P_{\hat{\mathbf{X}}}(z^t)[\hat{x}] = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{\lambda}^{x_t} \cdot P_{\hat{\mathbf{X}}}(z^t). \tag{7}$$

It is clear from (7) that the filtering loss on the individual pair (x^n, z^n), averaged with respect to the randomization, depends on the filter $\mathbf{X}$ only through $P_{\hat{\mathbf{X}}}$. Therefore, if only the expected performance of a filter is of interest (expectation with respect to the randomization), it is sufficient to specify a filter by identifying it with $P_{\hat{\mathbf{X}}}$ (as was done, e.g., in [24]). In our present work, however, we are ultimately going to be interested in addressing the actual, rather than the expected, loss, which is our reason for considering explicitly the dependence of the filter on its source of randomness. Note that for two filters $\hat{\mathbf{X}}$ and $\tilde{\mathbf{X}}$, $P_{\hat{\mathbf{X}}}(z^t) = P_{\tilde{\mathbf{X}}}(z^t)$ if and only if $\hat{X}_t(z^t, U_t) \overset{d}{=} \tilde{X}_t(z^t, U_t)$ and, consequently,

$$\{\hat{X}_t(z^t, U_t)\}_t \overset{d}{=} \{\tilde{X}_t(z^t, U_t)\}_t \ \forall \mathbf{z} \in \mathcal{Z}^\infty \quad \Leftrightarrow \quad P_{\hat{\mathbf{X}}}(z^t) = P_{\tilde{\mathbf{X}}}(z^t) \ \forall t, z^t, \tag{8}$$

where $\overset{d}{=}$ denotes equality in distribution. This observation motivates the following notion of equivalence.

Definition 1 *Two filters, $\hat{\mathbf{X}}$ and $\tilde{\mathbf{X}}$, will be said to be* equivalent *if $P_{\hat{\mathbf{X}}}(z^t) = P_{\tilde{\mathbf{X}}}(z^t)$ for all t and z^t.*

Thus, equivalence allows for the mappings $\hat{X}_t(z^t, \cdot)$ and $\tilde{X}_t(z^t, \cdot)$ to differ, provided they satisfy the equality in distribution on the left-hand side of (8).

Let now $h : \mathcal{Z} \to \mathbb{R}^{\mathcal{X}}$ have the property that, for $a, b \in \mathcal{X}$,

$$E_a h_b(Z) = \sum_{z \in \mathcal{Z}} h_b(z)\Pi(a, z) = \delta(a, b) \overset{\triangle}{=} \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

where E_a denotes expectation over the channel output Z given that the channel input is a, and $h_b(z)$ denotes the b-th component of $h(z)$. Let H denote the $|\mathcal{Z}| \times |\mathcal{X}|$ matrix whose z-th row is $h^T(z)$, i.e., $H(z, b) = h_b(z)$. To see that our assumption of a channel matrix with full row rank guarantees the existence of such an h note that (9) can equivalently be stated in matrix form as

$$\Pi H = I, \tag{10}$$

where I is the $|\mathcal{X}| \times |\mathcal{X}|$ identity matrix. Thus, e.g., any H of the form $H = \Gamma^T(\Pi\,\Gamma^T)^{-1}$, for any Γ such that $\Pi\,\Gamma^T$ is invertible, satisfies (10). In particular, $\Gamma = \Pi$ is a valid choice ($\Pi\,\Pi^T$ is invertible since Π is of full row rank) corresponding to the Moore-Penrose generalized inverse [21].

Ultimately, our interest is in the setting of a noiseless individual sequence, where the noisy (channel-corrupted) and randomization sequences are stochastic (and independent). Our main result, which pertains to this setting, is the following:

Theorem 1 *For every finite set of filters $\mathcal{G}$ there exists a filter $\hat{\mathbf{X}}$ (not necessarily in $\mathcal{G}$) such that for all $\mathbf{x} \in \mathcal{X}^\infty$, all n, and all $\varepsilon > 0$*

1.

$$EL_{\hat{\mathbf{X}}}(x^n, Z^n, U^n) - \min_{\hat{\mathbf{X}}' \in \mathcal{G}} EL_{\hat{\mathbf{X}}'}(x^n, Z^n, U^n) \leq C(|\mathcal{X}|\Lambda_{max}H_{max}, |\mathcal{G}|)/\sqrt{n}, \tag{11}$$

where $C(\alpha, \beta) = \alpha\sqrt{\frac{1}{2}\ln \beta}$.

2.

$$\begin{aligned} &P\left(L_{\hat{\mathbf{X}}}(x^n, Z^n, U^n) - \min_{\hat{\mathbf{X}}' \in \mathcal{G}} L_{\hat{\mathbf{X}}'}(x^n, Z^n, U^n) \geq \varepsilon + C(|\mathcal{X}|\Lambda_{max}H_{max}, |\mathcal{G}|)/\sqrt{n}\right) \\ \leq\ & 4(|\mathcal{G}|+1)\exp\left[-n\frac{\varepsilon^2}{8(|\mathcal{X}|H_{max}\Lambda_{max})^2}\right]. \end{aligned} \tag{12}$$

It will be seen in the proof of Theorem 1, given in Section 4, that the constant $C(|\mathcal{X}|\Lambda_{max}H_{max}, |\mathcal{G}|)$ in (11) and (12) can be improved (reduced) to $C(\ell_{max}, |\mathcal{G}|)$ where ℓ_{max} is the maximum value of a loss function for a prediction problem that will be specified (in display (36)). The quantity $|\mathcal{X}|\Lambda_{max}H_{max}$ is merely a crude upper bound on ℓ_{max}. Ultimately, the best possible constant that our results will imply is $C(\ell_{max}, |\mathcal{G}|)$, where ℓ_{max}, which will be seen to depend on H, will be minimized over all choices of H (that satisfy (10)). This perhaps also suggests the minimization of ℓ_{max} as a reasonable guideline for the choice of H.

Since for any real-valued random variable V we have $EV \leq \int_0^\infty P(V \geq x)dx$ (with equality for nonnegative random variables), the following corollary is a direct consequence of (12):

Corollary 1 *For every finite set of filters $\mathcal{G}$ there exists a filter $\hat{\mathbf{X}}$ (not necessarily in $\mathcal{G}$) such that for all $\mathbf{x} \in \mathcal{X}^\infty$ and all n,*

$$EL_{\hat{\mathbf{X}}}(x^n, Z^n, U^n) - E\min_{\hat{\mathbf{X}}' \in \mathcal{G}} L_{\hat{\mathbf{X}}'}(x^n, Z^n, U^n) \leq K/\sqrt{n}, \tag{13}$$

where K is a constant that depends on ℓ_{max} and $|\mathcal{G}|$.

Clearly, the benchmark against which the universal filter of Corollary 1 competes is more demanding than the one in (11): Part 1 of Theorem 1 corresponds to competing with a "genie" that selects the filter to be used based only on the underlying noiseless individual sequence (averaging over the channel noise and the randomization variable), whereas Corollary 1 corresponds to competing with the genie that selects the filter based not only on the noiseless sequence, but also on the channel and randomization variable realizations. However, the constant K in (13) is larger than $C(\ell_{max}, |\mathcal{G}|)$).

3 Prediction of Individual Sequences

Let the finite sets $\mathcal{Y}$, $\mathcal{A}$ be, respectively, a source alphabet and a prediction space (also referred to as the "action space"). A predictor, $F = \{F_t\}$, is a sequence of functions $F_t : \mathcal{Y}^{t-1} \to \mathcal{M}(\mathcal{A})$ with the interpretation that the prediction for time t is given by $a \in \mathcal{A}$ with probability $F_t(y^{t-1})[a]$. 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 conform to the standard practice of letting the prediction be a distribution on the prediction alphabet, with no access to external randomization. This definition will simplify the statement of the results below, and suffice for our later needs of transforming results from prediction to filtering. Assuming a given loss function $l : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$, for any n and $y^n \in \mathcal{Y}^n$ we define the *normalized cumulative loss* of the predictor F by[3]

$$L_F(y^n) = \frac{1}{n}\sum_{t=1}^{n}\sum_{a\in\mathcal{A}} l(y_t, a)F_t(y^{t-1})[a] = \frac{1}{n}\sum_{t=1}^{n}\mathcal{L}^{y_t}\cdot F_t(y^{t-1}), \tag{14}$$

where $\mathcal{L}^y$ denotes the y-th row of the matrix representing the loss function l. Note that this can be interpreted as the expected prediction loss on the individual sequence y^n, when averaging over the randomization. The following result is implicit in [8]:

Theorem 2 *For every finite set of predictors $\mathcal{F}$ there exists a predictor F (not necessarily in $\mathcal{F}$) such that for all $y^n \in \mathcal{Y}^n$*

$$L_F(y^n) - \min_{F'\in\mathcal{F}} L_{F'}(y^n) \le C(\ell_{max}, |\mathcal{F}|)/\sqrt{n}, \tag{15}$$

where $C(\cdot,\cdot)$ is the function from the right-hand side of (11).

Proof: For the predictor F defined by

$$F_t(y^{t-1}) = \frac{\sum_{F'\in\mathcal{F}} e^{-\eta L_{F'}(y^{t-1})} F'_t(y^{t-1})}{\sum_{F'\in\mathcal{F}} e^{-\eta L_{F'}(y^{t-1})}} \tag{16}$$

the proof of [8, Theorem 1], which addresses the binary case, carries over to imply

$$L_F(y^n) - \min_{F'\in\mathcal{F}} L_{F'}(y^n) \le \frac{\ln|\mathcal{F}|}{\eta} + \frac{\eta\ell_{max}^2}{8n}. \tag{17}$$

The right-hand side is minimized by taking $\eta = \frac{\sqrt{8n\ln|\mathcal{F}|}}{\ell_{max}}$, which gives the bound (15). □

There exist additional results in the literature on prediction of individual sequences implying similar bounds, e.g., [32, 7]. All are based on predictors similar to that in (16), which randomize between the

[3] The fact that we let L, subscripted by a scheme, denote both the loss of a filter and the loss of a predictor should not confuse since in the former case there are two or more arguments (noise-free, noisy, and, possibly, the randomization sequence) while the latter involves only one sequence.

predictors in the expert set by assigning weights, at each time point, that depend on their performance thus far.

For the special case of an expert set consisting of all the constant predictors, results in the spirit of Theorem 2 date back to Hannan's [18] and Blackwell's approachability theorem [6, 5]. We end this section by detailing Hannan's predictor, of which we will make later use when constructing a universal filter. For $\zeta \in \mathbb{R}^{\mathcal{Y}}$, let $U_l(\zeta)$ denote the *Bayes envelope* of the loss function l, defined by

$$U_l(\zeta) = \min_{a \in \mathcal{A}} \zeta^T \cdot \mathcal{L}_a, \tag{18}$$

$\mathcal{L}_a$ denoting the column of the matrix of the loss function l corresponding to the a-th action. We also let $b(\zeta)$ denote the achiever of the minimum in the right-hand side of (18), namely, the *Bayes Response*

$$b(\zeta) = \arg\min_{a \in \mathcal{A}} \zeta^T \cdot \mathcal{L}_a, \tag{19}$$

resolving ties lexicographically. The empirical distribution of a sequence $y^n \in \mathcal{Y}^n$ will be denoted by $p_{y^n} \in \mathcal{M}(\mathcal{Y})$, i.e., $p_{y^n}[y]$ is the fraction of appearances of y in y^n. Note that, for all n and $y^n \in \mathcal{Y}^n$,

$$\min_{F \in \mathcal{M}_0} L_F(y^n) = U_l(p_{y^n}), \tag{20}$$

where $\mathcal{M}_0$ denotes the class of constant predictors. Letting $B_l = \max_{y \in \mathcal{Y}, a \in \mathcal{A}}[l(y,a) - \min_{a'} l(y,a')]$, the following result was established in the seminal work [18]:

Theorem 3 *[18, Theorem 6] The predictor F defined by*

$$F_t(y^{t-1})[a] = \Pr\left\{ b\left((t-1)p_{y^{t-1}} + \mathbf{U}\sqrt{t} \right) = a \right\}, \tag{21}$$

where $\mathbf{U}$ is uniformly distributed on $\left[0, \sqrt{\frac{6}{|\mathcal{Y}|}}\right]^{\mathcal{Y}}$, satisfies for all n and $y^n \in \mathcal{Y}^n$

$$L_F(y^n) - \min_{F' \in \mathcal{M}_0} L_{F'}(y^n) = L_F(y^n) - U_l(p_{y^n}) \leq B_l \sqrt{\frac{6|\mathcal{Y}|}{n}}. \tag{22}$$

Note that the exponential weighting predictor of Theorem 2, when applied to the reference class $\mathcal{M}_0$, yields

$$L_F(y^n) - \min_{F' \in \mathcal{M}_0} L_{F'}(y^n) \leq \ell_{max} \sqrt{\frac{\ln |\mathcal{A}|}{2n}}, \tag{23}$$

since the size of the set of constant non-randomized predictors $\mathcal{M}_0$ equals the size of the action alphabet $\mathcal{A}$. Evidently, which of the bounds is better depends on the loss function and the cardinalities $|\mathcal{Y}|$ and $|\mathcal{A}|$. In Section 5 we will use the predictor of (21) as a building block in the construction of a universal predictor which, in turn, will lead to a universal filter.

4 Filtering as a Prediction Problem

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, $\mathcal{Y} = \mathcal{Z}$. The prediction alphabet we take to be $\mathcal{A} = \mathcal{S}$, where $\mathcal{S}$ is the (finite) set of mappings s that take $\mathcal{Z}$ into $\hat{\mathcal{X}}$, i.e., $\mathcal{S} = \left\{ s : \mathcal{Z} \to \hat{\mathcal{X}} \right\}$. 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 $\hat{\mathbf{X}}^F$ as follows:

$$\hat{X}_t^F(z^t, u_t) = \hat{x} \quad \text{if} \quad \sum_{\tilde{x}=0}^{\hat{x}-1} \sum_{s:s(z_t)=\tilde{x}} F_t(z^{t-1})[s] \leq u_t < \sum_{\tilde{x}=0}^{\hat{x}} \sum_{s:s(z_t)=\tilde{x}} F_t(z^{t-1})[s], \tag{24}$$

where here and throughout summation over the empty set is defined as zero. Thus

$$P_{\hat{\mathbf{X}}^F}(z^t)[\hat{x}] = \sum_{s:s(z_t)=\hat{x}} F_t(z^{t-1})[s]. \tag{25}$$

In words, $\hat{X}^F$ is defined such that the probability that $\hat{X}_t^F(z^t, U_t) = \hat{x}$ is the probability that the mapping S, generated according to the distribution $F_t(z^{t-1})$, maps z_t to $\hat{x}$. Given the predictor F, we shall refer to the filter $\hat{X}^F$ as the 'prediction-filtering transformation' of F. Conversely, for any filter $\hat{\mathbf{X}}$, we define the associated predictor $F^{\hat{\mathbf{X}}}$ by:

$$\begin{aligned} F_t^{\hat{\mathbf{X}}}(z^{t-1})[s] &= \Pr\left\{ \hat{X}_t(z^{t-1}z, U_t) = s(z) \ \ \forall z \in \mathcal{Z} \right\} & (26) \\ &= \Pr\left\{ \hat{X}_t(z^{t-1}\cdot, U_t) = s \right\} = \int_{u \in [0,1]: s = \hat{X}_t(z^{t-1}\cdot, u)} du. & (27) \end{aligned}$$

Given the filter $\hat{\mathbf{X}}$, we shall refer to the predictor $F^{\hat{\mathbf{X}}}$ as the 'filtering-prediction transformation' of $\hat{\mathbf{X}}$. We note the following two points:

1. The transformation from a predictor F to the filter $\hat{X}^F$ is *not* one to one, i.e., two different predictors can yield the same filter. For a simple example, take $\mathcal{X} = \mathcal{Z} = \hat{\mathcal{X}} = \{0, 1\}$ and consider the following four mappings

$$\begin{aligned} s_1(z) &\equiv 0 \\ s_2(z) &\equiv 1 \\ s_3(z) &= z \\ s_4(z) &= \overline{z}, \end{aligned} \tag{28}$$

where $\overline{z}$ denotes the binary complement of z. Let F, G be constant predictors given, for all t and z^{t-1}, by

$$F_t(z^{t-1})[s] = \begin{cases} 1/2 & \text{if } s = s_1 \text{ or } s = s_2 \\ 0 & \text{otherwise,} \end{cases} \qquad G_t(z^{t-1})[s] = \begin{cases} 1/2 & \text{if } s = s_3 \text{ or } s = s_4 \\ 0 & \text{otherwise.} \end{cases} \tag{29}$$

When transforming, $\hat{X}^F$ becomes the filter saying '0' with probability 1/2 and '1' with probability 1/2, without regard to the observations. $\hat{X}^G$, on the other hand, says 'what it sees' with probability 1/2 and the binary complement of 'what it sees' with probability 1/2. The net effect is the same filter given, for all t, z^t, u_t, by

$$\hat{X}_t^F(z^t, u_t) = \hat{X}_t^G(z^t, u_t) = \begin{cases} 0 & \text{if } u_t \in [0, 1/2) \\ 1 & \text{otherwise.} \end{cases} \tag{30}$$

2. The transformation from a filter $\hat{\mathbf{X}}'$ to the predictor $F^{\hat{\mathbf{X}}'}$ is one to one when identifying filters belonging to the same equivalence class as per Definition 1, i.e., for any filter $\hat{\mathbf{X}}'$, $P_{\hat{\mathbf{X}}^{(F^{\hat{\mathbf{X}}'})}} = P_{\hat{\mathbf{X}}'}$. Indeed,

$$P_{\hat{\mathbf{X}}^{(F^{\hat{\mathbf{X}}'})}}(z^t)[\hat{x}] = \sum_{s:s(z_t)=\hat{x}} F_t^{\hat{\mathbf{X}}'}(z^{t-1})[s] = \sum_{s:s(z_t)=\hat{x}} \Pr\left\{s = \hat{X}'_t(z^{t-1}\cdot, U_t)\right\} \tag{31}$$

$$= \Pr\left\{\hat{X}'_t(z^{t-1}z_t, U_t) = \hat{x}\right\} = P_{\hat{\mathbf{X}}'}(z^t)[\hat{x}], \tag{32}$$

where the first equality follows from (25), the one following it is due to (26), and the right-most equality follows from the definition of $P_{\hat{\mathbf{X}}'}$ (in (5)).

To state the main result of this section, we let, for any $s : \mathcal{Z} \to \hat{\mathcal{X}}$, $\rho(s)$ denote the column vector with xth component

$$\rho_x(s) = \sum_z \Lambda(x, s(z))\Pi(x, z). \tag{33}$$

In words, $\rho_x(s)$ is the expected loss when using the estimator $s(Z)$ while the underlying symbol is x (and is observed through the channel Π in a 'single-letter' problem). The above prediction-filtering correspondence is motivated by the following result.

Theorem 4 *For all n, $x^n \in \mathcal{X}^n$, and any predictor F*

[Unbiasedness:]

$$EL_{\hat{\mathbf{X}}^F}(x^n, Z^n) = EL_F(Z^n) \tag{34}$$

[Concentration:]

$$P\left(\left|L_{\hat{\mathbf{X}}^F}(x^n, Z^n) - L_F(Z^n)\right| \geq \varepsilon\right) \leq 2\exp\left(-n\frac{2\varepsilon^2}{L_{max}^2}\right), \tag{35}$$

with $L_F(z^n)$ denoting the normalized cumulative loss of the predictor F (as defined in (14)) for source alphabet $\mathcal{Y} = \mathcal{Z}$ and prediction space $\mathcal{A} = \mathcal{S}$ under the loss function

$$\ell(z, s) = h(z)^T \cdot \rho(s), \tag{36}$$

where h is any function satisfying (9) and $L_{max} = \max\{\Lambda_{max}, \ell_{max}\}$.

In words, for every predictor F the observable $L_F(Z^n)$ is an unbiased efficient estimate of $L_{\hat{\mathbf{X}}^F}(x^n, Z^n)$. Theorem 4 is a consequence of the following lemma, whose proof is given in Appendix A.

Lemma 2 *For all $\mathbf{x} \in \mathcal{X}^\infty$, and any predictor F, $\left\{n\left[L_{\hat{\mathbf{X}}^F}(x^n, Z^n) - L_F(Z^n)\right]\right\}_{n\geq 1}$ is a $\{Z_n\}$-martingale.*

Proof of Theorem 4: The first item immediately follows from Lemma 2. For the second item note that, by Lemma 2, $L_{\hat{\mathbf{X}}^F}(x^n, Z^n) - L_F(Z^n)$ is a normalized sum of martingale differences. These differences, as is evident from (A.9), are bounded by L_{max}. Inequality (35) then follows from the Hoeffding-Azuma inequality [12, Theorem 9.1]. □

Let now $\mathcal{G}$ be a given finite reference class of filters. Theorem 4 suggests the following recipe for construction of a competing filter:

- Transform each of the filters in $\mathcal{G}$ into its associated predictor to obtain the predictor set $\mathcal{F} = \left\{F^{\hat{\mathbf{X}}'} : \hat{\mathbf{X}}' \in \mathcal{G}\right\}$.
- Construct a predictor F that competes with $\mathcal{F}$ in the sense of Theorem 2.
- Take $\hat{\mathbf{X}}^F$ to be the competing filter.

It is this recipe which we use in proving the following theorem:

Theorem 5 *For every finite filter set $\mathcal{G}$ there exists a filter $\hat{\mathbf{X}}$ such that for all $\mathbf{x} \in \mathcal{X}^\infty$, all n, and all $\varepsilon > 0$*

1.

$$EL_{\hat{\mathbf{X}}}(x^n, Z^n) - \min_{\hat{\mathbf{X}}' \in \mathcal{G}} EL_{\hat{\mathbf{X}}'}(x^n, Z^n) \leq C(\ell_{max}, |\mathcal{G}|)/\sqrt{n}. \tag{37}$$

2.

$$P\left(L_{\hat{\mathbf{X}}}(x^n, Z^n) - \min_{\hat{\mathbf{X}}' \in \mathcal{G}} L_{\hat{\mathbf{X}}'}(x^n, Z^n) \geq \varepsilon + C(\ell_{max}, |\mathcal{G}|)/\sqrt{n}\right) \leq 2(|\mathcal{F}| + 1)\exp\left(-n\frac{\varepsilon^2}{2L_{max}^2}\right), \tag{38}$$

where ℓ_{max} pertains to the loss function in (36).

Proof: Let $\mathcal{F}$ be the predictor set defined by

$$\mathcal{F} = \left\{F^{\hat{\mathbf{X}}'} : \hat{\mathbf{X}}' \in \mathcal{G}\right\} \tag{39}$$

and F be the predictor that competes with this set in the sense of Theorem 2. Letting $\hat{\mathbf{X}} = \hat{\mathbf{X}}^F$,

$$\begin{aligned} EL_{\hat{\mathbf{X}}}(x^n, Z^n) - \min_{\hat{\mathbf{X}}' \in \mathcal{G}} EL_{\hat{\mathbf{X}}'}(x^n, Z^n) &= EL_F(Z^n) - \min_{\hat{\mathbf{X}}' \in \mathcal{G}} EL_{F^{\hat{\mathbf{X}}'}}(Z^n) && (40) \\ &= EL_F(Z^n) - \min_{F' \in \mathcal{F}} EL_{F'}(Z^n) && (41) \end{aligned}$$

$$\begin{aligned}
&\leq E\left[L_F(Z^n) - \min_{F'\in\mathcal{F}} L_{F'}(Z^n)\right] && (42)\\
&\leq C(\ell_{max}, |\mathcal{F}|)/\sqrt{n} && (43)\\
&\leq C(\ell_{max}, |\mathcal{G}|)/\sqrt{n}, && (44)
\end{aligned}$$

where (40) follows from (34), (41) follows from the definition of $\mathcal{F}$ in (39), (43) follows from Theorem 2, and the last equality follows since $|\mathcal{F}| \leq |\mathcal{G}|$.[4] This proves (37). For the second item note that for all x^n, z^n

$$\begin{aligned}
&\left|L_{\hat{\mathbf{X}}}(x^n, z^n) - \min_{\hat{\mathbf{X}}'\in\mathcal{G}} L_{\hat{\mathbf{X}}'}(x^n, z^n) - \left(L_F(z^n) - \min_{F'\in\mathcal{F}} L_{F'}(z^n)\right)\right|\\
= &\left|L_{\hat{\mathbf{X}}}(x^n, z^n) - \min_{F'\in\mathcal{F}} L_{\hat{\mathbf{X}}^{F'}}(x^n, z^n) - \left(L_F(z^n) - \min_{F'\in\mathcal{F}} L_{F'}(z^n)\right)\right| && (45)\\
\leq &\left|L_{\hat{\mathbf{X}}^F}(x^n, z^n) - L_F(z^n)\right| + \max_{F'\in\mathcal{F}} \left|L_{\hat{\mathbf{X}}^{F'}}(x^n, z^n) - L_{F'}(z^n)\right|, && (46)
\end{aligned}$$

where equality (45) follows from the fact that $\mathcal{G} = \left\{\hat{\mathbf{X}}^{F'} : F' \in \mathcal{F}\right\}$. It follows from (15), (35), (46), and a union bound that

$$\begin{aligned}
&P\left(L_{\hat{\mathbf{X}}}(x^n, Z^n) - \min_{\hat{\mathbf{X}}'\in\mathcal{G}} L_{\hat{\mathbf{X}}'}(x^n, Z^n) \geq \varepsilon + C(\Lambda_{max}, |\mathcal{G}|)/\sqrt{n}\right)\\
\leq\ &P\left(L_{\hat{\mathbf{X}}}(x^n, Z^n) - \min_{\hat{\mathbf{X}}'\in\mathcal{G}} L_{\hat{\mathbf{X}}'}(x^n, Z^n) \geq \varepsilon + L_F(Z^n) - \min_{F'\in\mathcal{F}} L_{F'}(Z^n)\right)\\
\leq\ &P\left(\left|L_{\hat{\mathbf{X}}^F}(x^n, Z^n) - L_F(Z^n)\right| \geq \varepsilon/2\right) + P\left(\max_{F'\in\mathcal{F}} \left|L_{\hat{\mathbf{X}}^{F'}}(x^n, Z^n) - L_{F'}(Z^n)\right| \geq \varepsilon/2\right)\\
\leq\ &2(|\mathcal{F}| + 1)\exp\left(-n\frac{\varepsilon^2}{2L_{max}^2}\right).
\end{aligned}$$

□

Theorem 5 is very close to our end goal, which is to prove Theorem 1. The only difference is that the filtering loss in the former is averaged over the randomization sequence. The concentration stated in Lemma 1, however, provides the link that allows us to deduce Theorem 1 from Theorem 5. Specifically:

Proof of Theorem 1: The first item follows directly from that of Theorem 5, combined with (3) and the fact that $\ell_{max} \leq |\mathcal{X}|\Lambda_{max}H_{max}$. For the second item, note first that, similarly as in (46),

$$\begin{aligned}
&\left|L_{\hat{\mathbf{X}}}(x^n, z^n) - \min_{\hat{\mathbf{X}}'\in\mathcal{G}} L_{\hat{\mathbf{X}}'}(x^n, z^n) - \left(L_{\hat{\mathbf{X}}}(x^n, z^n, u^n) - \min_{\hat{\mathbf{X}}'\in\mathcal{G}} L_{\hat{\mathbf{X}}'}(x^n, z^n, u^n)\right)\right| && (47)\\
\leq\ &\left|L_{\hat{\mathbf{X}}}(x^n, z^n) - L_{\hat{\mathbf{X}}}(x^n, z^n, u^n)\right| + \max_{\hat{\mathbf{X}}'\in\mathcal{G}} \left|L_{\hat{\mathbf{X}}'}(x^n, z^n) - L_{\hat{\mathbf{X}}'}(x^n, z^n, u^n)\right|. && (48)
\end{aligned}$$

[4] The second point noted at the beginning of the section implies that $|\mathcal{F}| = |\mathcal{G}|$ whenever each filter in $\mathcal{G}$ belongs to a different equivalence class. The situation $|\mathcal{F}| < |\mathcal{G}|$ will arise if $\mathcal{G}$ contains two different filters from the same equivalence class.

Hence, for all ε' and ε,

$$
\begin{aligned}
& P\left(L_{\hat{\mathbf{X}}}(x^n, Z^n, U^n) - \min_{\hat{\mathbf{X}}' \in \mathcal{G}} L_{\hat{\mathbf{X}}'}(x^n, Z^n, U^n) \geq \varepsilon + C(\ell_{max}, |\mathcal{G}|)/\sqrt{n}\right) && (49)\\
\leq\; & P\left(L_{\hat{\mathbf{X}}}(x^n, Z^n) - \min_{\hat{\mathbf{X}}' \in \mathcal{G}} L_{\hat{\mathbf{X}}'}(x^n, Z^n) \geq \varepsilon' + C(\ell_{max}, |\mathcal{G}|)/\sqrt{n}\right) && (50)\\
& + P\left(\left|L_{\hat{\mathbf{X}}}(x^n, Z^n) - L_{\hat{\mathbf{X}}}(x^n, Z^n, U^n)\right| \geq (\varepsilon - \varepsilon')/2\right) && (51)\\
& + P\left(\max_{\hat{\mathbf{X}}' \in \mathcal{G}} \left|L_{\hat{\mathbf{X}}'}(x^n, z^n) - L_{\hat{\mathbf{X}}'}(x^n, z^n, u^n)\right| \geq (\varepsilon - \varepsilon')/2\right) && (52)\\
\leq\; & 2(|\mathcal{G}|+1)\exp\left(-n\frac{\varepsilon'^2}{2L_{max}^2}\right) + 2(|\mathcal{G}|+1)\exp\left(-n\frac{(\varepsilon-\varepsilon')^2}{2\Lambda_{max}^2}\right), && (53)
\end{aligned}
$$

where the first inequality is due to (48), the triangle inequality, and (repeated use of) the union bound, while the second is due to (38), Lemma 1 and (repeated use of) the union bound. Taking $\varepsilon' = \varepsilon \frac{L_{max}}{\Lambda_{max}+L_{max}}$ we obtain

$$
\begin{aligned}
& P\quad\left(L_{\hat{\mathbf{X}}}(x^n, Z^n, U^n) - \min_{\hat{\mathbf{X}}' \in \mathcal{G}} L_{\hat{\mathbf{X}}'}(x^n, Z^n, U^n) \geq \varepsilon + C(\ell_{max}, |\mathcal{G}|)/\sqrt{n}\right)\\
\leq\; & 4(|\mathcal{G}|+1)\exp\left(-n\frac{\varepsilon^2}{2(\Lambda_{max}+L_{max})^2}\right). && (54)
\end{aligned}
$$

Inequality (12) follows by noting that $\Lambda_{max} + L_{max} \leq 2|\mathcal{X}|H_{max}\Lambda_{max}$ and that $\ell_{max} \leq |\mathcal{X}|\Lambda_{max}H_{max}$ (because $|\mathcal{X}|H_{max} \geq 1$). □

5 Competition with Finite-State Machines

5-A Finite-State Filterability

$\hat{\mathbf{X}}$ is a finite–state filter with finite state–space Ω if there exists a next–state function $g : \Omega \times \mathcal{Z} \to \Omega$, a reconstruction function $f : \Omega \times \mathcal{Z} \to \hat{\mathcal{X}}$, and an initial state $\omega \in \Omega$ such that, for $t \geq 1$,

$$\hat{X}_t(z^t) = f(\omega_t, z_t), \quad \omega_{t+1} = g(\omega_t, z_t), \quad \omega_1 = \omega. \tag{55}$$

Note that a finite–state filter is deterministic, as there is no dependence on a randomization sequence $\mathbf{U}$. It is not hard to see that there is no loss in restricting attention to non-randomized schemes in the sense that for every x^n, z^n, the finite–state filter with minimum expected loss (expectation with respect to randomization) among all (possibly randomized) schemes with a given number of states is deterministic. Let $\mathcal{G}_\Omega$ denote the class of all finite–state filters with state space Ω and let $\phi_\Omega(x^n, z^n)$ be the loss incurred by the best filter in $\mathcal{G}_\Omega$ for x^n, z^n:

$$\phi_\Omega(x^n, z^n) = \min_{\hat{\mathbf{X}} \in \mathcal{G}_\Omega} L_{\hat{\mathbf{X}}}(x^n, z^n). \tag{56}$$

For the individual pair $\mathbf{x}, \mathbf{z}$, let $\phi(\mathbf{x}, \mathbf{z})$ denote the *finite-state filterability* defined by

$$\phi(\mathbf{x}, \mathbf{z}) = \lim_{|\Omega| \to \infty} \limsup_{n \to \infty} \phi_\Omega(x^n, z^n). \tag{57}$$

Our goal, in this section, is to establish, for the semi-stochastic setting, existence of a filter $\hat{\mathbf{X}}$ which is universal in the strong sense of satisfying

$$\limsup_{n \to \infty} L_{\hat{\mathbf{X}}}(x^n, Z^n, U^n) \leq \phi(\mathbf{x}, \mathbf{Z}) \quad a.s. \ \ \forall \mathbf{x} \in \mathcal{X}^\infty. \tag{58}$$

In accord with the theme of this paper, we will achieve this goal by moving to the prediction domain, constructing a predictor which is universal with respect to the class of finite-state predictors, and transforming back to the filtering setting.

We note the easily verifiable fact that, for fixed $\mathbf{x}$, $\phi(\mathbf{x}, \mathbf{z})$ is invariant to a change in any finite number of components of $\mathbf{z}$. Combined with the Kolmogorov zero-one law (cf., e.g., [14]),[5] this observation implies the existence of a deterministic constant $\phi(\mathbf{x})$ such that

$$\phi(\mathbf{x}, \mathbf{Z}) = \phi(\mathbf{x}) \quad a.s. \tag{59}$$

This result is analogous to the noisy prediction [36, Theorem 15] and denoising [37, Claim 1] cases.

5-B Finite-State Predictability

Consider the generic prediction setting of Section 3. F is said to be a finite–state predictor with state space Ω if there exists a next state function $\hat{g} : \Omega \times \mathcal{Y} \to \Omega$, an action function $\hat{f} : \Omega \to \mathcal{A}$, and an initial state ω such that

$$F_t(y^{t-1}) = \delta_{\hat{f}(\omega_t)}, \quad \omega_{t+1} = \hat{g}(\omega_t, y_t), \quad \omega_1 = \omega, \tag{60}$$

where $\delta_a \in \mathcal{M}(\mathcal{A})$ denotes the degenerate simplex member assigning probability 1 to a and 0 to all the rest. Let $\mathcal{F}_\Omega$ denote the class of all finite–state predictors with state space Ω and define

$$\lambda_\Omega(y^n) = \min_{F \in \mathcal{F}_\Omega} L_F(y^n), \tag{61}$$

namely, the loss incurred by the best predictor in $\mathcal{F}_\Omega$ for y^n. As is well known, and easy to see, there is no loss in restricting attention, as we have, to a reference class of machines with deterministic predictions in the sense that the minimum in (56) would be achieved by a deterministic machine even had $\mathcal{F}_\Omega$ been defined to include machines with stochastic predictions. This is in analogy to the situation for the filtering problem, as mentioned previously. For the individual sequence $\mathbf{y}$ let now

$$\lambda(\mathbf{y}) = \lim_{|\Omega| \to \infty} \limsup_{n \to \infty} \lambda_\Omega(y^n). \tag{62}$$

[5] Specifically, we use the fact implied by the Kolmogorov zero-one law that if the function $f(a_1, a_2, \ldots)$ is invariant to changes in a finite number of a_i-s then $f(A_1, A_2, \ldots)$ is almost surely constant provided $A_1, A_2, \ldots$ are independent.

There are various results in universal prediction that imply the existence of a predictor P satisfying

$$\limsup_{n\to\infty} L_P(y^n) \le \lambda(\mathbf{y}) \quad \forall \mathbf{y} \in \mathcal{Y}^\infty. \tag{63}$$

One example is the following extension of the incremental parsing predictor in [17, Section V] which was designed for the binary case and Hamming loss: The predictor sequentially parses the sequence into distinct phrases, starting with the empty phrase, such that each phrase is the shortest string which is not a previously parsed phrase. Let $y^{t'}$ be the string obtained by concatenating all complete phrases in the parsing of y^t, $0 \le t' \le t$, and denote $q(y^t) = y_{t'+1} \cdots y_t$ (defined to be the empty string in case $t' = t$). We will refer to $q(y^t)$ as the *context* in which the symbol y_{t+1} occurs. We let $\mathbf{m}(y^{t-1}, q) \in \mathbb{R}^{\mathcal{Y}}$ denote the vector whose y-th component is the number of occurrences of symbol y in context q along y^{t-1}, and we further denote $\mathbf{w}(y^{t-1}) = \mathbf{m}(y^{t-1}, q(y^{t-1}))$. The prediction for time t is then given by

$$P_t(y^{t-1})[a] = \Pr\left\{ b\left(\mathbf{w}(y^{t-1}) + \mathbf{U}\sqrt{\sum_{y\in\mathcal{Y}} \mathbf{w}(y^{t-1})[y] + 1} \right) = a \right\}, \tag{64}$$

where $\mathbf{U}$ is uniformly distributed on $\left[0, \frac{6}{|\mathcal{Y}|}\right]^{|\mathcal{Y}|}$ and $b(\cdot)$ is the Bayes Response defined in (19). Clearly, this predictor can be efficiently implemented by growing a tree. In the appendix, we show that it satisfies (63). A predictor based on incremental parsing which, as this one, also uses ideas from [18], is proposed in [33].

5-C FS Filter is Mapped to FS Predictor with Same Number of States

Throughout the remainder of this section, when referring to the prediction setting, we take, as in Section 4, $\mathcal{Y} = \mathcal{Z}$, $\mathcal{A} = \mathcal{S}$, and $\ell(z, s) = h(z)^T \cdot \rho(s)$. We note first that if $\hat{\mathbf{X}} \in \mathcal{G}_\Omega$, with associated reconstruction and next-state functions f and g, respectively, then $F^{\hat{\mathbf{X}}} \in \mathcal{F}_\Omega$ with action and next-state functions $\hat{f}(\omega_t) = f(\omega_t, \cdot)$ and $\hat{g} = g$, respectively, and the same initial state. Conversely, $F \in \mathcal{F}_\Omega$ implies that $\hat{\mathbf{X}}^F \in \mathcal{G}_\Omega$, with the same correspondence between $f, g, \hat{f}$ and $\hat{g}$. This implies also that

$$\{\hat{\mathbf{X}}^F : F \in \mathcal{F}_\Omega\} = \mathcal{G}_\Omega. \tag{65}$$

Theorem 6 *For any predictor P satisfying (63),*

$$\limsup_{n\to\infty} L_{\hat{\mathbf{X}}^P}(x^n, Z^n, U^n) \le \phi(\mathbf{x}, \mathbf{Z}) \quad a.s. \;\; \forall \mathbf{x} \in \mathcal{X}^\infty. \tag{66}$$

In words, if P is a universal predictor then $\hat{\mathbf{X}}^P$ is a universal filter.

Proof of Theorem 6: Fixing a finite set of states Ω, it will suffice to show that if

$$\limsup_{n\to\infty} L_P(z^n) \le \limsup_{n\to\infty} \lambda_\Omega(z^n) \quad \forall \mathbf{z} \in \mathcal{Z}^\infty \tag{67}$$

then

$$\limsup_{n\to\infty} L_{\hat{\mathbf{X}}^P}(x^n, Z^n, U^n) \le \limsup_{n\to\infty} \phi_\Omega(x^n, Z^n) \quad a.s. \;\; \forall \mathbf{x} \in \mathcal{X}^\infty. \tag{68}$$

To this end, fix $\mathbf{x} \in \mathcal{X}^\infty$ and note that

$$\begin{aligned}
& P\left(|\phi_\Omega(x^n, Z^n) - \lambda_\Omega(Z^n)| \ge \varepsilon\right) && (69)\\
= \;& P\left(\left|\min_{\hat{\mathbf{X}}\in\mathcal{G}_\Omega} L_{\hat{\mathbf{X}}}(x^n, Z^n) - \min_{F\in\mathcal{F}_\Omega} L_F(Z^n)\right| \ge \varepsilon\right) && (70)\\
= \;& P\left(\left|\min_{F\in\mathcal{F}_\Omega} L_{\hat{\mathbf{X}}^F}(x^n, Z^n) - \min_{F\in\mathcal{F}_\Omega} L_F(Z^n)\right| \ge \varepsilon\right) && (71)\\
\le \;& P\left(\max_{F\in\mathcal{F}_\Omega} \left|L_{\hat{\mathbf{X}}^F}(x^n, Z^n) - L_F(Z^n)\right| \ge \varepsilon\right) && (72)\\
\le \;& 2|\mathcal{F}_\Omega| \exp\left(-n\frac{2\varepsilon^2}{L_{max}^2}\right), && (73)
\end{aligned}$$

where (71) follows by (65), and the last inequality by Theorem 4. Consequently, assuming (67), we have a.s.

$$\begin{aligned}
\limsup_{n\to\infty} L_{\hat{\mathbf{X}}^P}(x^n, Z^n, U^n) &= \limsup_{n\to\infty} L_{\hat{\mathbf{X}}^P}(x^n, Z^n) && (74)\\
&= \limsup_{n\to\infty} L_P(Z^n) && (75)\\
&\le \limsup_{n\to\infty} \lambda_\Omega(Z^n) && (76)\\
&= \limsup_{n\to\infty} \phi_\Omega(x^n, Z^n), && (77)
\end{aligned}$$

where, with repeated use of the Borel-Cantelli lemma, (74) follows from Lemma 1, (75) from Theorem 4, (76) follows from (67), and (77) from (73). □

5-D Markov Filters are as Good as Finite-State Filters

F will be said to be a *Markov* or *finite-memory predictor of order* k, if there exists a mapping $f : \mathcal{Y}^k \to \mathcal{M}(\mathcal{A})$ such that for all $t > k$ and $y^{t-1} \in \mathcal{Y}^{t-1}$

$$F_t(y^{t-1}) = f(y_{t-k}^{t-1}). \tag{78}$$

We let $\mathcal{M}_k$ denote the class of all Markov predictors of order k. In universal prediction, it is known since [17, 22] that the finite-state performance can be attained by predictors from the much smaller class of "finite-memory", or "Markov", predictors. As we now show, the prediction-filtering correspondence implies that an analogous fact is true for the filtering problem as well. A *Markov filter of order* k is a finite-state filter with state space $\Omega = \mathcal{Z}^k$ and $\omega_t = z_{t-k}^{t-1}$. We let $\mathcal{G}_k$ denote the class of finite-memory filters of order k and let $\mu_k(x^n, z^n)$ denote the loss incurred by the best k-th order Markov filter (when observing z^n while the underlying noiseless sequence is x^n), in analogy to (56):

$$\mu_k(x^n, z^n) = \min_{\hat{\mathbf{X}} \in \mathcal{G}_k} L_{\hat{\mathbf{X}}}(x^n, z^n). \tag{79}$$

Denoting

$$\mu(\mathbf{x}, \mathbf{z}) = \lim_{k \to \infty} \limsup_{n \to \infty} \mu_k(x^n, z^n), \tag{80}$$

we have the following result:

Theorem 7

$$\mu(\mathbf{x}, \mathbf{Z}) = \phi(\mathbf{x}, \mathbf{Z}) \quad a.s. \;\; \forall \mathbf{x} \in \mathcal{X}^\infty. \tag{81}$$

Proof: For any fixed k, the inequality

$$P\left(\left| \mu_k(x^n, Z^n) - \min_{F \in \mathcal{M}_k} L_F(Z^n) \right| \geq \varepsilon \right) \leq 2|\mathcal{G}_k| \exp\left(-n \frac{2\varepsilon^2}{L_{max}^2} \right) \tag{82}$$

follows analogously as in the chain leading to (73). For any finite set of states Ω, Theorem 2 of [22] shows the existence of a constant C depending only on the loss function l, such that for all k, n, and z^n

$$\min_{F \in \mathcal{M}_k} L_F(z^n) \leq \lambda_\Omega(z^n) + \left(\frac{2C \log |\Omega|}{k+1} \right)^{1/2}, \tag{83}$$

where $\lambda_\Omega(z^n)$ is the Ω-state predictability defined in (56).[6] Inequalities (82) and (83) imply

$$\limsup_{n \to \infty} \mu_k(x^n, Z^n) \leq \limsup_{n \to \infty} \lambda_\Omega(Z^n) + \left(\frac{C \log |\Omega|}{k+1} \right)^{1/2} = \limsup_{n \to \infty} \phi_\Omega(x^n, Z^n) + \left(\frac{C \log |\Omega|}{k+1} \right)^{1/2} \quad a.s., \tag{84}$$

where the equality follows from (73) and the Borel-Cantelli lemma. It follows that

$$\mu(\mathbf{x}, \mathbf{Z}) = \lim_{k \to \infty} \limsup_{n \to \infty} \mu_k(x^n, Z^n) \leq \limsup_{n \to \infty} \phi_\Omega(x^n, Z^n) \quad a.s., \tag{85}$$

implying $\mu(\mathbf{x}, \mathbf{Z}) \leq \phi(\mathbf{x}, \mathbf{Z})$ a.s. by the arbitrariness of Ω. Combined with the obvious fact that $\mu(\mathbf{x}, \mathbf{Z}) \geq \phi(\mathbf{x}, \mathbf{Z})$ (for all realizations), we obtain (81). □

5-E A Universal Incremental Parsing Filter

To conclude this section, we detail the form of the "incremental parsing filter" $\hat{\mathbf{X}}^P$ obtained when P is the incremental parsing predictor detailed in Subsection 5-B. To this end, recall that $\mathcal{S}$ is the (finite) set of mappings taking $\mathcal{Z}$ into $\hat{\mathcal{X}}$, i.e., $\mathcal{S} = \left\{ s : \mathcal{Z} \to \hat{\mathcal{X}} \right\}$. For $\xi \in \mathbb{R}^{\mathcal{Z}}$, let $B_H(\xi, \cdot) \in \mathcal{S}$ be defined by

$$B_H(\xi, z) = \arg\min_{\hat{x}} \xi^T \cdot H \cdot [\boldsymbol{\lambda}_{\hat{x}} \odot \boldsymbol{\pi}_z], \tag{86}$$

[6] For $p, q \in \mathcal{M}(\mathcal{Y})$, letting $\Delta(p\|q) = p^T \cdot \mathcal{L}_{b(q)} - p^T \cdot \mathcal{L}_{b(p)} = p^T \cdot \mathcal{L}_{b(q)} - U_l(p)$ (recall notation from Section 3), Theorem 2 of [22] asserts that $\min_{F \in \mathcal{M}_k} L_F(z^n) \leq \lambda_\Omega(z^n) + [(2C \log |\Omega|)/(k+1)]^{\delta/2}$ for δ satisfying $\Delta(p\|q) \leq C\|p-q\|_1^\delta$. Inequality (83) is obtained by taking $\delta = 1$, which is justified since in our finite-alphabet setting $\Delta(p\|q) \leq l_{max}\|p-q\|_1$ (cf., e.g., [37, Lemma 1] for a proof).

where, for vectors v_1 and v_2 of equal dimensions, $v_1 \odot v_2$ denotes the vector obtained by componentwise multiplication, $\boldsymbol{\pi}_z$ denotes the column of the channel matrix Π corresponding to channel output z, and $\boldsymbol{\lambda}_{\hat{x}}$ denotes the $\hat{x}$th column of the loss matrix Λ. Note that

$$B_H(\xi, \cdot) = \arg\min_{s \in \mathcal{S}} \sum_z \xi^T \cdot H \cdot [\boldsymbol{\lambda}_{s(z)} \odot \boldsymbol{\pi}_z], \tag{87}$$

since minimizing the sum on the right-hand side of (87) boils down to minimizing the summand with respect to $s(z)$ for each z. Thus,

$$\begin{aligned} B_H(\xi, \cdot) &= \arg\min_{s \in \mathcal{S}} \xi^T \cdot H \cdot \rho(s) = \arg\min_{s \in \mathcal{S}} \sum_z \xi(z)[h^T(z) \cdot \rho(s)] \\ &= \arg\min_{s \in \mathcal{S}} \sum_z \xi(z) l(z, s) = \arg\min_{s \in \mathcal{S}} \xi^T \cdot \mathcal{L}_s = b(\xi), \end{aligned} \tag{88}$$

where the first equality follows by (87) upon recalling (33) which implies $\rho(s) = \sum_z \boldsymbol{\lambda}_{s(z)} \odot \boldsymbol{\pi}_z$, the equality before the right-most one follows by recalling from Section 3 that $\mathcal{L}_s$ denotes the column of the matrix of the loss function ℓ corresponding to s, and where the right-most equality follows by the definition of the Bayes response $b(\xi)$ in (19) (for our particular source and action alphabets, and loss function). With this notation it is now easy to see, from the definition of the transformation from a predictor to a filter, that the filter obtained by transforming the incremental parsing predictor detailed in Subsection 5-B to the filtering domain is equivalent to the filter which parses the source sequence and assigns contexts and counts exactly as does the incremental parsing predictor. Since the predictor is the Bayes response to the perturbed $\mathbf{w}(z^{t-1})$ (recall Equation (64)), it follows from (88) that the reconstruction given by the corresponding filter at time t is

$$\hat{X}_t^P(z^t, U_t) = B_H\left(\left[\mathbf{w}(y^{t-1}) + T(U_t)\sqrt{\sum_{y \in \mathcal{Y}} \mathbf{w}(y^{t-1})[y] + 1}\right], z_t\right), \tag{89}$$

where $T(\cdot)$ is any transformation under which $T(U_t)$ is uniformly distributed on the cube[7] $\left[0, \sqrt{\frac{6}{|\mathcal{Z}|}}\right]^{|\mathcal{Z}|}$. By Theorem 6, this "incremental parsing filter" is universal in the strong sense of satisfying (66), since the predictor it is based on satisfies (63).

6 Discussion

The problem of competing with Markov filters of a given order in a semi-stochastic setting of an individual sequence corrupted by memoryless noise has been considered, prior to our work, under the

[7] Perhaps a slightly less convoluted presentation of this filter would have been to replace $T(U_t)$ with a random variable uniformly distributed on the cube $\left[0, \sqrt{\frac{6}{|\mathcal{Z}|}}\right]^{|\mathcal{Z}|}$. The presentation in (89), however, complies with our formal filtering problem definition, which assumes the randomization variable to be uniform on $[0, 1]$.

label of "the extended sequential compound decision problem" in [1, 2, 30, 31]. While the main focus of our work is the association between filtering and prediction, the results herein also generalize the extended sequential compound decision problem in several directions.

First, we note that Theorem 1 and Corollary 1 apply to a generic finite class of filters, with Markov filters of a given order as a special case. The setting of the extended sequential compound decision problem corresponds to Part 1 of Theorem 1 (as applied to Markov filters of a given order). Here, in Part 2 of the theorem, we also show measure concentration properties. Moreover, as noted at the end of Section 2, Corollary 1 corresponds to a stronger benchmark.

Second, the assumption in the extended compound decision problem was of a fixed and known Markov order k. Here, in Section 5, we consider competition against filters of any order, in the spirit of [40].

For a Markov filter of order k, one can generalize Hannan's predictor of Theorem 3 just as in Equation (64), where the context $q(y^{t-1})$ is replaced with z_{t-k}^{t-1}. The corresponding filter is thus given by Equation (89) with, again, $\mathbf{w}(y^{t-1})$ replaced with $\mathbf{m}(z^{t-1}, z_{t-k}^{t-1})$. Now, notice that in the case $|\mathcal{X}| = |\mathcal{Z}|$, Equation (10) necessitates $H = \Pi^{-1}$, implying $B_H(\xi, z) = \arg\min_{\hat{x}} \xi^T \cdot \Pi^{-1} \cdot [\boldsymbol{\lambda}_{\hat{x}} \odot \boldsymbol{\pi}_z]$. This observation shows, by comparing with display (12) of [37], that the DUDE algorithm (operating on kth-order double-sided contexts) is given by $\hat{X}_i = B_H\left(\mathbf{m}(z^n, z_{i-k}^{i-1}, z_{i+1}^{i+k}), z_i\right)$, where $\mathbf{m}(z^n, z_{i-k}^{i-1}, z_{i+1}^{i+k})$ is now the vector of counts associated with the number of occurrences of the various noisy symbols in the double-sided context $(z_{i-k}^{i-1}, z_{i+1}^{i+k})$ along the sequence z^n. Thus, up to the randomization, the filter is operating as the DUDE, but using the counts learned from z^{t-1} only (and in a one-sided manner). Notice that the above described filter does not utilize all the count information at our disposal at time t, which is summarized in the vector $\mathbf{m}(z^t, z_{t-k}^{t-1})$. Whether a filter based on the latter vector (which is intuitively more appealing) is universal, remains open.

The filter that our approach gives rise to is randomized, even when competing with a reference class of deterministic filters. This is because it inherits the randomization of its corresponding predictor (which competes with a set of predictors in an individual sequence setting and, hence, need be randomized). We conjecture, however, that in the filtering setting such randomization is not necessary since the channel noise can induce the required randomization. Specifically, our conjecture is that the filter obtained by transforming from the prediction domain the predictor which imitates, deterministically, the best predictor at each time-point, will successfully compete with any finite reference class of filters. Van-Ryzin's results in [27] imply that this conjecture holds for the reference class of constant "symbol by symbol" filters. Whether or not randomization is necessary for competing with a general finite reference class of filters remains open.

Similarly, the question of whether or not the incremental parsing filter of Subsection 5-E would remain universal without the random perturbation term in (89) is also open. In other words, would an incremental parsing filter that uses the "single-letter" denoising rule of the DUDE, using the counts $\mathbf{w}(z^{t-1})$ be universal? We note that in [24], building on the results of [27], a deterministic filter was constructed and shown to be universal. In that filter, the contexts are also determined by incremental parsing, but of the "counting sequence" instead of the noisy source sequence itself (and, hence, was dubbed as the "static IP filter"). This type of parsing guarantees that occurrences of the same context will not overlap. The idea of avoiding context overlap was pioneered in [30] to obtain a deterministic filter which competes with the class of Markov filters of a given order k. Based on that filter, it is possible to compete against Markov filters of *any* order by letting k increase slowly, as shown in [17] for the prediction setting. Theorem 7 then shows that this approach, while less elegant than the incremental parsing one, also competes against the class of FS filters.

7 Conclusion

We considered the problem of filtering a DMC-corrupted individual sequence with respect to an arbitrary reference class of filtering schemes. We established a correspondence between filtering and prediction that allowed us to capitalize on known results from prediction relative to a set of experts for constructing a filter competing with any finite reference class of filters. Thus, our work joins the results of [34] in showing that problems involving sequential decision making in the presence of noise can be reduced to the problem of noiseless prediction by appropriately modifying the loss function and the prediction space.

Our results also generalize the extended sequential compound decision problem in several directions: The special case of Markov filters of a given order is extended to a generic class of experts, we provide results on measure concentration which result in competition with a stronger benchmark, and we consider competition against filters of any order, in the spirit of [40].

We confined our interest in this work to the semi-stochastic setting of a noiseless individual sequence. We point out, however, that analogously as was shown to be the case for denoising in [37], optimality in the semi-stochastic setting implies optimality in the fully stochastic setting. It is straightforward to extend the proof ideas in Section VI of [37] to the sequential setting of the present work and show that any filter which is universal in the sense of satisfying (58) is also universal in the stochastic setting. More specifically, the asymptotic expected filtering loss of such a filter achieves the distribution-dependent optimum for any stationary noiseless process, and its actual (rather than expected) filtering loss attains the optimum performance with probability one if the process is also ergodic.

Appendix

A Proof of Lemma 2

Fix $x^n \in \mathcal{X}^n$. Consider

$$
\begin{aligned}
E\left[\sum_{\hat{x}} \Lambda(x_t,\hat{x}) P_{\hat{\mathbf{X}}^F}(Z^t)[\hat{x}] \middle| Z^{t-1}\right] &= \sum_z \sum_{\hat{x}} \Lambda(x_t,\hat{x}) P_{\hat{\mathbf{X}}^F}(Z^{t-1}z)[\hat{x}]\Pi(x_t,z) && \text{(A.1)}\\
&= \sum_z \Pi(x_t,z) \sum_{\hat{x}} \Lambda(x_t,\hat{x}) \left[\sum_{s:s(z)=\hat{x}} F_t(Z^{t-1})[s]\right] && \text{(A.2)}\\
&= \sum_z \Pi(x_t,z) \sum_s \Lambda(x_t,s(z)) F_t(Z^{t-1})[s] && \text{(A.3)}\\
&= \sum_s \rho_{x_t}(s) F_t(Z^{t-1})[s] && \text{(A.4)}\\
&= \sum_s [\boldsymbol{\delta}_{x_t}^T \cdot \rho(s)] F_t(Z^{t-1})[s] && \text{(A.5)}\\
&= E\left[\sum_s [h(Z_t)^T \cdot \rho(s)] F_t(Z^{t-1})[s] \middle| Z^{t-1}\right] && \text{(A.6)}\\
&= E\left[\sum_s \ell(Z_t,s) F_t(Z^{t-1})[s] \middle| Z^{t-1}\right], && \text{(A.7)}
\end{aligned}
$$

where $\boldsymbol{\delta}_x \in \mathbb{R}^{\mathcal{X}}$ denotes the column vector all of whose components are 0 except for the x-th one which is 1, equality (A.1) follows from the independence of Z_t and Z^{t-1}, equality (A.2) follows from (25), and equality (A.6) follows again by the independence of Z_t and Z^{t-1}, combined with $Eh(Z_t) = \boldsymbol{\delta}_{x_t}$ (which is the requirement in (9)). Subtracting (A.7) from the left hand side of (A.1) gives

$$
E\left[\sum_{\hat{x}} \Lambda(x_t,\hat{x}) P_{\hat{\mathbf{X}}^F}(Z^t)[\hat{x}] \middle| Z^{t-1}\right] - E\left[\sum_s \ell(Z_t,s) F_t(Z^{t-1})[s] \middle| Z^{t-1}\right] = 0. \qquad \text{(A.8)}
$$

Now, (7) implies

$$
\begin{aligned}
& n\left[L_{\hat{\mathbf{X}}^F}(x^n,Z^n) - L_F(Z^n)\right] - (n-1)\left[L_{\hat{\mathbf{X}}^F}(x^{n-1},Z^{n-1}) - L_F(Z^{n-1})\right]\\
= \; & \sum_{\hat{x}} \Lambda(x_n,\hat{x}) P_{\hat{\mathbf{X}}^F}(Z^n)[\hat{x}] - \sum_s \ell(Z_n,s) F_n(Z^{n-1})[s], \qquad \text{(A.9)}
\end{aligned}
$$

which when combined with (A.8) gives

$$
E\left[n\left[L_{\hat{\mathbf{X}}^F}(x^n,Z^n) - L_F(Z^n)\right] \middle| Z^{n-1}\right] = (n-1)\left[L_{\hat{\mathbf{X}}^F}(x^{n-1},Z^{n-1}) - L_F(Z^{n-1})\right]. \qquad \text{(A.10)}
$$

□

B Proof of Universality of the Incremental Parsing Predictor

We assume here the generic setting of Section 3.

Theorem 8 *Let P denote the incremental parsing predictor described in (64). Then, for all $\mathbf{y} \in \mathcal{Y}^\infty$, n, k,*

$$L_P(y^n) \leq \min_{F \in \mathcal{M}_k} L_F(y^n) + \frac{k \cdot c(y^n) \cdot \ell_{max}}{n} + B_l \sqrt{6|\mathcal{Y}|} \sqrt{\frac{c(y^n)}{n}}, \tag{A.11}$$

where $\mathcal{M}_k$ is the set of Markov predictors of order k (as defined in Subsection 5-D) and $c(y^n)$ denotes the number of phrases in the incremental parsing of y^n.

The proof of Theorem 8 consists of applying Theorem 3 in each possible context, similarly as was done in the proof of [17, Theorem 4].

Proof of Theorem 8: By the incremental parsing rule, each context must have occurred as a phrase in the parsing of the sequence. Thus, the number of different contexts occurring along y^n is $c = c(y^n)$. For $1 \leq j \leq c$, let $\mathbf{m}^j = \mathbf{m}(y^n, q_j)$, where q_j denotes the j-th context, and let $m^j = \sum_{y \in \mathcal{Y}} \mathbf{m}^j[y]$. We divide the sequence y^n into subsequences of symbols occurring at the same context ("bins"). Applying Theorem 3 to each one of the c bins and averaging over the bins, we obtain

$$\begin{aligned} L_P(y^n) &\leq \frac{1}{n} \sum_{j=1}^{c} \left[U_l(\mathbf{m}^j) + B_l \sqrt{6|\mathcal{Y}| m^j} \right] && \text{(A.12)} \\ &\leq \left(\frac{1}{n} \sum_{j=1}^{c} U_l(\mathbf{m}^j) \right) + B_l \sqrt{6|\mathcal{Y}|} \sqrt{\frac{c}{n}}, && \text{(A.13)} \end{aligned}$$

the second inequality following by Jensen's inequality and the concavity of the square root function. Now, for any $k \geq 0$, we can write

$$\sum_{j=1}^{c} U_l(\mathbf{m}^j) = \sum_{j \in J_1} U_l(\mathbf{m}^j) + \sum_{j \in J_2} U_l(\mathbf{m}^j), \tag{A.14}$$

where J_1 is the set of indices of contexts shorter than k, and J_2 is the set of indices of contexts of length k or longer. The fact that only the first k letters in each phrase are allocated to contexts pertaining to J_1 implies

$$\sum_{j \in J_1} U_l(\mathbf{m}^j) \leq k \cdot c \cdot \ell_{max}. \tag{A.15}$$

As for the second sum on the right-hand side of (A.14), it follows from the concavity of $U_l(\cdot)$ that $U_l(v + v') \geq U_l(v) + U_l(v')$ for any $v, v' \in \mathbb{R}^{\mathcal{Y}}$. Thus, since the contexts associated with J_2 serve as refinements to contexts of length k, we obtain

$$\sum_{j \in J_2} U_l(\mathbf{m}^j) \leq n \min_{F \in \mathcal{M}_k} L_F(y^n). \tag{A.16}$$

Inequality (A.11) follows by combining (A.13), (A.14), (A.15) and (A.16). □

Proof that (63) is satisfied by the incremental parsing predictor in (64): Theorem 8 and the fact that $\frac{1}{n}\max_{y^n} c(y^n) = O(1/\log n)$ (cf., e.g., [11, Lemma 12.10.1]) imply

$$\limsup_{n\to\infty} L_P(y^n) \le \lim_{k\to\infty} \limsup_{n\to\infty} \min_{F\in\mathcal{M}_k} L_F(y^n) = \lambda(\mathbf{y}), \tag{A.17}$$

where the equality is due to Theorem 2 of [22]. □

References

[1] R. J. Ballard. *Extended rules for the sequence compound decision problem with $m \times n$ component.* PhD thesis, Michigan State University, 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 University, 1974.

[3] A. Baruch. Universal algorithms for sequential decision in the presence of noisy observations. Master's thesis, Technion, Haifa, Israel, February 1999.

[4] A. Baruch and N. Merhav. Universal filtering and prediction of individual sequences corrupted by noise. *Proc. 37th Annu. Allerton Conf. Communication, Control, and Computing*, pages 470–479, September 1999.

[5] D. Blackwell. An analog of the minmax theorem for vector payoffs. *Pacific J. Math*, 6:1–8, 1956.

[6] D. Blackwell. Controlled random walks. *Proceedings International Congress of Mathematicians*, 3:336–338, 1956. Amsterdam: North Holland.

[7] N. Cesa-Bianchi, Y. Freund, D. P. Helmbold, D. Haussler, R. Schapire, and M. K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

[8] N. Cesa-Bianchi and G. Lugosi. On prediction of individual sequences. *Ann. Stat.*, 27(6):1865–1895, 1999.

[9] T. M. Cover. Behavior of sequential predictors of binary sequences. *Transactions of the Fourth Prague Conference on Information Theory*, September 1966. Prague.

[10] T. M. Cover and A. Shenhar. Compound Bayes predictors for sequences with apparent Markov structure. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-7(6):421–424, June 1977.

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

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

[13] R. L. Dobrushin and B. S. Tsybakov. Information transmission with additional noise. *IEEE Trans. Inform. Theory*, 8:293–304, September 1962.

[14] R. Durret. *Probability: Theory and Examples.* Duxbury Press, Belmont, California, 1991.

[15] Y. Ephraim and R. M. Gray. A unified approach for encoding clean and noisy sources by means of waveform and autoregressive model vector quantization. *IEEE Trans. Inform. Theory*, 34(4):826–834, July 1988.

[16] Y. Ephraim and N. Merhav. Hidden Markov processes. *IEEE Trans. Inform. Theory*, 48(6):1518–1569, June 2002.

[17] M. Feder, N. Merhav, and M. Gutman. Universal prediction of individual sequences. *IEEE Trans. Inform. Theory*, 38:1258–1270, July 1992.

[18] J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, III:97–139, 1957. Princeton, NJ.

[19] J. F. Hannan and H. Robbins. Asymptotic solutions of the compound decision problem for two completely specified distributions. *Ann. Math. Statist.*, 26:37–51, 1955.

[20] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Amer. Statist. Assoc.*, 58:13–30, 1963.

[21] P. Lancaster and M. Tismenetsky. *The Theory of Matrices.* Academic, Orlando, 1985.

[22] N. Merhav and M. Feder. Universal schemes for sequential decision from individual data sequences. *IEEE Trans. Inform. Theory*, 39(4):1280–1292, July 1993.

[23] N. Merhav and M. Feder. Universal prediction. *IEEE Trans. Inform. Theory*, 44(6):2124–2147, October 1998.

[24] E. Ordentlich, T. Weissman, M. J. Weinberger, A. Somekh-Baruch, and N. Merhav. Discrete universal filtering through incremental parsing. In *Proc. 2004 Data Compression Conference (DCC'04)*, pages 352–361, Snowbird, Utah, USA, March 2004.

[25] H. Robbins. Asymptotically subminimax solutions of compound statistical decision problems. *Proc. Second Berkeley Symp. Math. Statis. Prob.*, pages 131–148, 1951.

[26] J. Van Ryzin. The compound decision problem with $m \times n$ finite loss matrix. *Ann. Math. Statist.*, 37:412–424, 1966.

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

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

[29] C. E. Shannon. Channels with side information at the transmitter. *IBM J. Res. Dev.*, 2:289–293, 1958.

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

[31] S. B. Vardeman. Approximation to minimum k-extended Bayes risk in sequences of finite state decision problems and games. *Bulletin of the Institute of Mathematics Academia Sinica*, 10(1):35–52, March 1982.

[32] V. Vovk. Aggregating strategies. *Proc. 3rd Annu. Workshop on Computational Learning Theory*, pages 371–383, 1990. San Mateo, CA: Kaufmann.

[33] M. J. Weinberger and E. Ordentlich. On-line decision making for a class of loss functions via Lempel-Ziv parsing. In *Proc. 2000 Data Compression Conference (DCC'00)*, pages 163–172, Snowbird, Utah, USA, March 2000.

[34] T. Weissman and N. Merhav. Universal prediction of individual binary sequences in the presence of noise. *IEEE Trans. Inform. Theory*, 47(6):2151–2173, September 2001.

[35] T. Weissman and N. Merhav. On limited-delay lossy coding and filtering of individual sequences. *IEEE Trans. Inform. Theory*, 48(3):721–733, March 2002.

[36] T. Weissman, N. Merhav, and A. Baruch. Twofold universal prediction schemes for achieving the finite-state predictability of a noisy individual binary sequence. *IEEE Trans. Inform. Theory*, 47(5):1849–1866, July 2001.

[37] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger. Universal discrete denoising: Known channel. *IEEE Trans. Inform. Theory*, 51(1):5–28, January 2005.

[38] H. S. Witsenhausen. Indirect rate distortion problems. *IEEE Trans. Inform. Theory*, 26(5):518–521, September 1980.

[39] C. H. Zhang. Compound decision theory and empirical Bayes methods. *Annals of Statistics*, 31(2):379–390, 2003.

[40] J. Ziv and A. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Trans. Inform. Theory*, 24(5):530–536, September 1978.