COMPREHENSIVE LOWER BOUNDS ON SEQUENTIAL PREDICTION

N. Denizcan Vanli[*], *Muhammed O. Sayin*[*], *Salih Ergüt*[†], *and Suleyman S. Kozat*[*]

[*] Department of Electrical and Electronics Engineering
Bilkent University, Bilkent, Ankara 06800, Turkey
{vanli, sayin, kozat}@ee.bilkent.edu.tr

[†] AveaLabs, Istanbul, Turkey
salih.ergut@avea.com.tr

ABSTRACT

We study the problem of sequential prediction of real-valued sequences under the squared error loss function. While refraining from any statistical and structural assumptions on the underlying sequence, we introduce a competitive approach to this problem and compare the performance of a sequential algorithm with respect to the large and continuous class of parametric predictors. We define the performance difference between a sequential algorithm and the best parametric predictor as "regret", and introduce a guaranteed worst-case lower bounds to this relative performance measure. In particular, we prove that for any sequential algorithm, there always exists a sequence for which this regret is lower bounded by zero. We then extend this result by showing that the prediction problem can be transformed into a parameter estimation problem if the class of parametric predictors satisfy a certain property, and provide a comprehensive lower bound to this case.

Index Terms— Sequential prediction, lower bound, worst-case performance.

1. INTRODUCTION

In this paper, we investigate the generic sequential prediction problem under the squared error loss function, where we refrain from any statistical assumptions both on the algorithms and sequences [1–3]. We consider an arbitrary, deterministic, bounded and unknown signal $\{x[t]\}_{t\geq 1}$, where $|x[t]| < A < \infty$ and $x[t] \in \mathbb{R}$. In this sense, we define the performance of a sequential algorithm with respect to a comparison class and try to predict the sequence as well as the best predictor among the comparison class. In particular, we define this competitive performance metric as follows

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

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 only access to data from $x[1]$ to $x[t-1]$, 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 the competition class does not have any restrictions while making the prediction, e.g., this class may contain predictors that has access to entire sequence $\{x[t]\}_{t\geq 1}$ even before processing starts (i.e., batch predictors). In this sense, this competitive performance metric in (1) can in fact, be viewed as the "regret" of the sequential predictor for not knowing the future.

In order to obtain comprehensive results, we do not set a specific comparison class but parameterize the comparison classes such that the parameter set and functional form of these classes can be chosen as desired. Therefore, we uniquely identify the class of parametric predictors with their parameter vector of $\boldsymbol{w} \triangleq [w_1, \ldots, w_m]^T$, and denote the regret in (1) as follows [1]

$$\mathcal{R}(x_1^n) \triangleq \sum_{t=1}^{n}(x[t]-\hat{x}_s[t])^2 - \inf_{\boldsymbol{w}\in\mathbb{R}^m} \sum_{t=1}^{n}(x[t]-f(\boldsymbol{w}, x_{t-a}^{t-1}))^2, \qquad (2)$$

where $f(\boldsymbol{w}, x_{t-a}^{t-1})$ is a parametric function whose parameters $\boldsymbol{w}$ can be set prior to prediction, and a is an arbitrary integer representing the tap size of the predictor. We emphasize that even though the parameters of a parametric predictor can be set prior to prediction, it is still obligated to use the data x_{t-a}^{t-1} in order to predict $x[t]$.

Under this framework, we introduce the generalized lower bounds for sequential prediction by transforming the prediction problem to a well-known and widely studied statistical parameter learning problem [1–5]. Specifically, we show that there always exist a sequence $\{x[t]\}_{t\geq 1}$ such that the regret in (2) is lower bounded by zero. We push the analysis further and prove that there always exist a sequence for which this regret cannot be smaller than $O(\ln(n))$ if the parameter function is in a separable form, i.e.,

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

The organization of the paper is as follows. In Section 2, we present the lower bounds for a generic class of parametric

[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 $x_a^b \triangleq \{x[t]\}_{t=a}^{b}$.

predictors. In Section 3, we consider a specific type of parametric predictors, namely the separable ones (the meaning of "separable" will be cleared in the paper), and introduce a procedure to transform the prediction problem into a parameter estimation problem. We finalize our paper by pointing out several concluding remarks.

2. PARAMETRIC PREDICTORS

In this section, we investigate the worst-case performance of sequential algorithms compared to the generic class of parametric predictors in order to obtain guaranteed lower bounds on the regret. For any arbitrary data sequence $\{x[t]\}_{t\geq 1}$ with an arbitrary length n, we consider the optimal sequential predictor for that sequence and seek to find a lower bound on the following regret

$$\inf_{s\in\mathcal{S}} \sup_{x_1^n} \mathcal{R}(x_1^n), \tag{3}$$

where $\mathcal{S}$ is the class of all parametric predictors. For this formulation, we introduce the following theorem, which relates the performance of any sequential algorithm to the general class of parametric predictors.

Theorem 1: *Given a parametric class of predictors in the form $f(\boldsymbol{w}, x_{t-a}^{t-1})$, where $\boldsymbol{w}\in\mathbb{R}^m$, we have*

$$\inf_{s\in\mathcal{S}} \sup_{x_1^n} \mathcal{R}(x_1^n) \geq 0. \tag{4}$$

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.

Proof of Theorem 1: We begin our proof by noting that for an arbitrary sequence of x_1^n, the optimal sequential predictor may not be found straightforwardly. Yet, for a specific distribution on x_1^n, the best predictor is the conditional mean on x_1^n under the squared error [6]. For any distribution on x_1^n, we have

$$\inf_{s\in\mathcal{S}} \sup_{x_1^n} \mathcal{R}(x_1^n) \geq \inf_{s\in\mathcal{S}} E_{x_1^n}\left[\mathcal{R}(x_1^n)\right], \tag{5}$$

where expectation is taken with respect to this particular distribution. Hence, it is enough to lower bound the right hand side of (5) to get a final lower bound. By the linearity of the expectation, we obtain

$$\inf_{s\in\mathcal{S}} E_{x_1^n}\left[\mathcal{R}(x_1^n)\right] = \mathcal{L}_s(x_1^n) - \mathcal{L}_c(x_1^n), \tag{6}$$

where $\mathcal{L}_s(x_1^n)$ denotes the minimum loss that can be achieved with a sequential predictor for the sequence x_1^n, i.e.,

$$\mathcal{L}_s(x_1^n) \triangleq \inf_{s\in\mathcal{S}} E_{x_1^n}\left[\sum_{t=1}^{n}(x[t]-\hat{x}_s[t])^2\right],$$

and $\mathcal{L}_c(x_1^n)$ denotes the loss of the optimal predictor in the competition class, i.e.,

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

We now select a parametric distribution for x_1^n with parameter vector $\boldsymbol{\theta}=[\theta_1,\ldots,\theta_m]^T$. Then consider $\mathcal{L}_s(x_1^n)$ and $\mathcal{L}_c(x_1^n)$ terms separately.

The squared-error loss $E_{x_1^n}\left[(x[t]-\hat{x}_s[t])^2\right]$ is minimized with the well-known minimum mean squared error (MMSE) predictor given by [6]

$$\hat{x}_s[t] = E\left[x[t]\middle|x[t-1],\ldots,x[1]\right] = E\left[x[t]\middle|x_1^{t-1}\right], \tag{7}$$

where we drop the explicit x_1^n-dependence of the expectation to simplify notation. By expanding the expectation, we then obtain

$$\mathcal{L}_s(x_1^n) = E_{\boldsymbol{\theta}}\left[E_{x_1^n|\boldsymbol{\theta}}\left[\sum_{t=1}^{n}\left(x[t]-E\left[x[t]\middle|x_1^{t-1}\right]\right)^2\right]\right]. \tag{8}$$

Now turning our attention back to $\mathcal{L}_c(x_1^n)$, we expand the expectation and observe that

$$\mathcal{L}_c(x_1^n) \leq E_{\boldsymbol{\theta}}\left[\inf_{\boldsymbol{w}\in\mathbb{R}^m} E_{x_1^n|\boldsymbol{\theta}}\left[\sum_{t=1}^{n}(x[t]-f(\boldsymbol{w}, x_{t-a}^{t-1}))^2\right]\right]. \tag{9}$$

Hence, for a distribution on x_1^n such that

$$E\left[x[t]\middle|x_1^{t-1},\boldsymbol{\theta}\right] = a(\boldsymbol{\theta})\, h(\boldsymbol{\theta}, x_{t-a}^{t-1}), \tag{10}$$

with some functions $a(\cdot)$ and $h(\cdot,\cdot)$, if we can find a vector function $\boldsymbol{g}(\boldsymbol{\theta})$ such that

$$f(\boldsymbol{g}(\boldsymbol{\theta}), x_{t-a}^{t-1}) = a(\boldsymbol{\theta})\, h(\boldsymbol{\theta}, x_{t-a}^{t-1}),$$

then (9) can be written as

$$\mathcal{L}_c(x_1^n) \leq E_{\boldsymbol{\theta}}\left[E_{x_1^n|\boldsymbol{\theta}}\left[\sum_{t=1}^{n}\left(x[t]-E\left[x[t]\middle|x_1^{t-1},\boldsymbol{\theta}\right]\right)^2\right]\right]. \tag{11}$$

Combining (6) with (8) and (11), we obtain

$$\begin{aligned}\inf_{s\in\mathcal{S}} E_{x_1^n}\left[\mathcal{R}(x_1^n)\right] \geq{}& \\ & E_{\boldsymbol{\theta}}\left[E_{x_1^n|\boldsymbol{\theta}}\left[\sum_{t=1}^{n}\left(x[t]-E\left[x[t]\middle|x_1^{t-1}\right]\right)^2\right]\right] \\ & - E_{\boldsymbol{\theta}}\left[E_{x_1^n|\boldsymbol{\theta}}\left[\sum_{t=1}^{n}\left(x[t]-E\left[x[t]\middle|x_1^{t-1},\boldsymbol{\theta}\right]\right)^2\right]\right],\end{aligned} \tag{12}$$

which is by definition of the MMSE estimator is always lower bounded by zero, i.e.,

$$\inf_{s\in\mathcal{S}} E_{x_1^n}\left[\mathcal{R}(x_1^n)\right] \geq 0.$$

Hence, 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 the competition class of parametric predictors for some sequence x_1^n. This means that our proof follows if a suitable distribution on x_1^n can be found for a given $f(\boldsymbol{w}, x_{t-a}^{t-1})$ such that $f(\boldsymbol{g}(\boldsymbol{\theta}), x_{t-a}^{t-1}) = a(\boldsymbol{\theta})\, h(\boldsymbol{\theta}, x_{t-a}^{t-1})$ with a suitable transformation $\boldsymbol{g}(\boldsymbol{\theta})$.

We proceed by considering the following distribution on x_1^n. Suppose $f(\boldsymbol{w}, x_{t-a}^{t-1})$ is bounded by some $M \in R^+$ with $M < \infty$ 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

$$x[t] = \begin{cases} \frac{A}{M} f(\boldsymbol{w}, x_{t-a}^{t-1}) & \text{, with probability } \theta \\ -\frac{A}{M} f(\boldsymbol{w}, x_{t-a}^{t-1}) & \text{, with probability } 1-\theta \end{cases}.$$

Then

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

Hence, this concludes the proof of the Theorem 1. □

3. SEPARABLE PARAMETRIC PREDICTORS

In this section, we consider the restricted functional form $f(\boldsymbol{w}, x_{t-a}^{t-1})$ so that $f(\boldsymbol{w}, x_{t-a}^{t-1})$ is 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 some vector functions. Denoting $\boldsymbol{v} \triangleq \boldsymbol{f}_w(\boldsymbol{w})$, we obtain the regret compactly as follows

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

We emphasize that this restricted form can be considered as the super set of entire polynomial predictors, which are widely used in many signal processing applications to model nonlinearity such as Volterra filters [7]. This filtering technique is attractive when linear filtering techniques do not provide satisfactory results, and includes cross products of the input signals.

Similar to the previous section, for any arbitrary data sequence $\{x[t]\}_{t\geq 1}$ with an arbitrary length n, we consider the optimal sequential predictor for that sequence and seek to find a lower bound on the following regret

$$\inf_{s\in\mathcal{S}} \sup_{x_1^n} \mathcal{R}(x_1^n),$$

where $\mathcal{S}$ is the class of all parametric predictors.

In Section 2, we have proven that there always exists a sequence such that the performance of any sequential algorithm compared to the generic class of parametric predictors is lower bounded by zero. In the following theorem, we compare the performance of any sequential algorithm with respect to the class of separable parametric predictors and introduce the following theorem.

Theorem 2: *For any sequential algorithm, there always exist a sequence for which the performance of a sequential algorithm with respect to the class of separable parametric predictors will always be lower bounded by* $O(\ln(n))$*, i.e.,*

$$\inf_{s\in\mathcal{S}} \sup_{x_1^n} \mathcal{R}(x_1^n) \geq O(\ln(n)).$$

This theorem indicates that when the competition class only consists of separable parametric predictors, the prediction problem can be transformed into a parameter estimation problem. By doing so, we show that no matter how smart a sequential algorithm can be, it cannot possibly achieve a better learning rate than $O(\ln(n))$ for all sequences. The algorithms that are claimed to achieve a better learning rate are *certainly* based on some ad-hoc assumptions such as a priori knowledge on the underlying sequence and cannot be guaranteed to achieve the claimed learning rate for all sequences. In fact, if one finds an algorithm with an upper bound of $O(\ln(n))$, then the performance of that algorithm cannot be further improved for all sequences.

Proof of Theorem 2: Since we consider the class of separable parametric predictors, we have

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

We then generate the underlying sequence x_1^n as follows. Denoting

$$\boldsymbol{f}_x(x_{t-a}^{t-1}) \triangleq [f_1(x_{t-a}^{t-1}), \ldots, f_p(x_{t-a}^{t-1})]^T,$$

for some integer p, and 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

$$x[t] = \begin{cases} f_n(x_{t-a}^{t-1}) & \text{, with probability } \theta \\ -f_n(x_{t-a}^{t-1}) & \text{, with probability } 1-\theta \end{cases},$$

where

$$f_n(x_{t-a}^{t-1}) \triangleq \frac{A}{M} f_1(x_{t-r}^{t-1}),$$

i.e., the 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 $(1-\theta)$. We then have

$$E\left[x[t] \middle| x_1^{t-1}, \theta\right] = (2\theta - 1) f_n(x_{t-a}^{t-1}).$$

Since we have

$$\inf_{s\in\mathcal{S}} \sup_{x_1^n} \mathcal{R}(x_1^n) \geq \inf_{s\in\mathcal{S}} E_{x_1^n}\left[\mathcal{R}(x_1^n)\right],$$

we obtain the lower bound for the regret as follows

$$\begin{aligned}\inf_{s\in\mathcal{S}} E_{x_1^n}\left[\mathcal{R}(x_1^n)\right] &= E\left[(x[t]-(2\hat{\theta}-1)f_n(x_{t-a}^{t-1}))^2\right] \\ &\quad - E\left[(x[t]-(2\theta-1)f_n(x_{t-a}^{t-1}))^2\right],\end{aligned}$$

where we have the optimal sequential predictor in the following form

$$\hat{\theta} = E[\theta|x_1^{t-1}].$$

After some algebra we achieve

$$\begin{aligned}\inf_{s\in\mathcal{S}} E_{x_1^n}\left[\mathcal{R}(x_1^n)\right] &= -4E[\hat{\theta}x[t]f_n(x_{t-a}^{t-1})] \\ &\quad + 4E[\theta x[t]f_n(x_{t-a}^{t-1})] + E[(2\hat{\theta}-1)^2] - E[(2\theta-1)^2].\end{aligned} \tag{13}$$

Now considering the first term of (13), we observe that

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

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,

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

where the third line follows since

$$E[x[t]f_n(x_{t-a}^{t-1})] = E[(2\theta-1)A^2] = 0,$$

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

$$\begin{aligned}\inf_{s\in\mathcal{S}} E_{x_1^n}\left[\mathcal{R}(x_1^n)\right] &= -4\frac{t-2}{t-2+2C}E[\theta x(t)f_n(x_{t-a}^{t-1})] \\ &\quad + 4E[\theta x(t)f_n(x_{t-a}^{t-1})] + E[(2\hat{\theta}-1)^2] - E[(2\theta-1)^2].\end{aligned}$$

After this line the derivation follows similar lines to Theorem 3 of [3], which results in

$$\inf_{s\in\mathcal{S}} E_{x_1^n}\left[\mathcal{R}(x_1^n)\right] \geq O(\ln(n)).$$

This concludes the proof of Theorem 2. □

4. CONCLUDING REMARKS

In this paper, we consider the problem of sequential prediction from a mixture of experts perspective. We introduce 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 non-casually. We then consider a specific type of parametric predictors (i.e., separable parametric predictors), where we emphasize that this class of predictors are still a comprehensive one, e.g., all linear and polynomial predictors are subsets of separable parametric predictors. In this framework, we transform the prediction problem to a parameter estimation problem and show that there always exists a sequence such that the regret of a sequential predictor is lower bounded by $O(\ln(n))$.

REFERENCES

[1] A. C. Singer and M. Feder, "Universal linear prediction by model order weighting," *IEEE Transactions on Signal Processing*, vol. 47, no. 10, pp. 2685–2699, 1999.

[2] G. C. Zeitler and A. C. Singer, "Universal linear least-squares prediction in the presence of noise," in *IEEE/SP 14th Workshop on Statistical Signal Processing, 2007. SSP '07*, 2007, pp. 611–614.

[3] A. C. Singer, S. S. Kozat, and M. Feder, "Universal linear least squares prediction: upper and lower bounds," *IEEE Transactions on Information Theory*, vol. 48, no. 8, pp. 2354–2362, 2002.

[4] T. Weissman and N. Merhav, "Universal prediction of individual binary sequences in the presence of noise," *IEEE Transactions on Information Theory*, vol. 47, no. 6, pp. 2151–2173, 2001.

[5] V. Vovk, "Competitive on-line statistics," *International Statistical Review*, vol. 69, pp. 213–248, 2001.

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

[7] V. Mathews, "Adaptive polynomial filters," *Signal Processing Magazine, IEEE*, vol. 8, no. 3, pp. 10–26, 1991.