Correspondence

Universal Switching FIR Filtering

Taesup Moon

***Abstract*—We revisit recently considered universal finite-impulse-response (FIR) filtering problem and devise a scheme that asymptotically attains the expected mean-square error (MSE) of the best *switching* FIR filters for every underlying bounded, real-valued signal, provided that the switch rate of the best filters are sufficiently slow. As a performance metric, we consider adaptive expected regret, the maximum difference between the expected MSE of our filter and that of the best FIR filter over any contiguous time interval. Our algorithm is shown to have $O(\log^2 n)$ bound on the adaptive expected regret with $O(n^2)$ time-complexity, where n is the length of the signal; the bound implies that, regardless of the underlying signal, the expected MSE of our filter universally converges to that of the best switching FIR filters, if the number of switches is $o(\frac{n}{\log^2 n})$. The experimental results show that our filter outperforms its stationary counterpart particularly when the underlying signal has time-varying characteristics. We also show a heuristic scheme with $O(n)$ time-complexity works well without losing too much of the filtering performance.**

***Index Terms*—Competing with compound experts, FIR MMSE filtering, non-stationary signal, regret minimization, universal filter.**

I. Introduction

Estimating the real-valued signal based on its noisy observation is a fundamental problem in signal processing. Recently, [1] developed a universal finite-duration-impulse-response (FIR) causal estimation scheme, or *filter*,[1] which essentially attains the mean-square error (MSE) of the best FIR filter for every underlying clean signal. The algorithm only assumed the knowledge of the noise variance and the time-independence of the noise. The result in [1] is strong that their algorithm achieves the performance of the Bayes optimal solution when the underlying clean signal is a stationary process, and the algorithm may outperform conventional Bayesian schemes when the prior on the clean signal may be wrong or may not be available.

Although the result in [1] holds for every bounded underlying signal, the performance of the algorithm may be limited when the best FIR filter that the algorithm competes with does not perform well enough, for example, when the underlying signal characteristic changes over time. In order to cope with such nonstationarity of the signal, we extend the scheme in [1] to devise a universal filter that not only competes with the best *fixed* FIR filter, but also with the best *switching* FIR filters which have freedom to switch between filters that are tailored to specific segments of the signal to minimize the overall MSE, for every underlying signal. We show that if the switches between such filters do not occur too frequently, or more concretely, if the number of switches is $o\left(\frac{n}{\log^2 n}\right)$ where n is the signal length, our universal filter can still achieve the performance of the best switching filters as the signal length n increases.

We use the framework of competing with *compound experts*, which has obtained attentions from a number of different domains of research in order to combat with the nonstationarity of the data. Particularly, this work is closely related to the recent work in sequential prediction [2], [3] and online learning [4]; [2] devised an online least-squares prediction algorithm that competes with the best piecewise linear predictor for bounded real-valued signals, and [3] extended [2] to the noisy prediction problem in which the prediction on the clean signal is made based on the i.i.d. additive noise corrupted past symbols. Reference [4] considered a general online optimization problem and devised an algorithm that competes with hindsight optimal switching solutions for a certain class of convex loss functions.

In this correspondence, we combine tools and results of [1] and [4] and implement necessary variations for our filtering problem to compete with best switching FIR filters. We first note that, although developed from different principles, the main results of [2] and [4] are very similar except for the constant factors in the regret bound. Due to the simpler argument, we mainly adopt the work in [4], but the results in [2] can be used in the same way.

Moreover, the difference between filtering and noisy prediction problems, which is also pointed out in [1], requires variations for the scheme in [4] (or [2]); in noisy prediction, it is shown in [3] that although the clean signal is not observable, applying the same algorithm in [2] on the noisy signal as if the goal is to predict the noise-corrupted signal was good enough. However, the situation is different in filtering problem, in which the noisy symbols up to the time index t are used in order to estimate the clean symbol at t. Namely, while [1] casted filtering to a prediction problem with a new loss function (not the squared loss), it is not straightforward to apply the algorithm in [4] (or [2]) directly to the converted prediction problem. The reason is that the new loss function devised in [1] does not satisfy the condition, the exp-concavity, that is necessary for the algorithm to work. We thus implement a trick to work with blocks of the loss functions in order to satisfy the condition for the algorithm. The problem specific technical results needed in the course of our arguments are inherited from [1].

The rest of the correspondence is organized as follows. Section II gives brief notation and problem setting, Section III presents the main result of the correspondence and the analysis of it, then Section IV gives the experimental results for the proposed algorithm.

II. Notation and Problem Setting

A. Notations

We follow the notations in [1]. We let $\{x_t\}_{t=1}^n$, $\{N_t\}_{t=1}^n$, and $\{Y_t\}_{t=1}^n$ denote the real-valued clean, additive noise, and additive noise-corrupted signals with length n, respectively. We assume $x_t \in \mathcal{D} = [-K, K] \subset \mathbb{R}$ for some $K < \infty$ without any additional statistical assumptions, $\{N_t\}_{t=1}^n$ is also bounded and i.i.d. with mean zero and known variance σ^2, and

$$Y_t = x_t + N_t, \text{ for } 1 \le t \le n.$$

The bold face notations denote the d-dimensional column vector of d recent symbols, e.g., $\mathbf{Y}_t = [Y_t, \ldots, Y_{t-(d-1)}]^T$, where $(\cdot)^T$ is a transpose operator. Sometimes, $\cdot$ notation indicates an inner product

Manuscript received January 26, 2011; revised August 10, 2011; accepted November 10, 2011. Date of publication November 22, 2011; date of current version February 10, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Suleyman S. Kozat.

The author is with the Yahoo! Labs, Sunnyvale, CA 94089 USA (e-mail: tsmoon@ymail.com).

Color versions of one or more of the figures in this correspondence are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2011.2176931

[1]We refer filter to the causal estimator throughout the correspondence.

between two vectors. Also, denote $\mathbf{c} = [\sigma^2, 0, \ldots, 0]^T \in \mathbb{R}^d$. $\|\cdot\|$ denotes the standard Euclidean norm for vectors. $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ stand for the minimum and maximum eigenvalues for the argument matrices, respectively. $\sigma(\cdot)$ stands for the sigma-algebra generated from the argument random variables.

A filter $\hat{\mathbf{X}}$ is defined to be a sequence of mappings $\{\hat{X}_t(\cdot)\}_{t\geq 1}$, where $\hat{X}_t(\cdot) : \mathbb{R}^t \to \mathbb{R}$ and $\hat{X}_t(Y^t)$ is the causal estimator of x_t based on the noisy observation $Y^t = (Y_1, \ldots, Y_t)$. The performance of a filter is measured by the MSE,

$$\frac{1}{n}\sum_{t=1}^{n}\left(x_t - \hat{X}_t(Y^t)\right)^2,$$

and as in [1], we focus on the class of FIR filters of order d, i.e., the filters of the form $\hat{X}_t(Y^t) = \mathbf{u}^T\mathbf{Y}_t$, where $\mathbf{u} \in \mathbb{R}^d$.

B. Problem Setting

In [1], the main performance evaluation metric for a filter $\hat{\mathbf{X}}$ was the *expected regret*,

$$R(n) \triangleq E\left[\sum_{t=1}^{n}\left(x_t - \hat{X}_t(Y^t)\right)^2\right] - \min_{\mathbf{u}\in\mathbb{R}^d} E\left[\sum_{t=1}^{n}\left(x_t - \mathbf{u}^T\mathbf{Y}_t\right)^2\right], \tag{1}$$

the difference between the expected MSE of the universal filter and the best fixed FIR filter for the underlying signal $\{x_t\}_{t=1}^n$, where the expectation is with respect to the noise distribution. [1, Theorem 1(a)] shows that their scheme attains $R(n) = O(\log n)$ for all $x^n \in \mathcal{D}^n$ and n. In order to compete with the switching FIR filters as mentioned in the Section I Introduction, first consider the set of m switching time positions, $\mathcal{T}_{m,n} \triangleq \{(t_1, \ldots, t_{m-1}) | 1 = t_0 < t_1 < \cdots < t_{m-1} < t_m = n\}$, that divides length n sequence into m disjoint segments. Then, denote S_i as the ith segment defined by $\mathcal{T}_{m,n}$ and $\mathbf{u}_i$ as the FIR filter dedicated to S_i. Following these notations, our goal then would be to design a filter $\hat{\mathbf{X}}$ such that for all $x^n \in \mathcal{D}^n$,

$$E\left[\sum_{t=1}^{n}\left(x_t - \hat{X}_t(Y^t)\right)^2\right] - \min_{\mathcal{T}_{m,n}}\sum_{i=1}^{m}\min_{\mathbf{u}_i\in\mathbb{R}^d} E\left[\sum_{t\in S_i}\left(x_t - \mathbf{u}_i^T\mathbf{Y}_t\right)^2\right] \tag{2}$$

is sublinear in n for sufficiently small m, so that when n grows, the MSE of the filter $\hat{\mathbf{X}}$ would achieve that of the best m-switching FIR filters. The similar performance targets as (2) appear in [2] and [3] for linear prediction and linear noisy prediction problems, respectively. Competing with switching FIR filters obviously is a much more challenging task than competing with a single FIR filter since $|\mathcal{T}_{m,n}| = \binom{n}{m-1}$ is *exponential* in n, and the algorithm needs to sequentially learn both the appropriate switching points among $\mathcal{T}_{m,n}$ and the best FIR filter in each segment, solely based on the noisy signal observations.

In this correspondence, instead of working with the quantity in (2) directly, we work with a slightly more general performance metric called the *adaptive* expected regret [see (3) shown at the bottom of the page], which was first defined in [4]. Note that $R_{\mathrm{A}}(n)$ is a stronger notion than the ordinary regret $R(n)$ since $R(n) \leq R_{\mathrm{A}}(n)$ by definition. Moreover, we can easily see that (2)$\leq m \cdot R_{\mathrm{A}}(n)$; hence, if we show that $R_{\mathrm{A}}(n)$ grows sufficiently slowly with n for our algorithm, then combined with appropriate growth rate of m, we can show that (2) is also sublinear in n.

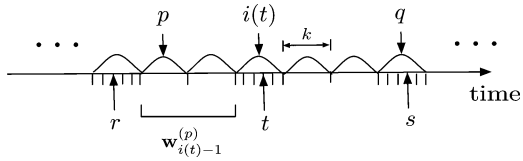


Fig. 1. Time axis divided into blocks of size k and the defined indices.

III. Main Results

A. Two Preliminary Lemmas

We first present some definitions and two necessary lemmas.

Definition 1: For any dth-order FIR filter coefficients $\mathbf{u} \in \mathbb{R}^d$, we define the true loss (squared error), estimated loss, and k-average estimated loss of $\mathbf{u}$ as follows (the first two are directly from [1, Definition 1]):

1) $\Lambda_t(\mathbf{u}) = (x_t - \mathbf{u}^T\mathbf{Y}_t)^2$;
2) $\ell_t(\mathbf{u}) = (Y_t - \mathbf{u}^T\mathbf{Y}_t)^2 + 2\mathbf{u}^T\mathbf{c}$;
3) $f_j(\mathbf{u}) = \frac{1}{k}\sum_{t=jk}^{(j+1)k-1}\ell_t(\mathbf{u})$. for some $k > 0$ and $j > 0$.

Remark: Recall that $\ell_t(\mathbf{u})$ does not depend on $\{x_t\}_{t\geq 1}$, but for any $\mathbf{u} \in \sigma(Y^{t-1})$, $E[\Lambda_t(\mathbf{u})|Y^{t-1}] = E[\ell_t(\mathbf{u}) - \sigma^2|Y^{t-1}]$ from the martingale relationship in [1, Lemma 1]. Similarly, the unbiased relationship for $f_j(\mathbf{u})$ and the k-average of the corresponding true losses can be derived as well.

Before presenting the main theorem, we first consider the quantity in (3) without the supremum, that is, a regular expected regret for some fixed r and s. For the convenience of our arguments below, consider positive integers k, r, and s such that $d \ll k \ll s - r$ and denote $|I| = s - r$. Moreover, given k, define $p = \lceil \frac{r}{k} \rceil$ and $q = \lfloor \frac{s}{k} \rfloor$. Also, denote $i(t) = \lfloor \frac{t}{k} \rfloor$. In other words, k is the block size, p is the first size-k block after r, q is the last size-k block that contains s, and $i(t)$ is the block index that contains t. Then, consider a filtering algorithm $\hat{\mathbf{X}}^{(r)} = \{\hat{X}_t^{(r)}(Y_r^t)\}_{t=r}^s$ that starts its observations from $t = r$, and the filter at t has the form $\hat{X}_t^{(r)}(Y_r^t) = \mathbf{w}_{i(t)-1}^{(p)} \cdot \mathbf{Y}_t$ as in [1], where the filter coefficients for ith size-k block is defined as

$$\mathbf{w}_{i-1}^{(p)} \triangleq \arg\min_{\mathbf{u}\in\mathbb{R}^d}\left\{\sum_{j=p}^{i-1} f_j(\mathbf{u}) + \|\mathbf{u}\|^2\right\} = \left(I + k\sum_{j=p}^{i-1}\sum_{t=jk}^{(j+1)k-1}\mathbf{Y}_t\mathbf{Y}_t^T\right)^{-1}\left(k\sum_{j=p}^{i-1}\sum_{t=jk}^{(j+1)k-1}(y_t\mathbf{Y}_t - \mathbf{c})\right). \tag{4}$$

Note that above filter does use Y_t to estimate x_t, but the filter coefficients are determined by only using σ^2 and noisy symbols up to the size-k blocks that occur before t. Thus, instead of updating the filter coefficients for every t as in [1], $\hat{\mathbf{X}}^{(r)}$ updates the coefficients after every k observations since it is using the k-average of the estimated losses, $\{f_j(\mathbf{u})\}$'s. The summary of defined indices and the data range that the filter coefficient $\mathbf{w}_{i(t)-1}^{(p)}$ is using for $i(t)$th block are depicted in Fig. 1. Now, we have the following lemma.

$$R_{\mathrm{A}}(n) \triangleq \sup_{I=[r,s]\subseteq[1,n]}\left\{E\left[\sum_{t=r}^{s}\left(x_t - \hat{X}_t(Y^t)\right)^2\right] - \min_{\mathbf{u}\in\mathbb{R}^d} E\left[\sum_{t=r}^{s}\left(x_t - \mathbf{u}^T\mathbf{Y}_t\right)^2\right]\right\} \tag{3}$$

Lemma 1: Suppose that $d \ll k \ll s-r$ and denote $|I| = s-r$. Then, for all $x^n \in \mathcal{D}^n$, the filtering algorithm $\hat{\mathbf{X}}^{(r)}$ defined in (4) satisfies

$$E\left[\sum_{t=r}^{s}\left(x_t - \hat{X}_t^{(r)}(Y_r^t)\right)^2\right] - \min_{\mathbf{u}\in\mathbb{R}^d} E\left[\sum_{t=r}^{s}\left(x_t - \mathbf{u}^T\mathbf{Y}_t\right)^2\right] \le \Theta\left(k\log\frac{|I|}{k}\right).$$

Proof: The proof is straightforward by using the filtering-prediction association developed in [1] and techniques used for proving [1, Theorem 1(a)]. By using the notations in Definition 1, for any $\mathbf{u}\in\mathbb{R}^d$, we have

$$E\left[\sum_{t=r}^{s}\Lambda_t(\mathbf{w}_{i(t)-1}^{(p)}) - \sum_{t=r}^{s}\Lambda_t(\mathbf{u})\right]$$
$$\le E\left[\sum_{t=r}^{s}\ell_t(\mathbf{w}_{i(t)-1}^{(p)}) - \sum_{t=r}^{s}\ell_t(\mathbf{u})\right] \quad (5)$$
$$= k\cdot E\left[\sum_{i=p}^{q} f_i(\mathbf{w}_{i-1}^{(p)}) - \sum_{i=p}^{q} f_i(\mathbf{u})\right] + 2kB \quad (6)$$

where (5) follows from $\mathbf{w}_{i(t)-1}^{(p)} \in \sigma(Y_r^{t-1})$ and the unbiased relationship between $\Lambda_t(\cdot)$ and $\ell_t(\cdot)$, and (6) follows from substituting the k-average of $\ell_t(\cdot)$'s with $f_j(\cdot)$'s and crudely bounding the edge instances of $qk \le t \le s$ and $r \le t < pk$ by $B < \infty$. Equation (6) suggests that we can think of the filtering problem as a prediction problem with the loss function $f_i(\cdot)$, similarly as in [1], where $\ell_t(\cdot)$ was used as a loss function. Now, the expectation term in (6) can be bounded by $\Theta(\log\frac{|I|}{k})$ by applying the same analysis as in [1, Theorem 1(a)], but replacing $\ell_t(\cdot)$ with $f_j(\cdot)$, which also has all properties needed for the upper bound. We omit the detailed proof for brevity. ∎

Remark: For a completely rigorous argument, we need to introduce the idea of projecting (4) to a bounded ball as in [1], (7), since the technical lemmas used in [1] only hold for bounded filter coefficients. However, we omit such technicality for simplicity and assume that, from now on, the filter coefficients $\{\mathbf{w}_{i(t)-1}^{(p)}\}$ are always bounded by some $B_{\mathbf{w}} < \infty$.

From Lemma 1, one may wonder why we introduce and use the k-average estimated loss $f_j(\mathbf{u})$ as in Definition 1(3) and in the filter definition (4), since we can get a tighter upper bound of $\Theta(\log|I|)$ by simply applying the algorithm in [1] from $t=r$. However, as is described in the following lemma, the procedure ensures the exp-concavity of the loss function $f_j(\mathbf{u})$ for appropriate k, which is a necessary condition for devising our universal switching filter.

Lemma 2: Consider filter coefficients $\mathbf{u}$ that are in $\mathcal{D}_{\mathbf{u}} = \{\mathbf{u} : \|\mathbf{u}\| \le B_{\mathbf{w}}\}$. Also, denote

$$B_f = \sup_{j,\mathbf{u}\in\mathcal{D}_{\mathbf{u}}}\{\lambda_{\max}(\nabla f_j(\mathbf{u})\nabla f_j(\mathbf{u})^T)\},$$

and let $\alpha_j = \frac{1}{B_f}\cdot\lambda_{\min}(\nabla^2 f_j(\mathbf{u}))$. Then, $f_j(\mathbf{u})$ is an α_j-exp-concave function with probability one (w.p.1) if and only if $k \ge d$.

Proof: A convex function $g(\mathbf{u})$ is β-exp-concave if there exists some $\beta > 0$ such that $\exp(-\beta g(\mathbf{u}))$ is concave for all $\mathbf{u}$ in the domain. An equivalent condition for the exp-concavity is that there exists $\beta > 0$ such that

$$\beta\nabla g(\mathbf{u})\cdot\nabla g(\mathbf{u})^T \preceq \nabla^2 g(\mathbf{u}).$$

Since $\nabla^2 f_j(\mathbf{u}) = \frac{1}{k}\sum_{t=jk}^{(j+1)k-1}\mathbf{Y}_t\mathbf{Y}_t^T$, is a sum of random rank-1 matrices, thus, positive semi-definite, we see that $\nabla^2 f_j(\mathbf{u})$ becomes full-rank and $\lambda_{\min}(\nabla^2 f_j(\mathbf{u})) > 0$ w.p.1[2] if and only if $k \ge d$. Thus, when $k \ge d$, $\alpha_j = \frac{1}{B_f}\cdot\lambda_{\min}(\nabla^2 f_j(\mathbf{u})) > 0$, and, by algebra, we can check $f_j(\mathbf{u})$ is α_j-exp-concave w.p.1 for all $\mathbf{u}\in B_{\mathbf{w}}$.

In contrast, when $k < d$ (including $k=1$, which stands for $f_j(\mathbf{u}) = \ell_j(\mathbf{u})$), $\lambda_{\min}(\nabla^2 f_j(\mathbf{u})) = 0$, and we can always find $\mathbf{u}\in\mathcal{D}_{\mathbf{u}}$ such that no $\beta > 0$ makes $\exp(-\beta f_j(\mathbf{u}))$ concave. ∎

B. Algorithm and Complexity

Equipped with the *static* filter definition in (4) and above two lemmas, we now define our universal switching FIR MMSE filter $\hat{\mathbf{X}}_{\text{univ}}^* = \{\hat{X}_t^*(Y^t)\}_{t=1}^n$ as following;

1) We initialize the block size for $f_j(\mathbf{u})$ as $k = \log n \ge d$ and $\mathbf{w}_0 = \mathbf{0}$.
2) At time t, we define $\hat{X}_t^*(Y^t) = \mathbf{W}_{i(t)-1}^*\cdot\mathbf{Y}_t$ where the filter coefficients for the ith block is
$$\mathbf{w}_{i-1}^* = \sum_{j=1}^{i-1} p_{i-1}^{(j)}\mathbf{w}_{i-1}^{(j)}, \quad (7)$$
and $\mathbf{p}_{i-1} = \{p_{i-1}^{(j)}\}_{j=1}^{i-1}$ is a probability distribution over the integers $1 \le j \le i-1$. Moreover, $\{\mathbf{w}_{i-1}^{(j)}\}_{j=1}^{i-1}$ is the set of filter coefficients defined in (4) with different starting blocks.
3) After filtering all the instances of the ith block using (7), the probability distribution $\mathbf{p}_{i-1}$ is updated to $\mathbf{p}_i = \{p_i^{(j)}\}_{j=1}^{i}$ multiplicatively as following; Set $\alpha = \frac{\sigma^2}{2B_f}$ and define
$$d\hat{p}_i^{(j)} = \frac{p_{i-1}^{(j)}e^{-\alpha f_i(\mathbf{w}_{i-1}^{(j)})}}{\sum_{j=1}^{i-1}p_{i-1}^{(j)}e^{-\alpha f_i(\mathbf{w}_{i-1}^{(j)})}}$$
for $1 \le j \le i-1$ using the incurred loss $f_i(\mathbf{w}_{i-1}^{(j)})$ for each $\mathbf{w}_{i-1}^{(j)}$. Then, set $p_i^{(i)} = \frac{1}{i}$, and $p_i^{(j)} = (1-\frac{1}{i})\hat{p}_i^{(j)}$ for $j \ne i$. The updated $\mathbf{p}_i$ is used for filtering the $(i+1)$th block.

Thus, the $\hat{\mathbf{X}}_{\text{univ}}^*$ uses a weighted average of FIR filters that started at different blocks, and the probabilities associated with each FIR filter is multiplicatively updated with the incurred loss values for each block, measured by the function $f_j(\cdot)$. Although above algorithm is similar to those given in [4] (or [2]), there are a few differences. First, similarly as in [1], we converted the filtering problem to a prediction problem by using a convex loss $f_j(\cdot)$, the k-average estimated loss with $k = \log n$ that only depends on $\{Y_t\}$ and σ^2, as a surrogate to the true loss. Unlike [1], however, we do not use $\ell_t(\cdot)$, the surrogate loss used in [1], and directly apply the algorithms from [4] (or [2]) since those functions are not exp-concave as shown in Lemma 2. Instead, we ensure to make $f_j(\cdot)$ exp-concave by choosing the block size $k = \log n \ge d$ so that we can leverage tools from [4] for the analysis of our scheme. The choice of $\log n$ will be justified in the proof below. Ensuring the loss functions to be exp-concave is subtle but important not only for the proof of the theorem, but also in practice, since we show by experiments in Section IV that using $f_j(\cdot)$ with $k < d$, would not work well in our switching filtering problem. Note that working with an unbiased surrogate loss for the unobservable true loss was also used in [3] for noisy prediction, but there the same squared loss, which is exp-concave for bounded signal, on the noisy symbols was valid to use. Therefore, the exact same algorithm from [2] on the noisy signal was sufficient there for competing with switching linear noisy predictors. Another difference with above algorithm and those in [4] (or [2]) is that $f_j(\cdot)$ with $k \ge d$ are exp-concave with different random constants $\{\alpha_j\}$'s defined in Lemma 2, unlike the case of fixed constants in [4] (or [2]).

The time complexity of the algorithm is $O(n^2)$ since we need to run the (growing) linear combinations as in (7). Although the complexity is

[2]From now on, all inequalities between random variables should be understood in w.p.1 sense.

efficient compared to the complexity of exhaustively searching the best filters over the reference class $|\mathcal{T}_{m,n}|$ of which size is exponential in n and m, the algorithm may still suffer from long running time when n is large. To alleviate this, we may maintain only finite number of block indices—*states*—with nonzero probability by keeping the indices with largest probabilities and setting else to zero as in [2, Sec. IV]. By this way, we can have $O(n)$ complexity algorithm although it lacks the theoretical guarantee. The effectiveness of this heuristic, however, is shown by the experimental results in the next section. We also note that data streaming based technique as in [4] can be also used to reduce the complexity to $O(n \log n)$ with a performance guarantee, but we omit such argument here since such derivation is straightforward.

C. Main Theorem

Theorem 1: For all $x^n \in \mathcal{D}^n$ and all n, $\hat{\mathbf{X}}^*_{\text{univ}}$ satisfies $R_{\text{A}}(n) \leq \Theta\Big(\log^2 n\Big)$.

Remark: The theorem asserts that if $m = o\Big(\frac{n}{\log^2 n}\Big)$, then $\frac{1}{n} \cdot (2) \to 0$ for $\hat{\mathbf{X}}^*_{\text{univ}}$ as $n \to \infty$.

Before the proof, we need one simple technical lemma.

Lemma 3: For all $\beta > 0$, define

$$g_i(\beta) = -\frac{1}{\beta} \log \left(\sum_{j=1}^{i-1} p_{i-1}^{(j)} e^{-\beta f_i(\mathbf{w}_{i-1}^{(j)})} \right).$$

Then,

a) $g_i(\beta)$ is a strictly decreasing function in β;

b) $|g_i(\beta)| \leq B_g$ for all i and β when $|f(\mathbf{w}_{i-1}^{(j)})| < B_g$ for all i, j.

Proof:

a) Let $\mathbf{q}_{i-1} = \{q_{i-1}^{(j)}\}_{j=1}^{i-1}$ denote a probability distribution with jth component

$$q_{i-1}^{(j)} = \frac{p_{i-1}^{(j)} e^{-\beta f_i(\mathbf{w}_{i-1}^{(j)})}}{\sum_{j=1}^{i-1} p_{i-1}^{(j)} e^{-\beta f_i(\mathbf{w}_{i-1}^{(j)})}}.$$

Then, by doing some algebra, we can verify

$$\frac{dg_i(\beta)}{d\beta} = -\frac{D(\mathbf{q}_{i-1} \| \mathbf{p}_{i-1})}{\beta^2} \leq 0$$

where $D(\mathbf{q}\|\mathbf{p})$ is the Kullback–Liebler divergence between the probability distributions $\mathbf{q}$ and $\mathbf{p}$. Since there exists at least one j such that $f_i(\mathbf{w}_{i-1}^{(j)})$ is not zero w.p.1, we know $D(\mathbf{q}_{i-1} \| \mathbf{p}_{i-1}) > 0$, hence, from $\beta > 0$, we see that $\frac{dg_i(\beta)}{d\beta} < 0$.

b) It is easy to verify that there exists some constant $B_g < \infty$ that satisfies $|f(\mathbf{w}_{i-1}^{(j)})| < B_g$ for all i, j from the boundedness assumptions. Then, by using L'Hospital's rule, we can verify that $\lim_{\beta \to 0} g_i(\beta) = \sum_{j=1}^{i-1} p_{i-1}^{(j)} f_i(\mathbf{w}_{i-1}^{(j)})$, and $\lim_{\beta \to \infty} g_i(\beta) = \min_{1 \leq j \leq i-1} f_i(\mathbf{w}_{i-1}^{(j)})$. Since the minimum is less than the expectation and the function $g_i(\beta)$ is strictly decreasing, if $|f_t(\mathbf{w}_{i-1}^{(j)})| \leq B_g$, we obtain $|g_t(\beta)| \leq B_g$. ∎

Proof of Theorem 1: We first show that for *all* r and s that satisfy the same condition as in Lemma 1, $\hat{\mathbf{X}}^*_{\text{univ}} = \{\hat{X}_t^*(Y^t)\}$ satisfies

$$E\left[\sum_{t=r}^{s} \Big(x_t - \hat{X}_t^*(Y^t)\Big)^2\right] - E\left[\sum_{t=r}^{s} \Big(x_t - \hat{X}_t^{(r)}(Y_r^t)\Big)^2\right] \quad (8)$$

$$\leq \Theta(\log^2 n). \quad (9)$$

Following the same steps in (6), we obtain

$$(8) \leq k \cdot E\left[\sum_{i=p}^{q} f_i(\mathbf{w}_{i-1}^*) - \sum_{i=p}^{q} f_i(\mathbf{w}_{i-1}^{(p)})\right] + 2kB.$$

From Lemma 2, since $k = \log n \geq d$, $f_i(\mathbf{u})$ are ensured to be α_i-exp-concave w.p.1, and therefore, we have

$$f_i(\mathbf{w}_{i-1}^*) \leq g_i(\alpha_i)$$

for $\mathbf{w}_{i-1}^*$, $g_i(\cdot)$, and α_i defined in (7), Lemma 3, and Lemma 2. From this critical step and the similar argument as in [4, Claim 3.2], we get

$$\begin{aligned} & f_i(\mathbf{w}_{i-1}^*) - f_i(\mathbf{w}_{i-1}^{(p)}) \\ & \quad \leq \frac{1}{\alpha} \log(e^{-\alpha f(\mathbf{w}_{i-1}^{(p)})}) + g_i(\alpha) + g_i(\alpha_i) - g_i(\alpha) \\ & \quad = \frac{1}{\alpha} \log \frac{\hat{p}_i^{(j)}}{p_{i-1}^{(j)}} + g_i(\alpha_i) - g_i(\alpha), \end{aligned} \quad (10)$$

with $\alpha = \frac{\sigma^2}{2B_f}$ defined in the second step of the algorithm. Then, by applying [4, Claim 3.1], we obtain $\sum_{i=p}^{q} \{f_i(\mathbf{w}_{i-1}^*) - f_i(\mathbf{w}_{i-1}^{(p)})\} \leq \frac{4}{\alpha} \log n + \sum_{ip}^{q} \{g_i(\alpha_i) - g_i(\alpha)\}$. By taking expectation on both side of the inequality, we have the following inequalities:

$$\begin{aligned} & E\left[\sum_{i=p}^{q} f_j(\mathbf{w}_{i-1}^*) - \sum_{i=p}^{q} f_j(\mathbf{w}_{i-1}^{(p)})\right] \\ & \quad \leq \frac{4}{\alpha} \log n + \sum_{i=p}^{q} E\left[g_i(\alpha_i) - g_i(\alpha) \middle| \alpha_i \leq \alpha\right] \Pr(\alpha_i \leq \alpha) \quad (11) \\ & \quad \leq \frac{4}{\alpha} \log n + B_g \cdot (q-p) \cdot \exp(-kC) \quad (12) \end{aligned}$$

where (11) follows from conditioning on the event $\{\alpha_i \leq \alpha\}$ and Lemma 3(a) that $g_i(\beta)$ is a strictly decreasing function of β; (12) follows from $\Pr(\alpha_i \leq \alpha) \leq \exp(-kC)$ for all i, which follows from [1, Lemma 3(b)], and $g_i(\beta)$ being bounded by B_g for all i and β as shown in Lemma 3(b). We omitted the exact exponent for the upper bound of $\Pr(\alpha_i \leq \alpha)$ since we are mainly caring about the convergence order in n. Moreover, the effect of constants in the exponent were verified in [1], which was not severe. Now, by substituting $k = \log n$, we get (9) since $B_g \cdot (q-p) \cdot \exp(-kC)$ becomes a constant factor. Note that if k is a constant independent of n, (12) can grow linearly in n. Given (9), we can now easily combine with Lemma 1 with $k = \log n$ to obtain $E[\sum_{t=r}^{s} \Big(x_t - \hat{X}_t^*(Y^t)\Big)^2] - \min_{\mathbf{u} \in \mathbb{R}^d} E[\sum_{t=r}^{s} \Big(x_t - \mathbf{u}^T \mathbf{Y}_t\Big)^2] \leq \Theta(\log^2 n)$ for *all* r and s. Thus, the theorem is proven. ∎

Remark: We also note that the high probability result as in [1, Theorem 1(b)] can be derived, but omit it here.

IV. Experiments

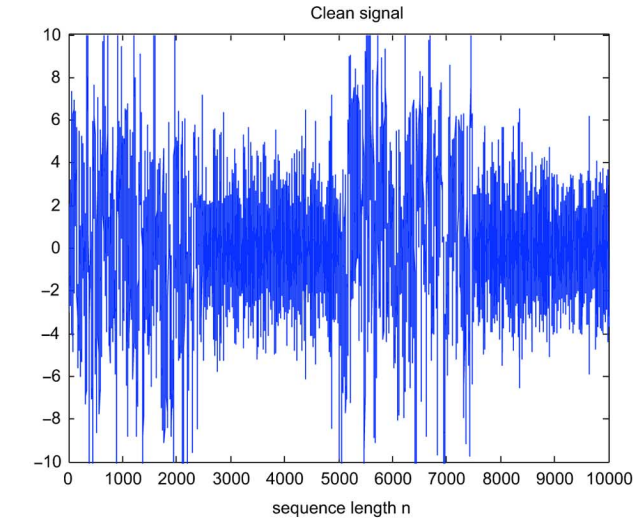
Here, we report the experimental results for our universal switching FIR filtering algorithm. We generated a piecewise stationary signal $\{X_t\}_{t \geq 1}$ of length $n = 10^4$, which alternates between

$$X_t = 1.4X_{t-1} - 0.45X_{t-2} + Z_t \quad (13)$$

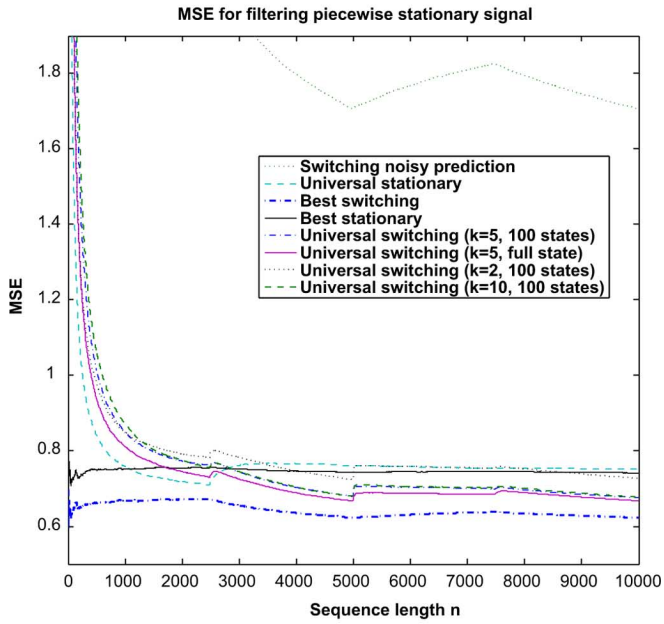
and

$$X_t = -0.7X_{t-1} + 0.18X_{t-2} + Z_t \quad (14)$$

every 2500 samples for $t \geq 3$, where $\{Z_t\}_{t \geq 3} \sim$ i.i.d. $\mathcal{N}(0,1)$. Also, X_1 and X_2 were i.i.d. $\mathcal{N}(0,1)$. The sample path of $\{X_t\}_{t \geq 1}$ is shown in Fig. 2(a). This signal was corrupted by the additive noise $\{N_t\}_{t \geq 1} \sim$ i.i.d. $\mathcal{N}(0,1)$, and the noisy signal $\{Y_t\}_{t \geq 1}$ was observed.



(a)



(b)

Fig. 2. Experimental results: (a) Sample path of $\{X_t\}_{t\geq 1}$ and (b) Average MSE results.

We fixed $d = 5$ and measured the MSEs, which are averaged over 50 sample paths, of several schemes and report the results in Fig. 2(b). The results show four variations of our $\hat{\mathbf{X}}^*_{\mathrm{univ}}$ defined in (7) by including the schemes that maintain finite number of nonzero probability states—i) $k = 5$ and full states, ii) $k = 2$ and 100 states, iii) $k = 5$ and 100 states, iv) $k = 10$ and 100 states. We also show the results of four baseline schemes—v) best switching FIR filters that are optimized for four stationary segments, vi) best fixed FIR filter, vii) universal stationary filter from [1], v and iii) switching linear noisy predictor in [3, Sec. II-C]. The following are the discussion points that we can make on the experimental results.

1) $\hat{\mathbf{X}}^*_{\mathrm{univ}}$ clearly outperforms the universal stationary filter in [1] for the piecewise stationary underlying signal as our $\{X_t\}$; The MSE of scheme i) was 0.6678, which reduces the average MSE of scheme vii), 0.7541, by more than 11%.
2) As noted in Section III-B, maintaining sufficient number of nonzero probability states, thus having $O(n)$ complexity algorithm, would be enough in practice compared to maintaining all of them; The difference of average MSE between the schemes i) and iii) are marginal.
3) The condition $k \geq d = 5$ for $f_j(\cdot)$ is crucial also in practice as shown in the results of schemes ii), iii), and iv); With $k = 2 < d$, the performance is no better than scheme vii) that does not take any switchings into account, whereas for $k \geq d = 5$, the performance improves. The performance for $k = 1$ or using $\ell_t(\mathbf{u})$ directly was even worse and we omitted it here.
4) As also pointed out in [1, Section VII-A], the switching noisy predictor performs much worse than other filtering schemes, including the simplest filter $\hat{X}_t(Y^t) = Y_t$ that has $\mathrm{MSE} = 1$, since the switching noisy predictor does not use Y_t for estimating X_t. This underscores the clear difference between the filtering problem and the noisy prediction problem in [3].
5) The convergence speed of $\hat{\mathbf{X}}^*_{\mathrm{univ}}$ is relatively slower than the stationary case; This is predictable since the regret bound we get for (2) is worse than that in [1, Theorem 1(a)].

V. Concluding Remarks

In this correspondence, we have devised a universal switching FIR filter that asymptotically attains the MSE of the best combination of switching FIR filters for every underlying bounded real-valued signal, provided that the rate of switches are not too frequent. In place of the true squared loss, we used the k-average estimated loss functions with $k = \log n$ so that we use the unbiased surrogate function which is exp-concave. One possible theoretical question to explore further would be to find out whether the $\Theta(\log^2 n)$ upper bound on the adaptive regret is tight and a matching lower bound exists.

References

[1] T. Moon and T. Weissman, "Universal FIR MMSE filtering," *IEEE Trans. Signal Process.*, vol. 57, no. 3, pp. 1068–1083, 2009.
[2] S. Kozat and A. Singer, "Universal switching linear least squares prediction," *IEEE Trans. Signal Process.*, vol. 56, no. 1, pp. 189–204, 2008.
[3] S. Kozat and A. Singer, "Competitive prediction under additive noise," *IEEE Trans. Signal Process.*, vol. 57, no. 9, pp. 3698–3703, 2009.
[4] E. Hazan and C. Seshadhri, "Efficient learning algorithms for changing environments," in *Proc. 26th Int. Conf. Mach. Learn. (ICML)*, 2009, pp. 393–400.