MIN-MAX OPTIMAL UNIVERSAL PREDICTION WITH SIDE INFORMATION

Suleyman S. Kozat and Andrew C. Singer

Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
Urbana, IL 61801 USA
Email: $\{$*kozat, singer* $\}$@*ifp.uiuc.edu*

ABSTRACT

We consider the problem of sequential prediction of arbitrary real-valued sequences with side information. We first construct a universal algorithm that asymptotically achieves the performance of the best side-information dependent constant predictor, uniformly for all data and side-information sequences. We then extend these results to linear predictors of some fixed order. We derive matching upper and lower bounds, and show that the algorithms are not only universal but they are also optimal such that no sequential algorithm can give better performance for all sequences.

1. INTRODUCTION

In this paper, we investigate the problem of predicting a sequence $x[t], t = 1, \ldots, n$ with an associated side-information sequence $s[t]$, $t = 1, \ldots, n$, as well as the best predictor out of a large class of predictors. The real valued data sequence $x^n = \{x[t]\}_{t=1}^n$ is bounded, $|x[t]| \leq A$, but otherwise arbitrary, and the side-information sequence takes values from a finite set $s^n = \{s[t] \in \{1, \ldots, K\}\}_{t=1}^n$. The side-information sequence is used to incorporate additional information for prediction and may depend on the entire x^n in an arbitrary manner. Rather than making stochastic assumptions on the data or side information sequences, we simply try to predict the sequence x^n as well as the best predictor out of a large fixed class of predictors for all sequences x^n and side-information sequences s^n.

We first consider the class of state-constant predictors such that the predictor predicts the same value $c_i \in R$ when $s[t] = i$, $i = 1, \ldots, K$, given the state at time t. Hence there exist only K degrees of freedom for the output of the predictor. When there is no side information, i.e. $K = 1$, the predictor is a simple constant predictor. Although, this is a rather limited class of prediction in forecasting ability against which to compete, we permit the constants $c_i \in R$, $i = 1, \ldots, K$, to be selected based on observing the entire sequence and the associated side information states in advance of any prediction to be made. As such we seek to minimize the following regret

$$\sup_{x^n, s^n} \left\{ l\left(x^n, x_a^n | s^n\right) - l^*\left(x^n, x_c^n | s^n\right) \right\}$$

where $l\left(x^n, x_a^n | s^n\right)$ is the aggregated loss of any sequential prediction algorithm $x_a[t]$ that we might employ which for the square-error loss function would be $l\left(x^n, x_a^n | s^n\right) = \sum_{t=1}^n (x[t] - x_a[t])^2$ and $l^*\left(x^n, x_c^n | s^n\right)$ is the aggregated loss of the best state-constant algorithm for the sequence x^n and the side-information state sequence s^n.

We will construct a particular sequential algorithm such that this regret is at most $KA^2 \ln(n)$ and then demonstrate a corresponding lower bound of the same order $(KA^2 \ln(n))$ for any sequential algorithm, indicating a form of min-max optimality.

We then proceed to consider the class of state-constant fixed-order linear predictors such that each competing predictor forms its prediction as a linear function of the past observation sequence depending on the side-information state sequence. For this paper we will consider fixed-order predictors with order m, where m is an integer. As such, the prediction at time t is given by $\sum_{k=1}^m w_{i,k} x[t-k]$ when $s[t] = i$ and $i = 1, \ldots, K$ where $w_{1,1}, \ldots, w_{K,m} \in R$. Depending on the state sequence and the particular value of state s_i, at each time there are m degrees of freedom for selecting the prediction parameters $w_{i,k}$. Hence, there are Km parameters that can be selected. For the determination of $l^*\left(x^n, x_c^n | s^n\right)$ the corresponding $w_{i,k}$'s can be chosen by observing the entire sequence x^n and side-information state sequence. Here, we seek to minimize

$$\sup_{x^n, s^n} \left\{ l\left(x^n, x_a^n | s^n\right) - l^*\left(x^n, x_{\vec{w}}^n | s^n\right) \right\},$$

where $l^*\left(x^n, x_{\vec{w}}^n | s^n\right)$ is the aggregated loss of the best state constant mth-order linear predictor for the sequence x^n with the side-information state sequence s^n. We construct a sequential algorithm such that this regret is at most $mKA^2 \ln(n)$ and also demonstrate a lower bound of order

This work was supported by the National Science Foundation under CCR-0092598 (CAREER), by the NSF ITR 00-85929 .

0-7803-8484-9/04/$20.00 ©2004 IEEE

$mKA^2 \ln(n)$ for any sequential algorithm, again demonstrating a form of min-max optimality.

The idea of using side-information is applied by Cover in [3] for universal sequential investment. In [3] the authors present a universal algorithm that achieves, to first order in the exponent, the wealth of the best side-information dependent investment strategy. We extend this formulation of the problem to data prediction and seek universal algorithms that are optimum in certain min-max sense. Our algorithms are not only universal, i.e. they achieve the performance of the best batch predictor sequentially, but also their regret with respect to the best batch algorithm cannot be exceeded by any algorithm for all sequences as given by the corresponding lower bounds. The approach taken in this paper for constant predictors is based on universal sequential probability assignment and is similar to Vovk's Aggregating Algorithm for linear regression with the square error loss function [1].

2. STATE CONSTANT PREDICTORS

We first study the class of constant predictors when there is no side-information, i.e. $K = 1$. That is, we wish to obtain a sequential predictor that can predict every sequence x^n as well as the best constant predictor for that sequence even when the constant predictor is selected by observing the entire sequence in advance.

Minimizing $\sum_{t=1}^{n} (x[t] - c)^2$ for a specific sequence x^n yields the well-known least squares optimal parameter, $c[n] = \frac{1}{n} \sum_{t=1}^{n} x[t]$ which is a function of the entire sequence. A slightly more general loss function is given by, $\min_c \sum_{t=1}^{n} (x[t] - c)^2 + \delta(c - c_0)^2$, where $\delta \geq 0$ and c_0 are given. Here, δ is used to incorporate the additional a priori knowledge c_0 concerning c in the problem statement. In this paper we will assume $c_0 = 0$ without loss of generality. From this loss function, we derive a universal algorithm by performance weighted combination of all constant predictors using similar ideas to those introduced in [2] yielding,

$$x_u[n] = \frac{1}{n + \delta} \sum_{t=1}^{n-1} x[t].$$

We next relate the performance of this universal algorithm, $l\left(x^n, x_u^n\right) \triangleq \sum_{t=1}^{n} (x[t] - x_u[t])^2$ to the best constant predictor.

Theorem 1: *Let $x[n]$ be a bounded, real-valued arbitrary sequence such that $|x[n]| < A$ for all n. Then, $l\left(x^n, x_u^n\right)$ satisfies,*

$$\frac{1}{n} l\left(x^n, x_u^n\right) \leq \frac{1}{n} \inf_c \sum_{t=1}^{n} (x[t] - c)^2 + \frac{A^2}{n}(1 + \ln(n+1)).$$

Theorem 1 states that the average squared prediction error of the universal predictor is within $O\left(n^{-1} \ln(n)\right)$ of the batch constant prediction algorithm, uniformly, for every individual sequence x^n. The proof of Theorem 1 follows that of Theorem 1 of [2] based on sequential probability assignment and is omitted here for brevity.

The predictor described in Theorem 1 is optimal in that no sequential predictor can do much better in a min-max sense. This is made precise in the following theorem reported by Vovk[1],

Theorem 2: *Let $x[n]$ be a bounded, real-valued arbitrary sequence such that $|x[n]| < A$ for all n. Let $x_a[t]$ be the prediction from any sequential algorithm. Then for any $\epsilon > 0$ there exists a constant G such that*

$$\inf_{a \in S} \sup_{x^n} \frac{1}{n} \left\{ \sum_{t=1}^{n} (x[t] - x_a[t])^2 - \inf_{c \in R} \sum_{t=1}^{n} (x[t] - c)^2 \right\}$$
$$\geq \frac{A^2(1 - \epsilon)}{n} \ln(n) - \frac{G}{n},$$

where S is the class of all sequential predictors.

Theorem 2 states that for any sequential algorithm, there exists a sequence such that the time-average squared prediction error is at least $O(n^{-1} \ln(n))$ worse than the constant predictor tuned for that sequence.

When there is side-information, $K > 1$, we seek to minimize the following regret

$$\sup_{x^n, s^n} \frac{1}{n} \left\{ \sum_{t=1}^{n} (x[t] - x_a[t])^2 - \inf_{c_1, \ldots, c_k \in R} \sum_{t=1}^{n} \left(x[t] - c_{s[t]}\right)^2 \right\}$$

where $x_a[t]$ is the prediction of any sequential algorithm.

From the results of Theorem 1, we derive a state-constant universal algorithm as

$$x_u[t] = \frac{1}{n_{s[t]} + \delta} \sum_{i=1}^{t-1} I\left(s[i] = s[t]\right) x[i],$$

where $n_{s[t]}$ is the number of occurrences of state $s[t]$ in $i = 1, \ldots, t-1$, and $I(.)$ is the indicator function. Multiple application of the Theorem 1 to sequence x^n with side-information sequence s^n yields,

Theorem 3: *Let x^n be a bounded, real-valued arbitrary sequence such that $|x[t]| < A$, with an associated side-information sequence s^n taking values from a finite set $s[t] \in \{1, \ldots, K\}$ for all t. Then,*

$$\sum_{t=1}^{n} (x[t] - x_u[t])^2 - \inf_{c_1, \ldots, c_K \in R} \sum_{t=1}^{n} \left(x[t] - c_{s[t]}\right)^2$$
$$\leq A^2 \sum_{j=1}^{K} (1 + \ln(n_j)) \leq KA^2 \ln(n) + O(1).$$

where n_j is the number of occurrences of state n_j in s^n.

Theorem 3 states that the average squared prediction error of the state-constant universal predictor with side information is within $O(Kn^{-1} \ln(n))$ of the batch state-constant prediction algorithm, uniformly, for every individual sequence x^n and state sequence s^n.

To lower bound the performance of any sequential algorithm with respect to best state-constant predictor, we prove the following theorem,

Theorem 4: *Let x^n be a bounded, real-valued arbitrary sequence such that $|x[t]| < A$, with an associated side-information sequence s^n taking values from a finite set $\{1, \ldots, K\}$ for all t. Let $x_a[t]$ be the prediction from any sequential algorithm. Then for any $\epsilon > 0$ there exists a constant G such that*

$$\inf_{a \in S} \sup_{x^n, s^n} \left\{ \sum_{t=1}^{n} (x[t] - x_a[t])^2 - \inf_{c_i \in R} \sum_{t=1}^{n} \left(x[t] - c_{s[t]} \right)^2 \right\}$$
$$\geq K A^2 (1 - \epsilon) \ln(n) - G,$$

where S is the class of all sequential predictors.

Outline of the Proof of the Theorem 4: For an arbitrary state sequence s^n and for any distribution on x^n,

$$\inf_{a \in S} \sup_{x^n, s^n} \{ l(x^n, x_a^n | s^n) - l^* (x^n, x_c^n | s^n) \}$$
$$\geq \inf_{a \in S} E_{x^n} \{ l(x^n, x_a^n | s^n) - l^* (x^n, x_c^n | s^n) \},$$

where $E_{x^n}(\cdot)$ is an expectation taken with respect to the distribution on x^n. We proceed to apply Theorem 2 repeatedly for sequence values with the same state label.

For any state label $i = 1, \ldots, K$, we consider the following distribution on the values of x^n with $s[t] = i$. Let θ_i be a random variable drawn from a beta distribution with parameters (C_i, C_i), such that

$$p(\theta_i) = \frac{\Gamma(2C_i)}{\Gamma(C_i)\Gamma(C_i)} \theta^{C_i - 1} (1 - \theta_i)^{C_i - 1},$$

where $C_i > 0$ is a constant and $\Gamma(\cdot)$ is the gamma function. For any state i, generate the sequence x^n having only two values, A and $-A$, such that $x[t] = A$ with probability θ_i when $s[t] = i$ and $x[t] = -A$ with probability $(1 - \theta_i)$. The θ_i's are selected independently and given all θ_i's, each $x[t]$ is independent. Then following the lines of Theorem 2, we conclude

$$\inf_{a \in S} \sup_{x^n, s^n} (l(x^n, x_a^n | s^n) - l^* (x^n, x_c^n | s^n))$$
$$\geq \max_{n_1, \ldots, n_K} \sum_{i=1}^{K} A^2 (1 - \epsilon) \ln(n_i) - G$$

where n_i is the number of occurrences of state i in s^n and $\sum_{i=1}^{K} n_i = n$. Maximizing this lower bound with respect to the n_i's give the corresponding lower bound of Theorem 4.

3. LINEAR CONSTANT PREDICTORS WITH SIDE INFORMATION

In this section we extend the previous results for constant predictors to linear predictors. We first report the results corresponding to the class of mth-order linear predictors when there is no side-information, i.e. $K = 1$. Here, we seek to minimize the regret

$$\sup_{x^n} \left\{ \sum_{t=1}^{n} (x[t] - x_a[t])^2 - \inf_{\vec{w} \in R^m} \sum_{t=1}^{n} \left(x[t] - \vec{w}^T \vec{x}[t-1] \right)^2 \right\},$$

where $x_a[t]$ is the prediction at time t of any sequential algorithm, $\vec{w} = [w_1, \ldots, w_m]^T$ and $\vec{x}[t-1] = [x[t-1], \ldots, x[t-m]]^T$. That is, we wish to obtain a sequential predictor that can predict every sequence x^n as well as the best batch mth-order linear predictor tuned for that sequence, even when the linear predictor is selected by observing the entire sequence in advance.

Minimizing the total prediction error, $l(x^n, x_{\vec{w}}^n) = \sum_{t=1}^{n} \left(x[t] - \vec{w}^T \vec{x}[t-1] \right)^2$ over a batch of data of length n yields the well-known least squares solution $\vec{w}[n] = (R_{\vec{x}\vec{x}}^n)^{-1} r_{x\vec{x}}^n$, when $R_{\vec{x}\vec{x}}^n = \sum_{t=1}^{n} \vec{x}[t-1] \vec{x}[t-1]^T$ and $r_{x\vec{x}}^n = \sum_{t=1}^{n} x[k] \vec{x}[t-1]$. In [2] for the loss function, $l(x^n, x_{\vec{w}}^n) + \delta \vec{w}^T \vec{w}$ where $\delta > 0$, a universal predictor $x_u[n]$ was constructed as

$$x_u[n] = \vec{w}_u[n-1]^T \vec{x}[n-1],$$

where $\vec{w}_u[n] = \left[R_{\vec{x}\vec{x}}^{n+1} + \delta I \right]^{-1} r_{x\vec{x}}^n$.

For the performance of this universal predictor, we have:

Theorem 5: *Let x^n be a bounded, but otherwise arbitrary sequence, such that $|x[t]| < A$ for all t. Then the total squared prediction error of the mth-order universal predictor satisfies*

$$\frac{1}{n} \sum_{t=1}^{n} (x[t] - x_u[t])^2 \leq$$
$$\inf_{\vec{w} \in R^m} \frac{1}{n} \left(l(x^n, x_{\vec{w}}^n) + \delta \vec{w}^T \vec{w} \right) + \frac{m A^2}{n} \ln \left(1 + \frac{A^2 n}{\delta} \right).$$

Theorem 5 tells us that the average squared prediction error of the mth-order universal predictor is within $O(m \ln(n)/n)$ of the best batch mth-order linear prediction algorithm, for every individual sequence x^n.

The predictor described in Theorem 5 is optimal in that no sequential predictor can do much better in a min-max sense. This is made precise in the following theorem [2],

Theorem 6: *Let $x[n]$ be a bounded, real-valued arbitrary sequence such that $|x[n]| < A$ for all n. Let $x_a[t]$ be the prediction from any sequential algorithm. Then for any $\epsilon > 0$ there exists a constant G such that*

$$\inf_{a \in S} \sup_{x^n} \frac{1}{n} \left\{ \sum_{t=1}^{n} (x[t] - x_a[t])^2 - \inf_{\vec{w} \in R^m} l(x^n, x_{\vec{w}}^n) \right\}$$
$$\geq \frac{m A^2 (1 - \epsilon)}{n} \ln(n) - \frac{G}{n},$$

where S is the class of all sequential predictors.

When there is side-information, $K > 1$, we seek to minimize

$$\sup_{x^n, s^n} \left\{ \sum_{t=1}^{n} (x[t] - x_a[t])^2 \right.$$
$$\left. - \inf_{\vec{w}_1,\ldots,\vec{w}_K \in R^m} \sum_{t=1}^{n} \left(x[t] - \vec{w}_{s[t-1]}^T \vec{x}[t-1] \right)^2 \right\}$$

where $x_a[t]$ is the prediction of any sequential algorithm.

From the results of Theorem 5, we derive a state-constant universal algorithm depending on the state at time n as

$$x_u[n] = \left[\sum_{k=1}^{n} I\left(s[k] = s[n]\right) \vec{x}[k-1] \vec{x}[k-1]^T + \delta I \right]^{-1}$$
$$\left[\sum_{k=1}^{n-1} I\left(s[k] = s[n]\right) x[k] \vec{x}[k-1] \right] \vec{x}[n-1],$$

Multiple application of Theorem 5 to the sequence x^n with side-information yields,

Theorem 7: *Let x^n be a bounded, real-valued arbitrary sequence such that $|x[t]| < A$, with an associated side-information sequence s^n taking values from finite set, i.e. $s[t] \in 1,\ldots,K$ for all t. Then,*

$$\sum_{t=1}^{n} (x[t] - x_a[t])^2$$
$$- \inf_{\vec{w}_i \in R^m} \left\{ \sum_{t=1}^{n} \left(x[t] - \vec{w}_{s[t-1]}^T \vec{x}[t-1] \right)^2 + \delta ||\vec{w}_{s[t-1]}||^2 \right\}$$
$$\leq \sum_{i=1}^{K} A^2 m \ln\left(1 + \frac{A^2 n_i}{\delta}\right) \leq K m A^2 \ln(n) + O(1),$$

where n_i is the number of occurrence of state i in s^n.

Theorem 6 states that the average squared prediction error of the state-constant universal predictor with side-information is within $O(mKn^{-1}\ln(n))$ of the batch state-constant prediction algorithm, uniformly, for every individual sequence x^n and state sequence s^n. The mK term can be recognized as the number of degrees of freedom in the batch algorithm.

To lower bound the performance of any sequential algorithm with respect to state-constant linear predictors, we prove the following theorem.

Theorem 8: *Let x^n be a bounded, real-valued arbitrary sequence such that $|x[t]| < A$, with an associated side-information sequence s^n taking values from finite set, i.e. $s[t] \in \{1,\ldots,K\}$ for all t. Then for any $\epsilon > 0$ there exists a constant G such that*

$$\inf_{a \in S} \sup_{x^n, s^n} \left\{ \sum_{t=1}^{n} (x[t] - x_a[t])^2 - \inf_{\vec{w}_i \in R^m} \sum_{t=1}^{n} (x[t] \right.$$
$$\left. - \vec{w}_{s[t-1]}^T \vec{x}[t-1] \Big)^2 \right\} \geq \sum_{i=1}^{K} m A^2 (1-\epsilon) \ln(n_i) - G$$
$$\geq m K A^2 \ln(n) - O(1),$$

where $x_a[t]$ is the prediction of any sequential algorithm and n_i is the number of occurrences of state i in s^n.

This lower bound matching the upper bound in Theorem 7 shows that the universal algorithm is also optimal in this min-max sense.

Outline of the Proof of Theorem 8: The proof of Theorem 8 follows along the lines of Theorem 6. For an arbitrary state sequence s^n and for any distribution on x^n,

$$\inf_{a \in S} \sup_{x^n, s^n} \left(l\left(x^n, x_a^n | s^n\right) - l^*\left(x^n, x_{\vec{w}}^n | s^n\right) \right) \geq$$
$$\inf_{a \in S} E_{x^n} \left[l\left(x^n, x_a^n | s^n\right) - l^*\left(x^n, x_{\vec{w}}^n | s^n\right) \right],$$

where $E_{x^n}(\cdot)$ is an expectation taken with respect to the distribution on x^n. Since this is true for all s^n, we select s^n as the concatenation of K repeating states sequences each with n_i repeated entries of state i and with a single transition between consecutive regions, where $\sum_{i=1}^{n} n_i = n$. For each region i, we independently draw m random variables θ_k, $k = 1,\ldots,m$, from a beta distribution. Then for each k the corresponding two state Markov chain is generated where $x[t] = x[t-1]$ with probability θ_k and $x[t] = -x[t-1]$ with probability $(1-\theta_k)$. These m independent Markov chains are interleaved to give x^n in ith region. Hence at any time t the predictor can use only the information coming from $t-m$th sample of the same state i due to the independence of the θ_k's. Hence for each time interval, application of Theorem 6 gives a lower bound of order $O(mA^2 \ln(n_i))$. After maximizing the final lower bound with respect to the n_i's, we obtain the corresponding result completing the proof of Theorem 8.

4. REFERENCES

[1] V. Vovk,“Competitive on-line statistics,” *Int. Statist. Rev.*, vol. 69, pp. 213-248, 2001

[2] A. C. Singer , S. S. Kozat, M. Feder, “Universal linear least squares prediction: upper and lower bounds,” *IEEE Trans. Info. Theory*, vol. 48, no.8, pp. 2354-2362, Aug. 2002

[3] T. M. Cover, E. Ordentlich, “Universal Portfolios with Side Information,” *IEEE Trans. Info. Theo.*, vol. 42, No. 2, pp. 348-363, March 1996

[4] A. C. Singer and M. Feder, “Universal linear prediction by model order weighting,” *IEEE Trans. Signal Proc.*, vol. 47, no. 10, pp. 2685-2700, Oct. 1999.