Competitive Prediction Under Additive Noise

Suleyman S. Kozat and Andrew C. Singer

***Abstract*—In this correspondence, we consider sequential prediction of a real-valued individual signal from its past noisy samples, under square error loss. We refrain from making any stochastic assumptions on the generation of the underlying desired signal and try to achieve uniformly good performance for any deterministic and arbitrary individual signal. We investigate this problem in a competitive framework, where we construct algorithms that perform as well as the best algorithm in a competing class of algorithms for each desired signal. Here, the best algorithm in the competition class can be tuned to the underlying desired clean signal even before processing any of the data. Three different frameworks under additive noise are considered: the class of a finite number of algorithms; the class of all pth order linear predictors (for some fixed order p); and finally the class of all switching pth order linear predictors.**

***Index Terms*—Additive noise, competitive, real valued, sequential decisions, universal prediction.**

I. Introduction

In this correspondence, we investigate "sequential" prediction of a real-valued and bounded individual sequence from its past noisy samples. Specifically, we consider the case when the corrupting noise is independent identically distributed (i.i.d.) and additive. Here, neither the desired clean signal nor its past samples are available for constructing predictions or training the underlying algorithm, yet, the goal is to predict the (unavailable) clean signal. This framework models the case in which the desired deterministic signal is observed through an additive white noise channel and, then, predicted using only the received past noise-corrupted output samples. The desired signal is represented by $x[t]$, where $|x[t]| \leq A_x$, $A_x \in \mathbb{R}^+$. Instead of directly observing $x[t]$, we observe *only* a noise-corrupted version of $x[t]$, i.e., $y[t] = x[t] + z[t]$. As the noise model, we take $z[t]$ as a zero mean i.i.d. random process, where $|z[t]| \leq A_z$, $A_z \in \mathbb{R}^+$. Although, we observe only the noisy signal $y[t]$ and the clean signal $x[t]$ is not available, the performance measure, including the loss function, is still taken with respect to the desired clean signal $x[t]$. We consider the square error loss function, however, our results can be generalized to several different loss functions, such as those considered in [1].

If the desired signal $x[t]$ and the noise process $z[t]$ are assumed to be random processes, the optimal predictor of $x[t]$ that minimizes the mean-square error (MSE) between the desired signal and the predictions is the conditional mean, $E[x[t]|y_1^{t-1}]$, $y_1^{t-1} \triangleq \{y[1], \ldots, y[t-1]\}$ [2]. This predictor is *optimal* on the *average* over the ensemble of outcomes (in MSE sense), however, calculation of the conditional mean requires the statistics of the underlying signals. First, the underlying signal $x[t]$ may not be well-modeled as a stochastic process. Second, the desired signal $x[t]$ is not directly observable, hence it may not be possible to estimate its statistics, if they existed in a meaningful sense. Approaching this problem from an adaptive prediction perspective has a number of issues, since one usually needs the error between the predictions and the desired signal $x[t]$ for training, which is not available. While blind adaptive prediction algorithms exist, such blind algorithms usually exploit certain statistics of the underlying signal $x[t]$, such as the kurtosis, to operate [2].

Hence, we refrain from making statistical assumptions on $x[t]$ and desire uniformly good performance for any deterministic and arbitrary signal $x[t]$, $t \geq 1$. Since we do not employ a statistical framework for $x[t]$, to define a performance measure, we investigate the prediction problem in a competitive algorithm framework [1], [3]. In this approach, we have a class of algorithms that we call the competition class. The algorithms in the competition class are all thought of as working in parallel to predict the next sample $x[t]$. Suppose there are m such algorithms, producing predictions, $\hat{x}_k[t]$, $k = 1, \ldots, m$. Then, each algorithm has an implicit accumulated squared prediction error, $\sum_{t=1}^{n}(x[t] - \hat{x}_k[t])^2$. We note that we do not have access to this accumulated loss since we are unable to observe the clean signal $x[t]$. Our goal is to introduce a sequential algorithm, say $\hat{x}_y[t]$, that observes only past corrupted samples $y[1], \ldots, y[t-1]$, and whose accumulated loss nearly achieves that of the best algorithm in this class, i.e.,

$$\frac{1}{n}\sum_{t=1}^{n}(x[t] - \hat{x}_y[t])^2 - \frac{1}{n}\min_k \sum_{t=1}^{n}(x[t] - \hat{x}_k[t])^2 \leq \frac{o(n)}{n} \qquad (1)$$

uniformly for all n and x_1^n. Here, $(o(n)/n) \to 0$ as $n \gg 1$. We stress that $\hat{x}_y[t]$ does not observe $x[t]$ or have access to its prediction performance with respect to $x[t]$. After making its prediction, $\hat{x}_y[t]$, it will only observe $y[t]$.

Such competitive framework for sequential prediction of deterministic sequences was investigated in [1] and [3] against a finite number of predictors; in [4] against the class of fixed-order linear models; and finally, in [5] and [6] against switching linear and certain nonlinear models, respectively. However, in these past approaches [1], [4]–[6], there is no consideration for noise. To make their predictions of $x[t]$ at time t, say $\hat{x}[t]$, these algorithms observe and make explicit use of the clean sequence $\{x[1], \ldots, x[t-1]\}$. After producing their prediction and observing the clean desired signal $x[t]$, they use the prediction error, e.g., $(x[t] - \hat{x}[t])$, to further train their parameters. Hence, these results cannot be generalized to our case, since, here, we observe only the noise corrupted version of the desired signal $y[t]$. To make predictions at time t on $x[t]$, say $\hat{x}_y[t]$, we only have access to $\{y[1], \ldots, y[t-1]\}$. Further, after the prediction, $\hat{x}_y[t]$, is produced, we can only use the prediction error $(y[t] - \hat{x}_y[t])$, albeit, our performance metric is still with respect to the original desired signal $x[t]$, e.g., $\sum_t (x[t] - \hat{x}_y[t])^2$.

The framework investigated in this correspondence, i.e., additive noise on an individual deterministic sequence, is introduced in [7] for binary prediction. The results in [7] are extended to the filtering problem in [8], where the underlying algorithm is allowed to use all $\{y[1], \ldots, y[t]\}$ (including $y[t]$) to make its decisions on $x[t]$. We are inspired by [7] and [8] to extend the results presented in [3], [5], and [6] to the noise-corrupted prediction problem. In the linear filtering approach introduced in [8], knowledge of certain statistics of the noise process are required. Here, we investigate deterministic real-valued sequences and our setup is prediction, not filtering. Some initial and partial results of this correspondence were introduced in [9] in the linear prediction context. However, we note that the competition class discussed in [9] is the "best" pth-order linear predictor (for some p) tuned to the sequence $y[t]$, $t \geq 1$. Hence, this "best" predictor is just a particular predictor that is tuned to the noise corrupted signal $y[t]$, not to $x[t]$. Here, we compete against all linear predictors that have the form $\boldsymbol{w}^T \boldsymbol{y}[t-1]$, $\boldsymbol{w} \in \mathbb{R}^p$, $\boldsymbol{y}[t-1] = y[t-1], \ldots, y[t-p]$, where

Manuscript received June 07, 2008; accepted March 02, 2009. First published May 05, 2009; current version published August 12, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Marcelo G. S. Bruno. This work is supported in part by TUBITAK Career Award, Contract No. 108E195.

S. S. Kozat is with the Electrical Engineeirng and Electronics Department, Koc University, Istanbul 34450, Turkey (e-mail: skozat@ku.edu.tr).

A. C. Singer is with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana IL 61801 USA (e-mail: acsinger@ad.uiuc.edu).

Digital Object Identifier 10.1109/TSP.2009.2022357

the linear weights $\boldsymbol{w}$ can be tuned even by observing the whole $x[t]$ and $y[t]$, $t \geq 1$, beforehand. Furthermore, even in this restricted case [9], only a probabilistic bound was given. We extend this result not only to general linear predictors, but also provide both MSE results as well as bounds on probability. In addition, we also study competition against the class of a finite number of predictors as well as the class of all switching linear predictors. When the competition class is a finite class of predictors, we require only a bound on $y[t]$ to construct the algorithm. When the competition class is the class of all pth-order linear predictors, unlike [8], we require neither bounds on $y[t]$, $x[t]$, $z[t]$ nor the variance of $z[t]$. To construct the sequential algorithm for switching pth-order linear predictors, we only require a bound on $y[t]$. Our performance results are guaranteed to hold without any further assumptions on $x[t]$. We only require that the noise process is i.i.d. and that the variance of $z[t]$ exists.

The organization of the correspondence is as follows. We first investigate sequential prediction when the competition class contains a finite number of algorithms. We then continue with pth-order linear predictors, for a given p, and then investigate switching pth-order linear predictors. The correspondence concludes with simulations of these algorithms in one-step-ahead prediction.

II. Prediction Under Noise

For a real-valued and bounded data sequence, $x[t]$, $t \geq 1$, $x[t] \in [-A_x, A_x]$, $A_x \in \mathbb{R}^+$, we observe a noise-corrupted version of $x[t]$, $y[t] = x[t] + z[t]$, where $z[t]$ is a bounded real-valued i.i.d. zero mean noise process such that $z[t] \in [-A_z, A_z]$, $A_z \in \mathbb{R}^+$. Hence, we have $|y[t]| \leq A_y$, where $A_y \triangleq A_x + A_z$. In this framework, we consider following problems.[1]

A. Finite Competition Class

At each time t, we observe outcomes from m different adaptive algorithms, producing predictions $\hat{x}_j[t]$, $j = 1, \ldots, m$, of $x[t]$. Each $\hat{x}_j[t]$ is sequential such that $\hat{x}_j[t]$ only depends on $\{y[1], \ldots, y[t-1]\}$, but nothing from the future. The accumulated square-error of each algorithm is given by $\sum_{t=1}^{n}(x[t] - \hat{x}_j[t])^2$ (which is not observable). At time t, our algorithm observes $\{\hat{x}_j[t]\}_{j=1}^{m}$ and y_1^{t-1}, and reveals its prediction of $x[t]$ as $\hat{x}_{y,1}[t]$. Then, $y[t]$ is revealed, however, our performance measure is with respect to $x[t]$, i.e., $\sum_{t=1}^{n}(x[t] - \hat{x}_{y,1}[t])^2$. For this setup, we investigate an updated version of the sequential algorithm introduced in [3] given as

$$\hat{x}_{y,1}[t] \triangleq \sum_{r=1}^{m} \mu_r[t]\, \tilde{x}_r[t] \tag{2}$$

with

$$\mu_r[t] = \frac{\exp\left(-\frac{1}{c}\sum_{l=1}^{t-1}\{y[l] - \tilde{x}_r[l]\}^2\right)}{\sum_{i=1}^{m}\exp\left(-\frac{1}{c}\sum_{l=1}^{t-1}\{y[l] - \tilde{x}_r[l]\}^2\right)} \tag{3}$$

where $c = 8A_y^2$ and $\tilde{x}_r[t] \triangleq (\hat{x}_r[t])^+$ is the clipped $\hat{x}_r[t]$ into the interval $[-A_y, A_y]$. Clearly, $\hat{x}_{y,1}[t]$ does not observe $x[t]$ and only has access to the past samples, y_1^{t-1}, and predictions $\{\hat{x}_r[l]\}_{l=1}^{t}$, $r = 1, \ldots, m$, for all t. Here, $\hat{x}_{y,1}[t]$ is a performance-based mixture of the constituent algorithms. We note that, although $\hat{x}_{y,1}[t]$ will be judged with respect to $x[t]$, it is only allowed to use the performance of each $\tilde{x}_r[t]$ on $y[t]$ (not on $x[t]$) to calculate its mixture weights, while combining the $\tilde{x}_r[t]$'s. For this algorithm, we have the following results.

Theorem 1: Let $x[t]$ be a real-valued and bounded sequence, $x[t] \in [-A_x, A_x]$, $A_x \in \mathbb{R}^+$, $y[t] = x[t] + z[t]$ be the observation sequence, $z[t] \in [-A_z, A_z]$, $A_z \in \mathbb{R}^+$, be an i.i.d. noise process with zero mean and $\{\hat{x}_j[t]\}_{j=1}^{m}$ are predictions of m adaptive algorithms. The sequential algorithm $\hat{x}_{y,1}[t]$ when applied to $y[t]$, $t \geq 1$, satisfies, for all n,

$$\frac{1}{n}E\left\{\sum_{t=1}^{n}(x[t] - \hat{x}_{y,1}[t])^2 - \sum_{t=1}^{n}(x[t] - \hat{x}_j[t])^2\right\} \leq 8A_y^2\frac{\ln(m)}{n} + O\left(\frac{1}{n}\right) \tag{4}$$

and for any small $\epsilon > 0$

$$\Pr\left\{\frac{1}{n}\sum_{t=1}^{n}(x[t] - \hat{x}_{y,1}[t])^2 - \frac{1}{n}\sum_{t=1}^{n}(x[t] - \hat{x}_j[t])^2 \leq 8A_y^2\frac{\ln(m)}{n} + O\left(\frac{\epsilon}{n}\right)\right\} \geq 1 - 2\exp(-n\epsilon^2 B), \tag{5}$$

for any $j = 1, \ldots, m$, where the expectation in (4) and probability in (5) are with respect to noise process. Here, $B = \left(1/4A^2\sigma_z^2\right)$, $A = 4A_y$, $0 < \epsilon < \left(2A\sigma_z^2/A_z\right)$ and σ_z^2 is the variance of $z[t]$.

Theorem 1 holds for any deterministic sequence $x[t]$ without any stochastic assumptions. It states that the performance of $\hat{x}_{y,1}[t]$ is within $O(\ln(m)/n)$ of the best algorithm in the competition class that can only be chosen in hindsight by observing x_1^n and y_1^n, for all n. The upper bounds in (4) and (5) can be improved to $2A_y^2\ln(m)$ (instead of $8A_y^2\ln(m)$) by using the Aggregating Algorithm of [1] instead of the convex combination of (3).

Proof of Theorem 1: The main idea of the proof of Theorem 1 is to transform the loss with respect to the clean signal $(x[t] - \hat{x}_j[t])^2$ to the loss with respect to the noisy signal $(y[t] - \hat{x}_j[t])^2$. For any sequential algorithm $\hat{x}_y[t]$, we observe that

$$\begin{aligned} E[(y[t] - \hat{x}_y[t])^2] &= E\left(y[t]^2 - 2y[t]\hat{x}_y[t] + \hat{x}_y[t]^2\right) \\ &= E\left\{(x[t] + z[t])^2 - 2(x[t] + z[t])\hat{x}_y[t] + \hat{x}_y[t]^2\right\} \\ &= E\left\{x^2[t] + \sigma_z^2 - 2x[t]\hat{x}_y[t] + \hat{x}_y[t]^2\right\} \\ &= E\left\{(x[t] - \hat{x}_y[t])^2\right\} + \sigma_z^2 \end{aligned} \tag{6}$$

where in the second line, we observe that $z[t]$ is independent of the past realizations $\{y[1], \ldots, y[t-1]\}$, $x[t]$ and $\hat{x}_y[t]$. Hence, the difference between the accumulated loss of any sequential algorithm and any constituent algorithm (that is clipped) can be written as

$$\begin{aligned} &\sum_{t=1}^{n}\left\{E[(x[t] - \hat{x}_y[t])^2] - E[(x[t] - \tilde{x}_j[t])^2]\right\} \\ &\quad= \sum_{t=1}^{n}\left\{E[(y[t] - \hat{x}_y[t])^2] - \sigma_z^2 - E[(y[t] - \tilde{x}_j[t])^2] + \sigma_z^2\right\} \\ &\quad= \sum_{t=1}^{n}\left\{E[(y[t] - \hat{x}_y[t])^2] - E[(y[t] - \tilde{x}_j[t])^2]\right\} \end{aligned} \tag{7}$$

where $j = 1, \ldots, m$. Thus, performance with respect to $x[t]$ can be transformed into performance with respect to $y[t]$ in an expected sense. However, when $\hat{x}_{y,1}[t]$ is applied to $y[t]$, $t \geq 1$, we have the following result from [3]:

$$\sum_{t=1}^{n}(y[t] - \hat{x}_{y,1}[t])^2 - \sum_{t=1}^{n}(y[t] - \tilde{x}_j[t])^2 \leq 8A_y^2\ln(m) + O(1) \tag{8}$$

[1]All vectors are column vectors and represented by lowercase bold letters. For a vector $\boldsymbol{w}$, $|\boldsymbol{w}|_1 \triangleq \sum_i |w_i|$ is the l_1 norm, $|\boldsymbol{w}|_2 \triangleq \sqrt{\sum_i \boldsymbol{w}_i^2}$ is the l_2 norm. For a real number a, $|a|$ is the absolute value and $\boldsymbol{w}^T$ is the transpose of $\boldsymbol{w}$. For a symmetric matrix $\boldsymbol{R} \in \mathbb{R}^{p\times p}$, $\lambda_i(\boldsymbol{R})$, $i = 1, \ldots, p$ are the eigenvalues sorted in a descended order, based on value. For a real number $x \in \mathbb{R}$, $(x)^+ = x$ if $|x| \leq A_y$, $(x)^+ = A_y$ if $x > A_y$ and $(x)^+ = -A_y$ if $x < -A_y$, i.e., $(\cdot)^+$ is clipping into the $[-A_y, A_y]$ interval.

for any $j = 1,\ldots,m$. Noting that clipping $\hat{x}_j[t]$ into $[-A_y, A_y]$ will only improve the prediction performance of $\hat{x}_j[t]$, since $x[t] \in [-A_x, A_x] \subset [-A_y, A_y]$ and using (8) in (7) yields

$$\begin{aligned} &E\left[\sum_{t=1}^{n}(x[t]-\hat{x}_{y,1}[t])^2\right] - E\left[\sum_{t=1}^{n}(x[t]-\hat{x}_j[t])^2\right] \\ &\quad\le E\left[\sum_{t=1}^{n}(x[t]-\hat{x}_{y,1}[t])^2\right] - E\left[\sum_{t=1}^{n}(x[t]-\tilde{x}_j[t])^2\right] \\ &\quad\le 8A_y^2\ln(m)+O(1). \end{aligned}$$

This completes the first part of proof of Theorem 1. □

To prove (5), for any sequential algorithm $\hat{x}_y[t]$, including $\hat{x}_{y,1}[t]$, we have

$$\begin{aligned} &(y[t]-\hat{x}_y[t])^2-(y[t]-\tilde{x}_j[t])^2 \\ &\quad=(x[t]+z[t]-\hat{x}_y[t])^2-(x[t]+z[t]-\tilde{x}_j[t])^2 \\ &\quad=(x[t]-\hat{x}_y[t])^2+2z[t](x[t]-\hat{x}_y[t])+z[t]^2 \\ &\qquad-(x[t]-\tilde{x}_j[t])^2-2z[t](x[t]-\tilde{x}_j[t])-z[t]^2 \\ &\quad=(x[t]-\hat{x}_y[t])^2-(x[t]-\tilde{x}_j[t])^2-2z[t](\hat{x}_y[t]-\tilde{x}_j[t]). \end{aligned}$$

This yields

$$\begin{aligned} &\sum_{t=1}^{n}(x[t]-\hat{x}_y[t])^2-\sum_{t=1}^{n}(x[t]-\tilde{x}_j[t])^2 \\ &\quad=\sum_{t=1}^{n}(y[t]-\hat{x}_y[t])^2-\sum_{t=1}^{n}(y[t]-\tilde{x}_j[t])^2 \\ &\qquad+2\sum_{t=1}^{n}z[t](\hat{x}_y[t]-\hat{x}_j[t]). \end{aligned} \tag{9}$$

We know from (8) that the first term in the right-hand side of (9) is bounded by $O(\ln(m))$, when $\hat{x}_y[t] = \hat{x}_{y,1}[t]$. For $\sum_{t=1}^{n} z[t](\hat{x}_y[t]-\tilde{x}_j[t])$, we have the following. Since $\tilde{x}_j[t] \in [-A_y, A_y]$, then $\hat{x}_{y,1}[t] \in [-A_y, A_y]$, due to convex combination in (2). Hence, $|(\tilde{x}_j[t]-\hat{x}_y[t])| \le 2A_y$ for all t. Since, clipping only improves the performance, this yields

$$\begin{aligned} &\sum_{t=1}^{n}(x[t]-\hat{x}_{y,1}[t])^2-\sum_{t=1}^{n}(x[t]-\hat{x}_j[t])^2 \\ &\quad\le\sum_{t=1}^{n}(x[t]-\hat{x}_{y,1}[t])^2-\sum_{t=1}^{n}(x[t]-\tilde{x}_j[t])^2 \\ &\quad\le 8A_y^2\ln(m)+2A_y\sum_{t=1}^{n}z[t]. \end{aligned} \tag{10}$$

Since $\sum_{t=1}^{n} z[t]$ is sum of n i.i.d. noise samples bounded by A_z, using the Chernoff bound in (10) on $\sum_{t=1}^{n} z[t]$ yields the second part of the result in Theorem 1. This completes the proof of Theorem 1. □

B. Linear Prediction

Here, the competition class is the class of all pth order fixed linear predictors, i.e., $\boldsymbol{w}^T\boldsymbol{y}[t-1]$, $\boldsymbol{w} \in \mathbb{R}^p$, for some p. The goal is then to find a sequential algorithm which depends only on y_1^{t-1} and achieves, for all n, performance of the best linear predictor that is tuned to $x[t]$ and $y[t]$, $t \ge 1$. For any $\boldsymbol{w}$ and n, we define the accumulated loss of a linear predictor as $\sum_{t=1}^{n}(x[t]-\boldsymbol{w}^T\boldsymbol{y}[t-1])^2+\delta|\boldsymbol{w}|_2^2$, for all $\boldsymbol{w} \in \mathbb{R}^p$, for all $x[t]$, $t \ge 1$ and $\delta > 0$. We included the additional term $\delta|\boldsymbol{w}|^2$ for regularization purposes and note that this modified loss is often called the ridge-regression loss [10].

For this framework, we apply the sequential algorithm [4]

$$\hat{x}_{y,2}[t] \triangleq \tilde{\boldsymbol{w}}^T[t-1]\boldsymbol{y}[t-1] \tag{11}$$

where

$$\begin{aligned} \tilde{\boldsymbol{w}}[t-1] &= \boldsymbol{R}_{yy}^{-1}[t-1]\boldsymbol{p}[t-1], \\ \boldsymbol{R}_{yy}[t-1] &\triangleq \left(\sum_{l=1}^{t}\boldsymbol{y}[l-1]\boldsymbol{y}[l-1]^T+\delta I\right) \\ \text{and } \boldsymbol{p}[t-1] &\triangleq \sum_{l=1}^{t-1}y[l]\boldsymbol{y}[l-1] \end{aligned} \tag{12}$$

I is a size $p \times p$ identity matrix, and $\delta \in \mathbb{R}^+$. Clearly, $\hat{x}_{y,2}[t]$ is sequential such that it only employs y_1^{t-1} to make its predictions on $x[t]$. In construction of $\hat{x}_{y,2}[t]$, we do not use A_x, A_z, A_y or σ_z^2. We observe that $\hat{x}_{y,2}[t]$ has a similar form to that of the well-known recursive least squares algorithm (RLS) [2], with δI as the initial value for the inverse correlation matrix and can be implemented with similar computational complexity. For this algorithm, we have the following result.

Theorem 2: Let $x[t]$ be a real valued sequence, $x[t] \in [-A_x, A_x]$, $A_x \in \mathbb{R}^+$, $y[t] = x[t]+z[t]$ be the observation sequence and $z[t] \in [-A_z, A_z]$, $A_z \in \mathbb{R}^+$ be an i.i.d. noise process with zero mean. For any $\delta > 0$, the sequential algorithm $\hat{x}_{y,2}[t]$ of (11), when applied to $y[t]$, satisfies, for all n,

$$\begin{aligned} &\frac{1}{n}E\left\{\sum_{t=1}^{n}(x[t]-\hat{x}_{y,2}[t])^2 \right. \\ &\qquad\left.-\left[\sum_{t=1}^{n}(x[t]-\boldsymbol{w}^T\boldsymbol{y}[t-1])^2+\delta|\boldsymbol{w}|_2^2\right]\right\} \\ &\quad\le pA_y^2\frac{\ln(n+1)}{n}+O\left(\frac{\delta}{n}\right) \end{aligned} \tag{13}$$

and for any small $\epsilon > 0$

$$\begin{aligned} &\Pr\left\{\sum_{t=1}^{n}(x[t]-\hat{x}_{y,2}[t])^2 \right. \\ &\qquad-\left[\sum_{t=1}^{n}(x[t]-\boldsymbol{w}^T\boldsymbol{y}[t-1])^2+\delta|\boldsymbol{w}|_2^2\right] \\ &\qquad\left.\le pA_y^2\ln(n+1)+O(\epsilon)+O(\delta)\right\} \\ &\quad\ge 1-2\exp(-n\epsilon^2B) \end{aligned} \tag{14}$$

for all x_1^n, all $\boldsymbol{w} \in \mathbb{R}^p$, where $\boldsymbol{y}[t-1] = [y[t-1],\ldots,y[t-p]]^T$ and σ_z^2 is the variance of $z[t]$. Here, $B = \left(1/4A^2\sigma_z^2\right)$, $A = 2\left(|\boldsymbol{w}|_1A_y+\left(p^2A_y^3/\lambda_\infty\right)\right)$ and $0 < \epsilon < \left(2A\sigma_z^2/A_z\right)$, where $\lambda_\infty = \min\{\lambda_p(\boldsymbol{R}_{yy}[l-1])\}_{l=1}^{n}$.

Theorem 2 states that the performance of $\hat{x}_{y,2}[t]$, when applied to $y[t]$, is asymptotically as good as the performance of any pth-order linear predictor including the best $\boldsymbol{w}$ that is tuned to the underlying signal in advance. For example, for any n, the optimal predictor that minimizes $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} E\left[\sum_{t=1}^{n}(x[t]-\boldsymbol{w}^T\boldsymbol{y}[t-1])^2\right]+\delta|\boldsymbol{w}|_2^2$ is given by

$$\boldsymbol{w}^* = \left[\sum_{l=1}^{n}(\boldsymbol{x}[l-1]\boldsymbol{x}[l-1]^T)+(\delta+\sigma_z^2)I\right]^{-1}\sum_{l=1}^{n}x[l]\boldsymbol{x}[l-1]. \tag{15}$$

This optimal linear predictor can only be calculated in hindsight by observing all x_1^n and also requires σ_z^2. The performance of this optimal linear predictor, $E\left[\sum_{t=1}^{n}(x[t]-\boldsymbol{w}^{*T}\boldsymbol{y}[t-1])^2\right]+\delta|\boldsymbol{w}|_2^2$, is asymptotically achieved by an algorithm that is sequential, with no knowledge of n, x_1^n or σ_z^2.

Sequential predictor:

Initialize $W_1(1) = 0$

Prediction of $x[t]$ for all $t > 1$:

$\hat{x}_{y,3}[t] = \sum_{s=1}^{t-1} \beta_t(s) \frac{t-1-s+1/2}{t-1-s+1} \left(\tilde{\boldsymbol{w}}_s^T[t-1]\boldsymbol{y}[t-1]\right)^+$,

where

$\beta_t(s) \triangleq \frac{W_{t-1}(s)}{\sum_{i=1}^{t-1} W_{t-1}(i)}$, $\tilde{\boldsymbol{w}}_s[t-1] \triangleq \left[\left(\sum_{l=s}^{t-1} \boldsymbol{y}[l-1]\boldsymbol{y}[r-1]^T\right) + \delta I\right]^{-1} \left[\sum_{l=s}^{t-1} y[l]\boldsymbol{y}[l-1]\right]$,

and for a real number $x \in \mathbb{R}$, $(x)^+ = x$ if $|x| \leq A_y$, $(x)^+ = A_y$ if $x > A_y$ and $(x)^+ = -A_y$ if $x < -A_y$.

After observing $y[t]$, recursive updates to construct prediction at $t+1$:

for $s = 1, \ldots, t-1$,

$W_t(s) = W_{t-1}(s) \frac{t-1-s+1/2}{t-1-s+1} \exp(-\frac{(y[t]-\tilde{\boldsymbol{w}}_s^T[t-1]\boldsymbol{y}[t-1])^2}{c})$

for $s = t$,

$W_t(t) = \sum_{j=1}^{t-1} W_{t-1}(j) \frac{1/2}{t-1-j+1} \exp(-\frac{y^2[t]}{c})$, and $c = 8A_y^2$.

Fig. 1. Description of the sequential algorithm of Theorem 3, i.e., $\hat{x}_{y,3}[t]$.

Proof of Theorem 2: Since $\hat{x}_{y,2}[t]$ and $\boldsymbol{w}^T\boldsymbol{y}[t-1]$, (for a fixed $\boldsymbol{w}$), are sequential and only depend on $\{y[1], \ldots, y[t-1]\}$, we can still use the identity (6), so that

$$E\left\{\sum_{t=1}^{n}(x[t]-\hat{x}_{y,2}[t])^2 - \sum_{t=1}^{n}(x[t]-\boldsymbol{w}^T\boldsymbol{y}[t-1])^2\right\} = E\left\{\sum_{t=1}^{n}(y[t]-\hat{x}_{y,2}[t])^2 - \sum_{t=1}^{n}(y[t]-\boldsymbol{w}^T\boldsymbol{y}[t-1])^2\right\}.$$

However, when applied to $y[t]$, we have the following result for $\hat{x}_{y,2}[t]$ from [4]:

$$\sum_{t=1}^{n}(y[t]-\hat{x}_{y,2}[t])^2 - \left[\sum_{t=1}^{n}(y[t]-\boldsymbol{w}^T\boldsymbol{y}[t-1])^2 + \delta|\boldsymbol{w}|_2^2\right] \leq pA_y^2\ln(n+1) + O(\delta). \quad (16)$$

uniformly for all y_1^n, n, $\delta \in \mathbb{R}^+$. Using (16) in (10) yields

$$E\left\{\sum_{t=1}^{n}(x[t]-\hat{x}_{y,2}[t])^2 - \left[\sum_{t=1}^{n}(x[t]-\boldsymbol{w}^T\boldsymbol{y}[t-1])^2 + \delta|\boldsymbol{w}|_2^2\right]\right\} \leq pA_y^2\ln(n+1) + O(\delta).$$

This completes the first part of Theorem 2. □

For the second part of the proof, since both $\hat{x}_{y,2}[t]$ and $\boldsymbol{w}^T\boldsymbol{y}[t-1] \triangleq \boldsymbol{w}^T\boldsymbol{y}[t-1]$, for fixed $\boldsymbol{w}$, are sequential, using (9)

$$\sum_{t=1}^{n}(x[t]-\hat{x}_{y,2}[t])^2 - \sum_{t=1}^{n}(x[t]-\boldsymbol{w}^T\boldsymbol{y}[t-1])^2 \quad (17)$$

$$= \sum_{t=1}^{n}(y[t]-\hat{x}_{y,2}[t])^2 - \sum_{t=1}^{n}(y[t]-\boldsymbol{w}^T\boldsymbol{y}[t-1])^2 + 2\sum_{t=1}^{n} z[t](\hat{x}_{y,2}[t]-\boldsymbol{w}^T\boldsymbol{y}[t-1]). \quad (18)$$

Since the second part of (18) is bounded by (16)

$$\sum_{t=1}^{n}(x[t]-\hat{x}_{y,2}[t])^2 - \sum_{t=1}^{n}(x[t]-\boldsymbol{w}^T\boldsymbol{y}[t-1])^2 \leq pA_y^2\ln(n+1) + 2\sum_{t=1}^{n} z[t](\hat{x}_{y,2}[t]-\boldsymbol{w}^T\boldsymbol{y}[t-1]) + O(\delta).$$

For $2|\hat{x}_{y,2}[t]-\boldsymbol{w}^T\boldsymbol{y}[t-1]|$, we observe that $|\boldsymbol{w}^T\boldsymbol{y}[t-1]| \leq |\boldsymbol{w}|_1 A_y$. For $\hat{x}_{y,2}[t]$, we use a bound from [9] such that $|\hat{x}_{y,2}[t]| \leq \left(p^2A_y^3/\lambda_p(\boldsymbol{R}_{yy}[t-1])\right)$, where $\lambda_p(\boldsymbol{R}_{yy}[t-1])$ is the smallest eigenvalue of $\boldsymbol{R}_{yy}[t-1]$. Hence, setting $\lambda_\infty = \min\{\lambda_d(\boldsymbol{R}_{yy}[t-1])\}_{t=1}^n$, yields $2|\hat{x}_{y,2}[t]-\boldsymbol{w}^T\boldsymbol{y}[t-1]| \leq 2\left(|\boldsymbol{w}|_1 A_y + \left(p^2A_y^3/\lambda_\infty\right)\right)$. Using Chernoff bound on $\sum_{t=1}^n z[t]$ yields the second part of Theorem 2. □

C. Switching Linear Prediction

Unlike the framework of Theorem 1, we now allow the pth-order predictors in the competition class to switch their parameters in time. We define the class of switching linear predictors as follows [5]. For any n, a partition of $\{1, \ldots, n\}$ into $r+1$ segment is represented by switching instants $\boldsymbol{t}_{r,n} = (t_1, \ldots, t_r)$, $1 < t_1 < t_2 < \cdots, t_r < n+1$, such that $1, \ldots, n$ can be represented as a concatenation of $\{1, \ldots, n\} = \{1, \ldots, t_1 - 1\}\{t_1, \ldots, t_2 - 1\} \ldots \{t_r, \ldots, n\}$. For notational simplicity, we take $t_0 = 1$ and $t_{r+1} = n+1$. Obviously, the number of switchings allowed is bounded by n, i.e., $r < n$. An algorithm in the class of switching linear predictors assigns a different linear predictor $\boldsymbol{w}_i \in \mathbb{R}^p$, to each region independently, $i = 1, \ldots, r+1$. The pair $\boldsymbol{t}_{r,n}$ and $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{r+1})$ forms a competing algorithm, for all $r = 1, \ldots, n-1$, $\boldsymbol{w}_i \in \mathbb{R}^p$, $i = 1, \ldots, r+1$ and all $t_1 < \cdots < t_r$. Clearly, for any n, one can choose from an exponential number of switching patterns and an infinite continuum of linear predictors for each segment. An algorithm in the competition class then produces predictions of $x[t]$ as $\hat{x}_{\boldsymbol{t}_{r,n}}[t] \triangleq \boldsymbol{w}_i^T\boldsymbol{y}[t-1]$ for $t_{i-1} \leq t < t_i$, $i = 1, \ldots, r+1$.

For this problem, we investigate $\hat{x}_{y,3}[t]$, which is a modified version of a sequential algorithm from [5] described in Fig. 1. Clearly, $\hat{x}_{y,3}[t]$ requires only y_1^{t-1} to produce its predictions. For the algorithm in Fig. 1, $\tilde{\boldsymbol{w}}_s[t-1]$ is the linear model from (12), trained on data samples $y[s], \ldots, y[t-1]$, where $s = 1, \ldots, t-1$. We observe that $\hat{x}_{y,3}[t]$ is in a certain sense a combined version of $\hat{x}_{y,1}[t]$ and $\hat{x}_{y,2}[t]$. At each time t, to produce its prediction, $\hat{x}_{y,3}[t]$ combines predictions of $t-1$ algorithms, i.e., $\left(\tilde{\boldsymbol{w}}_s^T[t-1]\boldsymbol{y}[t-1]\right)^+$, $s = 1, \ldots, t-1$, each weighted by $\beta_s(t)$, $s = 1, \ldots, t-1$. Each $\beta_s(t)$ measures the relative performance of $\left(\tilde{\boldsymbol{w}}_s^T[t-1]\boldsymbol{y}[t-1]\right)^+$, similar to (3). For this algorithm, we have the following result.

Theorem 3: Let $x[t]$ be a real valued sequence, $x[t] \in [-A_x, A_x]$, $A_x \in \mathbb{R}^+$, $y[t] = x[t] + z[t]$ be the observation sequence and $z[t] \in [-A_z, A_z]$, $A_z \in \mathbb{R}^+$ be an i.i.d. noise process with zero mean. For all n, $\hat{x}_{y,3}[t]$ in Fig. 1 satisfies

$$E\left\{\sum_{t=1}^{n}(x[t]-\hat{x}_{y,3}[t])^2 - \left[\sum_{i=1}^{r+1}\left(\sum_{t=t_{i-1}}^{t_i-1}(x[t]-\boldsymbol{w}_i^T\boldsymbol{y}[t-1])^2 + \delta|\boldsymbol{w}_i|_2^2\right)\right]\right\} \quad (19)$$

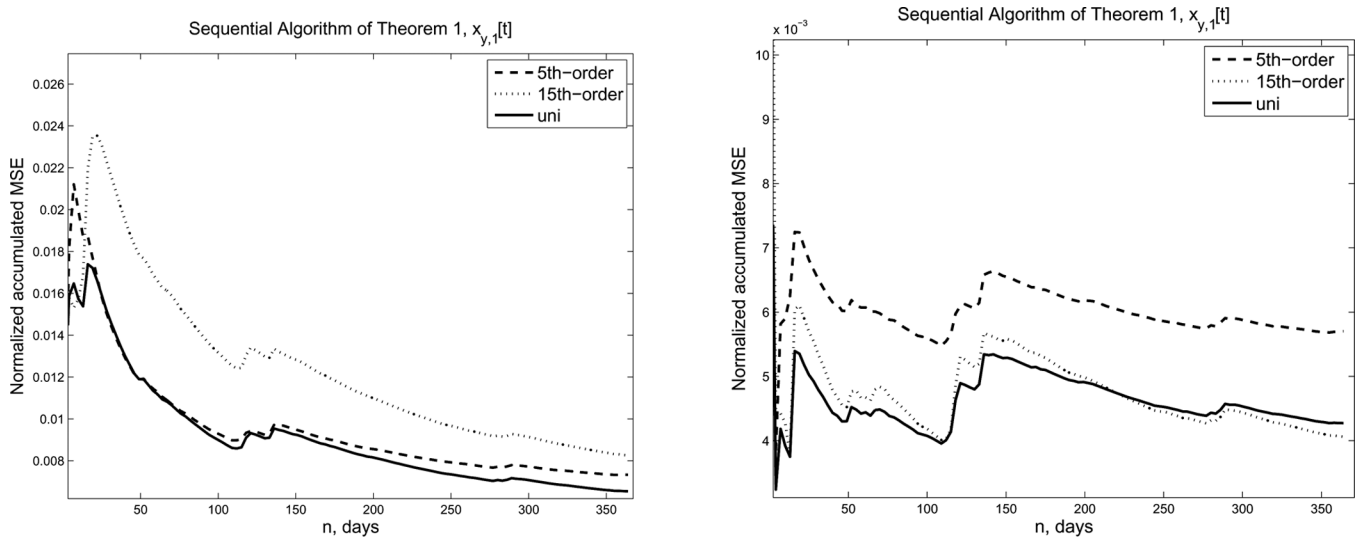

Fig. 2. Prediction results for the closing price of the Iroquois stock. NE-MSE for fifth-order linear model, fifteenth-order linear model and $\hat{x}_{y,1}[t]$ "uni". (a) All algorithms observe only $y[t]$. (b) Linear models observe the clean signal $x[t]$ and $\hat{x}_{y,1}[t]$ observes only $y[t]$.

$$\leq p(r+1)A_y^2\ln(n+1)+4A_y^2(3r+1)\ln(n)+O(\delta) \quad (20)$$

and for any small $\epsilon > 0$

$$\begin{aligned}\Pr\Bigg\{&\sum_{t=1}^{n}(x[t]-\hat{x}_{y,3}[t])^2 \\ &-\left[\sum_{i=1}^{r+1}\left(\sum_{t=t_i-1}^{t_i-1}(x[t]-\boldsymbol{w}_i^T\boldsymbol{y}[t-1])^2+\delta|\boldsymbol{w}_i|_2^2\right)\right] \\ &\leq p(r+1)A_y^2\frac{\ln(n+1)}{n}+4A_y^2(3r+1)\frac{\ln(n)}{n}+O\left(\frac{\delta}{n}\right)\Bigg\} \\ &\geq 1-2\exp(-n\epsilon^2 B) \quad (21)\end{aligned}$$

for any n, $\boldsymbol{w}_i \in \mathbb{R}_p^+$, $i=1,\ldots,r+1$, $r=1,\ldots,n-1$ and any $t_1 < \cdots < t_r$. Here, $B=\left(1/4A^2\sigma_z^2\right)$, $A=2(\max_i|\boldsymbol{w}|_i A_y+A_y)$ and $0<\epsilon<\left(2A\sigma_z^2/A_z\right)$, where σ_z^2 is the variance of $z[t]$.

Proof of Theorem 3: For any n and r, the partitioning of $1,\ldots,n$ into $r+1$ segments, i.e., $(t_1,\ldots,t_r)$ and assigning each segment a constant vector $\boldsymbol{w}_i$, $i=1,\ldots,r+1$ defines a predictor in the competition class. Here, the competition class is all such predictors for all $r=1,\ldots,n-1$ and $\boldsymbol{w}_i \in \mathbb{R}_p^+, i=1,\ldots,r+1$. Although, $\hat{x}_{y,3}[t]$ is strongly sequential, i.e., it does not depend on n, r or switching times, an algorithm in the competition class, $\hat{x}_{\boldsymbol{t}_{r,n}}[t]$, has access to n, r and $(t_1,\ldots,t_r)$ for all n. However, for any algorithm $\hat{x}_{\boldsymbol{t}_{r,n}}[t]$ in this competition class, we can still write

$$\begin{aligned}E\left[(x[t]-\hat{x}_{y,3}[t])^2-(x[t]-\hat{x}_{\boldsymbol{t}_{r,n}}[t])^2\right] \\ =E\left[(y[t]-\hat{x}_{y,3}[t])^2-(y[t]-\hat{x}_{\boldsymbol{t}_{r,n}}[t])^2\right]\end{aligned}$$

since, still, $E[z[t]\hat{x}_{\boldsymbol{t}_{r,n}}[t]]=0$ for all t, i.e., $z[t]$ has no correlation with $\hat{x}_{\boldsymbol{t}_{r,n}}[t]$. Hence

$$\begin{aligned}E\left[\sum_{t=1}^{n}\left\{(x[t]-\hat{x}_{y,3}[t])^2-(x[t]-\hat{x}_{\boldsymbol{t}_{r,n}}[t])^2\right\}\right] \\ =E\left[\sum_{t=1}^{n}\left\{(y[t]-\hat{x}_{y,3}[t])^2-(y[t]-\hat{x}_{\boldsymbol{t}_{r,n}}[t])^2\right\}\right]. \quad (22)\end{aligned}$$

Since clipping predictions $\boldsymbol{w}_s^T[t-1]\boldsymbol{y}[t-1]$ in each branch only improves prediction, we have the following result for $\hat{x}_{y,3}[t]$ from [5]:

$$\begin{aligned}\sum_{t=1}^{n}(y[t]-\hat{x}_{y,3}[t])^2-\sum_{k=1}^{r+1}\sum_{t_{k-1}}^{t_k-1}(x[t]-\boldsymbol{w}_k^T\boldsymbol{y}[t-1])^2+\delta|\boldsymbol{w}_i|_2^2 \\ \leq p(r+1)A_y^2\ln(n)+4A_y^2(3r+1)\ln(n)+O(\delta).\end{aligned}$$

Hence, applying the above equation in (22) gives the first part of Theorem 3. □

For the second part of Theorem 3, similar to (9), we have

$$\begin{aligned}&\sum_{t=1}^{n}(x[t]-\hat{x}_{y,3}[t])^2-\sum_{t=1}^{n}(x[t]-\hat{x}_{\boldsymbol{t}_{r,n}}[t])^2 \\ &\quad=\sum_{t=1}^{n}(y[t]-\hat{x}_{y,3}[t])^2-\sum_{t=1}^{n}(y[t]-\hat{x}_{\boldsymbol{t}_{r,n}}[t])^2 \\ &\qquad+2\sum_{t=1}^{z}z[t](\hat{x}_{y,3}[t]-\hat{x}_{\boldsymbol{t}_{r,n}}[t]).\end{aligned}$$

Hence, to get the result in Theorem 3, we need to bound $(\hat{x}_{y,3}[t]-\hat{x}_{\boldsymbol{t}_{r,n}}[t])$. Since, $\hat{x}_{\boldsymbol{t}_{r,n}}[t]$ is equal to $\boldsymbol{w}_k^T\boldsymbol{y}[t-1]$ for one $\boldsymbol{w}_i$, $i=1,\ldots,r+1$, then $|\boldsymbol{w}_i^T\boldsymbol{y}[t-1]| \leq \max_r|\boldsymbol{w}_r|_1 A_y$. Moreover, $|\hat{x}_{y,3}[t]| \leq A_y$ due to clipping to $[-A_y, A_y]$. Hence, $2|(\hat{x}_{y,3}[t]-\hat{x}_{\boldsymbol{t}_{r,n}}[t])| \leq 2(1+\max_r|\boldsymbol{w}_r|_1)A_y$. This completes the proof of Theorem 3. □

III. Simulations

In this section, we demonstrate the performance of each of the algorithms developed, in several different scenarios. As the first example, we apply our algorithms to historical data from the New York Stock Exchange. We predict the closing market price of the Iroquois stock, which is chosen because of its volatility. However, at each day, we only observe a noise-corrupted version of the desired signal, $x[t]$, i.e., $y[t]=x[t]+z[t]$, where $z[t]$ is i.i.d. and distributed uniformly between $[-0.25, 0.25]$. This added i.i.d. noise models the underlying intrinsic price fluctuations that are independent from the past observations. As the competing prediction algorithms, we use fifth-order (one week) and fifteenth-order (three weeks) linear models, where each model is

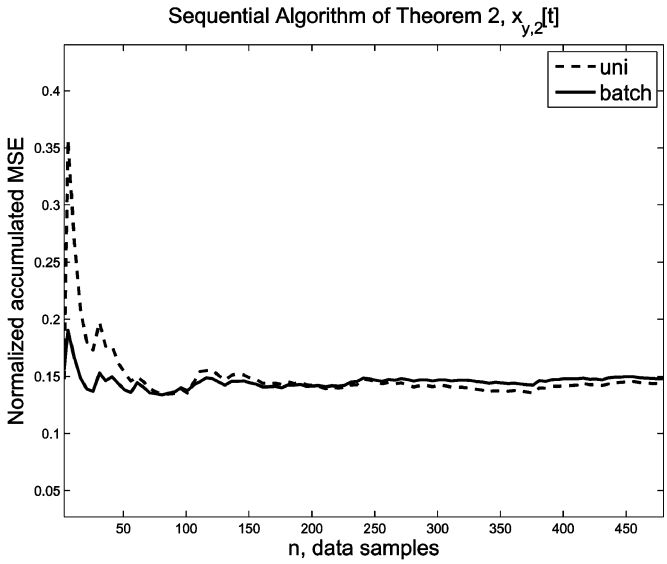

Fig. 3. Prediction result for a third-order AR process. The normalized MSE of $\hat{x}_{y,2}[t]$ and the batch predictor of (15).

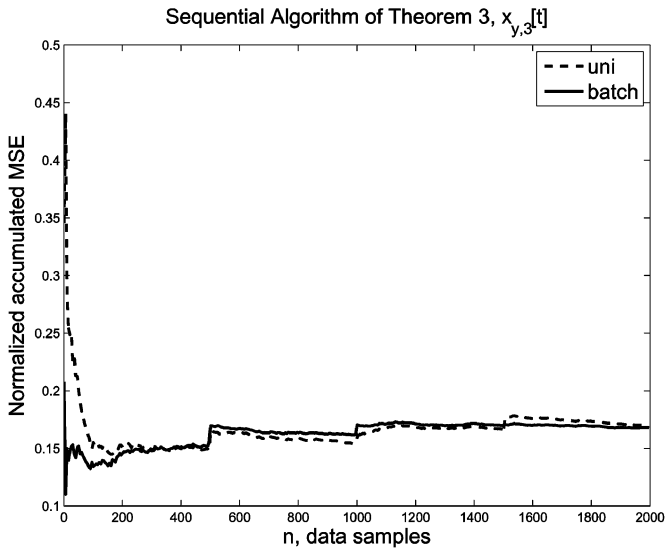

Fig. 4. Prediction result for a second-order AR process that changes its parameters every 500 samples. The normalized MSE of $\hat{x}_{y,2}[t]$ and the batch predictors from (15) that are tuned for each segment independently.

trained using the RLS algorithm with an effective window size of 30 days. These predictors are denoted as $\hat{x}_1[t]$ and $\hat{x}_2[t]$ respectively. Initially, these linear predictors solely work on the noisy stock prices $y[t]$. The output of these predictors are then combined to form $\hat{x}_{y,1}[t]$, using (3) to predict $x[t]$. Although all algorithms, $\hat{x}_1[t]$, $\hat{x}_1[t]$ and $\hat{x}_{y,1}[t]$, only observe $y[t]$, their performances are judge with respect to the clean signal $x[t]$. In Fig. 2(a), we plot the normalized accumulated MSE (NA-MSE) of these predictors, for 500 independent realization of the noise $z[t]$. We observe that $\hat{x}_{y,1}[t]$ follows $\hat{x}_1[t]$ in the start and favors $\hat{x}_2[t]$ later on, hence performs as good as the best algorithm that can only be chosen in hindsight. In Fig. 2(b), we simulate the same algorithms, however, $\hat{x}_1[t]$ and $\hat{x}_2[t]$ now observe and train on the clean signal $x[t]$. Here, $\hat{x}_{y,1}[t]$ still receives predictions from $\hat{x}_1[t]$ and $\hat{x}_2[t]$, however, trains on $y[t]$ as in (2). The losses of $\hat{x}_1[t]$, $\hat{x}_2[t]$ and $\hat{x}_{y,1}[t]$ are still with respect to $x[t]$. Even in this case, $\hat{x}_{y,1}[t]$ is able to perform a successful mixture based on judging the linear algorithms with respect to $y[t]$.

As the next set of experiments, we apply a third-order predictor $\hat{x}_{y,2}[t]$ from (11) to predict a sample function from the third-order autoregressive (AR) process, $x[t] = 0.9x[t-1] - 0.6x[t-2] + 0.5x[t-3] + a[t]$, where $a[t]$ is a Gaussian i.i.d. process with variance 0.1. We observe a noise-corrupted version of the desired signal $x[t]$, i.e., $y[t] = x[t] + z[t]$, where $z[t]$ is i.i.d. and distributed uniformly between $[-0.3, 0.3]$. In Fig. 3, we plot NA-MSE for $\hat{x}_{y,2}[t]$ for a single sample function of this third-order process and the batch predictor from (15) with a total of 100 sample functions of the noise process $z[t]$. Although $\hat{x}_{y,2}[t]$ relies only on the noisy observations, it is able to achieve the performance of the best batch predictor for increasing data lengths. Finally, we apply $\hat{x}_{y,3}[t]$ to a process that switches between different second-order AR processes for every 500 samples. Here, the process switches between $x[t] = -1.4x[t-1] + 0.74x[t-2] + a[t]$ and $x[t] = 1.4x[t-1] - 0.74x[t-2] + a[t]$, where $a[t]$ is i.i.d. Gaussian zero mean noise, with variance 0.1 and $z[t]$ is i.i.d. uniformly distributed between $[-0.3, 0.3]$. For $\hat{x}_{y,3}[t]$, third-order models $\boldsymbol{w}_s[t-1]$ are used in Fig. 1. In Fig. 4, we plot the NA-MSE of $\hat{x}_{y,3}[t]$ and that of the batch predictor. Here, the batch predictor knows *a priori* the switching pattern and uses (15) to select the best batch predictor independently in each segment by observing $x[t]$. However, $\hat{x}_{y,3}[t]$ observes only the noisy version $y[t]$ and has no knowledge of the switching pattern, the number of switchings or the length of the data. For this simulation, $\hat{x}_{y,3}[t]$ asymptotically achieves the performance of the batch algorithm and the difference between the two algorithms cannot be larger than the regret in Theorem 3.

IV. Conclusion

In this correspondence, we investigated sequential prediction of real-valued and bounded individual sequences that are corrupted by additive noise. Here, we introduced algorithms that are able to asymptotically achieve the performance of the best algorithm from a large class of competing algorithms that can only be chosen by observing the clean signal in hindsight. Our results are guaranteed to hold for any arbitrary, deterministic and bounded signal without any stochastic assumptions on the desired signal. We only assume that the noise is a zero mean, i.i.d. and bounded process.

References

[1] N. Bianchi and G. Lugosi, *Prediction, Learning and Games*. Cambridge, U.K.: Cambridge Univ. Press, 2006.
[2] S. Haykin, *Adaptive Filter Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1996.
[3] A. C. Singer and M. Feder, "Universal linear prediction by model order weighting," *IEEE Transactions on Signal Processing*, vol. 47, 1999.
[4] A. C. Singer, S. 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, Aug. 2002.
[5] S. S. Kozat and A. C. Singer, "Universal switching linear least squares prediction," *IEEE Trans. Signal Process.*, vol. 56, no. 1, pp. 189–204, Jan. 2008.
[6] S. S. Kozat, A. C. Singer, and G. Zeitler, "Universal piecewise linear prediction via context trees," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3730–3745, Jul. 2007.
[7] 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.
[8] T. Moon and T. Weissman, "Competitive online linear fir mmse filtering," in *Proc. ISIT*, 2007, pp. 1126–1130.
[9] G. C. Zeitler and A. C. Singer, "Universal linear least-squares prediction in the presence of noise," in *Proc. IEEE Workshop SSP*, 2007, pp. 611–614.
[10] A. N. Tikhonov, "On the stability of inverse problems," *Dokl. Akad. Nauk SSSR*, vol. 39, no. 5, pp. 195–198, 1943.