Competitive On-line Linear FIR MMSE Filtering

Taesup Moon
Information Systems Laboratory
Stanford University
Stanford, CA 94305
tsmoon@stanford.edu

Tsachy Weissman
Information Systems Laboratory
Stanford University
Stanford, CA 94305
tsachy@stanford.edu

***Abstract*— We consider the problem of causal estimation, i.e., *filtering*, of a real-valued signal corrupted by zero mean, i.i.d., real-valued additive noise under the mean square error (MSE) criterion. We build a *competitive* on-line filtering algorithm whose normalized cumulative MSE, for every bounded underlying signal, is asymptotically as small as the best linear finite-duration impulse response (FIR) filter of order d. We do not assume any stochastic mechanism in generating the underlying signal, and assume only the variance of the noise is known to the filter. The regret of our scheme is shown to decay in the order of $O(\log n/n)$, where n is the length of the signal. Moreover, we present a concentration of the average square error of our scheme to that of the best d-th order linear FIR filter. Our analysis combines tools from the problems of universal filtering and competitive on-line regression.**

I. Introduction

Estimating a real-valued signal corrupted by zero mean real-valued additive noise is a fundamental problem in signal processing and estimation theory. When the underlying signal is a stationary process, the usual criterion for the estimation is the mean square error (MSE), and much work on minimum MSE (MMSE) estimation has been done since Wiener [1]. Also, due to the ease of implementation, linear MMSE estimation has been popular for many decades [2]. There are noncausal and causal version of linear MMSE estimation, and in the signal processing literature, the term *filtering* is used for both cases. However, in this paper, we will only use that term for causal estimation and call that causal estimator a *filter*. The most common form of the linear MMSE filter is the finite-duration impulse response (FIR) filter, since the stability issue is not a problem and it is easy to implement.

In practice, there are two limitations in building the linear MMSE estimators. One is that we need a prior knowledge of the first and second moment of the signal which we usually do not have, and the other, which may be more stringent, is that we need stationarity assumptions on the underlying signal, whereas in practice it may be nonstationary, or even non-stochastic in many cases. In this paper, we will focus on the linear FIR MMSE filters, and try to tackle these limitations jointly.

A robust minimax approach [3][4][5] and adaptive filter approach [6] are the efforts that have been dedicated to deal with either of above limitations. The former aims to optimize for the worst case in the signal uncertainty set, to get a robust estimator. However, this approach ignores the fact that we can learn about the signal, and most of them allow large delay in estimation, i.e., noncausal estimation, which is not applicable in filtering problem that has strict causality constraint. On the other hand, the latter tries to build an FIR filter that sequentially updates its filter coefficients by learning from the noisy observation and a desired response, which the filter output needs to be close to. However, this is also not directly applicable to the case of filtering the underlying signal, since the desired response is not available to the filter. The unsupervised adaptive filtering [7] considered the case where there is no desired response, but certain statistical assumptions on the underlying signal were needed. Hence, when there is no knowledge about the statistical property of the underlying signal or when the underlying signal is not even a stochastic process, it is not clear how we can apply above two approaches for filtering the underlying signal.

Instead, we take a competitive on-line learning approach, whereby we do not assume any stochastic mechanism in generating the underlying signal. We additionally assume that the additive zero mean noise is i.i.d., bounded, and only the variance of the noise is known to the filter. This assumption on the noise is not too stringent because, in reality, it is easy to estimate the noise variance at the receiver. Since the underlying signal is not assumed to be random, we use the normalized cumulative MSE as a performance criterion. Then, we try to build a filter that performs essentially as well as the best linear FIR filter which is tuned to the actual underlying sequence, as the length of the observation increases regardless of what the underlying signal may be. By doing so, we can overcome the two limitations mentioned above while guaranteeing uniformly good performance for every possible underlying individual signal. A more precise problem formulation will be given in Section II.

Our competitive on-line learning approach for linear FIR MMSE filtering is intimately related to two lines of research in information theory and learning theory. One is the universal filtering problem, a.k.a. sequential compound decision problem, which is the problem of causally estimating the finite alphabet individual sequence based on the Discrete Memoryless Channel (DMC) corrupted noisy observation. This problem has been initiated and was the focus of much attention in 1950's and 1960's [8][9][10]. Recently, there has been resurgent interest in this area, and notable work was done in [11] where the connection of the universal filtering problem and universal

1-4244-1429-6/07/$25.00 ©2007 IEEE

prediction problem [12] was established. The other related problem area is the competitive on-line regression problem for real-valued data, which is the problem of estimating next signal component based on the past side information-signal pairs and current side information. [13] has developed on-line regressors for square error loss that compete with finite order linear regressors, and [14] extended this to the universal linear least squares prediction problem for real-valued data. Our work is an extension to both problems, i.e., an extension to the universal filtering problem to the case of real-valued individual sequences with square loss and linear experts, and an extension to the competitive on-line regression problem to the case where the clean signal is not available for learning, and yet to estimate that signal. Naturally, we try to merge the methods of [11] and [13] in developing our competitive on-line linear FIR MMSE filtering scheme.

In the rest of the paper, the formulation of the problem and the main result is in Section II, the proof in Section III, and conclusion and future follows in Section IV.

II. Problem Formulation and Main Result

Let $\{x_t\}_{t\geq 1}$ denote the real-valued signal that we want to estimate, and assume that for all t, x_t takes value in $\mathcal{D} = [-K, K] \subset \mathbb{R}$, for some $K < \infty$. We denote the signal with lower case, since we do not make any probabilistic assumption on the generation of x_t. Hence, $\{x_t\}_{t\geq 1}$ can be any arbitrary bounded individual sequence, even chaotic and adversarial. Suppose this signal goes through an additive channel, where the noise $\{N_t\}_{t\geq 1}$ is i.i.d., and $E(N_t) = 0, E(N_t^2) = \sigma^2$ for all t. Additionally, we assume that the noise is bounded, i.e., there exists an $M < \infty$, such that $|N_t| < M$ for all t, with probability one. This assumption simplifies our analysis but is not essential. We denote $\{Y_t\}_{t\geq 1}$ as the corresponding output of $\{x_t\}_{t\geq 1}$ from the additive channel, i.e.,

$$Y_t = x_t + N_t, \quad t = 1, 2, \cdots$$

The boldface notations will denote the d-dimensional column vector of d recent symbols, e.g., $\mathbf{Y}_t = [Y_t, Y_{t-1}, \cdots, Y_{t-(d-1)}]^T$, where $(\cdot)^T$ is a transpose operator. For completeness, we assign zeros to the elements of vectors whose indices are less than or equal to zero. Also, denote $\mathbf{c} = [\sigma^2, 0, \cdots, 0]^T \in \mathbb{R}^d$, and I as a d-by-d identity matrix. $\|\cdot\|$ denotes the Euclidean norm for vectors, and the operator norm (i.e., maximum singular value) for matrices. Also, for matrices, $\|\cdot\|_1$ denotes ℓ_1-norm, i.e., $\|A\|_1 = \sum_{i,j}|a_{ij}|$. $\lambda_1(\cdot)$ and $\lambda_d(\cdot)$ denote maximum and minimum eigenvalues of d-by-d matrices, respectively.

Generally, a filter $\hat{X}_t(Y^t)$ is defined to be a causal estimator of x_t based on the noisy observations $Y^t = (Y_1, Y_2, \cdots, Y_t)$. The performance of a filter for x_1^n is measured by the normalized cumulative MSE

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

where the expectation is with respect to the channel noise, N_t. Now, a linear FIR filter of order d, the focus of this paper, can be denoted as $\hat{X}_{\mathbf{u},t}(Y^t) = \mathbf{u}^T\mathbf{Y}_t$, where $\mathbf{u} \in \mathbb{R}^d$ is a vector of filter coefficients. We can also interpret $\mathbf{u}$ as a linear mapping that maps $\mathbb{R}^d$ to $\mathbb{R}$. Then, for each individual sequence x^n, there exists an optimal coefficient vector $\hat{\mathbf{u}}$ that achieves

$$\min_{\mathbf{u}\in\mathbb{R}^d} E\Big(\frac{1}{n}\sum_{t=1}^{n}(x_t - \mathbf{u}^T\mathbf{Y}_t)^2\Big), \tag{2}$$

and it is given by the well-known least squares solution, $\hat{\mathbf{u}} = \big(\sigma^2 I + \frac{1}{n}\sum_{t=1}^{n}\mathbf{x}_t\mathbf{x}_t^T\big)^{-1}\big(\frac{1}{n}\sum_{t=1}^{n}x_t\mathbf{x}_t\big)$. It is clear that this optimal $\hat{\mathbf{u}}$ depends on the entire unobserved sequence x^n.

Our goal is to construct a *competitive* filter whose normalized cumulative MSE asymptotically achieves (2) as n increases, uniformly for every possible individual sequences $x^n \in \mathcal{D}^n$. We will show that our competitive filter that meets this goal also has a form of linear mapping,

$$\hat{X}_t^*(Y^t) = \mathbf{w}_{t-1}^T\mathbf{Y}_t, \tag{3}$$

where $\mathbf{w}_{t-1}$ is defined to be

$$\mathbf{w}_{t-1} = \Big(I + \sum_{i=1}^{t-1}\mathbf{Y}_i\mathbf{Y}_i^T\Big)^{-1}\Big(\sum_{i=1}^{t-1}\{Y_i\mathbf{Y}_i - \mathbf{c}\}\Big) \in \mathbb{R}^d. \tag{4}$$

(3) is not linear in the noisy sequence $\{Y_t\}$'s, but we say it has a form of linear mapping, since the coefficient vector, that is updated sequentially, linearly combines the noisy observations to estimate x_t. Note that only Y^{t-1} and σ^2 are needed in defining $\mathbf{w}_{t-1}$, a coefficient vector for x_t, and neither the bound nor the distribution of the noise is required. The form of $\mathbf{w}_{t-1}$ resembles that of the RLS adaptive filter [6, Ch. 9], or the on-line ridge regressor [13]. The difference is that (4) is solely expressed with Y^{t-1}, whereas the other two need the desired response or the clean past signal components as well.

In addition to showing that (3) achieves (2), we can also show a much stronger concentration result. A precise statement of our result is summarized in the following theorem.

Theorem 1: Let $x^n \in \mathcal{D}^n$ be an arbitrary real-valued sequence with components in $\mathcal{D}$. Then, for all $\epsilon > 0$ and all $\mathbf{u} \in \mathbb{R}^d$, the filter $\hat{X}_t^*(Y^t)$ defined in (3) satisfies

1) $$E\Big(\frac{1}{n}\sum_{t=1}^{n}(x_t - \hat{X}_t^*(Y^t))^2\Big) - E\Big(\frac{1}{n}\sum_{t=1}^{n}(x_t - \mathbf{u}^T\mathbf{Y}_t)^2\Big) \leq O\Big(\frac{\log n}{n}\Big), \quad \text{and}$$

2) $$Pr\left(\frac{1}{n}\sum_{t=1}^{n}(x_t - \hat{X}_t^*(Y^t))^2 - \frac{1}{n}\sum_{t=1}^{n}(x_t - \mathbf{u}^T\mathbf{Y}_t)^2 \geq \epsilon + O\Big(\frac{\log n}{n}\Big)\right) \leq \alpha_1 \exp\big(-n\alpha_2\big)$$

where α_1,α_2 are positive and depend on ϵ,M,K, and d. [1]

III. Proof of the Main Result

The first part of the theorem implies that, as n increases, our scheme can perform essentially as well as any d-th order linear filter, including $\hat{\mathbf{u}}$, and thus achieves the benchmark (2). Moreover, the second part implies that the probability of our scheme's (time) average square error exceeding that of any

[1] Also, $O(\cdot)$ term in the theorem is independent of x^n.

d-th order linear filter is exponentially small in n. Although the second result is much stronger, we will focus on proving the first part, since the proof of the second part follows from the ingredients developed in this section. The detailed proof of the second part will be given in [19].

To prove the first part, we need three lemmas as intermediate steps. The main idea is to convert the problem into a prediction problem as in [11], and then upper bound the regret, i.e., the performance difference between our scheme and any fixed d-th order linear FIR filter. Lemma 1 enables this conversion with a parallel argument as in [11]. Lemma 2 and Lemma 3 give the explicit upper bound on the regret. Lemma 2 resembles the steps in [13] and [16, Chapter 11.7], and the use of the law of large numbers in Lemma 3 is similar to [10].

Before proceeding, we make following definition.

Definition 1: For any $\mathbf{u} \in \mathbb{R}^d$, define

(a) $\ell_t(\mathbf{u}) \triangleq (Y_t - \mathbf{u}^T\mathbf{Y}_t)^2 + 2\mathbf{u}^T\mathbf{c}$

(b) $L_0(\mathbf{u}) = \|\mathbf{u}\|^2, \quad L_t(\mathbf{u}) = L_{t-1}(\mathbf{u}) + \ell_t(\mathbf{u})$

We can interpret $\ell_t(\mathbf{u})$ as a loss incurred by a mapping $\mathbf{u}$ at time t. Then, consider our first lemma.

Lemma 1: For all $\mathbf{x}^\infty \in \mathcal{D}^\infty$, and for any weight vector $\mathbf{w}_{t-1} \in \mathbb{R}^d$, which is $\sigma(Y^{t-1})$-measurable,

$$\Big\{ \sum_{t=1}^{n}(x_t - \mathbf{w}_{t-1}^T\mathbf{Y}_t)^2 - \sum_{t=1}^{n}\{\ell_t(\mathbf{w}_{t-1}) - \sigma^2\} \Big\}_{n\geq 1}$$

is a $\{Y_n\}$-martingale

Proof: Fix $x^n \in \mathcal{D}^n$. Then, for all $1 \leq t \leq n$, [2]

$$\begin{aligned} &E\Big[(x_t - \mathbf{w}_{t-1}^T\mathbf{Y}_t)^2 | Y^{t-1}\Big] \\ =&E\Big[(x_t^2 - 2x_t\mathbf{w}_{t-1}^T\mathbf{Y}_t + \mathbf{w}_{t-1}^T\mathbf{Y}_t\mathbf{Y}_t^T\mathbf{w}_{t-1})\Big| Y^{t-1}\Big] \\ =&E\Big[\Big\{(Y_t - \mathbf{w}_{t-1}^T\mathbf{Y}_t)^2 + 2\mathbf{w}_{t-1}^T\mathbf{c} - \sigma^2\Big\}\Big| Y^{t-1}\Big], \end{aligned} \quad (5)$$

where (5) follows from the independence of Y_t and Y^{t-1}, and $\mathbf{w}_{t-1} \in \sigma(Y^{t-1})$. The condition $\mathbf{w}_{t-1} \in \sigma(Y^{t-1})$ is crucial to have the above equality. From (5), we can see that the sequence $\big\{(x_t - \mathbf{w}_{t-1}^T\mathbf{Y}_t)^2 - \{\ell_t(\mathbf{w}_{t-1}) - \sigma^2\}\big\}_{t\geq 1}$ is a martingale difference sequence with respect to $\{Y_t\}_{t\geq 1}$, and the lemma is proved. ■

From Lemma 1, for any filter of the form $\hat{X}_t(Y^t) = \mathbf{w}_{t-1}^T\mathbf{Y}_t$, where $\mathbf{w}_{t-1} \in \sigma(Y^{t-1})$,

$$E\Big(\sum_{t=1}^{n}(x_t - \hat{X}_t(Y^t))^2\Big) = E\Big(\sum_{t=1}^{n}\{\ell_t(\mathbf{w}_{t-1}) - \sigma^2\}\Big) \quad (6)$$

holds. Since Lemma 1 will also hold for any constant weight vector $\mathbf{u} \in \mathbb{R}^d$, we have for any $\mathbf{w}_{t-1} \in \sigma(Y^{t-1})$,

$$\begin{aligned} &E\Big(\sum_{t=1}^{n}(x_t - \hat{X}_t(Y^t))^2\Big) - E\Big(\sum_{t=1}^{n}(x_t - \mathbf{u}^T\mathbf{Y}_t)^2\Big) \\ =&E\Big(\sum_{t=1}^{n}\{\ell_t(\mathbf{w}_{t-1}) - \sigma^2\}\Big) - E\Big(\sum_{t=1}^{n}\{\ell_t(\mathbf{u}) - \sigma^2\}\Big) \\ =&E\Big(\sum_{t=1}^{n}\{\ell_t(\mathbf{w}_{t-1}) - \ell_t(\mathbf{u})\}\Big). \end{aligned} \quad (7)$$

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

Now, viewing from a prediction perspective as in [11], $\mathbf{w}_{t-1}$ can be thought of as a prediction of a linear mapping for time t based on Y^{t-1}, and $\ell_t(\mathbf{w}_{t-1})$ is the corresponding loss. Continuing this view, $\sum_{t=1}^{n}\{\ell_t(\mathbf{w}_{t-1}) - \ell_t(\mathbf{u})\}$ is a difference between the cumulative loss incurred by the predictor $\{\mathbf{w}_{t-1}\}_{t\geq 1}$ and a constant predictor $\mathbf{u}$. Hence, for fixed $\mathbf{u} \in \mathbb{R}^d$, if we find some specific predictor $\{\mathbf{w}_{t-1}\}_{t\geq 1}$ to bound the expectation of the difference, i.e., (7), with $O(\log n)$, then we are done. To do this, we consider

$$L_t(\mathbf{u}) = \mathbf{u}^T\Big(I + \sum_{i=1}^{t}\mathbf{Y}_i\mathbf{Y}_i^T\Big)\mathbf{u} - 2\mathbf{u}^T\Big(\sum_{i=1}^{t}\{Y_i\mathbf{Y}_i - \mathbf{c}\}\Big) + \sum_{i=1}^{t}Y_i^2,$$

which we can easily get from Definition 1. $L_t(\mathbf{u})$ is a strictly convex function in $\mathbf{u}$ for all t, since the Hessain matrix

$$A_t \triangleq \Big(I + \sum_{i=1}^{t}\mathbf{Y}_i\mathbf{Y}_i^T\Big) \in \mathbb{R}^{d\times d}, \quad \text{for } t \geq 0,$$

is positive definite for all t. Then, define our filtering coefficient $\mathbf{w}_t$, or the prediction for time $(t+1)$, as

$$\mathbf{w}_t \triangleq \arg\min_{u\in\mathbb{R}^d} L_t(\mathbf{u}) = A_t^{-1}\Big(\sum_{i=1}^{t}\{Y_i\mathbf{Y}_i - \mathbf{c}\}\Big), \quad (8)$$

which is the same as (4).[3] This looks similar to the follow-the-leader in the prediction literature [9][10] except for the *ridge* term I in A_t, which prevents A_t^{-1} from diverging. With this definition, here is the next step to bound (7).

Lemma 2: Consider the filtering coefficient in (8). Then,

(a) $\mathbf{w}_t$ satisfies the recursion

$$\mathbf{w}_t = \mathbf{w}_{t-1} - A_t^{-1}\{(\mathbf{w}_{t-1}^T\mathbf{Y}_t - Y_t)\mathbf{Y}_t + \mathbf{c}\}.$$

(b) For all $\mathbf{u} \in \mathbb{R}^d$, we have

$$\begin{aligned} &\sum_{t=1}^{n}\{\ell_t(\mathbf{w}_{t-1}) - \ell_t(\mathbf{u})\} \\ \leq&\|\mathbf{u}\|^2 + \sum_{t=1}^{n}(\mathbf{w}_{t-1} - \mathbf{w}_t)^T A_t(\mathbf{w}_{t-1} - \mathbf{w}_t). \end{aligned}$$

(c) Let $R_t = \mathbf{w}_{t-1}^T\mathbf{Y}_t - Y_t$. Then,

$$|R_t| \leq (1 + (t-1)\sigma^2)\lambda_1(A_{t-1}^{-1})\|\mathbf{Y}_t\|$$

Proof:

(a) This follows from the definition (8) and simple algebra.

(b) From Definition 1, $\ell_t(\mathbf{u}) = L_t(\mathbf{u}) - L_{t-1}(\mathbf{u}) = \{L_t(\mathbf{u}) - L_t(\mathbf{w}_t)\} + \{L_t(\mathbf{w}_t) - L_{t-1}(\mathbf{u})\}$ and $\ell_t(\mathbf{w}_{t-1}) = \{L_t(\mathbf{w}_{t-1}) - L_t(\mathbf{w}_t)\} + \{L_t(\mathbf{w}_t) - L_{t-1}(\mathbf{w}_{t-1})\}$. Hence,

$$\begin{aligned} &\ell_t(\mathbf{w}_{t-1}) - \ell_t(\mathbf{u}) \\ =&\{L_t(\mathbf{w}_{t-1}) - L_t(\mathbf{w}_t)\} - \{L_t(\mathbf{u}) - L_t(\mathbf{w}_t)\} \\ &+\{L_{t-1}(\mathbf{u}) - L_{t-1}(\mathbf{w}_{t-1})\}, \end{aligned}$$

[3]Calculating A_t^{-1} will take complexity of $O(d^2)$, not $O(d^3)$, due the matrix inversion lemma $A_t^{-1} = A_{t-1}^{-1} - (A_{t-1}^{-1}\mathbf{Y}_t)(A_{t-1}^{-1}\mathbf{Y}_t)^T/(1 + \mathbf{Y}_t^T A_{t-1}^{-1}\mathbf{Y}_t)$

which leads to

$$\begin{aligned}&\sum_{t=1}^{n}\{\ell_t(\mathbf{w}_{t-1})-\ell_t(\mathbf{u})\}\\ =&\{L_0(\mathbf{u})-L_0(\mathbf{w}_0)\}-\{L_n(\mathbf{u})-L_n(\mathbf{w}_n)\} \qquad (9)\\ &+\sum_{t=1}^{n}\{L_t(\mathbf{w}_{t-1})-L_t(\mathbf{w}_t)\}\\ \le&\|\mathbf{u}\|^2+\sum_{t=1}^{n}\{L_t(\mathbf{w}_{t-1})-L_t(\mathbf{w}_t)\}. \qquad (10)\end{aligned}$$

The inequality in (10) holds since $L_n(\mathbf{w}_n)\le L_n(\mathbf{u})$ for all $\mathbf{u}\in\mathbb{R}^d$. Now, since $L_t(\mathbf{u})$ is convex, and $\mathbf{w}_t$ is its minimizing argument, $\nabla L_t(\mathbf{w}_t)=0$. Following some algebra, we obtain

$$\begin{aligned}&L_t(\mathbf{w}_{t-1})-L_t(\mathbf{w}_t)\\ =&L_t(\mathbf{w}_{t-1})-L_t(\mathbf{w}_t)-(\mathbf{w}_{t-1}-\mathbf{w}_t)^T\nabla L_t(\mathbf{w}_t)\\ =&(\mathbf{w}_{t-1}-\mathbf{w}_t)^T A_t(\mathbf{w}_{t-1}-\mathbf{w}_t),\end{aligned}$$

which proves the lemma.[4]

(c) The bound can be obtained by following inequalities.

$$\begin{aligned}&|R_t|=|(\mathbf{w}_{t-1}-\mathbf{e}_1)^T\mathbf{Y}_t|\\ =&\Big|\big(A_{t-1}^{-1}\big(\sum_{i=1}^{t-1}\{\mathbf{Y}_i\mathbf{Y}_i^T-\sigma^2 I\}-A_{t-1}\big)\mathbf{e}_1\big)^T\mathbf{Y}_t\Big| \qquad (11)\\ \le&\big\|(1+(t-1)\sigma^2)A_{t-1}^{-1}\mathbf{e}_1\big\|\cdot\big\|\mathbf{Y}_t\big\| \qquad (12)\\ \le&(1+(t-1)\sigma^2)\big\|A_{t-1}^{-1}\big\|\cdot\big\|\mathbf{Y}_t\big\| \qquad (13)\\ =&(1+(t-1)\sigma^2)\lambda_1(A_{t-1}^{-1})\cdot\big\|\mathbf{Y}_t\big\|, \qquad (14)\end{aligned}$$

where (11) is from the definition (8), (12) is from Cauchy-Schwartz inequality, (13) is from the definition of matrix norm, and (14) is from the fact that A_{t-1}^{-1} is a symmetric matrix. ■

Using Lemma 2, we can upper bound (7) for $\mathbf{w}_t$ in (8) as,

$$\begin{aligned}(7)\le&\|\mathbf{u}\|^2+E\Big(\sum_{t=1}^{n}(\mathbf{w}_{t-1}-\mathbf{w}_t)^T A_t(\mathbf{w}_{t-1}-\mathbf{w}_t)\Big)\\ =&\|\mathbf{u}\|^2+E\Big(\sum_{t=1}^{n}\{R_t\mathbf{Y}_t+\mathbf{c}\}^T A_t^{-1}\{R_t\mathbf{Y}_t+\mathbf{c}\}\Big)\\ \le&\|\mathbf{u}\|^2+E\Big(\sum_{t=1}^{n}(|R_t|\|\mathbf{Y}_t\|+\sigma^2)^2\lambda_1(A_t^{-1})\Big) \qquad (15)\end{aligned}$$

Since $\|\mathbf{u}\|^2$ is a constant for fixed $\mathbf{u}$, now our goal becomes to show that the expectation in (15) is upper bounded by $O(\log n)$. By the inspection of (15), and from Lemma 2(c) and $\|\mathbf{Y}_t\|\le\sqrt{d}(K+M)$, we can see that this will be true if $\lambda_1(A_t^{-1})$ behaves in the order of $\frac{1}{t}$ with sufficiently high probability. To get this, we need following lemma.

Lemma 3: Denote $K_t=\sum_{i=1}^{t}\mathbf{Y}_i\mathbf{Y}_i^T$, and $K_{t/d}=\sum_{i=1}^{\lfloor t/d\rfloor}\mathbf{Y}_{1+di}\mathbf{Y}_{1+di}^T$. Then,

(a) $\lambda_d(K_t)\ge\lambda_d(K_{t/d})$ a.s.

(b) Let $t'=\lfloor t/d\rfloor$, $\tilde{\mathbf{Y}}_i=\mathbf{Y}_{1+id}$, and $\tilde{\mathbf{x}}_i=\mathbf{x}_{1+id}$. Then, for any $\epsilon>0$,

$$\begin{aligned}&Pr\Big(\Big\|\frac{1}{t'}\sum_{i=1}^{t'}\tilde{\mathbf{Y}}_i\tilde{\mathbf{Y}}_i^T-\Big(\sigma^2 I+\frac{1}{t'}\sum_{i=1}^{t'}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^T\Big)\Big\|_1>\epsilon\Big)\\ \le&2d^2\exp\big(-t'C\big)\end{aligned}$$

where $C=\frac{2\epsilon^2}{M^2(M+2K)^2d^4}$.

Proof:

(a) This follows immediately from the interlacing inequality for eigenvalues [18, Theorem 4.3.1], and the fact that $\mathbf{Y}_i\mathbf{Y}_i^T$ is a rank-1, positive semidefinite matrix.

(b) By the choice of the vector, every element in each vector $\tilde{\mathbf{Y}}_i$ is independent of every other element within the vector and also independent of elements in vectors with different indices i'. Thus, we can apply the law of large numbers for each element of the matrix $1/t'\sum_{i=1}^{t'}\tilde{\mathbf{Y}}_i\tilde{\mathbf{Y}}_i^T$, and use Hoeffding's inequality [16] to get an explicit probability bound.

Denoting the matrix

$$W_{t'}=\frac{1}{t'}\sum_{i=1}^{t'}\tilde{\mathbf{Y}}_i\tilde{\mathbf{Y}}_i^T-\Big(\sigma^2 I+\frac{1}{t'}\sum_{i=1}^{t'}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^T\Big),$$

we have

$$Pr\Big(\big\|W_{t'}\big\|_1>\epsilon\Big)\le\sum_{1\le a,b\le d}Pr\Big(\big|(W_{t'})_{ab}\big|>\frac{\epsilon}{d^2}\Big), \qquad (16)$$

where $(W_{t'})_{ab}$ denotes the ab-th entry of the matrix $W_{t'}$. Define also $\tilde{\mathbf{N}}_i=\mathbf{N}_{1+di}$. Then,

$$\frac{1}{t'}\sum_{i=1}^{t'}\tilde{\mathbf{Y}}_i\tilde{\mathbf{Y}}_i^T=\frac{1}{t'}\sum_{i=1}^{t'}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^T+\frac{1}{t'}\sum_{i=1}^{t'}\tilde{\mathbf{N}}_i\tilde{\mathbf{N}}_i^T+\frac{2}{t'}\sum_{i=1}^{t'}\tilde{\mathbf{x}}_i\tilde{\mathbf{N}}_i^T,$$

and by the law of large numbers,

$$\Big(\frac{1}{t'}\sum_{i=1}^{t'}\tilde{\mathbf{N}}_i\tilde{\mathbf{N}}_i^T\Big)_{ab}\to\left\{\begin{array}{ll}\sigma^2 & \text{a.s., if } a=b\\ 0 & \text{a.s., if } a\ne b\end{array}\right\}$$

$$\Big(\frac{1}{t'}\sum_{i=1}^{t'}\tilde{\mathbf{x}}_i\tilde{\mathbf{N}}_i^T\Big)_{ab}\to 0 \quad \text{a.s. for all } a,b$$

Since x_t's and N_t's are all assumed to be bounded, we can apply Hoeffding's inequality to get the bound,

$$\begin{aligned}&Pr\Big(\big|\big(\frac{1}{t'}\sum_{i=1}^{t'}\{\tilde{\mathbf{N}}_i\tilde{\mathbf{N}}_i^T+2\tilde{\mathbf{x}}_i\tilde{\mathbf{N}}_i^T\}\big)_{ab}-(\sigma^2 I)_{ab}\big|>\frac{\epsilon}{d^2}\Big)\\ \le&2\exp\Big(-\frac{2\epsilon^2}{M^2(M+2K)^2d^4}t'\Big),\end{aligned}$$

which, combined with (16), gives

$$\sum_{1\le a,b\le d}Pr\Big(\big|(W_{t'})_{ab}\big|>\frac{\epsilon}{d^2}\Big)\le 2d^2\exp\big(-t'C\big)$$

and completes the proof. ■

This lemma is helpful not only in getting the upper bound on the regret, but also in showing that, like [10] and unlike [11], the randomization of the filter is not necessary. Finally, we can prove our main theorem.

Proof of Theorem 1: In [17, (2.2)], we can find the inequality

$$\max_i\min_j|\lambda_j-\mu_i|\le\frac{d+2}{d}G_{AB}^{1-1/d}\|A-B\|_1$$

[4] The argument for this part almost coincides with that of [16, Chapter 11.7] which uses Bregman divergence and potential functions. The difference is the constant vector $\mathbf{c}$ in the definition of $\ell_t(\mathbf{u})$, but it hardly affects the argument.

where $\{\lambda_j\}_{1\le j\le d}$ and $\{\mu_i\}_{1\le i\le d}$ are the eigenvalues of d-by-d matrices A and B, respectively, and $G_{AB} = \max_{i,j}(|a_{ij}|, |b_{ij}|)$. From this, we can deduce that

$$|\lambda_d(A) - \lambda_d(B)| \le F_{AB}\|A - B\|_1, \tag{17}$$

where $F_{AB} = \frac{d+2}{d} G_{AB}^{1-1/d}$, i.e., the minimum eigenvalue is a Lipschitz continuous function of the elements of the matrix.

Now, let us denote the event $E_{t'} = \{\omega : \|W_{t'}\|_1 \le \epsilon\}$. Then, if $\omega \in E_{t'}$, we have

$$\frac{1}{t'}\lambda_d\Big(\sum_{i=1}^{t'} \tilde{\mathbf{Y}}_i \tilde{\mathbf{Y}}_i^T\Big) \ge \lambda_d\big(\sigma^2 I + \frac{1}{t'}\sum_{i=1}^{t'} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T\big) - \epsilon F \tag{18}$$

$$\ge \sigma^2 - \epsilon F \tag{19}$$

where $F = \frac{d+2}{d}\{\max(K^2+\sigma^2, M^2)\}^{1-1/d}$, (18) is from (17), and (19) is from the fact that $1/t' \sum_{i=1}^{t'} \mathbf{x}_i \mathbf{x}_i^T$ is positive semidefinite. Since F is finite, we can always make $\sigma^2 - \epsilon F > 0$ by choosing ϵ sufficiently small. Consequently, we can continue the chain of inequalities for the expectation in (15).

$$E\Big(\sum_{t=1}^{n}(|R_t|\|\mathbf{Y}_t\| + \sigma^2)^2 \lambda_1(A_t^{-1})\Big)$$

$$\le E\Big(\sum_{t=0}^{n}\Big\{\|\mathbf{Y}_t\|(1+t\sigma^2)\lambda_1(A_t^{-1}) + \sigma^2\Big\}^2 \lambda_1(A_t^{-1})\Big) \tag{20}$$

$$= E\Big(\sum_{t=0}^{n}\Big\{\sigma^2 + \frac{\|\mathbf{Y}_t\|(1+t\sigma^2)}{1+\lambda_d(K_t)}\Big\}^2 \frac{1}{1+\lambda_d(K_t)}\Big)$$

$$\le E\Big(\sum_{t=0}^{n}\Big\{\sigma^2 + \frac{\|\mathbf{Y}_t\|(1+t\sigma^2)}{1+\lambda_d(K_{t/d})}\Big\}^2 \frac{1}{1+\lambda_d(K_{t/d})}\Big) \tag{21}$$

$$\le \sum_{t'=0}^{\lfloor n/d \rfloor}\Big(2d^3\{\sigma^2 + \|\mathbf{Y}_t\|(1+d\sigma^2 t')\}^2\Big)\exp(-t'C) \tag{22}$$

$$+ \sum_{t'=0}^{\lfloor n/d \rfloor} d\Big\{\sigma^2 + \frac{\|\mathbf{Y}_t\|(1+d\sigma^2 t')}{1+(\sigma^2-\epsilon F)t'}\Big\}^2 \frac{1}{1+(\sigma^2-\epsilon F)t'} \tag{23}$$

where (20) is from Lemma 2(c), $\lambda_{\max}(A_{t-1}^{-1}) \ge \lambda_{\max}(A_t^{-1})$, and adding one more term in the end, (21) is from Lemma 3(a), and (22),(23) is from Lemma 3(b), (19), respectively. Since $\|\mathbf{Y}_t\| \le \sqrt{d}(K+M)$, and $\sum_{t=0}^{\infty} t^k e^{-t\alpha} < \infty$ for any $k \ge 0, \alpha > 0$, (22) is upper bounded by a constant. Also, the fact $\sum_{t=0}^{n} \frac{(b+ct)^2}{(1+at)^3} \le b + \int_1^n \frac{(b+cx)^2}{(1+ax)^3} dx \le O(\log n)$ for $a, b > 0$, upper bounds (23) by $O(\log n)$. Therefore, for all $\mathbf{u} \in \mathbb{R}^d$, we have

$$E\Big(\sum_{t=1}^{n}(x_t - \hat{X}_t^*(Y^t))^2\Big) - E\Big(\sum_{t=1}^{n}(x_t - \mathbf{u}^T\mathbf{Y}_t)^2\Big)$$

$$\le \|\mathbf{u}\|^2 + O(\log n), \tag{24}$$

and by dividing both sides of (24) by n, the first part of our theorem is proved.

The second part of the theorem can also be proved by careful application of the above lemmas, i.e, by combining the concentration of bounded martingales in Lemma 1 and the exponential decaying probability of Lemma 3. From the second part, we can also deduce

$$\limsup_{n\to\infty}\Big(\frac{1}{n}\sum_{t=1}^{n}(x_t - \hat{X}_t^*(Y^t))^2 - \min_{\mathbf{u}\in\mathbb{R}^d}\frac{1}{n}\sum_{t=1}^{n}(x_t - \mathbf{u}^T\mathbf{Y}_t)^2\Big) \le 0$$

almost surely. The details will appear in [19]. ■

IV. Conclusion and Future Work

We have constructed a competitive on-line filtering algorithm that competes with the best linear FIR MMSE filter for every bounded individual underlying signal with regret decaying in the order of $O(\log n/n)$. The concentration result of average square error of our scheme has also been established. Future work will be dedicated to extend our scheme to compete with a broader expert class, e.g., piecewise linear filters as in [15]. Also, seeking a more general relationship between filtering and prediction as in [11] for the case of real-valued observations is another natural direction for future work.

Acknowledgement

The authors are grateful to Olivier Lévêque and Amir Dembo for helpful discussions.

References

[1] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series, with Engineering Applications*, New York: Wiley 1949
[2] T. Kailath, A. Sayed, and B. Hassibi, *Linear Estimation*, Prentice-Hall, Upper Saddle River, 2000
[3] H.V. Poor, "On robust Wiener filtering," *IEEE Trans. Automat. Control*, AC-25(3):521-526, Jun 1980
[4] Y.C. Eldar and N. Merhav, "A competitive minimax approach to robust estimation and random parameters," *IEEE Signal Process.*, 52(7):1931-1946, Jul 2004
[5] Y.C. Eldar, A. Ben-Tal, and A. Nemirovski, "Linear minimax regreat estimation of deterministic parameters with bounded data uncertainties," *IEEE Trans. Signal Process.*, 52(8):2177-2188, Aug 2004
[6] S. Haykin, *Adaptive Filter Theory*, 4th ed., Prentice-Hall, Upper Saddle River, 2002
[7] S. Haykin, *Unsupervised adaptive filtering: Volume I,II*, Wiley, 2000
[8] H. Robbins, "Asymptotically subminimax solutions of compound statistical decision problems", *Proc. Second Berkeley Symp. Math. Statis. Prob.*, 131-148, 1951
[9] J. Hannan, "Approximation to Bayes risk in repeated play", *Contributions to the Theory of Games*, III:97-139,1957, Princeton, NJ
[10] J. Van Ryzin, "The sequential compound decision problem with $m \times n$ finite loss matrix," *Ann. Math. Statist.*, 37:954-975, 1966
[11] T. Weissman, E. Ordentlich, M. Weinberger, A. Somekh-Baruch, and N. Merhav, "Universal Filtering via Prediction" *IEEE Trans. Inform. Theory*, 53(4):1253-1264, Apr 2007
[12] N. Merhav and M. Feder, "Universal Prediction," *IEEE Trans. Inform. Theory*, 44(6):2124-2147, Oct 1998
[13] V. Vovk, "Competitive on-line statistics", *International Statistical Review*, 69:213-248, 2001
[14] A. Singer, S. Kozat, and M. Feder, "Universal Linear Least Squares Prediction: Upper and Lower Bounds", *IEEE Trans. Inform. Theory*, 48(8):2354-2362, Aug 2002
[15] A. Singer and S. Kozat, "Universal Context Tree Least Squares Prediction," *ISIT 2006*, Jul 2006
[16] N. Cesa-Bianchi and G. Lugosi, "Prediction, Learning, and Games," *Cambridge University Press*, 2006
[17] L. Elsner, "On the Variation of the Spectra of Matrices," *Linear Algebra and Its Applications*, 47, pp.127-138, 1982
[18] R. Horn and C. Johnson, "Matrix Analysis",*Cambridge University Press*, 1985
[19] T. Moon and T. Weissman, "Competitive On-line Linear FIR MMSE Filtering," *in preparation*