Universal FIR MMSE Filtering

Taesup Moon, *Member, IEEE*, and Tsachy Weissman, *Senior Member, IEEE*

***Abstract*—We consider the problem of causal estimation, i.e., *filtering*, of a real-valued signal corrupted by zero mean, time-independent, real-valued additive noise, under the mean-squared error (MSE) criterion. We build a *universal* filter whose per-symbol squared error, for every bounded underlying signal, is essentially as small as that of the best finite-duration impulse response (FIR) filter of a given order. We do not assume a stochastic mechanism generating the underlying signal, and assume only that the variance of the noise is known to the filter. The regret of the expected MSE of our scheme is shown to decay as $O(\log n/n)$, where n is the length of the signal. Moreover, we present a stronger concentration result which guarantees the performance of our scheme not only in expectation, but also with high probability. Our result implies a conventional stochastic setting result, i.e., when the underlying signal is a stationary process, our filter achieves the performance of the *optimal* FIR filter. We back our theoretical findings with several experiments showcasing the potential merits of our universal filter in practice. Our analysis combines tools from the problems of universal filtering and competitive on-line regression.**

***Index Terms*—FIR MMSE filtering, logarithmic regret, online learning, regret minimization, universal filtering, unsupervised adaptive filtering.**

I. Introduction

ESTIMATING the real-valued components of a 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]. Moreover, due to the ease of implementation, linear MMSE estimation has been popular for many decades [2]. There are noncausal and causal versions 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 refer to a causal estimator as a *filter*. The most common form of the linear MMSE filter is the finite-duration impulse response (FIR) filter, since stability is not an issue and it is easy to implement.

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

Robust minimax [3]–[5] and adaptive filtering [6] are approaches that have been taken to deal with the 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 problems that have strict causality constraints. On the other hand, adaptive filtering tries to build an FIR filter that sequentially updates its filter coefficients by learning from the noisy observation and a desired response signal, which the filter output aims to approach. However, this is also not directly applicable to our setting of filtering the underlying signal, since the desired response signal, which is the underlying signal itself, is not available to the filter. Unsupervised adaptive filtering [7] considered the case where the desired response signal is not available, 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 a stochastic process, it is not clear how we can apply the above approaches.

Instead, we take an on-line learning approach, whereby we do not assume any stochastic mechanism in generating the underlying signal. Unlike the underlying signal, we do make assumptions on the noise, i.e., we assume that the noise is additive zero mean, time-independent, bounded, and the variance of the noise is known to the filter. The assumption of known noise variance is not too stringent in practice given that the noise is time-independent. That is, by sending some training sequence before the filtering process begins, we can have a good estimate on the noise variance by taking the sample variance of the noise and assuming that the noise variance is known. Given above assumptions, we build a filter that performs essentially as well as the best FIR filter which is tuned to the actual underlying sequence, as the length of the observation sequence increases, regardless of what that underlying sequence may be. We obtain performance guarantees pertaining both to the expected and the actual MSEs. By doing so, we overcome the two limitations mentioned above, guaranteeing uniformly good performance for every possible underlying individual signal. This individual sequence setting result is strong enough to imply the conventional stochastic

Manuscript received April 07, 2008; revised October 19, 2008. First published November 21, 2008; current version published February 13, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mark J. Coates. This work was supported in part by the NSF by Grants 0546535 and 0729119, and by the Samsung Scholarship. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Nice, France, June 2007.

T. Moon was with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA. He is currently with Yahoo! Inc., Sunnyvale CA 94089 USA (e-mail: tsmoon@ymail.com).

T. Weissman is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA, and is also with the Department of Electrical Engineering, Technion, Haifa 32000, Israel (e-mail: tsachy@stanford.edu).

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

Digital Object Identifier 10.1109/TSP.2008.2009894

setting result as well, namely, when the underlying signal is assumed to be stationary, the performance of our filter achieves the performance of the *optimal* FIR filter. A more precise problem formulation will be given in Section II.

Our on-line learning approach for FIR MMSE filtering is intimately related to two lines of research in information theory and learning theory. One is the universal filtering problem, also known as 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]–[10]. Recently, there has been resurgent interest in this area. For example, [11] establishes a connection between universal filtering and universal prediction [12]. The other related problem area is the competitive on-line linear regression problem for real-valued data, which is the problem of estimating the signal components based on past side information-signal pairs and current side information. [13] has developed on-line linear 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 of both problems, i.e., an extension of the universal filtering problem to the case of real-valued individual sequences with squared error loss and linear experts, and an extension of the competitive on-line linear regression problem to the case where the clean signal is not available for learning. Naturally, we try to merge the methods of [11] and [13] in developing our universal FIR MMSE filter.

The rest of the paper is organized as follows. The formulation of the problem and the main result are given in Section II. We derive our universal filter in Section III, and prove the main theorem in Section IV. The stochastic setting result follows in Section V, and several discussions are given in Section VI. Section VII presents five different experiment sets that showcases the potential merits of our universal filter in practice. Finally, concluding remarks and future work are given in Section VIII. Proofs of lemmas are moved to the Appendix to allow for a smooth flow of the arguments.

II. Problem Formulation, Filter Description, and Main Result

A. Problem Formulation

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 $\mathcal{D} = [-B_X, B_X] \subset \mathbb{R}$, for some $B_X < \infty$. 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 independent over t, and $E(N_t) = 0$, $E(N_t^2) = \sigma^2$ for all t. Thus, the noise at each time t is not necessarily identically distributed, but we require the variance to be equal for all time.[1] Additionally, we assume that the noise is bounded almost surely, i.e., there exists a $B_N < \infty$, such that $|N_t| \leq B_N$ for all t, with probability one. The bounded noise assumption simplifies our analysis but is not essential. We denote $\{Y_t\}_{t\geq 1}$ as the output of the additive noisy channel whose input is $\{x_t\}_{t\geq 1}$, i.e.,

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

The boldface notations will denote the d-dimensional column vector of d recent symbols, i.e., $\mathbf{x}_t = [x_t, x_{t-1}, \cdots, x_{t-(d-1)}]^T$, $\mathbf{N}_t = [N_t, N_{t-1}, \cdots, N_{t-(d-1)}]^T$, and $\mathbf{Y}_t = [Y_t, Y_{t-1}, \cdots, Y_{t-(d-1)}]^T$, where $(\cdot)^T$ is a transposition operator. For completeness, we assign zeros to the elements of vectors whose indices are less than or equal to zero. We denote $x^n = (x_1, \cdots, x_n)$ and $Y^n = (Y_1, \cdots, Y_n)$. Also, we denote $\mathbf{c} = [\sigma^2, 0, \cdots, 0]^T \in \mathbb{R}^d$. $\|\cdot\|$ denotes the Euclidean norm if it is used for vectors, and operator norm (i.e., maximum singular value) if used for matrices. Also, for matrices, $\|\cdot\|_1$ denotes ℓ_1-norm, i.e., $\|A\|_1 = \sum_{i,j} |a_{ij}|$.

Generally, a filter $\hat{\mathbf{X}}$ is 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 $Y^t = (Y_1, Y_2, \cdots, Y_t)$. The performance of a filter for x^n is measured by the normalized cumulative squared error or, equivalently, the mean-squared error (MSE)[2]

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

Now, an 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. Then, for each individual sequence x^n and noisy sequence realization Y^n, the best FIR filter coefficients $\mathbf{u}^*$ that achieves

$$\min_{\mathbf{u}\in\mathbb{R}^d} \frac{1}{n}\sum_{t=1}^{n}(x_t - \mathbf{u}^T\mathbf{Y}_t)^2 \tag{3}$$

is given as $\mathbf{u}^* = ((1/n)\sum_{t=1}^{n}\mathbf{Y}_t\mathbf{Y}_t^T)^{-1}((1/n)\sum_{t=1}^{n}x_t\mathbf{Y}_t)$. Therefore, for given clean and noisy signal realization of length n, the best FIR filter of order d is obtained from the complete knowledge of (x^n, Y^n).

In this paper, we devise a filter $\hat{\mathbf{X}}^* \triangleq \{\hat{X}_t^*(Y^t)\}_{t\geq 1}$ that only depends on the noisy signal $\{Y_t\}_{t\geq 1}$ and the noise variance σ^2, whose MSE asymptotically achieves (3) for every underlying signal $\{x\}_{t\geq 1}$, as n becomes large. A more precise description of the performance guarantee will be presented in our main theorem. As mentioned in the Introduction, this universal FIR MMSE filtering problem is more challenging than the on-line linear least-squares regression problem [13], [14], since the filter cannot observe the clean signal, but only observes its noisy observation. Therefore, the filter needs to combat not only the

[1] In fact, the equal variance assumption for the noise components is not crucial, but it was assumed for the simplicity of the argument. Our scheme and results would naturally generalize to the case of $E(N_t^2) = \sigma_t^2$, where σ_t^2 is bounded away from zero for all t, provided that the variance sequence $\{\sigma_t^2\}_{t\geq 1}$ is known to the filter.

[2] In conventional signal processing literature where the underlying signal is usually a stationary stochastic process, MSE means the expected squared-error at certain time t, where the expectation is with respect to the signal and the noise stationary distributions. In our setting, however, since we do not make any assumption on the distribution of the underlying signal, we use the empirical average of squared-errors and refer MSE to that quantity. As shown in Section V, this performance measure is more general than the conventional one, since our result implies a result for stochastic setting with conventional MSE.

arbitrariness of the underlying signal, but also the randomness of the noise. A similar setting of the linear least-squares prediction with noisy observations has been considered in [15]. The difference between our filter and the noisy predictor in [15] is that, by definition, our filter utilizes the noisy observation Y_t for estimating x_t, whereas the noisy predictor does not have the access to Y_t. This difference is a crucial one since Y_t is the most important observation for estimating x_t, and it will result in a significant performance gap between the two schemes. Several experiments in Section VII will stress this point. Furthermore, the result in [15] was obtained directly from the prediction result in [14] and a concentration of the sum of noise symbols, namely, the noisy predictor for x_t simply tries to predict Y_t based on Y^{t-1}, whereas our result is attained by adopting a more involved prediction-filtering association developed in [11] and applying probabilistic arguments. Hence, similarly as [15] is an extension of [16] from finite-alphabet to the continuous-valued setting in the prediction context, our work can be considered as an extension of [11] in the same direction for the filtering context.

B. Description of Our Filter

Here, we describe our filter. A detailed derivation of the filter will be given in Section III. First, we define a positive definite matrix $A_t \triangleq (I + \sum_{i=1}^{t} \mathbf{Y}_t \mathbf{Y}_t^T) \in \mathbb{R}^{d\times d}$, and a preliminary filter coefficient vector

$$\mathbf{w}_{t-1}^* \triangleq A_{t-1}^{-1} \left(\sum_{i=1}^{t-1} \{Y_i \mathbf{Y}_i - \mathbf{c}\} \right) \tag{4}$$

for each t. We also define a ball of filter coefficients

$$\mathcal{U} \triangleq \left\{ \mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\| \le R \right\} \tag{5}$$

where $R = \max\{(2/\sigma^2)\sqrt{d} B_X (B_X + B_N), 4\}$, and a projection to the ball

$$\Pi_{\mathcal{U}}(\mathbf{u}) \triangleq \left\{ \begin{array}{ll} \mathbf{u}, & \text{if } \mathbf{u} \in \mathcal{U} \\ R\frac{\mathbf{u}}{\|\mathbf{u}\|}, & \text{if } \mathbf{u} \notin \mathcal{U} \end{array} \right\} \tag{6}$$

for any $\mathbf{u} \in \mathbb{R}^d$. The value of R will be justified later. Then, our filter at time t is given as

$$\hat{X}_t^*(Y^t) = \widehat{\mathbf{w}}_{t-1}^{*T} \mathbf{Y}_t \tag{7}$$

where $\widehat{\mathbf{w}}_{t-1}^* = \Pi_{\mathcal{U}}(\mathbf{w}_{t-1}^*)$, a projection of $\mathbf{w}_{t-1}^*$ to $\mathcal{U}$. Note that this filter is not linear in the noise sequence $\{Y_t\}_{t\ge 1}$, but, for given $\widehat{\mathbf{w}}_{t-1}^*$, it linearly combines the noisy components $\mathbf{Y}_t$ to estimate x_t. A discussion of an algorithmic aspect of our filter will be given in Section VI. The definition of our filter (7) also requires the knowledge of signal and noise bounds, B_X and B_N, in addition to the noise variance σ^2. This is a requirement to bound our filter coefficient $\widehat{\mathbf{w}}_{t-1}^*$ for all t, and is needed for proving our high probability results below. However, in Section VI-B, we argue that this requirement is not necessary in any meaningful practical scenarios, and only the knowledge about $\{Y_t\}_{t\ge 1}$ and σ^2 are enough in building our universal filter. Furthermore, one may be intrigued by the exclusion of Y_t in determining the filter coefficients at time t since Y_t is the most important observation in estimating x_t. Although this may seem counterintuitive, it is a necessary requirement for our analysis that will become clear in Section III. Nonetheless, a reader should not be confused with the fact that our scheme indeed is a filter since it *does* use Y_t by combining components of $\mathbf{Y}_t$ in estimating x_t. It will also become clear that the exclusion of Y_t in determining the filter coefficients does not affect the filter performance much as we present our simulation results in Section VII.

Following subsection presents our main result of this paper.

C. Main Result

Theorem 1: Consider a filter $\hat{\mathbf{X}}^* = \{\hat{X}_t^*(Y^t)\}_{t\ge 1}$ as defined in (7). Then, we have following two theorems.

(a) For all $x^n \in \mathcal{D}^n$ and all n

$$E\left(\frac{1}{n} \sum_{t=1}^{n} \left(x_t - \hat{X}_t^*(Y^t) \right)^2 \right) - \min_{\mathbf{u}\in\mathbb{R}^d} E\left(\frac{1}{n} \sum_{t=1}^{n} (x_t - \mathbf{u}^T \mathbf{Y}_t)^2 \right) \le \Theta\left(\frac{\log n}{n} \right).$$

(b) For all $x^n \in \mathcal{D}^n$, all $\epsilon > 0$, and sufficiently large n,

$$P\left(\frac{1}{n} \sum_{t=1}^{n} \left(x_t - \hat{X}_t^*(Y^t) \right)^2 - \min_{\mathbf{u}\in\mathbb{R}^d} \frac{1}{n} \sum_{t=1}^{n} (x_t - \mathbf{u}^T \mathbf{Y}_t)^2 > \epsilon + \Theta\left(\frac{\log n}{n} \right) \right) \le \exp\left(-\Theta(n^{1/3}) \right).$$

Remark: Note that we have suppressed all the constants in the bound with $\Theta(\cdot)$ notation. To state the dependencies on constants qualitatively, the bound in Part (a) depends polynomially on B_X, B_N, σ^2, and d, and the bound in Part (b) depends polynomially on B_X, B_N, σ^2, and $1/\epsilon$, and exponentially on d. However, we omit these dependencies on constants in stating the theorem to avoid unnecessarily complicated expression of the theorem and to highlight the dependence of the bound on the sequence length n. Instead, we examine the effect of constants on the convergence rate via various experimentations given in Section VII, which will show that the effects are not as severe as we see on the complicated upper bound expressions. Part (a) of the theorem asserts the logarithmic decay rate of the *regret* of the expected MSE of our filter, where the expectation is with respect to the noise distribution. Note that this logarithmic decay rate parallels that of the results in [13] and [14]. Part (b) gives a much stronger result than Part (a), i.e., it shows that, as n grows, not only the expected MSE of our filter gets close to the minimum expected MSE $\min_{\mathbf{u}\in\mathbb{R}^d} E((1/n)\sum_{t=1}^{n}(x_t - \mathbf{u}^T\mathbf{Y}_t)^2)$, but also the actual MSE of our filter is guaranteed to be no larger than the minimum actual MSE $\min_{\mathbf{u}\in\mathbb{R}^d}(1/n)\sum_{t=1}^{n}(x_t - \mathbf{u}^T\mathbf{Y}_t)^2$, with high probability. It is worth noting that, while in most statistical signal processing contexts with a stochastic setting, it is usually satisfactory and informative enough to make statements regarding the expected performance of a filter, this is not the case in the individual sequence setting considered here. The whole point of the individual sequence setting is to have a complete picture of what is really happening (actual rather than expected MSE) for every possible sequence. This is why we obtain the high probability result, which guarantees the actual

performance of the filter, in addition to Part (a). Finally, note that from Part (b), we easily obtain the almost sure convergence

$$\limsup_{n\to\infty}\left[\frac{1}{n}\sum_{t=1}^{n}\left(x_t-\hat{X}_t^*(Y^t)\right)^2-\min_{\mathbf{u}\in\mathbb{R}^d}\frac{1}{n}\sum_{t=1}^{n}(x_t-\mathbf{u}^T\mathbf{Y}_t)^2\right]\le 0 \text{ a.s.}$$

by the fact that $\exp(-\Theta(n^{1/3}))$ is summable and applying the Borel-Cantelli lemma.[3]

III. Derivation of the Universal Filter

In this section, we derive our universal filter based on a similar argument as in [11]. We first introduce the following definition to further simplify our notation.

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

(a) $\Lambda_t(\mathbf{u})\triangleq(x_t-\mathbf{u}^T\mathbf{Y}_t)^2$;

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

Remark: When we think of $\mathbf{u}$ as a filter coefficient of order d, $\Lambda_t(\mathbf{u})$ denotes the squared loss incurred by a filter $\hat{X}_t(Y^t)=\mathbf{u}^T\mathbf{Y}_t$. Note that, although suppressed in the notation, $\Lambda_t(\mathbf{u})$ depends not only on $\mathbf{Y}_t$, but also on x_t. In contrast, $\ell_t(\mathbf{u})$ denotes the *estimated loss* of $\hat{X}_t(Y^t)=\mathbf{u}^T\mathbf{Y}_t$ based on $\{Y_i\}_{i\le t}$, the meaning of which will become clear in what follows. Unlike $\Lambda_t(\mathbf{u})$, $\ell_t(\mathbf{u})$ does not depend on x_t and hence is observable.

Equipped with this notation, we have the following martingale lemma, which is inspired by [11].

Lemma 1: Consider a sequence of random vectors $\{\mathbf{w}_{t-1}\}_{t\ge1}$, where each $\mathbf{w}_{t-1}\in\mathbb{R}^d$. Suppose $\mathbf{w}_{t-1}$ is $\sigma(Y^{t-1})$-measurable for all t. Then, for all $\mathbf{x}^\infty\in\mathcal{D}^\infty$

$$\left\{\left[\sum_{t=1}^{n}\Lambda_t(\mathbf{w}_{t-1})-\sum_{t=1}^{n}\left\{\ell_t(\mathbf{w}_{t-1})-\sigma^2\right\}\right]\right\}_{n\ge1}$$

is a $\{Y_n\}$-martingale

Proof: See Appendix A. ∎

Now, consider a class of filters of the form $\hat{X}_t(Y^t)=\mathbf{w}_{t-1}^T\mathbf{Y}_t$, where $\mathbf{w}_{t-1}\in\sigma(Y^{t-1})$. Then, since Lemma 1 also holds for any constant weight vector $\mathbf{u}\in\mathbb{R}^d$, we have

$$\begin{aligned}&E\left(\sum_{t=1}^{n}\left(x_t-\hat{X}_t(Y^t)\right)^2\right)-E\left(\sum_{t=1}^{n}(x_t-\mathbf{u}^T\mathbf{Y}_t)^2\right)\\&=E\left(\sum_{t=1}^{n}\{\Lambda_t(\mathbf{w}_{t-1})-\Lambda_t(\mathbf{u})\}\right)\\&=E\left(\sum_{t=1}^{n}\{\ell_t(\mathbf{w}_{t-1})-\ell_t(\mathbf{u})\}\right)\end{aligned}\tag{8}$$

for all $\mathbf{u}\in\mathbb{R}^d$, where (8) is from the martingale result established in Lemma 1. Hence, the observable $\sum_{t=1}^{n}\{\ell_t(\mathbf{w}_{t-1})-\ell_t(\mathbf{u})\}$ is an unbiased estimate of $\sum_{t=1}^{n}\{\Lambda_t(\mathbf{w}_{t-1})-\Lambda_t(\mathbf{u})\}$. This is the reason why we referred to $\ell_t(\mathbf{u})$ as an estimated loss in Definition 1. One important thing to note is that, from the relationship in (8), we can replace the sum, $\sum_{t=1}^{n}\{\Lambda_t(\mathbf{w}_{t-1})-\Lambda_t(\mathbf{u})\}$, that depends both on x^n and Y^n with its unbiased estimate that only depends on Y^n. We attempt to build our universal filter that, by definition, should only depend on the noisy observation causally, based on these unbiased estimates of the squared-error losses. This approach of working with an unbiased estimate to circumvent the difficulty of not observing the underlying clean signal has also been utilized in various previous research papers such as wavelet-based denoising [18], parameter estimation [19], discrete denoising [20], [21] and universal filtering of finite-alphabet signals [10], [11], [22].

To derive our universal filter, we follow the perspective of prediction-filtering association developed in [11]. Namely, we can think of the filter coefficient $\mathbf{w}_{t-1}$, which is based on Y^{t-1}, as a prediction of a linear mapping for time t that maps a vector $\mathbf{Y}_t$ into $\mathbb{R}$. Then, $\ell_t(\mathbf{w}_{t-1})$ can be thought of as the corresponding loss incurred at time t by that prediction. Conversely, whenever we have a sequence of predictors $\{\mathbf{w}_{t-1}\}_{t\ge1}$ in the above sense, we can associate an FIR filter by merely defining $\hat{X}_t(Y^t)=\mathbf{w}_{t-1}^T\mathbf{Y}_t$. As in [11], we continue to adhere to the prediction viewpoint in further development of our filter. Note the difference that we are trying to predict *a linear mapping to apply at time* t, unlike the scheme in [15] which tries to predict Y_t. The sum $\sum_{t=1}^{n}\{\ell_t(\mathbf{w}_{t-1})-\ell_t(\mathbf{u})\}$ can then be interpreted as a difference between the cumulative loss incurred by the sequence of predictors $\{\mathbf{w}_{t-1}\}_{t\ge1}$ and that of a constant predictor $\mathbf{u}$. Our approach is to come up with a sequence of predictors $\{\mathbf{w}_{t-1}\}_{t\ge1}$ that makes the cumulative loss of the predictors close to that of the best constant predictor, and then show that the associated filter indeed is defined as (7) and has the properties presented in Theorem 1.

In solving the above prediction problem, by recognizing $\ell_t(\mathbf{u})$ as a convex function in $\mathbf{u}$, one may be tempted to use algorithms that are developed in the context of online convex optimization [23] in the learning theory community. That is, to obtain the logarithmic decay rate of Part (a), we can proceed as in [11] by treating $\{Y_t\}_{t\ge1}$ as an individual sequence and simply apply the algorithms in [23] to the prediction problem inside the expectation in (8), and get the logarithmic regret even before taking the expectation. A slower rate than the logarithmic rate, e.g., $O(1/\sqrt{n})$, can indeed be attained this way by applying general online gradient descent algorithms as in [24]. However, for the logarithmic rate, the subtle point is that, due to $\mathbf{Y}_t$ being random, the induced loss function $\ell_t(\mathbf{u})$ does not satisfy the conditions required by the algorithms in [23]: being *exp-concave*[4] with some constant $\alpha>0$ for all t. Therefore, we cannot directly apply the algorithms developed in [23]. Instead, we derive our predictor in a rather intuitive way, and carefully analyze the behavior of our associated filter's performance by taking into account the randomness of $\{Y_t\}_{t\ge1}$. A detailed analysis will follow in the next section.

[3] A part of this result was presented in [17].

[4] A convex function $f(\mathbf{x})$ is an exp-concave function with parameter $\alpha>0$, if $\exp(-\alpha f(\mathbf{x}))$ is a concave function in $\mathbf{x}$.

Before obtaining our filter, we consider our estimator for the (regularized) cumulative loss up to time t, which we define to be

$$L_t(\mathbf{u}) \triangleq \|\mathbf{u}\|^2 + \sum_{i=1}^{t} \ell_i(\mathbf{u}) = \mathbf{u}^T \left(I + \sum_{i=1}^{t} \mathbf{Y}_i \mathbf{Y}_i^T \right) \mathbf{u} - 2\mathbf{u}^T \left(\sum_{i=1}^{t} \{ Y_i \mathbf{Y}_i - \mathbf{c} \} \right) + \sum_{i=1}^{t} Y_i^2, \quad (9)$$

where I is the d-by-d identity matrix. Note that $A_t \triangleq (I + \sum_{i=1}^{t} \mathbf{Y}_i \mathbf{Y}_i^T) \in \mathbb{R}^{d \times d}$, defined in Section II-B, is the Hessian of $L_t(\mathbf{u})$ and is positive definite for all t. Then, it is clear to realize that $\mathbf{w}_{t-1}^*$ defined in (4) is a unique minimizer of $L_{t-1}(\mathbf{u})$, the cumulative estimated losses up to time $t-1$. Note that depending on Y^{t-1}, $\|\mathbf{w}_{t-1}^*\|$ can grow without bound as t becomes large. However, as shown in the next section, the best FIR filter coefficient $\mathbf{u}^*$ that achieves (3) is bounded with high probability, and we would only need to consider the filter coefficients that are bounded, i.e., coefficients in $\mathcal{U}$. Therefore, by projecting $\mathbf{w}_{t-1}^*$ onto $\mathcal{U}$, we obtain our prediction $\widehat{\mathbf{w}}_{t-1}^*$ for time t which is always in $\mathcal{U}$ and $\sigma(Y^{t-1})$-measurable. This predictor can be thought of as a follow-the-leader type predictor in [9], [10] except for the *ridge* term I in A_t that prevents A_t^{-1} from diverging. Finally, following the prediction-filtering association mentioned above, we define our filter at time t as

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

which is also given in (7). Since $\widehat{\mathbf{w}}_{t-1}^*$ is $\sigma(Y^{t-1})$-measurable, (8) remains valid with $\widehat{\mathbf{w}}_{t-1}^*$ replacing $\mathbf{w}_{t-1}$. The form of our filter resembles that of the Recursive Least Square (RLS) adaptive filter [6, Ch. 9] or the on-line ridge regressor [13]. The difference is that (7) is solely expressed with the noisy signals and the noise variance, whereas the other two need to know a desired response or the clean past signal components. We now move on to prove that our filter (7) satisfies the properties stated in Theorem 1.

IV. Analysis

We first present two lemmas needed for the proof of Part (a) of our theorem. Lemma 2, which resembles the steps in [13] and [25, Ch. 11.7], collects properties of $\mathbf{w}_{t-1}^*$ and $\widehat{\mathbf{w}}_{t-1}^*$. Lemma 3 asserts a key concentration result and borrows a law of large numbers argument from [10].

Lemma 2: Consider $\mathbf{w}_t^*$ and $\widehat{\mathbf{w}}_t^*$ defined in Section II-B.

(a) $\mathbf{w}_t^*$ satisfies[5]

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

and

$$\|\mathbf{w}_t^*\| \le 1 + (1 + t\sigma^2) \lambda_{\max} \left(A_t^{-1} \right) \quad (10)$$

where $\lambda_{\max}(A_t^{-1})$ is the maximum eigenvalue of A_t^{-1}.

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

$$|R_t| \le \left(1 + (t-1)\sigma^2 \right) \lambda_{\max} \left(A_{t-1}^{-1} \right) \|\mathbf{Y}_t\|.$$

(c) For all $\mathbf{u} \in \mathbb{R}^d$

$$\sum_{t=1}^{n} \left\{ \ell_t \left(\widehat{\mathbf{w}}_{t-1}^* \right) - \ell_t(\mathbf{u}) \right\} \le \|\mathbf{u}\|^2 + \sum_{t=1}^{n} \left(\widehat{\mathbf{w}}_{t-1}^* - \mathbf{w}_t^* \right)^T A_t \left(\widehat{\mathbf{w}}_{t-1}^* - \mathbf{w}_t^* \right).$$

Proof: Part (a) and (b) follow from manipulations of the definition of $\mathbf{w}_t^*$. Part (c) builds a telescoping sum and uses the convexity of $L_t(\mathbf{u})$. See Appendix B for a detailed proof. ■

Lemma 3: Denote $K_t = \sum_{i=1}^{t} \mathbf{Y}_i \mathbf{Y}_i^T$. Then,

(a) For any $\epsilon > 0$,

$$P \left(\left\| \frac{1}{t} K_t - \left(\sigma^2 I + \frac{1}{t} \sum_{i=1}^{t} \mathbf{x}_i \mathbf{x}_i^T \right) \right\|_1 > \epsilon \right) \le 2d^2 \exp \left(-\frac{2\epsilon^2}{C} t \right),$$

where $C = B_N^2 (B_N + 2B_X)^2 d^4$.

(b) Let $\lambda_{\min}(K_t)$ be the minimum eigenvalue of the random matrix K_t. Then,

$$P \left(\lambda_{\min}(K_t) \ge \frac{\sigma^2 t}{2} \right) \ge 1 - 2d^2 \exp \left(-\frac{\sigma^4}{2CF^2} t \right)$$

where $F = \frac{d+2}{d} (B_X + B_N)^{2(1-1/d)}$.

Proof: Part (a) is based on the concentration of the sum of bounded martingale differences. Part (b) uses the fact that the minimum eigenvalue of a matrix is a continuous function of the elements of the matrix. See Appendix C for a detailed proof. ■

Remark: Part (b) of the lemma shows that, as t grows, the minimum eigenvalue of K_t will grow linearly in t with high probability. This property plays a central role in the proof of our theorem.

Equipped with the above two lemmas, we now prove Part (a) of our theorem.

Proof of theorem 1(a): First, note that

$$\hat{\mathbf{u}} = \left(\sigma^2 + \frac{1}{n} \sum_{t=1}^{n} \mathbf{x}_t \mathbf{x}_t^T \right)^{-1} \left(\frac{1}{n} \sum_{t=1}^{n} x_t \mathbf{x}_t \right)$$

achieves $\min_{\mathbf{u} \in \mathbb{R}^d} E((1/n) \sum_{t=1}^{n} (x_t - \mathbf{u}^T \mathbf{Y}_t)^2)$ and $\|\hat{\mathbf{u}}\| \le \sqrt{d} B_X^2 / \sigma^2 \le R$. Hence, it is enough to only consider the filter coefficients in $\mathcal{U}$ and show

$$E \left(\frac{1}{n} \sum_{t=1}^{n} \left(x_t - \hat{X}_t^*(Y^t) \right)^2 \right) - \min_{\mathbf{u} \in \mathcal{U}} E \left(\frac{1}{n} \sum_{t=1}^{n} (x_t - \mathbf{u}^T \mathbf{Y}_t)^2 \right) \le \Theta \left(\frac{\log n}{n} \right) \quad (11)$$

[5]Here and throughout, equalities and inequalities between random variables, when not explicitly mentioned, are to be understood in the almost sure sense.

to prove Part (a) of our theorem. To show this, for our filter $\hat{X}_t^*(Y^t)$ defined in (7) and for all $\mathbf{u} \in \mathcal{U}$, we begin with the following inequality:

$$\begin{aligned} &E\left(\sum_{t=1}^{n}\left(x_t - \hat{X}_t^*(Y^t)\right)^2\right) - E\left(\sum_{t=1}^{n}(x_t - \mathbf{u}^T\mathbf{Y}_t)^2\right) - \|\mathbf{u}\|^2 \\ &= E\left(\sum_{t=1}^{n}\Lambda_t\left(\widehat{\mathbf{w}}_{t-1}^*\right) - \sum_{t=1}^{n}\Lambda_t(\mathbf{u})\right) - \|\mathbf{u}\|^2 \\ &= E\left(\sum_{t=1}^{n}\left\{\ell_t\left(\widehat{\mathbf{w}}_{t-1}^*\right) - \ell_t(\mathbf{u})\right\}\right) - \|\mathbf{u}\|^2 \qquad (12) \\ &\le E\left(\sum_{t=1}^{n}\left(\widehat{\mathbf{w}}_{t-1}^* - \mathbf{w}_t^*\right)^T A_t\left(\widehat{\mathbf{w}}_{t-1}^* - \mathbf{w}_t^*\right)\right) \qquad (13) \end{aligned}$$

where (12) follows from (8) and definition of $\widehat{\mathbf{w}}_{t-1}^*$, and (13) follows from Lemma 2(c). To proceed, consider

$$\begin{aligned} &\left(\widehat{\mathbf{w}}_{t-1}^* - \mathbf{w}_t^*\right)^T A_t\left(\widehat{\mathbf{w}}_{t-1}^* - \mathbf{w}_t^*\right) \\ &= \left(\mathbf{w}_{t-1}^* - \mathbf{w}_t^*\right)^T A_t\left(\mathbf{w}_{t-1}^* - \mathbf{w}_t^*\right) \\ &\quad + \left(\widehat{\mathbf{w}}_{t-1}^* - \mathbf{w}_{t-1}^*\right)^T A_t\left(\widehat{\mathbf{w}}_{t-1}^* - \mathbf{w}_{t-1}^*\right) \\ &\quad + 2\left(\widehat{\mathbf{w}}_{t-1}^* - \mathbf{w}_{t-1}^*\right)^T A_t\left(\mathbf{w}_{t-1}^* - \mathbf{w}_t^*\right) \\ &\le (R_t\mathbf{Y}_t + \mathbf{c})^T A_t^{-1}(R_t\mathbf{Y}_t + \mathbf{c}) + \|A_t\|\left\|\widehat{\mathbf{w}}_{t-1}^* - \mathbf{w}_{t-1}^*\right\|^2 \\ &\quad + 2\|R_t\mathbf{Y}_t + \mathbf{c}\|\left\|\widehat{\mathbf{w}}_{t-1}^* - \mathbf{w}_{t-1}^*\right\|, \qquad (14) \end{aligned}$$

where (14) follows from applying $\mathbf{w}_{t-1}^* - \mathbf{w}_t^* = A_t^{-1}(R_t\mathbf{Y}_t + \mathbf{c})$ obtained in Lemma 2(a) and the Cauchy-Schwartz inequality. We now continue (13) separately on each term of (14). The expected sum of the first term in (14) becomes

$$\begin{aligned} &E\left(\sum_{t=1}^{n}(R_t\mathbf{Y}_t + \mathbf{c})^T A_t^{-1}(R_t\mathbf{Y}_t + \mathbf{c})\right) \\ &\le E\left(\sum_{t=1}^{n}\left(|R_t|\|\mathbf{Y}_t\| + \sigma^2\right)^2\lambda_{\max}\left(A_t^{-1}\right)\right) \qquad (15) \\ &\le E\left(\sum_{t=1}^{n}\left\{b_1\left(1 + (t-1)\sigma^2\right)\lambda_{\max}\left(A_{t-1}^{-1}\right) + \sigma^2\right\}^2\right. \\ &\qquad \left.\times\lambda_{\max}\left(A_t^{-1}\right)\right) \qquad (16) \\ &\le E\left(\sum_{t=0}^{n}\left\{b_1(1 + t\sigma^2)\lambda_{\max}\left(A_t^{-1}\right) + \sigma^2\right\}^2\lambda_{\max}\left(A_t^{-1}\right)\right) \qquad (17) \\ &= E\left(\sum_{t=0}^{n}\left\{\sigma^2 + b_1\frac{1 + t\sigma^2}{1 + \lambda_{\min}(K_t)}\right\}^2\frac{1}{1 + \lambda_{\min}(K_t)}\right) \qquad (18) \end{aligned}$$

where (15) follows from the fact that A_t^{-1} is symmetric and $\|A_t^{-1}\| = \lambda_{\max}(A_t^{-1})$; (16) follows from Lemma 2(b) and setting $b_1 = \max\|\mathbf{Y}_t\|^2 = d(B_X + B_N)^2$; (17) follows from $\lambda_{\max}(A_{t-1}^{-1}) \ge \lambda_{\max}(A_t^{-1})$ by interlacing inequality [26, Theorem 4.3.1] and adding the $(n+1)$th term in the end, and (18) follows from the fact $A_t = I + K_t$. Now, by applying Lemma 3(b) again, we know that with probability of at least $1 - 2d^2\exp(-(\sigma^4/2CF^2)t)$, the event

$$\begin{aligned} &\left\{\sigma^2 + b_1\frac{1 + \sigma^2 t}{1 + \lambda_{\min}(K_t)}\right\}^2\cdot\frac{1}{1 + \lambda_{\min}(K_t)} \\ &\qquad \le \left\{\sigma^2 + b_1\frac{1 + \sigma^2 t}{1 + (\sigma^2/2)t}\right\}^2\cdot\frac{1}{1 + (\sigma^2/2)t} \qquad (19) \end{aligned}$$

hold. Therefore, by conditioning on this event and its complement, we can continue to upper bound (18) as

$$\begin{aligned} (18) &\le \sum_{t=0}^{n}\left\{\sigma^2 + b_1(1 + \sigma^2 t)\right\}^2 \\ &\quad\cdot 2d^2\exp\left(-\frac{\sigma^4}{2CF^2}t\right) \qquad (20) \\ &\quad + \sum_{t=0}^{n}\left\{\sigma^2 + b_1\frac{1+\sigma^2 t}{1+(\sigma^2/2)t}\right\}^2\cdot\frac{1}{1+(\sigma^2/2)t}. \qquad (21) \end{aligned}$$

Since $\sum_{t=0}^{\infty}\alpha_1 t^2 e^{-t\alpha_2} < \infty$ for any $\alpha_1, \alpha_2 > 0$, we know that (20) is upper bounded by a constant. Furthermore, since the bound $\sum_{t=0}^{n}\frac{(b+ct)^2}{(1+at)^3} \le b^2 + \int_1^n\frac{(b+cx)^2}{(ax)^3}dx \le b^2 + \frac{(b+c)^2}{a^3}\cdot\log n$ holds for any $a, b > 0$, we conclude that (21) is $\Theta(\log n)$.

We can apply a similar technique to bound the expected sum of the second and the third term in (14). From Lemma 2(a) and Lemma 3(b), we can see that for $t > 1 + (2/\sigma^2)$, with probability of at least $1 - 2d^2\exp(-(\sigma^4/2CF^2)(t-1))$, we have $\|\mathbf{w}_{t-1}^*\| \le 4 \le R$ and thus, $\|\widehat{\mathbf{w}}_{t-1}^* - \mathbf{w}_{t-1}^*\| = 0$. Therefore, by conditioning on this event and its complement, we have

$$\begin{aligned} E\left(\sum_{t=1}^{n}\|A_t\|\left\|\widehat{\mathbf{w}}_{t-1}^* - \mathbf{w}_{t-1}^*\right\|^2\right) &\le \sum_{t=1}^{n}\left(1 + t(B_X + B_N)^2\right) \\ &\quad\times(2 + t\sigma^2)^2\exp\left(-\frac{\sigma^4}{2CF^2}(t-1)\right) \qquad (22) \end{aligned}$$

where (22) follows from $\|A_t\| \le t(B_X + B_N)^2$ and $\|\widehat{\mathbf{w}}_{t-1}^* - \mathbf{w}_{t-1}^*\| \le \|\mathbf{w}_{t-1}^*\| \le 2 + t\sigma^2$. Thus, we conclude that (22) is upper bounded by a constant. Similarly, we have

$$\begin{aligned} &E\left(\sum_{t=1}^{n}2\|R_t\mathbf{Y}_t + \mathbf{c}\|\left\|\widehat{\mathbf{w}}_{t-1}^* - \mathbf{w}_{t-1}^*\right\|\right) \\ &\le E\left(\sum_{t=1}^{n}2\left(b_1\left(1 + (t-1)\sigma^2\right)\lambda_{\max}\left(A_{t-1}^{-1}\right) + \sigma^2\right)\right. \\ &\qquad \left.\times\left\|\widehat{\mathbf{w}}_{t-1}^* - \mathbf{w}_{t-1}^*\right\|\right) \\ &\le \sum_{t=1}^{n}2\left(b_1\left(1 + (t-1)\sigma^2\right) + \sigma^2\right)(2 + t\sigma^2) \\ &\quad\times\exp\left(-\frac{\sigma^4}{2CF^2}(t-1)\right) \qquad (23) \end{aligned}$$

and see that (23) is again upper bounded by a constant. Therefore, by combining the bounds on (20), (21), (22), and (23), we continue from (13) and obtain

$$\begin{aligned} &E\left(\sum_{t=1}^{n}\left(x_t - \hat{X}_t^*(Y^t)\right)^2 - \sum_{t=1}^{n}(x_t - \mathbf{u}^T\mathbf{Y}_t)^2\right) \\ &\qquad \le \|\mathbf{u}\|^2 + \Theta(\log n). \qquad (24) \end{aligned}$$

for all $\mathbf{u} \in \mathcal{U}$. Since $\mathcal{U}$ is a bounded set, we have proved (11) and Part (a) of the theorem. ■

To prove Part(b) of Theorem 1, we need two additional lemmas. Lemma 4 below shows that when the probability of each random variable indexed by t being positive has an upper bound that exponentially decreases in t, the probability of the average being positive also has a bound that is summable in n. Lemma 5 gives a result paralleling (24) for the high probability setting.

Lemma 4: Let $\{Z_t\}_{t\geq 1}$ be a sequence of random variables satisfying $0 \leq Z_t \leq c_1(c_2+t)^2$ a.s. for some positive constants c_1 and c_2 and, for each t, $P(Z_t > 0) \leq 2d^2 \exp(-tC)$ for some positive constant C. Then

$$P\left(\frac{1}{n}\sum_{t=1}^{n} Z_t > \epsilon\right) \leq \frac{2d^2}{1-\exp(-C)} \times \exp\left(-\left\lfloor\left(\frac{3n\epsilon}{c_1}\right)^{1/3} - c_2\right\rfloor C\right). \quad (25)$$

Proof: The lemma follows from successive applications of the union bound. See Appendix D for a detailed proof. ■

Remark: As aforementioned, the key point of this lemma is that the right-hand side of (25) decays fast enough with n so that it ensures $\sum_{n=1}^{\infty} P((1/n)\sum_{t=1}^{n} Z_t > \epsilon) < \infty$.

Lemma 5: Fix $\epsilon > 0$. Then, for all n, all $x^n \in \mathcal{D}^n$, and for any fixed $\mathbf{u} \in \mathcal{U}$, our filter $\hat{X}_t^*(Y^t)$ defined in (7) satisfies

$$P\left(\frac{1}{n}\sum_{t=1}^{n}\left(x_t - \hat{X}_t^*(Y^t)\right)^2 - \frac{1}{n}\left\{\sum_{t=1}^{n}\Lambda_t(\mathbf{u}) + \|\mathbf{u}\|^2\right\} > \epsilon + \Theta\left(\frac{\log n}{n}\right)\right) \leq \exp\left(-\Theta(n^{1/3})\right).$$

Proof: The proof follows from the martingale result in Lemma 1 and the result of Lemma 4. See Appendix E for a detailed proof. ■

Now, we can prove the second part of our theorem.

Proof of theorem 1(b): Recall from Section II-A that the best FIR filter coefficients that achieves (3) is given as

$$\mathbf{u}^* = \arg\min_{\mathbf{u}\in\mathbb{R}^d}\sum_{t=1}^{n}\Lambda_t(\mathbf{u}) = \left(\frac{1}{n}\sum_{t=1}^{n}\mathbf{Y}_t\mathbf{Y}_t^T\right)^{-1}\left(\frac{1}{n}\sum_{t=1}^{n}x_t\mathbf{Y}_t\right).$$

Lemma 3(b) shows that with probability of at least $1 - 2d^2\exp(-(\sigma^4/2CF^2)n)$, the maximum eigenvalue of $((1/n)\sum_{t=1}^{n}\mathbf{Y}_t\mathbf{Y}_t)^{-1}$ is less than or equal to $2/\sigma^2$, hence, $\|\mathbf{u}^*\| \leq (2/\sigma^2)\sqrt{d}B_X(B_X + B_N)$ and $\mathbf{u}^* \in \mathcal{U}$. This shows the reason why we set the value of R as in Section II-B. From this observation, we know that

$$\begin{aligned} &P\left(\min_{\mathbf{u}\in\mathcal{U}}\left\{\frac{1}{n}\sum_{t=1}^{n}\Lambda_t(\mathbf{u})\right\} - \min_{\mathbf{u}\in\mathbb{R}^d}\left\{\frac{1}{n}\sum_{t=1}^{n}\Lambda_t(\mathbf{u})\right\} > 0\right) \\ &\leq 2d^2\exp\left(-\frac{\sigma^4}{2CF^2}n\right) \\ &= \exp\left(-\Theta(n)\right) \end{aligned} \quad (26)$$

and similarly as in Part (a), it suffices to only consider the filter coefficients in $\mathcal{U}$ and prove

$$P\left(\frac{1}{n}\sum_{t=1}^{n}\left(x_t - \hat{X}_t^*(Y^t)\right)^2 - \min_{\mathbf{u}\in\mathcal{U}}\left\{\frac{1}{n}\sum_{t=1}^{n}\Lambda_t(\mathbf{u})\right\} > \epsilon + \Theta\left(\frac{\log n}{n}\right)\right) \leq \exp\left(-\Theta(n^{1/3})\right). \quad (27)$$

Since $\mathcal{U}$ is compact and $\{\Lambda_t(\mathbf{u})\}_{t\geq 1}$ are bounded for all $\mathbf{u} \in \mathcal{U}$, we can easily verify that $(1/n)\{\sum_{t=1}^{n}\Lambda_t(\mathbf{u}) + \|\mathbf{u}\|^2\}$ is a Lipschitz continuous function on $\mathcal{U}$. Now, let $\mathcal{U}_\delta$ be a finite set that is obtained by uniformly quantizing $\mathcal{U}$ with resolution δ. Then, from Lipschitz continuity, we can find a constant G such that

$$\min_{\mathbf{u}\in\mathcal{U}_\delta}\left\{\frac{1}{n}\left\{\sum_{t=1}^{n}\Lambda_t(\mathbf{u}) + \|\mathbf{u}\|^2\right\}\right\} \leq \min_{\mathbf{u}\in\mathcal{U}}\left\{\frac{1}{n}\left\{\sum_{t=1}^{n}\Lambda_t(\mathbf{u}) + \|\mathbf{u}\|^2\right\}\right\} + G\delta, \quad (28)$$

where G is a constant independent of δ. Note that $|\mathcal{U}_\delta| \leq (R/\delta)^d$. Furthermore, for given $\epsilon > 0$, there exists some sufficiently large N_ϵ such that for all $n \geq N_\epsilon$,

$$\min_{\mathbf{u}\in\mathcal{U}}\left\{\frac{1}{n}\left\{\sum_{t=1}^{n}\Lambda_t(\mathbf{u}) + \|\mathbf{u}\|^2\right\}\right\} \leq \min_{\mathbf{u}\in\mathcal{U}}\left\{\frac{1}{n}\sum_{t=1}^{n}\Lambda_t(\mathbf{u})\right\} + \frac{\epsilon}{4} \quad (29)$$

since $\|\mathbf{u}\| \leq R$ for all $\mathbf{u} \in \mathcal{U}$. Now, fix $\epsilon > 0$, and let $\delta = \epsilon/4G$. Then, we have

$$\begin{aligned} &P\left(\frac{1}{n}\sum_{t=1}^{n}\Lambda_t(\widehat{\mathbf{w}}_{t-1}^*) - \min_{\mathbf{u}\in\mathcal{U}}\left\{\frac{1}{n}\sum_{t=1}^{n}\Lambda_t(\mathbf{u})\right\} > \epsilon + \Theta\left(\frac{\log n}{n}\right)\right) \\ &\leq P\left(\frac{1}{n}\sum_{t=1}^{n}\Lambda_t(\widehat{\mathbf{w}}_{t-1}^*) - \min_{\mathbf{u}\in\mathcal{U}_\delta}\left\{\frac{1}{n}\left\{\sum_{t=1}^{n}\Lambda_t(\mathbf{u}) + \|\mathbf{u}\|^2\right\}\right\} + G\delta + \frac{\epsilon}{4} > \epsilon + \Theta\left(\frac{\log n}{n}\right)\right) \quad (30) \\ &\leq P\left(\frac{1}{n}\sum_{t=1}^{n}\Lambda_t(\widehat{\mathbf{w}}_{t-1}^*) - \min_{\mathbf{u}\in\mathcal{U}_\delta}\left\{\frac{1}{n}\left\{\sum_{t=1}^{n}\Lambda_t(\mathbf{u}) + \|\mathbf{u}\|^2\right\}\right\} > \frac{\epsilon}{2} + \Theta\left(\frac{\log n}{n}\right)\right) \quad (31) \\ &\leq \sum_{\mathbf{u}\in\mathcal{U}_\delta} P\left(\frac{1}{n}\sum_{t=1}^{n}\Lambda_t(\widehat{\mathbf{w}}_{t-1}^*) - \left\{\frac{1}{n}\left\{\sum_{t=1}^{n}\Lambda_t(\mathbf{u}) + \|\mathbf{u}\|^2\right\}\right\} > \frac{\epsilon}{2} + \Theta\left(\frac{\log n}{n}\right)\right) \quad (32) \\ &\leq \left(\frac{4RG}{\epsilon}\right)^d \cdot P\left(\frac{1}{n}\sum_{t=1}^{n}\Lambda_t(\widehat{\mathbf{w}}_{t-1}^*) - \left\{\frac{1}{n}\left\{\sum_{t=1}^{n}\Lambda_t(\mathbf{u}) + \|\mathbf{u}\|^2\right\}\right\} > \frac{\epsilon}{2} + \Theta\left(\frac{\log n}{n}\right)\right) \quad (33) \end{aligned}$$

where (30) follows from (29); (31) follows from $G\delta < \epsilon/4$; (32) follows from the union bound, and (33) follows from $|\mathcal{U}_\delta| \leq (R/\delta)^d$. Now, applying Lemma 5 asserts that (33) is upper bounded by $\exp(-\Theta(n^{1/3}))$. Therefore, (27) is proved and Part(b) of the theorem follows. ■

V. Stochastic Setting

The individual sequence setting result of Theorem 1 ensures a conventional stochastic setting result as well. Namely, when the underlying signal is a bounded, real-valued, stationary stochastic process, our universal filter achieves the performance of the *optimal* FIR filter, or the Wiener FIR filter. This result is analogous to the stochastic setting result for the finite-alphabet underlying signals in [27].

Suppose the underlying signal $\{X_t\}_{t\geq 1}$ is now a stationary stochastic process, independent of the noise process, and denote $P_{\mathbf{X}}$ as its probability distribution. Without loss of generality, we assume $E(X_t) = 0$ for all t. Then, we can denote the minimum MSE (MMSE) attained by the Wiener FIR filter as $\mathbb{D}_d(P_{\mathbf{X}}, \sigma^2) = \min_{\mathbf{u}\in\mathbb{R}^d} E(X_t - \mathbf{u}^T\mathbf{Y}_t)^2 = \sigma_X^2 - \Sigma_{X\mathbf{Y}}^T \Sigma_{\mathbf{Y}}^{-1} \Sigma_{X\mathbf{Y}}$, where $\sigma_X^2 = \mathrm{Var}(X_t), \Sigma_{X\mathbf{Y}} = E[X_t\mathbf{Y}_t]$, and $\Sigma_{\mathbf{Y}} = E[\mathbf{Y}_t\mathbf{Y}_t^T]$. The following corollary asserts our stochastic setting result.

Corollary 1: Suppose the underlying signal $\{X_t\}_{t\geq 1}$ is a stationary stochastic process. Then the filter $\hat{\mathbf{X}}^* = \{\hat{X}_t^*(Y^t)\}$ defined in (7) satisfies

$$E\left(\frac{1}{n}\sum_{t=1}^{n}\left(X_t - \hat{X}_t^*(Y^t)\right)^2\right) - \mathbb{D}_d(P_{\mathbf{X}}, \sigma^2) \leq \Theta\left(\frac{\log n}{n}\right).$$

Proof: The proof follows from applying Part (a) of Theorem 1. Note that from the stationarity,

$$\begin{aligned}\mathbb{D}_d(P_{\mathbf{X}}, \sigma^2) &= \min_{\mathbf{u}\in\mathbb{R}^d} E(X_t - \mathbf{u}^T\mathbf{Y}_t)^2 \\ &= \min_{\mathbf{u}\in\mathbb{R}^d} E\left(\frac{1}{n}\sum_{t=1}^{n}(X_t - \mathbf{u}^T\mathbf{Y}_t)^2\right). \quad (34)\end{aligned}$$

Therefore, we have

$$\begin{aligned}&E\left(\frac{1}{n}\sum_{t=1}^{n}\left(X_t - \hat{X}_t^*(Y^t)\right)^2\right) - \mathbb{D}_d(P_{\mathbf{X}}, \sigma^2) \\ &= E\left(\frac{1}{n}\sum_{t=1}^{n}\left(X_t - \hat{X}_t^*(Y^t)\right)^2\right) \\ &\quad - \min_{\mathbf{u}\in\mathbb{R}^d} E\left(\frac{1}{n}\sum_{t=1}^{n}(X_t - \mathbf{u}^T\mathbf{Y}_t)^2\right) \quad (35) \\ &\leq E\left(E\left(\frac{1}{n}\sum_{t=1}^{n}\left(X_t - \hat{X}_t^*(Y^t)\right)^2 | X^n\right)\right. \\ &\quad \left. - \min_{\mathbf{u}\in\mathcal{U}} E\left(\frac{1}{n}\sum_{t=1}^{n}(X_t - \mathbf{u}^T\mathbf{Y}_t)^2 | X^n\right)\right) \quad (36) \\ &\leq \Theta\left(\frac{\log n}{n}\right) \quad (37)\end{aligned}$$

where (35) follows from (34), (36) follows from exchanging expectation with minimum, and (37) follows from applying Part (a) of Theorem 1 for each conditioned sequence $X^n = x^n$. ■

VI. Discussion

A. Algorithmic Description

As shown in the definition of our filter, the main requirement in implementing our filter is to calculate the preliminary filter coefficient $\mathbf{w}_{t-1}^*$ for each time t. Lemma 2(a) and the matrix inversion lemma $A_t^{-1} = A_{t-1}^{-1} - ((A_{t-1}^{-1}\mathbf{Y}_t)(A_{t-1}^{-1}\mathbf{Y}_t)^T/(1 + \mathbf{Y}_t^T A_{t-1}^{-1}\mathbf{Y}_t))$ shows that $\mathbf{w}_{t-1}^*$ can be recursively updated with complexity of $O(d^2)$, instead of $O(d^3)$, which a naive inversion of A_t will require. Therefore, the total complexity of our filter for given sequence length n and filter order d is $O(nd^2)$.

B. Requirement of the Knowledge on Bounds of Signal and Noise

As we mentioned in Section II-B, implementation of our filter coefficient $\widehat{\mathbf{w}}_{t-1}^*$ requires the knowledge of the signal and noise bounds, B_X and B_N. This was necessary in proving Lemma 5 where we needed to make sure that the martingale differences $\{\Lambda_t(\widehat{\mathbf{w}}_{t-1}^*) - \ell_t(\widehat{\mathbf{w}}_{t-1}^*) + \sigma^2\}_{t\geq 1}$ are bounded for all t. However, in any practical scenarios, we claim that this requirement is not necessary since all possible implementable filter coefficients that we are competing with, including the best implementable FIR filter coefficient, should be bounded anyway. More specifically, when we build the FIR filters with Digital Signal Processor (DSP) chips, any possible filter coefficients should have bounded norms due to the memory limits of the processors. Suppose B_{DSP}, which is independent of B_X and B_N, is the maximum bound on the coefficients that a DSP chip can support. Then, it is clear that the norm of the best FIR filter coefficient that is implementable with the DSP is less than or equal to B_{DSP}. Therefore, when we set the bound of $\mathcal{U}$ in (5) as $R = \max\{B_{\mathrm{DSP}}, 4\}$, all the analysis that we gave will still hold. Hence, in most practical scenarios, we would not need to know the bounds B_X and B_N explicitly. Instead, the knowledge of the predetermined parameter of a DSP chip B_{DSP}, which we know from the specification of the DSP chips, the noise variance σ^2, and the noisy signal $\{Y_t\}_{t\geq 1}$ would suffice to implement our universal filter $\{\hat{X}_t^*(Y^t)\}_{t\geq 1}$.

C. Comments on the Expectation Result

In Theorem 1(a), we focused on the regret of the expected MSEs

$$E\left(\frac{1}{n}\sum_{t=1}^{n}\left(x_t - \hat{X}_t^*(Y^t)\right)^2\right) - \min_{\mathbf{u}\in\mathcal{U}} E\left(\frac{1}{n}\sum_{t=1}^{n}(x_t - \mathbf{u}^T\mathbf{Y}_t)^2\right) \quad (38)$$

and showed that this regret goes to zero at rate $\Theta(\log n/n)$. In fact, we can consider an even stronger notion, the *expectation of the actual regret*,

$$E\left(\frac{1}{n}\sum_{t=1}^{n}\left(x_t - \hat{X}_t^*(Y^t)\right)^2 - \min_{\mathbf{u}\in\mathcal{U}}\frac{1}{n}\sum_{t=1}^{n}(x_t - \mathbf{u}^T\mathbf{Y}_t)^2\right). \quad (39)$$

Clearly, (39) is an upper bound on (38), and we do not know how to attain the logarithmic decay rate for (39). However, with additional complexity of the filtering scheme, we can upper bound (39) by $\Theta((\log n)^2/n)$. The trick would be to consider the noisy signal components with blocks (of length k) so that concentration of the block-sum of the estimated losses can happen to ensure the exp-concavity of the loss functions with sufficiently high probability, and use the result of [23]. This trick would lose additional $\log n$ factor due to treating estimated losses with blocks. Although this gives a meaningful bound for the stronger measure (39), we omit a detailed analysis.

VII. Simulation Results

In this section, we demonstrate the performance of our universal filter with several experiments.

A. Linear, Stochastic Signal

Our first example considers the case where the underlying signal is a stationary, first order autoregressive signal. More specifically, the clean signal $\{X_t\}_{t\geq 1}$ evolves as

$$X_t = \alpha X_{t-1} + Z_t, \quad \alpha = 0.9, \quad t = 1, 2, \cdots \tag{40}$$

where $\{Z_t\}_{t\geq 1}$ is iid $\sim \mathcal{N}(0,1)$, and $X_0 \sim N(0, \frac{1}{1-\alpha^2})$ to assure the stationarity of $\{X_t\}_{t\geq 1}$. The noisy signal $\{Y_t\}_{t\geq 1}$ is obtained from passing the clean signal through the additive channel (1), where $\{N_t\}_{t\geq 1}$ is iid $\sim \mathcal{N}(0,1)$, independent of $\{X_t\}_{t\geq 1}$. Note that we assumed the signal and the noise are Gaussian processes, although we required them to be bounded in the analysis of the theorem. However, for any finite n, the signal and the noise are bounded by $B_X = \max_{1\leq t\leq n} |X_t|$ and $B_N = \max_{1\leq t\leq n} |N_t|$, which are both finite. Therefore, the analysis of our theorem still holds. Moreover, in the practical scenario as discussed in Section VI-B, our universal filter $\{\hat{X}_t^*(Y^t)\}_{t\geq 1}$ in (7) can still be implemented without any knowledge of B_X or B_N. That is, we assumed that the limit of a DSP chip D_{DSP} is sufficiently large, and we used the raw $\mathbf{w}_{t-1}^*$ for our filter coefficients.

We implemented our universal filter of order $d = 5$, and experimented with the sequence length $n = 10^4$. For comparison purpose, we implemented the noisy predictor in [15] and a filter that can be induced by applying the online gradient descent algorithm in [24], both with the same order. The noisy predictor in [15] is given as $\hat{X}_{t,\text{ZS}}(Y^{t-1}) = \mathbf{w}_{t-1,\text{ZS}}^T \mathbf{Y}_{t-1}$, where

$$\mathbf{w}_{t-1,\text{ZS}} \triangleq A_{t-1}^{-1} \left(\sum_{i=1}^{t-1} Y_i \mathbf{Y}_{i-1} \right) \tag{41}$$

and A_t, $\mathbf{Y}_t$ are defined in the same way as our filter. (41) looks very similar to (4), but $\hat{X}_{t,\text{ZS}}(Y^{t-1})$ is clearly a predictor and not a filter, since it does not utilize Y_t in estimating X_t. The gradient-descent filter obtained by applying the online gradient descent algorithm in [24], as described in Section III, is given as $\hat{X}_{t,\text{GD}}(Y^t) = \mathbf{w}_{t-1,\text{GD}}^T \mathbf{Y}_t$, where

$$\begin{aligned} &\mathbf{w}_{t-1,\text{GD}} \\ &\quad = \Pi_{\mathcal{U}} \left(\mathbf{w}_{t-2,\text{GD}} - \eta_t \nabla \ell_{t-1}(\mathbf{w}_{t-2,\text{GD}}) \right) \\ &\quad = \Pi_{\mathcal{U}} \left(\mathbf{w}_{t-2,\text{GD}} - \eta_t \left(2 \left(\mathbf{c} - \left(y_{t-1} - \mathbf{w}_{t-2,\text{GD}}^T \mathbf{Y}_{t-1} \right) \right. \right. \right. \\ &\qquad \left. \left. \left. \cdot \mathbf{Y}_{t-1} \right) \right) \right). \end{aligned} \tag{42}$$

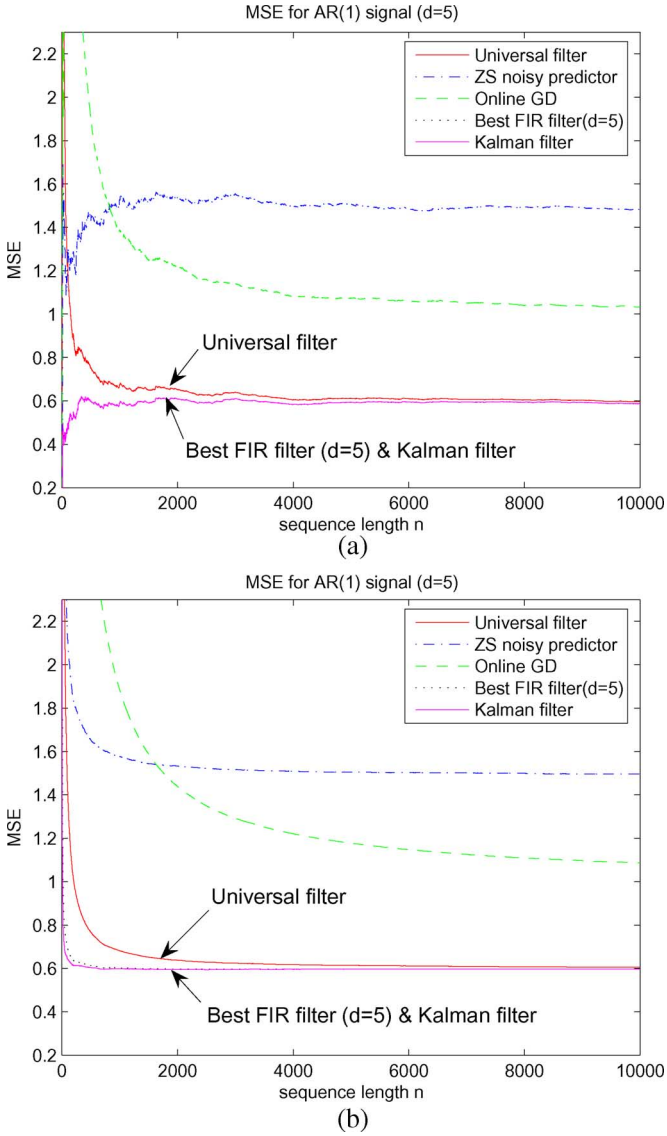

Fig. 1. MSEs for AR(1) signal (40). 1(a) is for a single sample path, and 1(b) is for the average of 100 experiments.

$\Pi_{\mathcal{U}}(\cdot)$ is again the projection function in (6), and η_t is the learning rate. Since the estimated loss function $\ell_t(\mathbf{u})$ is convex for all t and $\mathbf{u}$, when $\eta_t = 1/\sqrt{t}$, [24] assures the same asymptotical optimality of $\{\hat{X}_{t,\text{GD}}(Y^t)\}_{t\geq 1}$ as our filter, but with a slower convergence rate $O(1/\sqrt{n})$. As an ultimate comparison scheme, we also implemented Kalman filter, which is the *optimal* filter for above Gaussian signal and noise. Note that although Kalman filter is also a linear filter, the order is not finite, but is increasing with t.

Fig. 1(a) shows the MSE results of our universal filter $\hat{X}_t^*(Y^t)$, the noisy predictor $\hat{X}_{t,\text{ZS}}(Y^{t-1})$, the gradient-descent filter $\hat{X}_{t,\text{GD}}(Y^t)$, the best FIR filter $\mathbf{u}^{*T}\mathbf{Y}_t$ that achieves (3), and Kalman filter, for a *single* realization of the signal and the noise. Since the convergence of $\hat{X}_{t,\text{GD}}(Y^t)$ with $\eta_t = 1/\sqrt{t}$ was extremely slow in our experiment, we instead plot with a faster learning rate $\eta_t = 1/t$. First thing to note in Fig. 1(a) is that the performance of the best FIR filter of order $d = 5$ nearly overlaps with the performance of the optimal Kalman

filter. This may be due to the diminishing dependency of noisy signal on the past and enhances justification of our focus on the finite-order filters. Now, from the MSE curve of our universal filter, we can clearly see that our filter, which only observes the noisy signals causally together with the knowledge of the noise variance, successfully attains the performance of the best FIR filter with the same order, which is determined by a complete knowledge on (X^n, Y^n), as guaranteed by our Theorem 1. In addition, from above observation, we notice that our filter nearly attains the *optimal* performance of Kalman filter with almost negligible margin as the sequence length increases. Moreover, we observe that the convergence rate of our filter is much faster than the gradient-descent filter $\hat{X}_{t,\mathrm{GD}}(Y^t)$, which is again predicted by our theorem. Thus, although the gradient-descent filter may have the same asymptotically optimal performance as our filter, it performs poorly in practice with finite-length signal. It is also obvious from the figure that the noisy predictor $\hat{X}_{t,\mathrm{ZS}}(Y^{t-1})$ is not able to achieve the performance of the best FIR filter. This is because the noisy predictor does not have an access to Y_t in estimating X_t, whereas Y_t is the most important observation for X_t. Therefore, this experiment demonstrates that our universal filter successfully generalizes the noisy predictor in [15] to the filtering setting. Fig. 1(b) presents the result of an average performance of 100 different sample paths of the signal and the noise. We observe that the performance and convergence rate of each scheme for a single sample path is consistent with the average performance. This asserts the high probability result of our theorem.

B. Nonlinear, Stochastic Signal

Our next example considers the case where the clean signal involves nonlinear terms. That is, we consider the following underlying nonlinear signal

$$X_t = 0.1X_{t-1} - 0.5\cos(3X_{t-1}) + 0.4\sin(X_{t-2}) + 0.1X_{t-2} + Z_t, \quad t = 2, 3, \cdots \tag{43}$$

where $\{Z_t\}_{t\geq 0}$ is iid $\sim \mathcal{N}(0,1)$ and $X_t = Z_t$ for $t = 0,1$, which also appears in [28, Sec. VI]. We pass this signal again through the additive channel (1) with $\{N_t\}_{t\geq 1}$ iid $\sim \mathcal{N}(0,1)$, independent of $\{X_t\}_{t\geq 1}$. We again experimented with $d = 5$ for our filter and $n = 10^4$. Unlike the autoregressive signal case, Kalman filter is neither optimal nor implementable for this signal. Instead, we compare our filter with the extended Kalman filter [29], which is commonly used in practice for filtering nonlinear signals of known statistics. Note that, however, the extended Kalman filter is *not* an optimal filter, but just one heuristic that approximates the nonlinear terms with the first order Taylor expansions. Therefore, the extended Kalman filter would not necessarily perform better than our universal filter. Fig. 2 shows the MSE results of our filter, the noisy predictor, the gradient-descent filter, the best FIR filter, and the extended Kalman filter for the nonlinear signal (43). The single sample path result in Fig. 2(a) again shows the similar result as the autoregressive signal case in Fig. 1(a). The most notable point of this experiment is that our filter outperforms the performance of

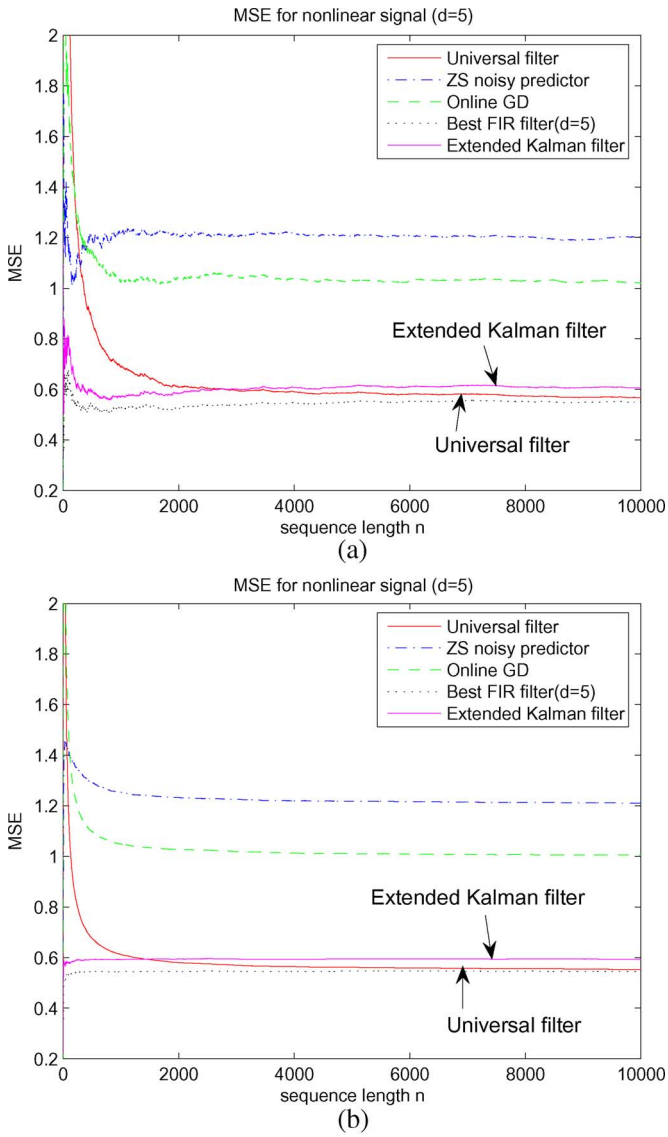

Fig. 2. MSEs for nonlinear signal (43). 2(a) is for a single sample path, and 2(b) is for the average of 100 experiments.

the extended Kalman filter. That is, although our filter only competes with linear filters with finite order, since the performance target of our filter is the best FIR filter that is determined by the actual realization of the signal and the noise, it can outperform the extended Kalman filter which is nonlinear and knows the signal model (43). Again, Fig. 2(b) shows the average performance which is consistent with the single sample path result.

C. Universality of Our Filter

The above two examples show that our filter, which does not know about the underlying signal model, can learn about the signal and perform as well as or better than the schemes that rely on the exact knowledge of the signal model. The third example stresses this powerful universality of our filter. We again experiment with the first order autoregressive signal and the nonlinear signal, but with different models. That is,

$$X_t = \alpha X_{t-1} + Z_t, \quad \alpha = 0.1, \quad t = 1, 2, \cdots \tag{44}$$

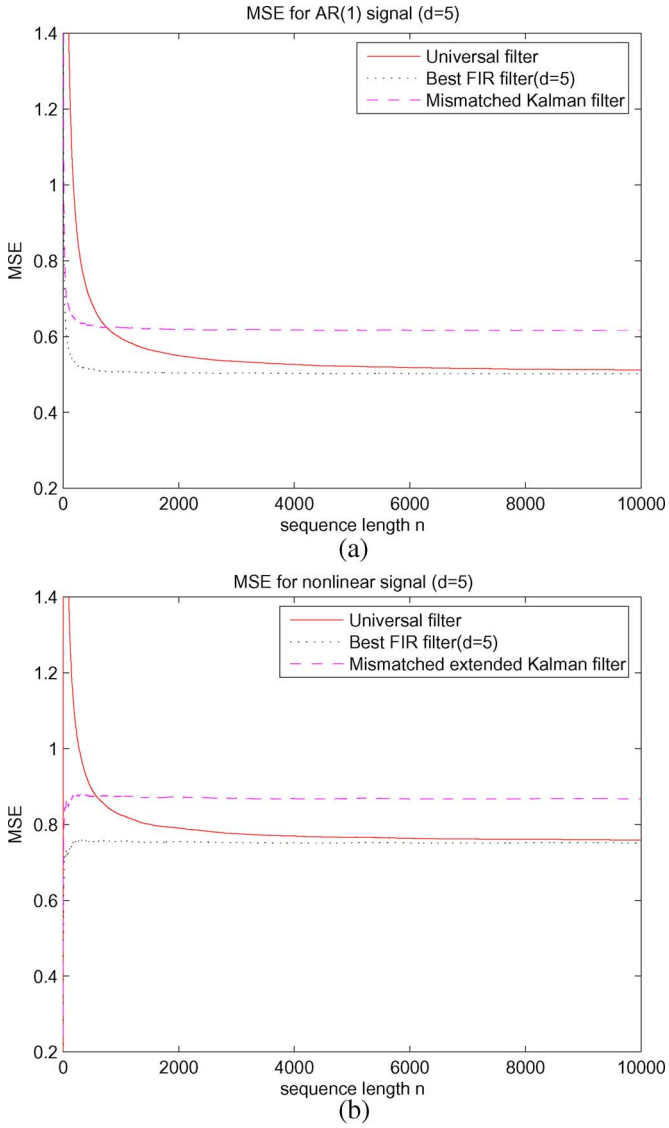

Fig. 3. Average MSEs for signals (44) and (45). Kalman filter and the extended Kalman filter used here are matched to wrong signals (40) and (43), respectively. (a) Average MSE result for autoregressive signal (44); (b) average MSE result for nonlinear signal (45).

and

$$X_t = 0.1X_{t-1} - 2\cos(5X_{t-1}) + \sin(0.1X_{t-2}) + 0.1X_{t-2} + Z_t, \quad t = 2, 3 \cdots \tag{45}$$

with the same initial conditions as (40) and (43), respectively, are now the inputs to the additive channel (1) with $\{N_t\}_{t\geq 1}$ iid $\sim \mathcal{N}(0,1)$, independent of $\{X_t\}_{t\geq 1}$. Since our filter does not depend on the signal model, the exact same scheme as what we used for the above two experiments is again applied for filtering both (44) and (45). For comparison schemes, we use the Kalman filter that is matched to (40) for (44) and the extended Kalman filter that is matched to (43) for (45) to see the sensitivity of those schemes to the underlying signal models. Fig. 3 shows the average MSE results of 100 experiments with $d = 5$ for our filter and sequence length $n = 10^4$. We observe that

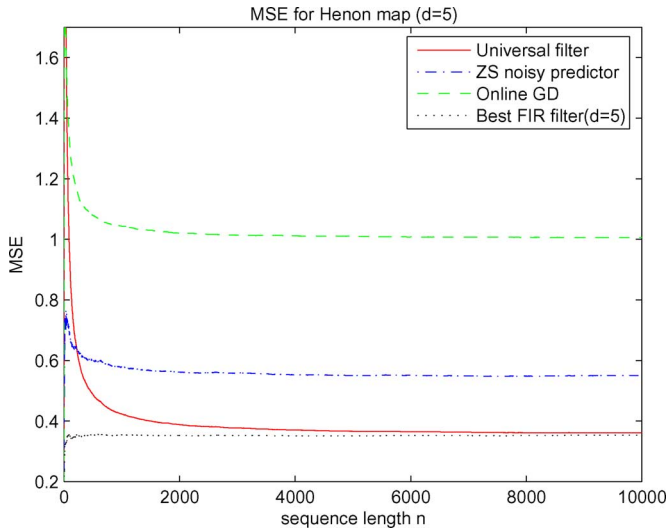

Fig. 4. MSE results averaged over 100 experiments for Henon map (46).

our filter outperforms the mismatched Kalman and extended Kalman filter for both cases with significant margins. These experiments plainly show that our filter *universally* attains the performance of the best FIR filter regardless of the signal models, whereas schemes that heavily depend on the knowledge of the signal models are very sensitive to the assumed models. Therefore, when there are uncertainties in the signal model, which is usually the case in practice, our universal filter clearly has a potential in improving on conventional filtering schemes that require knowledge of signal models.

D. Filtering Deterministic Signal

We next consider the case where the underlying signal is the Henon map,

$$X_t = 1 - 1.4X_{t-1}^2 + 0.3X_{t-2}, \quad t = 2, 3, \cdots \tag{46}$$

with $X_0 = X_1 = 0$, which is deterministic but known to exhibit chaotic behavior. For the demonstration of the chaotic behavior of Henon map, refer to [28, Sec. VI, Fig. 8]. Again, this signal is corrupted by the channel (1) with $\{N_t\}_{t\geq 1}$ iid $\sim \mathcal{N}(0,1)$. Now, since the underlying signal is deterministic, a filtering scheme that relies on the knowledge of the signal model does not make sense in this case, because knowing the model is equivalent to knowing the signal completely. Therefore, it is not clear what conventional schemes to apply for filtering the above Henon map-generated signal. However, we can still apply our universal filter since it does not depend on the underlying signal. Fig. 4 again shows the average MSE results of our filter, the noisy predictor, the gradient-descent filter, and the best FIR filter with $d = 5$ and $n = 10^4$. We observe that our filter reduces MSE significantly from the noise variance 1, which is the MSE of *saying-what-you-see* filter $\hat{X}_t(Y^t) = Y_t$, and outperforms both the noisy predictor and the gradient-descent filter significantly.

E. Effect of Constants on the Convergence Rate of Regret

Finally, our next example illustrates the effect of constants to the convergence rate of our filter, which was suppressed in the

presentation of our theorem. We set the nonlinear signal (43) as the underlying signal and measured the impact of three constants: the signal bound B_X, the noise variance σ^2, and the filter order d. We omitted to vary the bound on the noisy signal B_N since it is closely related to σ^2, and varying B_N would show the similar behavior as varying σ^2. Moreover, instead of varying B_X of the signal directly, we varied the variance of the innovation $\{Z_t\}_{t\geq 1}$ denoted as σ_Z^2 for the sake of simple simulation. Clearly, the effect of varying B_X is tied up with that of varying σ_Z^2. Fig. 5 summarizes the results of the experiments. First, note that instead of MSEs, the regrets

$$\frac{1}{n}\sum_{t=1}^{n}\left(x_t-\hat{X}_t^*(Y^t)\right)^2-\min_{\mathbf{u}\in\mathcal{U}}\frac{1}{n}\sum_{t=1}^{n}(x_t-\mathbf{u}^T\mathbf{Y}_t)^2$$

are plotted, and the scale of y axis of the plots are slightly different. The plots are again averages of 100 realizations with sequence length $n=10^4$. Fig. 5(a) shows the effect of the bound on the signal by experimenting with varying σ_Z^2. The noise variance σ^2 was fixed to 1, and the filter order was $d=5$. We can observe that as signal amplitude becomes large, the convergence of regret gets attenuated, but not so severely. On the contrary, Fig. 5(b) shows the effect of the noise variance and the bound on the noise by experimenting with varying σ^2. In this case, the innovation variance σ_Z^2 was fixed to 1, and the filter order was again $d=5$. The figure shows that larger noise variance, or smaller signal-to-noise ratio have an impact in attenuating the convergence rate of the regret more severely than the vice versa case. Fig. 5(c) shows the effect of the filter order d on the convergence rate of regret. The innovation variance σ_Z^2 and the noise variance σ^2 were all set to 1 in this experiment. We observe that, although the dependency on d was exponential in our upper bound (33), the slowdown of the convergence rate is not so severe in this case. Overall, although a qualitative statement, we state that despite the complex constant expressions in our analysis, the effect of those constants are not as severe as we got in our bound. Indeed, we believe this tendency of the dependency on the constants would mostly be the case in practice, since many of the constant bounds are obtained from the worst case scenario, i.e., signal or noise having always the maximum amplitudes.

From this representative set of simulations, we observe that our simple universal filter provides considerable performance gains in filtering noisy signals, especially when there are uncertainties in the underlying signal models.

VIII. Concluding Remarks and Future Work

We have devised a filtering scheme that, for every bounded underlying signal, performs essentially as well as the best FIR filter without any knowledge of the underlying signal and with only the knowledge of the first and second moments of the noise, under the MSE criterion. We showed that the regret vanishes in both expectation and high probability, and the decay rate of the regret of the expected MSE was shown to be logarithmic in n. The logarithmic regret was not straightforward to achieve due to the fact that the estimated loss functions $\{\ell_t(\mathbf{u})\}_{t\geq 1}$ are not always exp-concave functions. We also presented several

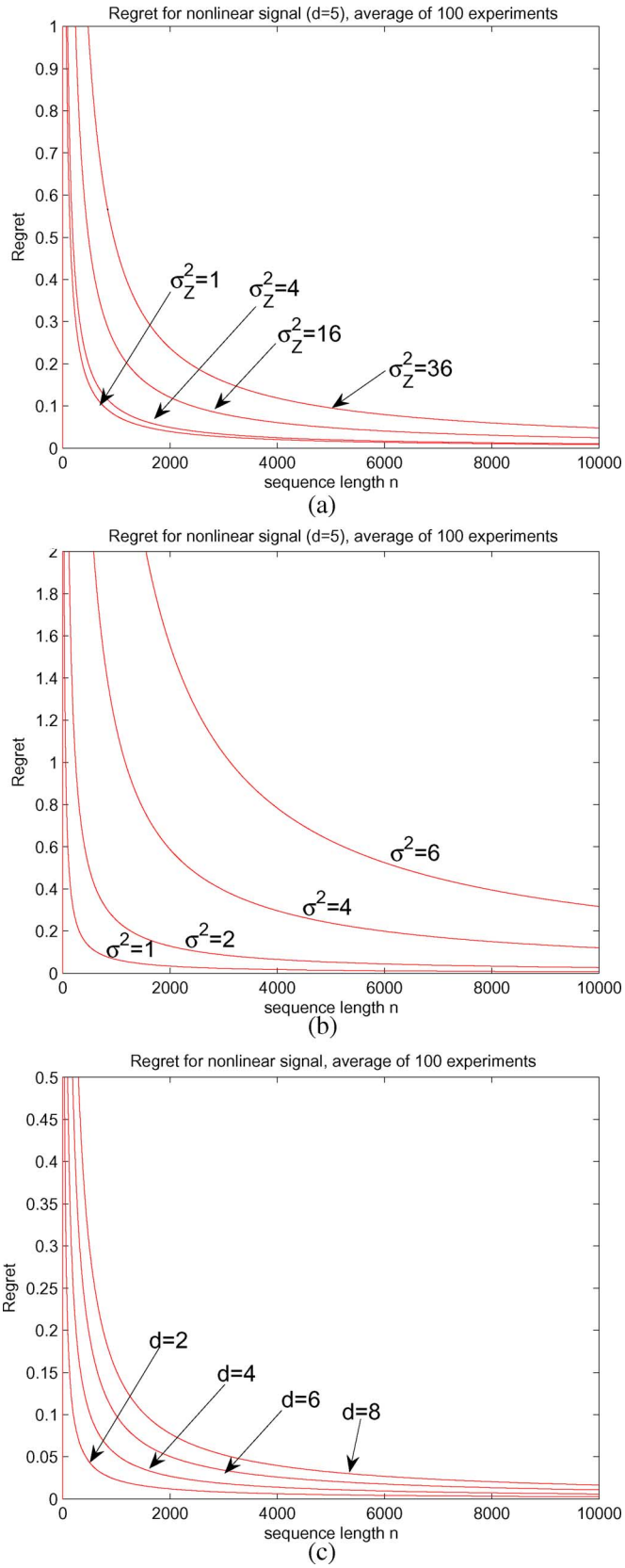

Fig. 5. Regrets averaged over 100 experiments for nonlinear signal (43) with varying parameters. (a) Regret when the innovation variance σ_Z^2 is varying; (b) regret when the noisy variance σ^2 is varying; (c) regret when the filter order d is varying.

simulation results that support our theoretical guarantees and show the potential merits of applying our filter in practice.

Although the dependency of the bounds on d was suppressed in our result, we can increase d with n with sufficiently slow speed and still guarantee the asymptotically optimal performance. We omit a detailed mathematical argument here, but, for example, for each sequence length n, if we set the order of our filter as $d_n = O(\log n)$, the regret of our filter to any FIR filter would still go to zero as n goes to infinity. This scheme resembles the schemes devised for the universal compression [30], prediction [31], and filtering [11] problems for finite-alphabet signals, that successfully compete with any order of Markov schemes. Our above scheme assumes the knowledge of n at the beginning of the filtering process to determine d_n, which implies that the scheme is not strongly sequential and depends on the horizon. However, it is straightforward to construct a strongly sequential scheme from above scheme by using techniques that are by now standard, e.g., doubling tricks in [25, Ch. 2.3]. That is, we can divide the sequence into blocks with exponentially growing length, and apply above scheme separately by blocks to make the regret to any FIR filter vanish as the sequence length increases.

As for future work, we can extend our scheme to compete with reference classes that are larger than the class of FIR filters in order to further minimize the MSE. One possible such extension is to devise a scheme that competes with the class of switching FIR filters that parallels the switching predictors in [32] and switching denoisers in [21]. Again, obtaining the expected regret of rate $O(m \log n/n)$ as in [32], where m is the number of switches, may not be straightforward since the loss function is not always exp-concave, a condition required for the scheme in [32]. Another direction is to compete with the class of general nonlinear schemes as has been done for the denoising (noncausal estimation) case in [33]. However, in this case, the procedure to obtain an unbiased estimate of the true MSE would not be as simple as that in this paper, since the martingale relationship in Lemma 1 relied heavily on the linearity of the filter. Instead, the channel inversion process developed in [33] may be a necessary component.

APPENDIX

A. Proof of Lemma 1

Proof: Fix $x^n \in \mathcal{D}^n$. Consider

$$\begin{aligned}
&E\left[\left(x_t - \mathbf{w}_{t-1}^T \mathbf{Y}_t\right)^2 |Y^{t-1}\right] \\
&\quad = E\left[\left(x_t^2 - 2x_t \mathbf{w}_{t-1}^T \mathbf{Y}_t + \mathbf{w}_{t-1}^T \mathbf{Y}_t \mathbf{Y}_t^T \mathbf{w}_{t-1}\right) |Y^{t-1}\right] \\
&\quad = E\left[\left\{\left(Y_t^2 - \sigma^2\right) - \left(2Y_t \mathbf{w}_{t-1}^T \mathbf{Y}_t - 2\mathbf{w}_{t-1}^T \mathbf{c}\right)\right.\right. \\
&\qquad \left.\left. + \mathbf{w}_{t-1}^T \mathbf{Y}_t \mathbf{Y}_t^T \mathbf{w}_{t-1}\right\} |Y^{t-1}\right] \\
&\quad = E\left[\left\{\left(Y_t - \mathbf{w}_{t-1}^T \mathbf{Y}_t\right)^2 + 2\mathbf{w}_{t-1}^T \mathbf{c} - \sigma^2\right\} |Y^{t-1}\right] \quad (47)
\end{aligned}$$

where (47) follows since $E(x_t^2|Y^{t-1}) = x_t^2 = E(Y_t^2 - \sigma^2) = E(Y_t^2 - \sigma^2|Y^{t-1})$ and $E(x_t \mathbf{w}_{t-1}^T \mathbf{Y}_t|Y^{t-1}) = \mathbf{w}_{t-1}^T E(x_t \mathbf{Y}_t|Y^{t-1}) = \mathbf{w}_{t-1}^T E(Y_t \mathbf{Y}_t - \mathbf{c}|Y^{t-1}) = E(Y_t \mathbf{w}_{t-1}^T \mathbf{Y}_t - \mathbf{w}_{t-1}^T \mathbf{c}|Y^{t-1})$. Therefore, for all $t \geq 1$

$$E\left[\Lambda_t(\mathbf{w}_{t-1}) - \left\{\ell_t(\mathbf{w}_{t-1}) - \sigma^2\right\} |Y^{t-1}\right] = 0. \quad (48)$$

Note that $\mathbf{w}_{t-1} \in \sigma(Y^{t-1})$ is crucial to have the above equality. Hence, $\{\Lambda_t(\mathbf{w}_{t-1}) - \{\ell_t(\mathbf{w}_{t-1}) - \sigma^2\}\}_{t \geq 1}$ is a martingale difference, and, therefore, $\{[\sum_{t=1}^n \Lambda_t(\mathbf{w}_{t-1}) - \sum_{t=1}^n \{\ell_t(\mathbf{w}_{t-1}) - \sigma^2\}]\}_{n \geq 1}$ is a martingale. ∎

B. Proof of Lemma 2

Proof: The argument for Part (a) and Part (b) almost coincides with that of [25, Ch. 11.7] except for the constant vector $\mathbf{c}$ in the definition of $\ell_t(\mathbf{u})$. But, that difference hardly affects the argument.

(a) From (4),

$$\begin{aligned}
\mathbf{w}_t^* &= A_t^{-1} \left(\sum_{i=1}^t \{Y_i \mathbf{Y}_i - \mathbf{c}\}\right) \\
&= A_t^{-1} \left(A_{t-1} \mathbf{w}_{t-1}^* + Y_t \mathbf{Y}_t - \mathbf{c}\right) \\
&= A_t^{-1} \left(A_t \mathbf{w}_{t-1}^* - \mathbf{Y}_t \mathbf{Y}_t^T \mathbf{w}_{t-1}^* + Y_t \mathbf{Y}_t - \mathbf{c}\right) \\
&= \mathbf{w}_{t-1}^* - A_t^{-1} \left\{\left(\mathbf{w}_{t-1}^{*T} \mathbf{Y}_t - Y_t\right) \mathbf{Y}_t + \mathbf{c}\right\}
\end{aligned}$$

Also

$$\begin{aligned}
\|\mathbf{w}_t^*\| &= \left\| A_t^{-1} \left(\sum_{i=1}^t \left\{Y_i \mathbf{Y}_i^T - \mathbf{c}\right\}\right) \right\| \\
&= \left\| A_t^{-1} \left(\sum_{i=1}^t \left\{\mathbf{Y}_i \mathbf{Y}_i^T - \sigma^2 I\right\} \mathbf{e}_1\right) \right\| \\
&= \left\| A_t^{-1} (A_t - I - t\sigma^2 I) \mathbf{e}_1 \right\| \\
&\leq \left\| I - (1 + t\sigma^2) A_t^{-1} \right\| \leq 1 + (1 + t\sigma^2) \lambda_{\max}\left(A_t^{-1}\right)
\end{aligned}$$

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

$$\begin{aligned}
&\left|\mathbf{w}_{t-1}^{*T} \mathbf{Y}_t - Y_t\right| \\
&\quad = \left|\left(\mathbf{w}_{t-1}^* - \mathbf{e}_1\right)^T \mathbf{Y}_t\right| \\
&\quad = \left|\left(A_{t-1}^{-1} \left(\sum_{i=1}^{t-1} \left\{\mathbf{Y}_i \mathbf{Y}_i^T - \sigma^2 I\right\} - A_{t-1}\right) \mathbf{e}_1\right)^T \mathbf{Y}_t\right| \quad (49) \\
&\quad \leq \left\|\left(1 + (t-1)\sigma^2\right) A_{t-1}^{-1} \mathbf{e}_1\right\| \cdot \|\mathbf{Y}_t\| \quad (50) \\
&\quad \leq \left(1 + (t-1)\sigma^2\right) \left\|A_{t-1}^{-1}\right\| \cdot \|\mathbf{Y}_t\| \quad (51) \\
&\quad = \left(1 + (t-1)\sigma^2\right) \lambda_{\max}\left(A_{t-1}^{-1}\right) \cdot \|\mathbf{Y}_t\|, \quad (52)
\end{aligned}$$

where (49) is from the definition (4), (50) is from Cauchy-Schwartz inequality, (51) is from the definition of matrix norm, and (52) is from the fact that A_{t-1}^{-1} is a symmetric matrix.

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

$$\begin{aligned}
\ell_t\left(\widehat{\mathbf{w}}_{t-1}^*\right) - \ell_t(\mathbf{u}) &= \left\{L_t\left(\widehat{\mathbf{w}}_{t-1}^*\right) - L_t\left(\mathbf{w}_t^*\right)\right\} \\
&\quad - \left\{L_t(\mathbf{u}) - L_t\left(\mathbf{w}_t^*\right)\right\} \\
&\quad + \left\{L_{t-1}(\mathbf{u}) - L_{t-1}\left(\widehat{\mathbf{w}}_{t-1}^*\right)\right\} \\
&\leq \left\{L_t\left(\widehat{\mathbf{w}}_{t-1}^*\right) - L_t\left(\mathbf{w}_t^*\right)\right\} \\
&\quad - \left\{L_t(\mathbf{u}) - L_t\left(\mathbf{w}_t^*\right)\right\} \\
&\quad + \left\{L_{t-1}(\mathbf{u}) - L_{t-1}\left(\mathbf{w}_{t-1}^*\right)\right\} \quad (53)
\end{aligned}$$

where (53) holds since $L_{t-1}(\mathbf{w}_{t-1}^*) \le L_{t-1}(\widehat{\mathbf{w}}_{t-1}^*)$ from definition in (4). Therefore, summing over t leads to

$$\begin{aligned}\sum_{t=1}^{n} &\left\{\ell_t\left(\widehat{\mathbf{w}}_{t-1}^*\right) - \ell_t(\mathbf{u})\right\} \\ &\le \left\{L_0(\mathbf{u}) - L_0\left(\mathbf{w}_0^*\right)\right\} - \left\{L_n(\mathbf{u}) - L_n\left(\mathbf{w}_n^*\right)\right\} \\ &\quad + \sum_{t=1}^{n}\left\{L_t\left(\widehat{\mathbf{w}}_{t-1}^*\right) - L_t\left(\mathbf{w}_t^*\right)\right\} \\ &\le \|\mathbf{u}\|^2 + \sum_{t=1}^{n}\left\{L_t\left(\widehat{\mathbf{w}}_{t-1}^*\right) - L_t\left(\mathbf{w}_t^*\right)\right\}. \end{aligned} \tag{54}$$

The inequality in (54) 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

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

which proves the lemma. ∎

C. Proof of Lemma 3

Proof:

(a) From the union bound,

$$P\left(\left\|\frac{1}{t}\sum_{i=1}^{t}\mathbf{Y}_i\mathbf{Y}_i^T - \left(\sigma^2 I + \frac{1}{t}\sum_{i=1}^{t}\mathbf{x}_i\mathbf{x}_i^T\right)\right\|_1 > \epsilon\right) \tag{55}$$

$$\le \sum_{1\le a,b\le d} P\left(\left|\left(\frac{1}{t}\sum_{i=1}^{t}\mathbf{Y}_i\mathbf{Y}_i^T\right)_{ab} - \left(\sigma^2 I + \frac{1}{t}\sum_{i=1}^{t}\mathbf{x}_i\mathbf{x}_i^T\right)_{ab}\right| > \frac{\epsilon}{d^2}\right) \tag{56}$$

where $(A)_{ab}$denotes the abth entry of the matrix A. Since $\sum_{i=1}^{t}\mathbf{Y}_i\mathbf{Y}_i^T = \sum_{i=1}^{t}\mathbf{x}_i\mathbf{x}_i^T + 2\sum_{i=1}^{t}\mathbf{x}_i\mathbf{N}_i^T + \sum_{i=1}^{t}\mathbf{N}_i\mathbf{N}_i^T$, we consider the concentration of $(1/t)(2\sum_{i=1}^{t}\mathbf{x}_i\mathbf{N}_i^T + \sum_{i=1}^{t}\mathbf{N}_i\mathbf{N}_i^T)$. We note that

$$\left(2\sum_{i=1}^{t}\mathbf{x}_i\mathbf{N}_i^T + \sum_{i=1}^{t}\mathbf{N}_i\mathbf{N}_i^T\right)_{ab} = 2\sum_{i=1}^{t}x_{i+1-a}N_{i+1-b} + \sum_{i=1}^{t}N_{i+1-a}N_{i+1-b},$$

and consider the cases when $a = b$ and $a \ne b$, separately. When $a = b$, we can verify that the sequence $\{2x_{i+1-b}N_{i+1-b} + N_{i+1-b}^2 - \sigma^2\}_{i\ge1}$ is a martingale difference with respect to $\{N_{i+1-b}\}_{i\ge1}$, since $\{N_i\}$ are assumed to be independent with $EN_i = 0$, $EN_i^2 = \sigma^2$ for all i. When $a \ne b$, without loss of generality, we can assume $a > b$. Then, we can again verify that $\{2x_{i+1-a}N_{i+1-b} - N_{i+1-a}N_{i+1-b}\}_{i\ge1}$ is a martingale difference with respect to $\{N_{i+1-b}\}_{i\ge1}$, since N_{i+1-a} and N_{i+1-b} are zero mean and independent. Therefore, since we assumed that $\{x_t\}$s and $\{N_t\}$s are all bounded, we can apply Hoeffding-Azuma inequality [25, Sec. A.1.3] to get the bound

$$P\left(\left|\left(\frac{1}{t}\sum_{i=1}^{t}\left\{\mathbf{N}_i\mathbf{N}_i^T + 2\mathbf{x}_i\mathbf{N}_i^T\right\}\right)_{ab} - (\sigma^2 I)_{ab}\right| > \frac{\epsilon}{d^2}\right) \le 2\exp\left(-\frac{2\epsilon^2}{B_N^2(B_N+2B_X)^2d^4}t\right) \tag{57}$$

which, combined with (56), proves part (a).

(b) In [34, (2.2)], we find

$$\max_i \min_j |\lambda_j - \mu_i| \le \frac{d+2}{d}G_{AB}^{1-1/d}\|A-B\|_1 \tag{58}$$

where $\{\lambda_j\}_{1\le j\le d}$ and $\{\mu_i\}_{1\le i\le d}$ are the eigenvalues of d-by-d matrix A and B, respectively, and $G_{AB} = \max_{i,j}(|a_{ij}|, |b_{ij}|)$. Let us denote $F_{AB} = ((d+2)/d)G_{AB}^{1-1/d}$. Then, since (58) is symmetric in A and B, the inequality $\max_j \min_i |\mu_i - \lambda_j| \le F_{AB}\|A-B\|_1$ is also true. Now, we observe that

$$|\lambda_{\min}(A) - \lambda_{\min}(B)| \le \max\left\{\max_i\min_j|\lambda_j-\mu_i|, \max_j\min_i|\mu_i-\lambda_j|\right\}$$

due to the symmetry of $|\lambda_{\min}(A) - \lambda_{\min}(B)|$ in A and B, and, thus, deduce that

$$|\lambda_{\min}(A) - \lambda_{\min}(B)| \le F_{AB}\|A-B\|_1 \tag{59}$$

i.e., the minimum eigenvalue is a Lipschitz continuous function of the elements of the matrix. Now, denote the event $E_t = \{\omega : \|(1/t)\mathbf{K}_t - (\sigma^2 I + (1/t)\sum_{i=1}^{t}\mathbf{x}_i\mathbf{x}_i^T)\|_1 \le \epsilon\}$. Then, if $\omega \in E_t$, we have

$$\frac{1}{t}\lambda_{\min}(\mathbf{K}_t) \ge \lambda_{\min}\left(\sigma^2 I + \frac{1}{t}\sum_{i=1}^{t}\mathbf{x}_i\mathbf{x}_i^T\right) - \epsilon F \tag{60}$$

$$\ge \sigma^2 - \epsilon F \tag{61}$$

where $F = (d+2)/d(B_X+B_N)^{2(1-1/d)}$, (60) is from (59), and (61) is from the fact that $(1/t)\sum_{i=1}^{t}\mathbf{x}_i\mathbf{x}_i^T$ is positive semidefinite. Since $F > 0$, by choosing $\epsilon = \sigma^2/2F$, part (b) is proven by applying the result of part (a). ∎

D. Proof of Lemma 4

Proof: Note first that, by the union bound, for any m

$$P\left(\bigcup_{t=m}^{\infty}\{Y_t > 0\}\right) \tag{62}$$

$$\le 2d^2\sum_{t=m}^{\infty}\exp(-tC) = \frac{2d^2}{1-\exp(-C)}\exp(-mC). \tag{63}$$

Fix $m < (3n\epsilon/c_1)^{1/3} - c_2 - 1$ for sufficiently large n. Then, consider

$$\begin{aligned}
&P\left(\frac{1}{n}\sum_{t=1}^{n} Y_t > \epsilon\right) \\
&= P\left(\sum_{t=1}^{n} Y_t > n\epsilon\right) = P\left(\sum_{t=1}^{m} Y_t + \sum_{t=m+1}^{n} Y_t > n\epsilon\right) \\
&\le P\left(\sum_{t=m+1}^{n} Y_t > n\epsilon - c_1\sum_{t=1}^{m}(c_2+t)^2\right) && (64)\\
&\le P\left(\sum_{t=m+1}^{n} Y_t > n\epsilon - \frac{c_1(m+c_2+1)^3}{3}\right) && (65)\\
&\le P\left(\sum_{t=m+1}^{n} Y_t > 0\right) && (66)\\
&\le P\left(\bigcup_{t=m+1}^{n}\{Y_t > 0\}\right) && (67)\\
&\le P\left(\bigcup_{t=m+1}^{\infty}\{Y_t > 0\}\right) \\
&\le \frac{2d^2}{1-\exp(-C)}\exp\left(-(m+1)C\right) && (68)
\end{aligned}$$

where (64) follows from the given bound $Y_t \le c_1(c_2+t)^2$; (65) follows from $\sum_{t=1}^{m}(c_2+t)^2 \le \sum_{t=1+c_2}^{m+c_2} t^2 \le \sum_{t=1}^{m+c_2} t^2 \le ((m+c_2+1)^3/3)$; (66) follows from the condition on m; (67) follows from the union of events, and (68) follows from (63). In particular, taking $m = \lfloor (3n\epsilon/c_1)^{1/3} - c_2 \rfloor - 1$ gives

$$\begin{aligned}
P\left(\frac{1}{n}\sum_{t=1}^{n} Y_t > \epsilon\right) \le \frac{2d^2}{1-\exp(-C)} \\
\times \exp\left(-\left\lfloor\left(\frac{3n\epsilon}{c_1}\right)^{1/3} - c_2\right\rfloor C\right)
\end{aligned}$$

and proves the lemma. ■

E. Proof of Lemma 5

To simplify the notation, we will use the $\Lambda_t(\mathbf{u})$ notation in Definition 1. First, note that we have following decomposition:

$$\begin{aligned}
&\sum_{t=1}^{n}\Lambda_t\left(\widehat{\mathbf{w}}_{t-1}^*\right) - \left\{\sum_{t=1}^{n}\Lambda_t(\mathbf{u}) + \|\mathbf{u}\|^2\right\} \\
&= \underbrace{\sum_{t=1}^{n}\left\{\Lambda_t\left(\widehat{\mathbf{w}}_{t-1}^*\right) - \left\{\ell_t\left(\widehat{\mathbf{w}}_{t-1}^*\right) - \sigma^2\right\}\right\}}_{(a)} \\
&\quad + \underbrace{\sum_{t=1}^{n}\left\{\ell_t\left(\widehat{\mathbf{w}}_{t-1}^*\right) - \ell_t(\mathbf{u})\right\} - \|\mathbf{u}\|^2}_{(b)} \\
&\quad \underbrace{- \sum_{t=1}^{n}\left\{\Lambda_t(\mathbf{u}) - \left\{\ell_t(\mathbf{u}) - \sigma^2\right\}\right\}}_{(c)}. && (69)
\end{aligned}$$

Then, from the union bound, we have

$$\begin{aligned}
&P\left(\frac{1}{n}\sum_{t=1}^{n}\Lambda_t\left(\widehat{\mathbf{w}}_{t-1}^*\right) - \frac{1}{n}\left\{\sum_{t=1}^{n}\Lambda_t(\mathbf{u}) + \|\mathbf{u}\|^2\right\}\right. \\
&\quad \left. > \epsilon + \Theta\left(\frac{\log n}{n}\right)\right) \\
&\le P\left(\frac{1}{n}(a) \ge \frac{\epsilon}{3}\right) + P\left(\frac{1}{n}(b) \ge \frac{\epsilon}{3} + \Theta\left(\frac{\log n}{n}\right)\right) \\
&\quad + P\left(-\frac{1}{n}(c) \ge \frac{\epsilon}{3}\right). && (70)
\end{aligned}$$

Since $\{\widehat{\mathbf{w}}_{t-1}^*\}_{t\ge 1}$ and $\mathbf{u}$ are bounded, and (a) and (c) are bounded martingale from Lemma 1. Thus, we can use the Hoeffding-Azuma inequality [25, Lemma A.7] to bound the first and third term of (70) as

$$\begin{aligned}
P\left(\frac{1}{n}(a) \ge \frac{\epsilon}{3}\right) &\le \exp\left(-n\frac{2\epsilon^2}{9L_{\max}^2}\right) \quad \text{and} && (71)\\
P\left(\frac{1}{n}(c) \le -\frac{\epsilon}{3}\right) &\le \exp\left(-n\frac{2\epsilon^2}{9L_{\max}^2}\right) && (72)
\end{aligned}$$

where

$$\begin{aligned}
L_{\max} \triangleq \max_{x_t\in\mathcal{D},\mathbf{Y}_t\in[-(B_X+B_N),(B_X+B_N)]^d,\mathbf{u}\in\mathcal{U}} \\
\times\left\{\Lambda_t(\mathbf{u}) - \ell_t(\mathbf{u}) + \sigma^2\right\}. && (73)
\end{aligned}$$

It is obvious that (71) and (72) vanish much faster than $\exp(-\Theta(n^{1/3}))$, and thus, the remaining property we need is

$$P\left(\frac{1}{n}(b) > \frac{\epsilon}{3} + \Theta\left(\frac{\log n}{n}\right)\right) \le \exp\left(-\Theta(n^{1/3})\right). \quad (74)$$

To show this, recall (14) and (19) and define

$$\begin{aligned}
Z_{1t} &\triangleq \left\{\sigma^2 + b_1\frac{1+\sigma^2 t}{1+\lambda_{\min}(K_t)}\right\}^2 \cdot \frac{1}{1+\lambda_{\min}(K_t)} \\
&\quad + \|A_t\|\left\|\widehat{\mathbf{w}}_{t-1}^* - \mathbf{w}_{t-1}^*\right\|^2 \\
&\quad + 2\|R_t\mathbf{Y}_t + \mathbf{c}\|\left\|\widehat{\mathbf{w}}_{t-1}^* - \mathbf{w}_{t-1}^*\right\| \\
Z_{2t} &\triangleq \left\{\sigma^2 + b_1\frac{1+\sigma^2 t}{1+(\sigma^2/2)t}\right\}^2 \cdot \frac{1}{1+(\sigma^2/2)t}.
\end{aligned}$$

Then, by denoting $Z_t = Z_{1t} - Z_{2t}$, again from Lemma 2(a) and Lemma 3(b), we have $P(Z_t > 0) \le 2d^2\exp(-(\sigma^4/2CF^2)t)$. Since K_t is positive semi-definite and $Z_{2t} \ge 0$, $Z_t \le Z_{1t} \le (\sigma^2 + b_1(1+\sigma^2 t))^2 = (b_1\sigma^2)^2((1/b_1) + (1/\sigma^2) + t)^2$. Hence, we can apply Lemma 4 and show

$$\begin{aligned}
&\exp\left(-\Theta(n^{1/3})\right) \\
&\quad \ge P\left(\frac{1}{n}\sum_{t=1}^{n} Z_t > \frac{\epsilon}{3}\right) && (75)
\end{aligned}$$

$$= P\left(\frac{1}{n}\sum_{t=1}^{n} Z_{1t} > \frac{\epsilon}{3} + \frac{1}{n}\sum_{t=1}^{n} Z_{2t}\right) \tag{76}$$

$$\geq P\left(\frac{1}{n}\left\{\sum_{t=1}^{n}\left\{\ell_t\left(\mathbf{w}_{t-1}^{*}\right) - \ell_t(\mathbf{u})\right\} - \|\mathbf{u}\|^2\right\} > \frac{\epsilon}{3} + \Theta\left(\frac{\log n}{n}\right)\right) \tag{77}$$

where (75) follows from Lemma 4; (76) follows from $Z_t = Z_{1t} - Z_{2t}$, and (77) follows from identical steps as in (12)–(18), the fact that $Z_{10} = Z_{20}$, and $\mathbf{w}_{-1}^{*} \triangleq 0$. Therefore, (74) and the lemma are proved. ■

ACKNOWLEDGMENT

The authors are grateful to Professor A. Dembo, Dr. O. Lévêque, and Professor A. Singer 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*. Upper Saddle River, NJ: Prentice-Hall, 2000.
[3] H. Poor, "On robust Wiener filtering," *IEEE Trans. Autom. Control*, vol. AC-25, no. 3, pp. 521–526, 1980.
[4] Y. Eldar and N. Nerhav, "A competitive minimax approach to robust estimation and random parameters," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1931–1946, 2004.
[5] Y. Eldar, A. Ben-Tal, and A. Nemirovski, "Linear minimax regreat estimation of deterministic parameters with bounded data uncertainties," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2177–2188, 2004.
[6] S. Haykin, *Adaptive Filter Theory*, 4th ed. Upper Saddle River, NJ: Prentice-Hall, 2002.
[7] S. Haykin, *Unsupervised Adaptive Filtering: Volume I, II*. New York: Wiley, 2000.
[8] H. Robbins, "Asymptotically subminimax solutions of compound statistical decision problems," in *Proc. 2nd Berkeley Symp. Math. Statis. Prob.*, 1951, pp. 131–148.
[9] J. Hannan, "Approximation to bayes risk in repeated play," *Contrib. Theory of Games*, vol. III, pp. 97–139, 1957.
[10] J. V. Ryzin, "The sequential compound decision problem with $m \times n$ finite loss matrix," *Ann. Math. Statist.*, vol. 37, pp. 954–975, 1966.
[11] T. Weissman, E. Ordentlich, M. Weinberger, A. Somekh-Baruch, and N. Merhav, "Universal filtering via prediction," *IEEE Trans. Inf. Theory*, vol. 53, no. 4, pp. 1253–1264, 2007.
[12] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2124–2147, 1998.
[13] V. Vovk, "Competitive on-line statistics," *Int. Statist. Rev.*, vol. 69, pp. 213–248, 2001.
[14] A. Singer, S. Kozat, and M. Feder, "Universal linear least squares prediction: Upper and lower bounds," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2354–2362, 2002.
[15] G. Zeitler and A. Singer, "Universal linear least-squares prediction in the presence of noise," in *Proc. IEEE/SP 14th Workshop on Statist. Signal Process.*, Aug. 2007, pp. 611–614.
[16] T. Weissman and N. Merhav, "Universal prediction of individual binary sequences in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 47, no. 6, pp. 2151–2173, 2001.
[17] T. Moon and T. Weissman, "Competitive on-line linear FIR MMSE filtering," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2007, pp. 1126–1130.
[18] D. Donoho and I. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Statist. Assoc.*, vol. 90, no. 432, pp. 1200–1224, 1995.
[19] W. James and C. Stein, "Estimation with quadratic loss," in *Proc. 4th Berkeley Symp. Math. Statist. Prob.*, 1961, vol. 1, pp. 311–319.
[20] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger, "Universal discrete denoising: Known channel," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 5–28, 2005.
[21] T. Moon and T. Weissman, "Discrete denoising with shifts," submitted to *IEEE Trans. Inf. Theory,* Aug. 2007 [Online]. Available: http://arxiv.org/abs/0708.2566v1
[22] S. Vardeman, "Admissible solutions of k-extended finite state set and the sequence compound decision problmes," *J. Multiv. Anal.*, vol. 10, pp. 426–441, 1980.
[23] E. Hazan, A. Agarwal, and S. Kale, "Logarithmic regret algorithms for online convex optimization," *Mach. Learn.*, vol. 69, no. 2–3, pp. 169–192, 2007.
[24] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *Proc. 20th Int. Conf. (ICML)*, 2003, pp. 928–936.
[25] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge, U.K.: Cambridge Univ. Press, 2006.
[26] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.
[27] T. Moon and T. Weissman, "Universal filtering via hidden Markov modeling," *IEEE Trans. Inf. Theory*, vol. 54, no. 2, pp. 692–708, 2008.
[28] S. Kozat, A. Singer, and G. Zeitler, "Universal piecewise linear prediction via context trees," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3730–3745, 2007.
[29] Stanford EE 363 Lecture Note 8 [Online]. Available: http://www.stanford.edu/class/ee363/ekf.pdf
[30] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inf. Theory*, vol. 24, no. 5, pp. 5530–5536, 1978.
[31] M. Feder, N. Merhav, and M. Gutman, "Universal prediction for individual sequences," *IEEE Trans. Inf. Theory*, vol. 38, no. 4, pp. 1258–1270, 1992.
[32] S. Kozat and A. Singer, "Universal switching linear least squares prediction," *IEEE Trans. Signal Process.*, vol. 56, no. 1, pp. 189–204, 2008.
[33] K. Sivaramakrishnan and T. Weissman, "Universal denoising of discrete-time continuous-amplitude signals," *IEEE Trans. Inf. Theory*, vol. 54, no. 12, pp. 5632–5660, Dec. 2008.
[34] L. Elsner, "On the variation of the spectra of matrices," *Linear Algebra and Its Applications*, vol. 47, pp. 127–138, 1982.

Taseup Moon (S'04–M'08) received the B.S. degree in electrical engineering from Seoul National University, Seoul, Korea, in 2002, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 2004 and 2008, respectively.

He joined Yahoo! Inc., Sunnyvale, CA, as a research scientist in 2008. His research interests are in information theory, statistical signal processing, machine learning, and information retrieval.

Dr. Moon was awarded the Samsung Scholarship and a fellowship from the Korean Foundation of the Advanced Studies.

Tsachy Weissman (S'99–M'02–SM'07) received the B.Sc. and Ph.D. degrees in electrical engineering from the Technion, Israel, in 1997 and 2001, respectively.

He has held postdoctoral appointments with the Statistics Department, Stanford University, Stanford, CA, and with Hewlett-Packard Laboratories, Palo Alto, CA. Currently he is with the Departments of Electrical Engineering, Stanford University, and at the Technion. His research interests span information theory and its applications, and statistical signal processing. His papers thus far have focused mostly on data compression, communications, prediction, denoising, and learning. He is also inventor or coinventor of several patents in these areas and has been involved in a number of high-tech companies as a researcher or member of the technical board.

Dr. Weissman has received the NSF CAREER award and a Horev fellowship for leaders in Science and Technology. He is a Robert N. Noyce Faculty Scholar of the School of Engineering at Stanford University, and a recipient of the 2006 IEEE joint IT/COM societies Best Paper Award.