Brief Papers

A Unified Approach to Universal Prediction: Generalized Upper and Lower Bounds

Nuri Denizcan Vanli and Suleyman S. Kozat,
Senior Member, IEEE

Abstract—We study sequential prediction of real-valued, arbitrary, and unknown sequences under the squared error loss as well as the best parametric predictor out of a large, continuous class of predictors. Inspired by recent results from computational learning theory, we refrain from any statistical assumptions and define the performance with respect to the class of general parametric predictors. In particular, we present generic lower and upper bounds on this relative performance by transforming the prediction task into a parameter learning problem. We first introduce the lower bounds on this relative performance in the mixture of experts framework, where we show that for any sequential algorithm, there always exists a sequence for which the performance of the sequential algorithm is lower bounded by zero. We then introduce a sequential learning algorithm to predict such arbitrary and unknown sequences, and calculate upper bounds on its total squared prediction error for every bounded sequence. We further show that in some scenarios, we achieve matching lower and upper bounds, demonstrating that our algorithms are optimal in a strong minimax sense such that their performances cannot be improved further. As an interesting result, we also prove that for the worst case scenario, the performance of randomized output algorithms can be achieved by sequential algorithms so that randomized output algorithms do not improve the performance.

Index Terms—Online learning, sequential prediction, worst-case performance.

I. INTRODUCTION

In this brief, we investigate the generic sequential (online) prediction problem from an individual sequence perspective using tools of computational learning theory, where we refrain from any statistical assumptions either in modeling or on signals [1]–[4]. In this approach, we have an arbitrary, deterministic, bounded, and unknown signal $\{x[t]\}_{t \geq 1}$, where $|x[t]| < A < \infty$, and $x[t] \in \mathbb{R}$. Since we do not impose any statistical assumptions on the underlying data, we, motivated by recent results from sequential learning [1]–[4], define the performance of a sequential algorithm with respect to a comparison class, where the predictors of the comparison class are formed by observing the entire sequence in hindsight,

under the squared error loss, that is

$$\sum_{t=1}^{n} (x[t] - \hat{x}_s[t])^2 - \inf_{c \in \mathcal{C}} \sum_{t=1}^{n} \left(x[t] - \hat{x}_c[t]\right)^2$$

for an arbitrary length of data n, and for any possible sequence $\{x[t]\}_{t \geq 1}$, where $\hat{x}_s[t]$ is the prediction at time t of any sequential algorithm that has access data from $x[1]$ up to $x[t-1]$ for prediction, and $\hat{x}_c[t]$ is the prediction at time t of the predictor c such that $c \in \mathcal{C}$, where $\mathcal{C}$ represents the class of predictors we compete against. We emphasize that since the predictors $\hat{x}_c[t]$, $c \in \mathcal{C}$ have the access to the entire sequence before the processing starts, the minimum squared prediction error that can be achieved with a sequential predictor $\hat{x}_s[t]$ is equal to the squared prediction error of the optimal batch predictor $\hat{x}_c[t]$, $c \in \mathcal{C}$. Here, we call the difference in the squared prediction error of the sequential algorithm $\hat{x}_s[t]$ and the optimal batch predictor $\hat{x}_c[t]$, $c \in \mathcal{C}$ as the regret of not using the optimal predictor (or equivalently, not knowing the future). Therefore, we seek for sequential algorithms $\hat{x}_s[t]$ that minimize this regret or loss for any possible $\{x[t]\}_{t \geq 1}$. We emphasize that this regret definition is for the accumulated sequential cost, instead of the batch cost.

Instead of fixing a comparison class of predictors, we parameterize the comparison classes such that the parameter set and functional form of these classes can be chosen as desired. In this sense, in this brief, we consider the most general class of parametric predictors as our class of predictors $\mathcal{C}$ such that the regret for an arbitrary length of data n is given by

$$\sum_{t=1}^{n} (x[t] - \hat{x}_s[t])^2 - \inf_{\boldsymbol{w} \in \mathbb{R}^m} \sum_{t=1}^{n} \left(x[t] - f\left(\boldsymbol{w}, x_{t-a}^{t-1}\right)\right)^2 \quad (1)$$

where $f(\boldsymbol{w}, x_{t-a}^{t-1})$ is a parametric function whose parameters $\boldsymbol{w} = [w_1, \ldots, w_m]^T$ can be set prior to prediction, and this function uses the data x_{t-a}^{t-1}, $t - a \geq 1$ for prediction for some arbitrary integer a, which can be viewed as the tap size of the predictor.[1] Although the parameters of the parametric prediction function $f(\boldsymbol{w}, x_{t-a}^{t-1})$ can be set arbitrarily, even by observing all the data $\{x[t]\}_{t \geq 1}$ *a priori*, the function is naturally restricted to use only the sequential data x_1^{t-1} in prediction [5]–[7].

Since we have no statistical assumptions on the underlying data, the corresponding lower and upper bounds on the regret in (1) in this sense provide the ultimate measure of the learning performance for any sequential predictor. We emphasize that lower bounds not only provide the worst-case performance of

---

[1]All vectors are column vectors and denoted by boldface lower case letters. For a vector $\boldsymbol{u}$, $\boldsymbol{u}^T$ is the ordinary transpose. We denote $x_a^b \triangleq \{x[t]\}_{t=a}^{b}$.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

2   IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

an algorithm, but also quantify the prediction power of the parametric class. As such, a positive lower bound guarantees the existence of a data sequence having an arbitrary length such that no matter how smart the learning algorithm is, the performance of this smart algorithm on this sequence will be worse than the class of parametric predictors by at least an order of the lower bound. Hence, if an algorithm is found such that the upper bound of the regret of that algorithm matches with the lower bound, then that algorithm is optimal in a strong minimax sense such that the actual convergence performance cannot be further improved [7]. To this end, the minimax sense optimality of different parametric learning algorithms, such as the well-known prediction algorithms, least mean squares (LMSs) [8], recursive least squares (RLSs) [8], and online sequential extreme learning machine of [1] can be determined using the lower bounds provided in this brief. In this sense, the rates of the corresponding upper and lower bounds are analogous to the VC dimension [9] of classifiers and can be used to quantify the learning performance [1]–[3], [10].

Various sequential learning algorithms have been proposed in [1], [7], [8], [10]–[12], and [13] in order to efficiently learn the relationship between the observations and the desired data. One of the simplest methods is to linearly model this relationship, i.e., $f(\boldsymbol{w}[t], x_{t-a}^{t-1}) = \boldsymbol{w}[t]^T x_{t-a}^{t-1}$, and then update $\boldsymbol{w}[t]$ using the well-known algorithms, such as the LMS or RLS algorithms [1], [8]. In more recent studies [7], [12], universal algorithms have been proposed that achieve the performance of the optimal weighting vector without any statistical assumptions. Kivinen and Warmuth [10] have proposed a multiplicative update of the weights and provided guaranteed upper bounds on the performance of the proposed algorithm. On the other hand, in order to introduce a nonlinear modeling, similar learning methods are usually extended by either mapping the observations to higher dimensions as in polynomial and Volterra filters [11] or partitioning the observation space and fitting linear models in each partition, i.e., piecewise linear modeling [13].

In order to derive upper and lower bounds on the performance of such learning algorithms, the mixture of experts framework is usually used. As an example, linear prediction [5], [7], [12], nonlinear models based on piecewise linear approximations [13], and the learning of an individual noise-corrupted deterministic sequence [14] are studied. These results are then extended to the filtering problems [15], [16]. In this brief, on the other hand, we consider a holistic approach and provide upper and lower bounds for the general framework, which was previously missing in the literature.

Our main contribution in this brief is to obtain the generalized lower bounds for a variety of prediction frameworks by transforming the prediction problem to a well known and studied statistical parameter learning problem [1], [4]–[7]. By doing so, we prove that for any sequential algorithm there always exists some data sequence over any length such that the regret of the sequential algorithm is lower bounded by zero. We further derive lower bounds for important classes of predictors heavily investigated in machine learning literature,

including univariate polynomial, multivariate polynomial, and linear predictors [4]–[7], [10]–[12], [14]. We also provide a universal sequential prediction algorithm and calculate upper bounds on the regret of this algorithm, and show that we obtain matching lower and upper bounds in some scenarios. As an interesting result, we also show that given the regret in (1) as the performance measure, there is no additional gain achieved by using randomized algorithms in the worst-case scenario.

The rest of this brief is organized as follows. In Section II, we first present general lower bounds, and then analyze couple of specific scenarios. We then introduce a universal prediction algorithm and calculate the upper bounds on its regret in Section III. In Section IV, we show that in the worst-case scenario, the performance of randomized algorithms can be achieved by sequential algorithms. Finally, conclusions are drawn in Section V.

II. Lower Bounds

In this section, we investigate the worst-case performance of sequential algorithms to obtain guaranteed lower bounds on the regret. Hence, for any arbitrary length of data n, $\{x[t]\}_{t \geq 1}$, we are trying to find a lower bound on the following:

$$\sup_{x_1^n} \left\{ \sum_{t=1}^n (x[t] - \hat{x}_s[t])^2 - \inf_{\boldsymbol{w} \in \mathbb{R}^m} \sum_{t=1}^n \left(x[t] - f\left(\boldsymbol{w}, x_{t-a}^{t-1}\right) \right)^2 \right\}. \tag{2}$$

For this regret, we have the following theorem that relates the performance of any sequential algorithm to the general class of parametric predictors. While proving this theorem, we also provide a generic procedure to find lower bounds on the regret in (2) and later use this method to derive lower bounds for parametric classes, including the classes of univariate polynomial, multivariate polynomial, and linear predictors [4]–[7], [10]–[12], [14].

Theorem 1: There is no best sequential algorithm for all sequences for any class in the parametric form $f(\boldsymbol{w}, x_{t-a}^{t-1})$, where $\boldsymbol{w} \in \mathbb{R}^m$. Given a parametric class there exists always a sequence such that the regret in (2) is always lower bounded by some nonnegative value.

This theorem implies that no matter how smart a sequential algorithm is or how naive the competition class is, it is not possible to outperform the competition class for all sequences. As an example, this result demonstrates that even competing against the class of constant predictors, i.e., the most naive competition class, where $\hat{x}_c[t]$ always predicts a constant value, any sequential algorithm, no matter how smart, cannot outperform this class of constant predictors for all sequences. We emphasize that in this sense, the lower bounds provide the prediction and modeling power of the parametric class.

Proof of Theorem 1: We begin our proof by pointing out that finding the best sequential predictor for an arbitrary and unknown sequence of x_1^n is not straightforward. Yet, for a specific distribution on x_1^n, the best predictor is the conditional mean on x_1^n under the squared error [17]. Therefore, by this clever transformation, we are able to calculate the regret in (2) in the expectation sense and prove this theorem.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS
3

Since the supremum in (2) is taken over all x_1^n, for any distribution x_1^n, the regret is lower bounded by

$$\sup_{x_1^n}\left(\sum_{t=1}^{n}(x[t]-\hat{x}_s[t])^2 - \inf_{\boldsymbol{w}\in\mathbb{R}^m}\sum_{t=1}^{n}\left(x[t]-f\left(\boldsymbol{w},x_{t-a}^{t-1}\right)\right)^2\right)$$
$$\geq \underbrace{E_{x_1^n}\left[\sum_{t=1}^{n}(x[t]-\hat{x}_s[t])^2 - \inf_{\boldsymbol{w}\in\mathbb{R}^m}\sum_{t=1}^{n}\left(x[t]-f\left(\boldsymbol{w},x_{t-a}^{t-1}\right)\right)^2\right]}_{\triangleq L(n)}$$

where expectation is taken with respect to this particular distribution. Hence, it is enough to lower bound $L(n)$ to get a final lower bound. By the linearity of the expectation

$$L(n) = E_{x_1^n}\left[\sum_{t=1}^{n}(x[t]-\hat{x}_s[t])^2\right]$$
$$-E_{x_1^n}\left[\inf_{\boldsymbol{w}\in\mathbb{R}^m}\sum_{t=1}^{n}\left(x[t]-f\left(\boldsymbol{w},x_{t-a}^{t-1}\right)\right)^2\right]. \quad (3)$$

The squared-error loss $E[(x[t]-\hat{x}_s[t])^2]$ is minimized with the well-known minimum mean squared error (MMSE) predictor given by [17]

$$\hat{x}_s[t] = E\left[x[t]\big|x[t-1],\ldots,x[1]\right] = E\left[x[t]\big|x_1^{t-1}\right] \quad (4)$$

where we drop the explicit x_1^n-dependence of the expectation to simplify the presentation.

Suppose we select a parametric distribution for x_1^n with parameter vector $\boldsymbol{\theta} = [\theta_1,\ldots,\theta_m]$. Then, for the second term in (3), we use the following inequality:

$$E_{\boldsymbol{\theta}}\left[E_{x_1^n|\boldsymbol{\theta}}\left[\inf_{\boldsymbol{w}\in\mathbb{R}^m}\sum_{t=1}^{n}\left(x[t]-f\left(\boldsymbol{w},x_{t-a}^{t-1}\right)\right)^2\right]\right]$$
$$\leq E_{\boldsymbol{\theta}}\left[\inf_{\boldsymbol{w}\in\mathbb{R}^m}E_{x_1^n|\boldsymbol{\theta}}\left[\sum_{t=1}^{n}\left(x[t]-f\left(\boldsymbol{w},x_{t-a}^{t-1}\right)\right)^2\right]\right]. \quad (5)$$

By using (4) and (5), and expanding the expectation, we can lower bound $L(n)$ as

$$L(n) \geq E_{\boldsymbol{\theta}}\left[E_{x_1^n|\boldsymbol{\theta}}\left[\sum_{t=1}^{n}\left(x[t]-E\left[x[t]\big|x_1^{t-1}\right]\right)^2\right]\right]$$
$$-E_{\boldsymbol{\theta}}\left[\inf_{\boldsymbol{w}\in\mathbb{R}^m}E_{x_1^n|\boldsymbol{\theta}}\left[\sum_{t=1}^{n}\left(x[t]-f\left(\boldsymbol{w},x_{t-a}^{t-1}\right)\right)^2\right]\right]. \quad (6)$$

The inequality in (6) is true for any distribution on x_1^n. Hence, for a distribution on x_1^n such that

$$E\left[x[t]\big|x_1^{t-1},\boldsymbol{\theta}\right] = h\left(\boldsymbol{\theta},x_{t-a}^{t-1}\right) \quad (7)$$

with some function h, if we can find a vector function $\boldsymbol{g}(\boldsymbol{\theta})$ satisfying $f(\boldsymbol{g}(\boldsymbol{\theta}),x_{t-a}^{t-1}) = h(\boldsymbol{\theta},x_{t-a}^{t-1})$ then the last term in (6) yields

$$E_{\boldsymbol{\theta}}\left[\inf_{\boldsymbol{w}\in\mathbb{R}^m}E_{x_1^n|\boldsymbol{\theta}}\left[\sum_{t=1}^{n}\left(x[t]-f\left(\boldsymbol{w},x_{t-a}^{t-1}\right)\right)^2\right]\right]$$
$$= E_{\boldsymbol{\theta}}\left[E_{x_1^n|\boldsymbol{\theta}}\left[\sum_{t=1}^{n}\left(x[t]-h\left(\boldsymbol{\theta},x_{t-a}^{t-1}\right)\right)^2\right]\right].$$

Thus, (6) can be written as

$$L(n) \geq E_{\boldsymbol{\theta}}\left[E_{x_1^n|\boldsymbol{\theta}}\left[\sum_{t=1}^{n}\left(x[t]-E\left[x[t]\big|x_1^{t-1}\right]\right)^2\right]\right]$$
$$-E_{\boldsymbol{\theta}}\left[E_{x_1^n|\boldsymbol{\theta}}\left[\sum_{t=1}^{n}\left(x[t]-E\left[x[t]\big|x_1^{t-1},\boldsymbol{\theta}\right]\right)^2\right]\right]$$

which is by definition of the MMSE estimator is always lower bounded by zero, i.e., $L(n) \geq 0$. By this inequality, we conclude that for predictors of the form $f(\boldsymbol{w},x_{t-a}^{t-1})$ for which this special parametric distribution, i.e., $\boldsymbol{w} = \boldsymbol{g}(\boldsymbol{\theta})$ exists, the best sequential predictor will be always outperformed by some predictor in this class for some sequence x_1^n. Hence, there is no best algorithm for all sequences for any class in this parametric form. The question arises if a suitable distribution on x_1^n can be found for a given $f(\boldsymbol{w},x_{t-a}^{t-1})$ such that $f(\boldsymbol{g}(\boldsymbol{\theta}),x_{t-a}^{t-1}) = h(\boldsymbol{\theta},x_{t-a}^{t-1})$ with a suitable transformation $\boldsymbol{g}(\boldsymbol{\theta})$.

Suppose $f(\boldsymbol{w},x_{t-a}^{t-1})$ is bounded by some $0 < M < \infty$ for all $|x[t]| \leq A$, i.e., $|f(\boldsymbol{w},x_{t-a}^{t-1})| \leq M$. Then, given θ from a beta distribution with parameters (C,C), $C \in R^+$, we generate a sequence x_1^n such that $x[t] = A/M f(\boldsymbol{w},x_{t-a}^{t-1})$ with probability θ and $x[t] = -A/M(f(\boldsymbol{w},x_{t-a}^{t-1}))$ with probability $(1-\theta)$. Then

$$E\left[x[t]\big|x_1^{t-1},\theta\right] = \frac{A}{M}(2\theta-1)f\left(\boldsymbol{w},x_{t-a}^{t-1}\right).$$

Hence, this concludes the proof of the Theorem 1. $\square$

As an important special case, if we use the restricted functional form $f(\boldsymbol{w},x_{t-a}^{t-1})$ so that $f(\boldsymbol{w},x_{t-a}^{t-1})$ is separable, then the prediction problem is transformed to a parameter estimation problem. The separable form is given by

$$f\left(\boldsymbol{w},x_{t-a}^{t-1}\right) = \boldsymbol{f}_w(\boldsymbol{w})^T \boldsymbol{f}_x(x_{t-a}^{t-1})$$

where $\boldsymbol{f}_w(\boldsymbol{w})$ and $\boldsymbol{f}_x(x_{t-a}^{t-1})$ are vector functions of size $m \times 1$ for some integer m. Then, (7) can be written as

$$E\left[x[t]\big|x_1^{t-1},\boldsymbol{\theta}\right] = \boldsymbol{f}_w(\boldsymbol{g}(\boldsymbol{\theta}))^T \boldsymbol{f}_x\left(x_{t-a}^{t-1}\right)$$

where $\boldsymbol{f}_w(\boldsymbol{g}(\boldsymbol{\theta})) = A/M(2\theta-1)\boldsymbol{f}_w(\boldsymbol{w})$. Denoting $\boldsymbol{f}_n(\boldsymbol{w}) \triangleq A/M\boldsymbol{f}_w(\boldsymbol{w})$ as the normalized prediction function, and after some algebra (6) is obtained as

$$L(n) \geq E_{\boldsymbol{\theta}}\left[E_{x_1^n|\boldsymbol{\theta}}\left[\sum_{t=1}^{n}\left(x[t]-E\left[(2\theta-1)\big|x_1^{t-1}\right]^T\right.\right.\right.$$
$$\left.\left.\left.\times \boldsymbol{f}_n(\boldsymbol{w})^T\boldsymbol{f}_x\left(x_{t-a}^{t-1}\right)\right)^2\right]\right]$$
$$-E_{\boldsymbol{\theta}}\left[E_{x_1^n|\boldsymbol{\theta}}\left[\sum_{t=1}^{n}\left(x[t]-(2\theta-1)\right.\right.\right.$$
$$\left.\left.\left.\times \boldsymbol{f}_n(\boldsymbol{w})^T\boldsymbol{f}_x\left(x_{t-a}^{t-1}\right)\right)^2\right]\right]$$

so that the regret of the sequential algorithm over the best prediction function is due to the regret attained by the sequential algorithm while learning the parameters of the prediction function, i.e., the parameters of the underlying distribution. To illustrate this procedure, we investigate the regret given in (2)

for three candidate function classes that are widely studied in computational learning theory.

A. mth-Order Univariate Polynomial Prediction

For an mth-order polynomial in $x[t-1]$, the regret is given by

$$\sup_{x_1^n} \left\{ \sum_{t=1}^n (x[t] - \hat{x}_s[t])^2 - \inf_{\boldsymbol{w} \in \mathbb{R}^m} \sum_{t=1}^n \left(x[t] - \sum_{i=1}^p w_i x^i[t-1] \right)^2 \right\} \tag{8}$$

where $\hat{x}_s[t]$ is the prediction at time t of any sequential algorithm that has access data from $x[1]$ up to $x[t-1]$ for prediction, $\boldsymbol{w} = [w_1, \ldots, w_m]^T$ is the parameter vector, and $x^i[t-1]$ is the ith power of $x[t-1]$.

Since $\sum_{i=1}^m w_i x^i[t-1] = w_1 x[t-1]$ with appropriate selection of $\boldsymbol{w}$, considering the following distribution on x_1^n, we can lower bound the regret in (8). Given θ from a beta distribution with parameters (C, C), $C \in R^+$, we generate a sequence x_1^n having only two values, A and $-A$ such that $x[t] = x[t-1]$ with probability θ and $x[t] = -x[t-1]$ with probability $(1-\theta)$. Then, $E[x[t]|x_1^{t-1}, \theta] = (2\theta - 1)x[t-1]$, giving $h(\boldsymbol{\theta}, x_{t-a}^{t-1}) = (2\theta - 1)x[t-1]$. Since the MMSE given θ is linear in $x[t-1]$, the optimum $\boldsymbol{w}$ that minimizes the accumulated error for this distribution is $\boldsymbol{w} = [(2\theta-1), 0, \ldots, 0]^T$. After following the lines in [5], we obtain a lower bound of the form $O(\ln(n))$.

B. Multivariate Polynomial Prediction

Suppose the prediction function is given by $\boldsymbol{w}^T \boldsymbol{f}_x(x_{t-a}^{t-1}) = \sum_{k=1}^m w_k f_k(x_{t-r}^{t-1})$, where each $f_k(x_{t-r}^{t-1})$ is a multivariate polynomial function (as an example $f_k(x_{t-r}^{t-1}) = x[t-1]x^2[t-2]/x[t-3]$), and regret is taken over all $\boldsymbol{w} = [w_1, \ldots, w_m]^T \in \mathbb{R}^m$, that is

$$\sup_{x_1^n} \left\{ \sum_{t=1}^n (x[t] - \hat{x}_s[t])^2 - \inf_{\boldsymbol{w} \in \mathbb{R}^m} \sum_{t=1}^n \left(x[t] - \boldsymbol{w}^T \boldsymbol{f}_x(x_{t-a}^{t-1}) \right)^2 \right\}$$

where $\hat{x}_s[t]$ is the prediction at time t of any sequential algorithm that has access data from $x[1]$ up to $x[t-1]$ for prediction, and $\boldsymbol{w}$ is the parameter for prediction.

We emphasize that this class of predictors are not only the super set of univariate polynomial predictors, but also widely used in many signal processing applications to model nonlinearity, such as Volterra filters [11]. This filtering technique is attractive when linear filtering techniques do not provide satisfactory results, and includes cross products of the input signals.

Since $\sum_{k=1}^m w_k f_k(x_{t-r}^{t-1}) = w_1 f_1(x_{t-r}^{t-1})$ with an appropriate selection of $\boldsymbol{w}$ and redefinition of $f_1(x_{t-r}^{t-1})$, we define the following parametric distribution on x_1^n to obtain a lower bound. Given θ from a beta distribution with parameters (C, C), $C \in R^+$, we generate a sequence x_1^n having only two values, A and $-A$, such that $x[t] = f_n(x_{t-a}^{t-1})$ with probability θ and $x[t] = -f_n(x_{t-a}^{t-1})$ with probability $(1-\theta)$, where $f_n(x_{t-a}^{t-1}) = Af_1(x_{t-r}^{t-1})/M$, i.e., normalized version of $f_1(x_{t-r}^{t-1})$. Thus, given θ, x_1^n forms a two-state Markov

chain with transition probability $(1 - \theta)$. Hence, we have $E[x[t]|x_1^{t-1}, \theta] = (2\theta - 1)f_n(x_{t-a}^{t-1})$. The lower bound for the regret is given by

$$L(n) = E\left[\left(x[t] - (2\hat{\theta} - 1)f_n(x_{t-a}^{t-1}) \right)^2 \right]$$
$$- E\left[\left(x[t] - (2\theta - 1)f_n(x_{t-a}^{t-1}) \right)^2 \right]$$

where $\hat{\theta} = E[\theta|x_1^{t-1}]$. After some algebra, we achieve

$$L(n) = -4E\left[\hat{\theta}x[t]f_n(x_{t-a}^{t-1}) \right] + 4E\left[\theta x[t]f_n(x_{t-a}^{t-1}) \right]$$
$$+ E\left[(2\hat{\theta} - 1)^2 \right] - E\left[(2\theta - 1)^2 \right].$$

It can be deduced that

$$\hat{\theta} = E\left[\theta|x_1^{t-1} \right] = \frac{t - 2 - F_{t-2} + C}{t - 2 + 2C}$$

where F_{t-2} is the total number of transitions between the two states in a sequence of length $(t-1)$, i.e., $\hat{\theta}$ is ratio of number of transitions to time period. Hence

$$E\left[\hat{\theta}x[t]f_n(x_{t-a}^{t-1}) \right]$$
$$= E\left[\frac{t - 2 - F_{t-2} + C}{t - 2 + 2C} x[t]f_n(x_{t-a}^{t-1}) \right]$$
$$= \frac{(t - 2 + C)E\left[x[t]f_n(x_{t-a}^{t-1}) \right] - E\left[F_{t-2}x[t]f_n(x_{t-a}^{t-1}) \right]}{t - 2 + 2C}$$
$$= -\frac{1}{t - 2 + 2C}E\left[(1 - \theta)(t - 2)x[t]f_n(x_{t-a}^{t-1}) \right]$$
$$= \frac{t - 2}{t - 2 + 2C}E\left[\theta x[t]f_n(x_{t-a}^{t-1}) \right]$$

where the third line follows from $E[x[t]f_n(x_{t-a}^{t-1})] = E[(2\theta - 1)A^2] = 0$ and $E[F_{t-2}|x[t]f_n(x_{t-a}^{t-1})] = (t-2)(1-\theta)$ since F_{t-2} is a binomial random variable with parameters $(1 - \theta)$ and size $(t - 2)$. Thus, we obtain

$$L(n) = -4\frac{t - 2}{t - 2 + 2C}E\left[\theta x(t)f_n(x_{t-a}^{t-1}) \right] + 4E\left[\theta x(t)f_n(x_{t-a}^{t-1}) \right]$$
$$+ E\left[(2\hat{\theta} - 1)^2 \right] - E\left[(2\theta - 1)^2 \right].$$

After this line, the derivation follows similar lines to [7], giving a lower bound of the form $O(\ln(n))$ for the regret.

C. k-Ahead mth-Order Linear Prediction

The regret in (2) for k-ahead mth-order linear prediction is given by

$$\sup_{x_1^n} \left\{ \sum_{t=1}^n (x[t] - \hat{x}_s[t])^2 - \inf_{\boldsymbol{w} \in \mathbb{R}^m} \sum_{t=1}^n (x[t] - \boldsymbol{w}^T \boldsymbol{x}[t-k])^2 \right\} \tag{9}$$

where $\hat{x}_s[t]$ is the prediction at time t of any sequential algorithm that has access data from $x[1]$ up to $x[t-k]$ for prediction for some integer k, $\boldsymbol{w} = [w_1, \ldots, w_m]^T$ is the parameter vector, and $\boldsymbol{x}[t-k] = [x[t-k], \ldots, x[t-k-m+1]]^T$.

We first find a lower bound for k-ahead first-order prediction, where $\boldsymbol{w}^T \boldsymbol{x}[t-k] = wx[t-k]$. For this purpose, we define the following parametric distribution on x_1^n as in [5].

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

5

Given θ from a beta distribution with parameters (C, C), $C \in R^+$, we generate a sequence x_1^n having only two values, A and $-A$, such that $x[t] = x[t - k]$ with probability θ and $x[t] = -x[t - k]$ with probability $(1 - \theta)$. Thus, given θ, x_1^n forms a two-state Markov chain with transition probability $(1 - \theta)$. Then, $E[x[t]|x_1^{t-k}, \theta] = (2\theta - 1)x[t - k]$, giving $h(\theta, x_{t-a}^{t-1}) = (2\theta - 1)x[t - k]$ and $g(\theta) = (2\theta - 1)$. After this point, the derivation exactly follows the lines in [5] resulting a lower bound of the form $O(\ln(n))$.

For k-ahead mth-order prediction, we generalize the lower bound obtained for k-ahead first-order prediction and following the lines in [5], we obtain a lower bound of the form $O(m \ln(n))$.

III. Comprehensive Approach to Regret Minimization

In this section, we introduce a method which can be used to predict a bounded, arbitrary, and unknown sequence. We derive the upper bounds of this algorithm such that for any sequence x_1^n, our algorithm will not perform worse than the presented upper bounds. In some cases, by achieving matching upper and lower bounds, we prove that this algorithm is optimal in a strong minimax sense such that the worst-case performance cannot be further improved.

We restrict the prediction functions to be separable, i.e., $f(\boldsymbol{w}, x_{t-a}^{t-1}) = \boldsymbol{f}_w(\boldsymbol{w})^T \boldsymbol{f}_x(x_{t-a}^{t-1})$, where $\boldsymbol{f}_w(\boldsymbol{w})$ and $\boldsymbol{f}_x(x_{t-a}^{t-1})$ are vector functions of size $m \times 1$ for some m integer. To avoid any confusion, we simply denote $\boldsymbol{\beta} \triangleq \boldsymbol{f}_w(\boldsymbol{w})$, where $\boldsymbol{\beta} \in \mathbb{R}^m$. Hence, the same prediction function can be written as $f(\boldsymbol{w}, x_{t-a}^{t-1}) = \boldsymbol{\beta}^T \boldsymbol{f}_x(x_{t-a}^{t-1})$.

If the parameter vector $\boldsymbol{\beta}$ is selected such that the total squared prediction error is minimized over a batch of data of length n, then the coefficients are given by

$$\boldsymbol{\beta}^*[n] = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^m} \sum_{t=1}^{n} \left(x[t] - \boldsymbol{\beta}^T \boldsymbol{f}_x(x_{t-a}^{t-1}) \right)^2.$$

The well-known least-squares solution to this problem is given by $\boldsymbol{\beta}^*[n] = (R_{ff}^n)^{-1} r_{xf}^n$, where

$$R_{ff}^n \triangleq \sum_{t=1}^{n} \boldsymbol{f}_x(x_{t-a}^{t-1}) \boldsymbol{f}_x(x_{t-a}^{t-1})^T$$

is invertible and

$$r_{xf}^n \triangleq \sum_{t=1}^{n} x[t] \boldsymbol{f}_x(x_{t-a}^{t-1}).$$

When R_{ff}^n is singular, the solution is no longer unique, however, a suitable choice can be made using, e.g., pseudoinverses.

We also consider the more general least-squares (ridge regression) problem that arises in many signal processing problems, and whose total squared prediction error is minimized over a batch of data of length n with

$$\boldsymbol{\beta}^*[n] = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^m} \left\{ \sum_{t=1}^{n} \left(x[t] - \boldsymbol{\beta}^T \boldsymbol{f}_x(x_{t-a}^{t-1}) \right)^2 + \delta ||\boldsymbol{\beta}||^2 \right\}$$

$$= \left[R_{ff}^n + \delta \boldsymbol{I} \right]^{-1} r_{xf}^n.$$

We define a universal predictor $\tilde{x}_u[n]$, as

$$\tilde{x}_u[n] = \boldsymbol{\beta}_u[n - 1]^T \boldsymbol{f}(x_{n-a}^{n-1})$$

where

$$\boldsymbol{\beta}_u[n] = \boldsymbol{\beta}^*[n] = \left[R_{ff}^n + \delta I \right]^{-1} r_{xf}^n$$

and $\delta > 0$ is a positive constant.

Theorem 2: The total squared prediction error of the mth-order universal predictor for any bounded arbitrary sequence of $\{x[t]\}_{t \geq 1}$, $|x[t]| \leq A$, having an arbitrary length of n satisfies

$$\sum_{t=1}^{n} (x[t] - \tilde{x}_u[t])^2 \leq \min_{\boldsymbol{\beta} \in R^m} \left\{ \sum_{t=1}^{n} \left(x[t] - \boldsymbol{\beta}^T \boldsymbol{f}_x(x_{t-a}^{t-1}) \right)^2 + \delta ||\boldsymbol{\beta}||^2 \right\}$$
$$+ A^2 \ln |I + R_{ff}^n \delta^{-1}|.$$

Theorem 2 indicates that the total squared prediction error of the mth-order universal predictor is within $O(m \ln(n))$ of the best batch mth-order parametric predictor for any individual sequence of $\{x[t]\}_{t \geq 1}$. This result implies that in order to learn m parameters, the universal algorithm pays a regret of $O(m \ln(n))$, which can be viewed as the parameter regret. After we prove Theorem 2, we apply Theorem 2 to the competition classes discussed in Section II.

Proof of Theorem 2: We prove this result for a scalar prediction function such that $\boldsymbol{f}_x(x_{t-a}^{t-1}) = f_x(x_{t-a}^{t-1})$ to avoid any confusions. Yet for a vector prediction function of $\boldsymbol{f}_x(x_{t-a}^{t-1})$, one can follow the exact same steps in this proof with vector extensions of the Gaussian mixture.

The derivations follow similar lines to [5] and [10], hence only main points are presented. We first define a function of the loss, namely the probability for a predictor having parameter β as follows:

$$P_\beta(x_1^n) = \exp\left(-\frac{1}{2h} \sum_{k=1}^{n} \left(x[k] - \beta f_x(x_{t-a}^{t-1}) \right)^2 \right)$$

which can be viewed as a probability assignment of the predictor with parameter β to the data $x[t]$, for $1 \leq t \leq n$, induced by performance of β on the sequence x_1^n. We then construct a universal estimate of the probability of the sequence x_1^n, as an *a priori* weighted mixture among all of the probabilities, i.e., $P_u(x_1^n) = \int_{-\infty}^{\infty} p(\beta) P_\beta(x_1^n) d\beta$, where $p(\beta)$ is an *a priori* weight assigned to the parameter β, and is selected as Gaussian in order to obtain a closed form bounds, i.e., $p(\beta) = 1/(2\pi)^{1/2} \sigma \exp\{-\beta^2/2\sigma^2\}$.

Following similar lines to [7] with a predictor of $\beta f_x(x_{t-a}^{t-1})$, we obtain:

$$P_u(x_n|x^{n-1}) = \gamma \exp\left\{ \frac{-1}{2h} \gamma^2 \left(x[n] - \beta[n - 1] f(x_{n-a}^{n-1}) \right)^2 \right\}$$

where $\gamma \triangleq \left((R_{ff}^{n-2} + \delta)/(R_{ff}^{n-1} + \delta) \right)^{1/2}$. If we could find another Gaussian satisfying $\tilde{P}_u(x^n) \geq P_u(x^n)$, then it would complete the proof of the theorem.

After some algebra, we find that the universal predictor is given by

$$\tilde{x}_u[n] = \gamma^2 \beta^*[n - 1] f(x_{n-a}^{n-1}) = \frac{r_{xf}^{n-1}}{R_{ff}^{n-1} + \delta} f(x_{n-a}^{n-1}).$$

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

6

IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

Now, we can select the smallest value of h over the region $[-A, A]$, $\tilde{P}_u(x_n|x^{n-1})$ is larger than $P_u(x_n|x^{n-1})$, that is

$$A \leq \frac{\sqrt{2h\ln(\gamma)(\gamma^2 - 1) + \gamma^2 \hat{x}_u[n]^2(1 - \gamma^2)}}{(1 - \gamma^2)}$$

which must hold for all values of $\hat{x}_u[n] \in [-A, A]$. Therefore, $h \geq A^2(1 - \gamma^2)/-2\ln(\gamma)$, where $\gamma < 1$. Note that for $0 < \gamma < 1$ we have $0 < (1 - \gamma^2)/-2\ln\gamma < 1$, which implies that we must have $h \geq A^2$ to ensure that $\tilde{P}_u \geq P_u$. In fact, since this bound on the value of h depends upon the value of γ and $\hat{x}_u[n]$, and is only tight for $\gamma \rightarrow 1$, and $\hat{x}_u[n] = 0$, then the restriction that $|x[n]| < A$ can actually be occasionally violated, as long as $\tilde{P}_u \geq P_u$ still holds. $\square$

To illustrate this procedure, we investigate the upper bound for the regret in (2) for the same candidate function classes as we also investigated in Section II.

A. mth-Order Univariate Polynomial Predictor

For an mth-order polynomial in $x[t - 1]$, the prediction function is given by $f(\boldsymbol{w}, x_{t-a}^{t-1}) = \boldsymbol{\beta}^T \boldsymbol{f}_x(x_{t-a}^{t-1}) = \boldsymbol{\beta}^T \boldsymbol{m}[t-1]$, where $\boldsymbol{m}[t - 1] = [x[t - 1], \ldots, x^m[t - 1]]^T$, i.e., the vector of powers of $x[t - 1]$. After replacing $R_{ff}^n = R_{mm}^n = \sum_{t=1}^{n} \boldsymbol{m}[t-1]\boldsymbol{m}[t-1]^T$ and $r_{xf}^n = r_{xm}^n = \sum_{t=1}^{n} x[t]\boldsymbol{m}[t-1]$, we obtain an upper bound

$$\sum_{t=1}^{n} (x[t] - \tilde{x}_u[t])^2 \leq \min_{\boldsymbol{\beta} \in \mathbb{R}^m} \left\{ \sum_{t=1}^{n} \left(x[t] - \boldsymbol{\beta}^T \boldsymbol{m}[t - 1] \right)^2 + \delta ||\boldsymbol{\beta}||^2 \right\}$$
$$+ \underbrace{A^2 \ln \left| I + R_{mm}^n \delta^{-1} \right|}_{\leq A^2 m \ln(1 + A^2 n/\delta)}.$$

B. Multivariate Polynomial Prediction

The upper bound for a multivariate polynomial prediction function $\boldsymbol{f}_x(x_{t-a}^{t-1})$ exactly follows the upper bound derivation of mth-order univariate polynomial predictor giving an upper bound:

$$\sum_{t=1}^{n} (x[t] - \tilde{x}_u[t])^2 \leq \min_{\boldsymbol{\beta} \in \mathbb{R}^m} \left\{ \sum_{t=1}^{n} \left(x[t] - \boldsymbol{\beta}^T \boldsymbol{f}_x \left(x_{t-a}^{t-1} \right) \right)^2 + \delta ||\boldsymbol{\beta}||^2 \right\}$$
$$+ A^2 m \ln \left(1 + \frac{A^2 n}{\delta} \right).$$

C. k-Ahead mth-Order Linear Prediction

For k-ahead mth-order prediction, the prediction class is given by $f(\boldsymbol{w}, x_{t-a}^{t-1}) = \boldsymbol{\beta}^T \boldsymbol{f}_x(x_{t-a}^{t-1}) = \boldsymbol{\beta}^T \boldsymbol{x}[t - k]$ where $\boldsymbol{x}[t - k] = [x[t - k], \ldots, x[t - k - m + 1]]^T$ as before. After replacing $R_{ff}^n = R_{xx}^n = \sum_{t=1}^{n} \boldsymbol{x}[t - k]\boldsymbol{x}[t - k]^T$ and $r_{xf}^n = r_{xx}^n = \sum_{t=1}^{n} x[t]\boldsymbol{x}[t - k]$ with suitable limits, we obtain an upper bound

$$\sum_{t=1}^{n} (x[t] - \tilde{x}_u[t])^2 \leq \min_{\boldsymbol{\beta} \in \mathbb{R}^m} \left\{ \sum_{t=1}^{n} (x[t] - \boldsymbol{\beta}^T \boldsymbol{x}[t - k])^2 + \delta ||\boldsymbol{\beta}||^2 \right\}$$
$$+ A^2 m \ln \left(1 + \frac{A^2 n}{\delta} \right).$$

IV. RANDOMIZED OUTPUT PREDICTIONS

In this section, we investigate the performance of randomized output algorithms for the worst-case scenario with respect to linear predictors with using the same regret measure in (2). We emphasize that the randomized output algorithms are a super set of the deterministic sequential predictors and the derivations here can be readily generalized to include any prediction class. In particular, we consider randomized output algorithms $f(\boldsymbol{\theta}(x_1^{t-1}), x_1^{t-1})$ such that the randomization parameters $\boldsymbol{\theta} \in R^m$ can be a function of the whole past. Hence, a randomized sequential algorithm introduce randomization or uncertainty in its output such that the output also depends on a random element. Note that such methods are widely used in applications involving security considerations. As an example, suppose there are m prediction algorithms running in parallel to predict the observation sequence $\{x[t]\}_{t \geq 1}$ sequentially. At each time t, the randomized output algorithm selects one of the constituent algorithms randomly such that the algorithm k is selected with probability $p_k[t]$. By definition $\sum_{k=1}^{m} p_k[t] = 1$ and $p_k[t]$ may be generated as the combination of the past observation samples x_1^{t-1} and a seed independent from the observations.

For such randomized output prediction algorithms, we consider the following time-accumulated prediction error over a deterministic sequence $\{x[t]\}_{t \geq 1}$ as the prediction error:

$$P_{\text{rand}}(n) = \sum_{t=1}^{n} E_{\boldsymbol{\theta}} \left[\left(x[t] - f \left(\boldsymbol{\theta}(x_1^{t-1}), x_1^{t-1} \right) \right)^2 \right]. \quad (10)$$

This expectation is taken over all the randomization due to independent or dependent seeds. Hence, our general regret can be extended to include this performance measure

$$\sup_{x_1^n} \left\{ P_{\text{rand}}(n) - \min_{\boldsymbol{w} \in \mathbb{R}^m} \sum_{t=1}^{n} \left(x[t] - \boldsymbol{w}^T \boldsymbol{x}[t - 1] \right)^2 \right\}. \quad (11)$$

Expanding (10), we obtain

$$P_{\text{rand}}(n) = \sum_{t=1}^{n} \left\{ \left(x[t] - E_{\boldsymbol{\theta}} \left[f \left(\boldsymbol{\theta}(x_1^{t-1}), x_1^{t-1} \right) \right] \right)^2 \right.$$
$$\left. + \text{Var}_{\boldsymbol{\theta}} \left(f \left(\boldsymbol{\theta}(x_1^{t-1}), x_1^{t-1} \right) \right) \right\}$$

noting that $x[t]$ is independent of the randomization. Since $E_{\boldsymbol{\theta}}[f(\boldsymbol{\theta}(x_1^{t-1}), x_1^{t-1})]$ is a sequential function of x_1^{t-1} and $\text{Var}_{\boldsymbol{\theta}}(f(\boldsymbol{\theta}(x_1^{t-1}), x_1^{t-1}))$ is always nonnegative, the performance of a randomized output algorithm can be reached by a deterministic sequential algorithm.

Since deterministic algorithms are subclass of randomized output algorithms, upper bounds we derived for k-ahead mth-order prediction in (9) also hold for (11). Since we also proved that the lower bound for such linear predictions of mth order are in the form of $O(m\ln(n))$, the lower and upper bounds are tight and of the form $O(m\ln(n))$.

V. CONCLUSION

In this brief, we consider the problem of sequential prediction from a mixture of experts perspective. We have introduced comprehensive lower bounds on the sequential learning

framework by proving that for any sequential algorithm, there always exists a sequence for which the sequential predictor cannot outperform the class of parametric predictors, whose parameters are set noncasually. The lower bounds for important parametric classes, such as univariate polynomial, multivariate polynomial, and linear predictor classes, are derived in detail. We then introduced a universal sequential prediction algorithm and investigated the upper bound on the regret of this algorithm. We also derived the upper bounds in detail for the same important classes that we discussed for lower bounds, where we further showed that this algorithm is optimal in a strong minimax sense for some scenarios. Finally, we have proven that for the worst-case scenario, randomized output algorithms cannot provide any improvement in the performance compared with the sequential algorithms.

REFERENCES

[1] N.-Y. Liang, G.-B. Huang, P. Saratchandran, and N. Sundararajan, "A fast and accurate online sequential learning algorithm for feedforward networks," *IEEE Trans. Neural Netw.*, vol. 17, no. 6, pp. 1411–1423, Nov. 2006.

[2] L. Devroye, T. Linder, and G. Lugosi, "Nonparametric estimation and classification using radial basis function nets and empirical risk minimization," *IEEE Trans. Neural Netw.*, vol. 7, no. 2, pp. 475–487, Mar. 1996.

[3] A. Krzyzak and T. Linder, "Radial basis function networks and complexity regularization in function learning," *IEEE Trans. Neural Netw.*, vol. 9, no. 2, pp. 247–256, Mar. 1998.

[4] N. Cesa-Bianchi, P. M. Long, and M. K. Warmuth, "Worst-case quadratic loss bounds for prediction using linear functions and gradient descent," *IEEE Trans. Neural Netw.*, vol. 7, no. 3, pp. 604–619, May 1996.

[5] A. C. Singer and M. Feder, "Universal linear prediction by model order weighting," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2685–2699, Oct. 1999.

[6] G. C. Zeitler and A. Singer, "Universal linear least-squares prediction in the presence of noise," in *Proc. IEEE/SP 14th Workshop SSP*, Aug. 2007, pp. 611–614.

[7] 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.

[8] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2000.

[9] V. Cherkassky, X. Shao, F. M. Mulier, and V. N. Vapnik, "Model complexity control for regression using VC generalization bounds," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 1075–1089, Sep. 1999.

[10] J. Kivinen and M. K. Warmuth, "Exponentiated gradient versus gradient descent for linear predictors," *J. Inf. Comput.*, vol. 132, no. 1, pp. 1–63, 1997.

[11] V. J. Mathews, "Adaptive polynomial filters," *IEEE Signal Process. Mag.*, vol. 8, no. 3, pp. 10–26, Jul. 1991.

[12] V. Vovk, "Competitive on-line statistics," *Int. Statist. Rev.*, vol. 69, no. 2, pp. 213–248, 2001.

[13] S. S. Kozat, A. C. Singer, and G. C. Zeitler, "Universal piecewise linear prediction via context trees," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3730–3745, Jul. 2007.

[14] 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, Sep. 2001.

[15] T. Moon and T. Weissman, "Universal FIR MMSE filtering," *IEEE Trans. Signal Process.*, vol. 57, no. 3, pp. 1068–1083, Mar. 2009.

[16] T. Moon and T. Weissman, "Competitive on-line linear FIR MMSE filtering," in *Proc. IEEE ISIT*, Jun. 2007, pp. 1126–1130.

[17] H. Stark and J. Woods, *Probability, Random Processes, and Estimation Theory for Engineers*. Upper Saddle River, NJ, USA: Prentice-Hall, 1994.