Correspondence

Universal Linear Least Squares Prediction: Upper and Lower Bounds

Andrew C. Singer, *Member, IEEE*, Suleyman S. Kozat, and Meir Feder, *Fellow, IEEE*

***Abstract*—We consider the problem of sequential linear prediction of real-valued sequences under the square-error loss function. For this problem, a prediction algorithm has been demonstrated [1]–[3] whose accumulated squared prediction error, for every bounded sequence, is asymptotically as small as the best fixed linear predictor for that sequence, taken from the class of all linear predictors of a given order p. The redundancy, or excess prediction error above that of the best predictor for that sequence, is upper-bounded by $A^2 p \ln(n)/n$, where n is the data length and the sequence is assumed to be bounded by some A. In this correspondence, we provide an alternative proof of this result by connecting it with universal probability assignment. We then show that this predictor is optimal in a min–max sense, by deriving a corresponding lower bound, such that no sequential predictor can ever do better than a redundancy of $A^2 p \ln(n)/n$.**

***Index Terms*—Min–max, prediction, sequential probability assignment, universal algorithms.**

I. Introduction

In this correspondence, we consider the problem of predicting a sequence $x^n = \{x[t]\}_{t=1}^n$ as well as the best linear predictor out of a large, continuous class of linear predictors. The real-valued sequence x^n is assumed to be bounded, in that $|x[t]| < A$ for some $A < \infty$, and for all t. Rather than assuming a statistical ensemble of sequences, and attempting to achieve good expected performance, the goal of this game is to predict the sequence as well as the best predictor out of a large class of predictors for every possible sequence x^n. As such, we seek to minimize the following form of regret:

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

where $\hat{x}_a[t]$ is the prediction at time t of a sequential algorithm and $\hat{x}_c[t]$ is the prediction at time t of a predictor in the class $\mathcal{C}$ of predictors.

We first consider the class of first-order linear predictors, such that the competing class of predictors $\mathcal{C} = R$ has elements $w \in R$, which form predictions as $\hat{x}_w[t] = w\,x[t-1]$ for each sample of the sequence x^n. For linear predictors, we assume predictions $\hat{x}_w[1] = 0$, i.e., that $x[t] = 0$, for $t \leq 0$. While this class of predictors is rather limited in forecasting ability, we permit the constant w to be selected based on observing the entire sequence x^n in advance. As we will show, there does not exist a sequential algorithm that can outperform the best predictor from this class for all sequences. In this correspondence, we present an algorithm for which this regret is at most $A^2 \ln(n)$ and also demonstrate that there is a lower bound of $A^2 \ln(n) - G$ for any sequential algorithm, and for some constant G. This algorithm was first shown by Vovk [4], and later by Azoury and Warmuth [2]. Our approach is based on sequential probability assignment, and is motivated by recent results in the universal source coding literature [5]–[11].

We then consider the class of pth-order linear predictors, such that the competing class of predictors $\mathcal{C} = R^p$ has elements $\hat{x}_{\vec{w}}$, $\vec{w} \in R^p$, which form predictions as a linear function of the past p samples, i.e.,

$$\hat{x}_{\vec{w}}[n] = \sum_{k=1}^{p} w_k\, x[n-k].$$

We again permit the parameter vector $\vec{w}$ to be selected based on observing the entire sequence x^n in advance. We will show an algorithm for which the regret in (1) is at most $A^2 p \ln(n)$. We then demonstrate that there exists a corresponding lower bound of the form $A^2 p \ln(n) - G$ for any sequential algorithm.

In [1], Vovk considers the regret in (1) for the problem of linear regression. That is, for

$$\hat{y}[t] = \sum_{k=1}^{p} w_k\, x_k[t]$$

where $y[t]$ and $\vec{x}[t]$ are bounded scalar and vector sequences, respectively. He demonstrates corresponding upper and lower bounds to those obtained in this correspondence for linear prediction, for linear regression. Specifically, for

$$|x_k[t]| < A_x \quad \text{and} \quad |y[t]| < A_y$$

he presents an algorithm for which (1) is upper-bounded by approximately $A_y^2 p \ln(1 + nA_x^2/\delta)$, for some constant δ. He then demonstrates a stochastic construction of sequences $y[t]$, $\vec{x}[t]$ such that (1) is lower-bounded in expectation by approximately $(p-\epsilon)A_y^2 \ln(n) - \delta p A_y^2 - C$, for suitable constants δ and C, and for any ϵ. As discussed later in this correspondence, this stochastic construction implies a form of min–max optimality. While our upper bounds for linear prediction can be derived as corollaries of those obtained in [1], we show that there are a number of important differences between the regression and prediction problems. The requirement that the samples and the labels must satisfy $x_k[t] = y[t-k]$ turns out to be particularly strong. For example, the prediction algorithm presented here will produce bounded predictions for a bounded input, however, the algorithm in [1] will not necessarily produce bounded regressions. Further, the lower bounds for regression in [1] cannot be applied to the linear prediction problem. We therefore build upon the results of [1], and extend them to the specific problem of linear prediction.

II. Scalar Linear Predictors

We begin with the class of scalar linear predictors and seek to minimize the following regret:

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

Manuscript received August 24, 2000; revised March 6, 2002. This work was supported in part by the National Science Foundation under Grants CCR-0092598 (CAREER), CCR 99-79381, and ITR 00-85929, and by the Office of Naval Research under Award N000140110117.

A. C. Singer and S. S. Kozat are with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: acsinger@uiuc.edu; kozat@ifp.uiuc.edu).

M. Feder is with the Department of Electrical Engineering–Systems, Tel-Aviv University, Ramat-Aviv, Tel–Aviv 69978, Israel (e-mail: meir@eng.tau.ac.il).

Communicated by G. Lugosi, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Publisher Item Identifier 10.1109/TIT.2002.800489.

where $\hat{x}_a[t]$ is the prediction at time t of any sequential algorithm. That is, we wish to obtain a sequential predictor that can predict every sequence x^n as well as the best fixed linear predictor for that sequence, even when the linear predictor is selected by observing the entire sequence in advance.

Minimizing $\sum_{t=1}^{n}(x[t] - wx[t-1])^2$ for a specific sequence x^n yields the well-known equation for the least squares optimal parameter

$$w[n] = \frac{\sum_{t=1}^{n} x[t]x[t-1]}{\sum_{t=1}^{n} x[t-1]^2} \tag{2}$$

$$= \frac{R_{xx}^n[-1]}{R_{xx}^{n-1}[0]} \tag{3}$$

where

$$R_{ab}^N[m] = \sum_{n=1}^{N} a[n]b[n+m].$$

Note that $w[n]$ this is a function of the entire sequence, and cannot be obtained until the whole sequence has been observed.

A slightly more general loss function which often arises in many signal processing problems is

$$\min_w \sum_{t=1}^{n} (x[t] - wx[t-1])^2 + \delta(w - w_0)^2$$

where $\delta \geq 0$, and w_0 is given. Choosing $\delta = 0$ yields the original least squares expression. Here, δ is typically used to incorporate additional *a priori* knowledge concerning w [12]. In this correspondence, we will assume that $w_0 = 0$, which could also be obtained through a suitable change of variables. The minimizing value of w for this problem is given by

$$w^*[n] = \frac{\sum_{t=1}^{n} x[t]x[t-1]}{\sum_{t=1}^{n} x[t-1]^2 + \delta} = \frac{R_{xx}^n[-1]}{R_{xx}^{n-1}[0] + \delta}.$$

We next describe a universal prediction algorithm whose accumulated average square error is as small, to within a negligible term, as that of a linear predictor that was preset to the best value given the sequence in advance. We can write

$$\tilde{x}_u[n] = w_u[n-1]x[n-1]$$

where

$$w_u[n] = \frac{R_{xx}^n[-1]}{R_{xx}^n[0] + \delta}$$

and $\delta > 0$ is a constant.

The following theorem, which is proven in the Appendix relates the performance of the universal predictor

$$l(x^n, \tilde{x}_u^n) = \sum_{t=1}^{n} (x[t] - \tilde{x}_u[t])^2$$

to that of the best batch predictor.

Theorem 1: Let x^n be a bounded, real-valued arbitrary sequence, such that $|x[t]| < A$, for all t. Then $l(x^n, \tilde{x}_u^n)$ satisfies

$$\frac{1}{n} l(x^n, \tilde{x}_u^n) \leq \frac{1}{n} \min_w \left\{ l(x^n, \hat{x}_w^n) + \delta w^2 \right\} + \frac{A^2}{n} \ln\left(1 + \frac{nA^2}{\delta}\right).$$

Theorem 1 states that the average squared prediction error of the universal predictor is within $O(n^{-1} \ln(n))$ of the best batch scalar linear prediction algorithm, uniformly, for every individual sequence x^n.

A. Outline of Proof of Theorem 1

The proof of the theorem is based on sequential probability assignment. Given a continuum of predictors, each with a different value of the parameter w, denoted, $\hat{x}_w[t] = wx[t-1]$, then for each of the predictors, a measure of their sequential prediction performance, or loss, is constructed

$$l(x^n, \hat{x}_w^n) = \sum_{t=1}^{n} (x[t] - wx[t-1])^2$$

Also, define a function of the loss, namely the "probability"

$$P_w(x^n) = \exp\left(-\frac{1}{2h} l(x^n, \hat{x}_w^n)\right)$$

which can be viewed as a probability assignment of the predictor with parameter w to the data x^n induced by performance of w on the sequence. We refer to such exponential functions of the loss as probabilities in analogy to problems in sequential data compression. We construct a universal estimate of the probability of the sequence x^n, as an *a priori* weighted combination, or mixture, among all of the probabilities

$$P_u(x^n) = \int_{-\infty}^{\infty} p(w) P_w(x^n)\, dw \tag{4}$$

where $p(w)$ is an *a priori* weighting assigned to the parameter w.

Since the assigned probabilities for the square-error loss are Gaussian in form, the Gaussian prior enables the integration of probabilities assigned to the sequence. We let

$$p(w) = \exp\{-w^2/2\sigma^2\}/(\sqrt{2\pi}\sigma).$$

The universal probability assignment can thus be obtained in closed form.

As shown in the Appendix, this universal probability is as large as the probability assigned to the sequence by the predictor with the smallest prediction error, i.e., the largest probability among the continuum of probabilities $P_w(x^n)$. We now must relate this universal probability to an actual prediction. We note that the universal probability is Gaussian, but not in the form of an assigned probability, i.e., with the loss of a particular predictor in the exponent. As such, we then find *another* Gaussian, expressed in predictor form which is *larger* than the universal probability, for all sequences of interest. Taking the negative logarithm of this probability then provides the loss of this universal predictor and completes the proof of the theorem.

B. Bounded Predictions

One interesting difference between the prediction and regression problems relates to the performance of the universal algorithm on bounded data.

Theorem 2: Let x^n be a bounded, real-valued arbitrary sequence, such that $|x[t]| < A$. Then the predictor $\tilde{x}_u[t]$ also satisfies $|\tilde{x}_u[t]| < A$.

This theorem is proven in the Appendix. We also note that when applied to the regression problem, as in [1], the corresponding universal regression algorithm does not share this property. In the Appendix, we also provide an example of bounded sequences $\vec{x}^n$ and y^n for which the associated universal regression algorithm does not produce bounded regressions.

III. Lower Bound

In this section, we will demonstrate that the predictor described in Theorem 1 is nearly optimal in that no sequential predictor can do much better, in a min–max sense. This is made precise in the following theorem.

Theorem 3: Let x^n be a bounded, real-valued arbitrary sequence such that $|x[t]| < A$ for all t. Let $\hat{x}_a^n$ be the predictions from any

sequential prediction algorithm. Then, for any $\epsilon > 0$, there exists a constant G such that $l(x^n, \hat{x}_a^n)$ satisfies

$$\inf_{a\in\mathcal{A}} \sup_{x^n} \left\{ l(x^n, \hat{x}_a^n) - \inf_{w\in R} l(x^n, \hat{x}_w^n) \right\} \geq \frac{A^2(1-\epsilon)}{n} \ln(n) - \frac{G}{n}$$

where $\mathcal{A}$ is the class of all sequential predictors.

Theorem 3 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 best fixed linear predictor for that sequence.

A. Proof of Theorem 3

We begin by noting that for any distribution on x^n

$$\inf_{a\in\mathcal{A}} \sup_{x^n} \left(l(x^n, \hat{x}_a^n) - \inf_w l(x^n, \hat{x}_w^n) \right) \geq \inf_{a\in\mathcal{A}} E_{x^n} \left(l(x^n, \hat{x}_a^n) - \inf_w l(x^n, \hat{x}_w^n) \right)$$

where $E_{x^n}(\cdot)$ is an expectation taken with respect to the distribution on x^n. Thus, it is enough to lower-bound

$$L(n) \triangleq \inf_{a\in\mathcal{A}} E_{x^n} \left(l(x^n, \hat{x}_a^n) - \inf_w l(x^n, \hat{x}_w^n) \right) \tag{5}$$

to obtain a lower bound on the total regret.

We proceed by considering the following distribution on x^n. Let θ be a random variable drawn from a beta distribution with parameters (C, C), such that

$$p(\theta) = \frac{\Gamma(2C)}{\Gamma(C)\Gamma(C)} \theta^{C-1}(1-\theta)^{C-1}$$

where $C > 0$ is a constant and $\Gamma(\cdot)$ is the gamma function. Generate the sequence x^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 $(1-\theta)$. Thus, given θ, any sequence x^n forms a two-state Markov chain with transition probability $(1-\theta)$. We select $x[t] = A$ and $x[t] = -A$ in the two corresponding states of the Markov chain. Note that given θ, each transition is independent from any other transition in the chain. By assuming that x^n is a segment of a stationary Markov sequence, generated from $-\infty$ to ∞, we avoid any subtleties induced by initialization at $t = 1$.

Given this distribution, we now compute a lower bound for (5). By the linearity of the expectation, (5) becomes

$$L(n) = \inf_{a\in\mathcal{A}} E[l(x^n, \hat{x}_a^n)] - E\left[\inf_w l(x^n, \hat{x}_w^n)\right] \tag{6}$$

where we drop the explicit x^n-dependence of the expectations to simplify notation.

Each term in (6) can now be calculated separately.

B. $\inf_{a\in\mathcal{A}} E[l(x^n, \hat{x}_a^n)]$

For the square-error loss function, $\inf_{a\in\mathcal{A}} E[l(x^n, \hat{x}_a^n)]$ is minimized with the well known minimum mean-squared error (MMSE) predictor, given by [13]

$$\hat{x}_{\mathcal{A}}[t] = E[x[t]|x[t-1], \ldots, x[1]].$$

By expanding the expectation

$$\hat{x}_{\mathcal{A}}[t] = E[E[x[t]|x[t-1], \ldots, x[1], \theta]|x[t-1], \ldots, x[1]].$$

Since the underlying process is a two state Markov chain

$$\hat{x}_{\mathcal{A}}[t] = E[E[x[t]|x[t-1], \theta]|x[t-1], \ldots, x[1]].$$

Given $x[t-1]$ and θ

$$E[x[t]|x[t-1], \theta] = \theta x[t-1] + (1-\theta)(-x[t-1]) = (2\theta-1)x[t-1].$$

Thus,

$$\begin{aligned}\hat{x}_{\mathcal{A}}[t] &= E[(2\theta-1)x[t-1]|x[t-1], \ldots, x[1]] \\ &= x[t-1]E[(2\theta-1)|x[t-1], \ldots, x[1]].\end{aligned} \tag{7}$$

To evaluate $E[\theta|x[t-1], \ldots, x[1]]$, we compute

$$p(\theta|x[t-1], \ldots, x[1]) = \frac{p(x[t-1], \ldots, x[1]|\theta)p(\theta)}{p(x[t-1], \ldots, x[1])}.$$

Given θ, the probability of any sequence x^{t-1} is equal to

$$p(x[t-1], \ldots, x[1]|\theta) = K(1-\theta)^{F_{t-2}}\theta^{t-2-F_{t-2}}$$

where F_{t-2} is the total number of transitions between the two states in a sequence of length $(t-1)$ and K is a constant. Given θ, F_{t-2} is a binomial random variable with parameter $(1-\theta)$ and size $(t-2)$. The constant K is the probability of $x[1]$.

We obtain

$$p\left(x^{t-1}\right) = \int_{\theta=0}^{1} K(1-\theta)^{F_{t-2}+C-1}\theta^{t-2-F_{t-2}+C-1} \frac{\Gamma(2C)}{\Gamma^2(C)} d\theta$$

and

$$p(\theta|x[t-1], \ldots, x[1]) = \frac{(1-\theta)^{F_{t-2}+C-1}\theta^{t-2-F_{t-2}+C-1}}{\int_{\theta=0}^{1}(1-\theta)^{F_{t-2}+C-1}\theta^{t-2-F_{t-2}+C-1} d\theta}.$$

Thus, the conditional expectation is given by,

$$E[\theta|x[t-1], \ldots, x[1]] = \frac{\int_{\theta=0}^{1}(1-\theta)^{F_{t-2}+C-1}\theta^{t-1-F_{t-2}+C-1} d\theta}{\int_{\theta=0}^{1}(1-\theta)^{F_{t-2}+C-1}\theta^{t-2-F_{t-2}+C-1} d\theta}.$$

Due to the well-known properties of the beta distribution, the preceding expectation becomes

$$\begin{aligned}E[\theta|x[t-1], \ldots, x[1]] &= \frac{\frac{\Gamma(F_{t-2}+C)\Gamma(t-1-F_{t-2}+C)}{\Gamma(F_{t-2}+C+t-1-F_{t-2}+C)}}{\frac{\Gamma(F_{t-2}+C)\Gamma(t-2-F_{t-2}+C)}{\Gamma(F_{t-2}+C+t-2-F_{t-2}+C)}} \\ &= \frac{t-2-F_{t-2}+C}{t-2+2C}.\end{aligned}$$

By this result, the MMSE prediction (7) is given by

$$\begin{aligned}\hat{x}_{\mathcal{A}}[t] &= \left(2\left(\frac{t-2-F_{t-2}+C}{t-2+2C}\right)-1\right)x[t-1] \\ &= \frac{t-2-2F_{t-2}}{t-2+2C}\, x[t-1].\end{aligned}$$

Thus, for the first term in the lower bound in (5), we have

$$\begin{aligned}\inf_{a\in\mathcal{A}} E[l(x^n, \hat{x}_a^n)] &= E\left[\sum_{t=1}^{n}(x[t]-\hat{x}_{\mathcal{A}}[t])^2\right], \\ &= E\left[\sum_{t=1}^{n}\left(x[t]-\left(\frac{t-2-2F_{t-2}}{t-2+2C}\, x[t-1]\right)\right)^2\right].\end{aligned}$$

This expectation can then be expanded and is evaluated in the following lemma.

Lemma 1:

$$\inf_{a\in\mathcal{A}} E[l(x^n, \hat{x}_a^n)] = \sum_{t=1}^{n}\left(A^2 - 2\,\frac{t-2}{(2C+1)(t-2+2C)}\,A^2 + \frac{A^2}{(t-2+2C)^2}\left(\frac{(t-2)^2}{2C+1} + \frac{2C}{2C+1}(t-2)\right)\right).$$

Proof: Given in the Appendix.

C. $E[\inf_w l(x^n, \hat{x}_w^n)]$

For the second term in (6), we need to calculate the following expectation:

$$E\left[\inf_w l(x^n, \hat{x}_w^n)\right] = E\left[\inf_w \sum_{t=1}^{n}(x[t]-\hat{x}_w[t])^2\right].$$

The variable $\hat{x}_w[t]$ is the best (in terms of square-error loss) first-order linear predictor which has access to the whole sequence x^n. With square-error loss, this is a well-known least squares problem. The desired predictor is given by

$$\hat{x}_w[t] = \frac{\sum_{t=1}^{n} x[t]x[t-1]}{\sum_{t=1}^{n-1} x[t]x[t]}\, x[t-1]$$

which is highly nonlinear and for which it is hard to calculate an expectation for a general sequence x^n. Nevertheless, with the selection of our special distribution, the corresponding terms become

$$\sum_{t=1}^{n-1} x[t]x[t] = (n-1)A^2$$

and

$$\sum_{t=1}^{n} x[t]x[t-1] = (n-2F_n)A^2.$$

As before, F_n is the number of transitions between the two states in a sequence of size n. This yields a simple form for the predictor

$$\hat{x}_w[t] = \frac{n-2F_n}{n-1}\, x[t-1]$$

which enables evaluation of the second term in (6) as described in the following lemma.

Lemma 2:

$$E\left[\inf_w\, l(x^n,\, \hat{x}_w^n)\right] = \sum_{t=1}^{n}\left(A^2 - 2\left(\frac{A^2}{2C+1} + \frac{A^2}{n-1}\right) + \frac{A^2}{(n-1)^2}\left(\frac{n^2}{2C+1} + \frac{2Cn}{2C+1}\right)\right).$$

Proof: Given in the Appendix.

Thus, using *Lemmas 1* and *2*, the overall lower bound $L(n)$ can be computed and is given by the following lemma.

Lemma 3:

$$L(n) = A^2 \sum_{t=1}^{n}\left\{\frac{2C}{2C+1}\,\frac{1}{t-2+2C}\right\} + O(1).$$

Proof: Given in the Appendix.

By lower-bounding the harmonic series with its integral, and setting $G = \ln(2C-1)$, then for any ϵ, we can find a constant C such that

$$\inf_{a\in\mathcal{A}} \sup_{x^n}\left\{\sum_{t=1}^{n}(x[t]-\hat{x}_{\mathcal{A}}[t])^2 - \inf_{w\in R}\sum_{t=1}^{n}(x[t]-wx[t-1])^2\right\} \geq A^2(1-\epsilon)\,\ln(n) - G \quad (8)$$

completing the proof.

IV. pth-Order Linear Prediction

In this section, we consider the problem of linear prediction with a predictor of fixed-order p. The predictor is now parameterized by the vector $\vec{w} = [w_1, \ldots, w_p]^T$, and the predicted value can be written $\hat{x}_{\vec{w}}[n] = \vec{w}^T\vec{x}[n]$, where, $\vec{x}[n] = [x[n-1], \ldots, x[n-p]]^T$. If the parameter vector $\vec{w}$ 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

$$\vec{w}_n = \arg\min_{\vec{w}} \sum_{k=1}^{n}\left(x[k] - \vec{w}^T\vec{x}[k]\right)^2.$$

The well-known least squares solution to this problem is given by $\vec{w}_n = (R^n_{\vec{x}\vec{x}})^{-1} r^n_{x\vec{x}}$, when $R^n_{\vec{x}\vec{x}} = \sum_{k=1}^{n}\vec{x}[k]\vec{x}[k]^T$ is invertible and where $r^n_{x\vec{x}} = \sum_{k=1}^{n} x[k]\vec{x}[k]$. When $R^n_{\vec{x}\vec{x}} = \sum_{k=1}^{n}\vec{x}[k]\vec{x}[k]^T$ is singular, the solution is no longer unique, however a suitable choice can often be made using, e.g., pseudoinverses.

We will also consider the more general least squares (ridge regression) problem

$$\vec{w}^*[n] = \arg\min_{\vec{w}}\left\{l(x^n,\, \hat{x}^n_{\vec{w}}) + \delta|\vec{w}|^2\right\} = [R^n_{\vec{x}\vec{x}} + \delta I]^{-1}\, r^n_{x\vec{x}},$$

where $l(x^n,\, \hat{x}^n_{\vec{w}})$ is the running total squared prediction error for a linear predictor with coefficients $\vec{w}$.

We now construct a universal pth-order linear predictor using a mixture over all predictors $\vec{w}$. The following theorem extends Theorem 1 using a vector version of the mixture approach.

Let $\hat{x}_w[n]$ be the output of a pth-order linear predictor with parameter vector $\vec{w}$, and $l(x^n,\, \hat{x}^n_{\vec{w}})$ be the running total squared prediction error. Define a universal predictor $\tilde{x}_u[n]$ as

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

where

$$\vec{w}_u[n] = \left[R^{n+1}_{\vec{x}\vec{x}} + \delta I\right]^{-1} r^n_{x\vec{x}}$$

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

Theorem 4: 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 pth-order universal predictor satisfies

$$l(x^n,\, \tilde{x}^n_u) \leq \min_{\vec{w}}\left\{l(x^n,\, \hat{x}^n_{\vec{w}}) + \delta\|\vec{w}\|^2\right\} + A^2\,\ln\left|I + R^n_{\vec{x}\vec{x}}\delta^{-1}\right|$$

and therefore,

$$\frac{1}{n}\, l(x^n, \tilde{x}^n_u) \leq \min_{\vec{w}}\frac{1}{n}\left\{l(x^n, \hat{x}^n_{\vec{w}}) + \delta\|\vec{w}\|^2\right\} + \frac{A^2 p}{n}\,\ln\left(1 + \frac{A^2 n}{\delta}\right).$$

Theorem 4 tells us that the average squared prediction error of the pth-order universal predictor is within $O(p\,\ln(n)/n)$ of the best batch pth-order linear prediction algorithm, for every individual sequence $x[n]$. This result can be compared with Foster's result for binary data and predictors in the simplex $\sum_i a_i = 1$, yielding regret of $[2 + p\log(p(n+1))]/n$ [14]. For $\delta = 2$, our bound yields $(2\|a\|^2 + A_x^2 p\log(1 + A_x^2 n/2))/n$, which for $A_x = \frac{1}{2}$, i.e., data on an interval of length 1, yields, $(2\|a\|^2 + (p/4)\log(1+n/8))/n$. The proof of Theorem 4 follows that of Theorem 1, with vector extensions of the Gaussian mixture and is omitted for brevity.

A. Lower Bound for pth-Order Linear Prediction

The lower bound obtained for first-order linear prediction can be generalized to the pth-order linear prediction case as described in the following theorem.

Theorem 5: Let x^n be a bounded, real-valued arbitrary sequence such that $|x[t]| < A$ for all t. Let $\hat{x}_a[n]$ be the predictions from any sequential prediction algorithm. Then for any $\epsilon > 0$ there exists a constant G such that $l(x^n,\, \hat{x}^n_a)$ satisfies

$$\inf_{a\in\mathcal{A}} \sup_{x^n}\left\{l(x^n, \hat{x}^n_a) - \inf_{\vec{w}\in R^p} l(x^n, \hat{x}^n_{\vec{w}})\right\} \geq \frac{A^2 p(1-\epsilon)}{n}\,\ln(n) - \frac{G}{n}$$

where $\mathcal{A}$ is the class of all sequential predictors.

We again focus on the lower bound

$$L(n) \triangleq \inf_{a\in\mathcal{A}} E_{x^n}\left[l(x^n,\, \hat{x}^n_a) - \inf_{\vec{w}\in R^p} l(x^n,\, \hat{x}^n_{\vec{w}})\right] \quad (9)$$

to get a lower bound on the total regret.

We consider the following distribution on x^n, which is constructed by interleaving p first-order Markov sequences. First, independently

draw p random variables θ_i, $i = 1, \ldots, p$, from a beta distribution. For each θ_i, the corresponding two-state Markov chains are interleaved, to create the sequence $x[n]$. Thus, for any n, $x[n]$ and $x[n-p]$ are from the same original two-state Markov chain.

With the expectation taken over this distribution, we can proceed to calculate the lower bound. Since each Markov chain is independent, the derivations follow the first-order case. The MMSE prediction is given by

$$\hat{x}_{\mathcal{A}}[t] = \frac{t^* - 2 - 2F_{t^*-2}}{t^* - 2 + 2C}\, x[t-p]$$

where t^* is the largest integer satisfying $t^* \le (t/p)$. With this, the first sum in the lower bound becomes

$$E\left(\sum_{t=1}^{n}(x[t]-\hat{x}_{\mathcal{A}}[t])^2\right) = \sum_{t=1}^{n}\left(A^2 - 2\,\frac{t^*-2}{(2C+1)(t^*-2+2C)}\,A^2 + \frac{A^2}{(t^*-2+2C)^2}\left(\frac{(t^*-2)^2}{2C+1} + \frac{2C}{2C+1}\,(t^*-2)\right)\right).$$

For the second term in (9), we need to calculate

$$E\left[\inf_{\vec{w}}\, l(x^n,\, \hat{x}^n_{\vec{w}})\right] = E\left[\inf_{\vec{w}}\, \sum_{t=1}^{n}(x[t]-\hat{x}_{\vec{w}}[t])^2\right].$$

The sequence $\hat{x}^n_{\vec{w}}$ is the best set (in terms of square error) of pth-order linear predictions which has access to the whole sequence x^n. With square error loss, this predictor is the well-known least squares predictor. However, the expected loss for this predictor is difficult to compute, even for our distribution. The following inequality will prove useful in this regard:

$$E_\theta\left[E_{x^n|\theta}\left[\inf_{\vec{w}}\, \sum_{t=1}^{n}(x[t]-\hat{x}_{\vec{w}}[t])^2\right]\right] \le E_\theta\left[\inf_{\vec{w}}\, E_{x^n|\theta}\left[\sum_{t=1}^{n}(x[t]-\hat{x}_{\vec{w}}[t])^2\right]\right]$$

where $E_{x^n|\theta}$ is the conditional expectation conditioned on all p values of θ. Therefore, the lower bound in (9) is lower-bounded by

$$L(n) \ge \inf_{a\in\mathcal{A}}\, E_{x^n}[l(x^n,\, \hat{x}^n_a)] - E_\theta\left[\inf_{\vec{w}}\, E_{x^n|\theta}[l(x^n,\, \hat{x}^n_{\vec{w}})]\right]. \quad (10)$$

The term $\inf_{\vec{w}} E_{x^n|\theta}[L(x^n,\, x^n_{\vec{w}})]$ is the MMSE and given by [13]

$$\inf_{\vec{w}}\, E_{x^n|\theta}[l(x^n,\, x^n_{\vec{w}})] = \sigma^2 - \underline{k}^T R^{-1}\underline{k},$$

where $\sigma^2 = A^2$ is the variance of the sequence given all θ_i, $\underline{k} = E[x[t]\vec{x}[t-1]|\theta]$ is the cross-correlation vector, and $R = E[\vec{x}[t-1]\vec{x}^T[t-1]|\theta]$ is the correlation matrix. Since the interleaved Markov chains are independent

$$\underline{k} = [0, \ldots, 0,\, E[x[t]x[t-p]|\theta]]^T$$
$$\underline{k} = [0, \ldots, 0,\, (2\theta_p - 1)A^2]^T]$$

and $R = A^2 I$ where I is a p-dimensional identity matrix, where $\theta_p = \theta_n \bmod p$. This results in

$$\inf_{\vec{w}}\, E_{x^n|\theta}[l(x^n,\, x^n_{\vec{w}})] = A^2 - (2\theta_p - 1)^2 A^2.$$

The second term in (10) yields

$$E_\theta\left[\inf_{\vec{w}}\, E_{x^n|\theta}[l(x^n,\, x^n_{\vec{w}})]\right] = A^2 - \frac{1}{2C+1}\,A^2.$$

Combining this result with those for scalar prediction, we obtain the lower bound as

$$\begin{aligned} L(n) &\ge \sum_{t=1}^{n}\left\{\left(A^2 - 2\,\frac{t^*-2}{(2C+1)(t^*-2+2C)}\,A^2 + \frac{A^2}{(t^*-2+2C)^2}\left(\frac{(t^*-2)^2}{2C+1} + \frac{2C}{2C+1}\,(t^*-2)\right)\right) - \left(A^2 - \frac{1}{2C+1}\,A^2\right)\right\}, \\ &= A^2\sum_{t=1}^{n}\left\{1 - 2\,\frac{(t^*-2+2C)}{(2C+1)(t^*-2+2C)^2} + \frac{4C}{(2C+1)(t^*-2+2C)} + \frac{(t^*-2+2C)^2}{(2C+1)(t^*-2+2C)^2} - \frac{4C(t^*-2+2C)}{(2C+1)(t^*-2+2C)^2} + \frac{4C^2}{(2C+1)(t^*-2+2C)^2} + \frac{2C(t^*-2-2C)}{(2C+1)(t^*-2+2C)^2} + \frac{4C^2}{(2C+1)(t^*-2+2C)^2} - 1 + \frac{1}{2C+1}\right\}, \\ &= A^2\sum_{t=1}^{n}\left\{\frac{2C}{2C+1}\,\frac{1}{t^*-2+2C}\right\} + O(1). \end{aligned}$$

Thus, after replacing t^* with its definition, we conclude that for any given ϵ, there exists a constant G such that

$$\inf_{a\in A}\sup_{x^n}\left\{\sum_{t=1}^{n}(x[t]-\hat{x}_a[t])^2 - \inf_{\vec{w}\in R^p}\sum_{t=1}^{n}(x[t]-\vec{w}^T\vec{x}[t-1])^2\right\} \ge A^2(1-\epsilon)p\,\ln(n) - G$$

completing the proof of the theorem.

APPENDIX

A. Proof of Theorem 1

The universal probability assignment can be obtained in closed form; integrating (4)

$$P_u(x^n) = \frac{1}{\sqrt{\delta^{-1}R_{xx}^{n-1}[0]+1}} \cdot \exp\left\{-\frac{1}{2h}\left(\frac{R^n_{xx}[0]R^{n-1}_{xx}[0] + \delta R^n_{xx}[0] - (R^n_{xx}[-1])^2}{R^{n-1}_{xx}[0]+\delta}\right)\right\} \quad (11)$$

where $\delta = h/\sigma^2$.

We would like to have the universal probability be as large as the probability assigned to the sequence by the predictor with

$$w^*[n] = \arg\min_w \left\{l(x^n,\, \hat{x}^n_w) + \delta w^2\right\}.$$

For this value of $w = w^*[n]$, after comparing with (4) and after some algebra, we obtain

$$-2h\,\ln(P_u(x^n)) = \min_w\left\{l(x^n,\, \hat{x}^n_w) + \delta w^2\right\} + h\,\ln\left(1 + R^{n-1}_{xx}[0]\delta^{-1}\right). \quad (12)$$

We now have a method of assigning a universal probability to the sequence that achieves, to first order in the exponent, the same sequential probability as the best predictor. We now must relate this universal probability to an actual prediction.

Since each of the predictors assigns a probability that is exponential in the prediction error for that predictor, we look to the exponent of $P_u(x^n)$ for the predictor. Specifically, we have

$$P_w\left(x_n\middle|x^{n-1}\right) = \exp\left\{-\frac{1}{2h}\left(x[n] - w\,x[n-1]\right)^2\right\}$$

relating the prediction error at time n to the probability $P_w(x_n|x^{n-1})$. Similarly, we expect to obtain an expression of this form for $P_u(x_n|x^{n-1})$. From (11), we obtain

$$P_u\left(x_n\middle|x^{n-1}\right) = \sqrt{\frac{R_{xx}^{n-2}[0]+\delta}{R_{xx}^{n-1}[0]+\delta}}\,\exp\left\{\frac{-1}{2h}\left(\frac{R_{xx}^{n-2}[0]+\delta}{R_{xx}^{n-1}[0]+\delta}\right)\cdot\left(x[n] - \frac{R_{xx}^{n-1}[-1]}{R_{xx}^{n-2}[0]+\delta}\,x[n-1]\right)^2\right\}.$$

Although Gaussian (quadratic exponential), $P_u(x_n|x^{n-1})$ cannot be expressed in the same form as $P_w(x_n|x^{n-1})$, i.e., quadratic exponential in the loss at time n.

However, after some algebra, we see that it is almost in this form

$$\begin{aligned}P_u\left(x_n\middle|x^{n-1}\right) &= \sqrt{\frac{R_{xx}^{n-2}[0]+\delta}{R_{xx}^{n-1}[0]+\delta}}\,\exp\left\{\frac{-1}{2h}\left(\frac{R_{xx}^{n-2}[0]+\delta}{R_{xx}^{n-1}[0]+\delta}\right)\cdot\left(x[n] - \frac{R_{xx}^{n-1}[-1]}{R_{xx}^{n-2}[0]+\delta}\,x[n-1]\right)^2\right\}\\ &= \alpha\exp\left\{\frac{-1}{2h}\,\alpha^2\left(x[n] - \frac{R_{xx}^{n-1}[-1]}{R_{xx}^{n-2}[0]+\delta}\,x[n-1]\right)^2\right\}\\ &= \alpha\exp\left\{\frac{-1}{2h}\,\alpha^2\left(x[n] - w^*[n-1]x[n-1]\right)^2\right\}\end{aligned}$$

where

$$\alpha = \sqrt{(R_{xx}^{n-2}[0]+\delta)/(R_{xx}^{n-1}[0]+\delta)}.$$

If we could find *another* Gaussian, which were expressed in the form

$$\begin{aligned}\tilde{P}_u\left(x_n\middle|x^{n-1}\right) &= \exp\left\{\frac{-1}{2h}\left(x[n] - \tilde{x}_u[n]\right)^2\right\}\\ &\geq \alpha\exp\left\{\frac{-1}{2h}\,\alpha^2\left(x[n] - w^*[n-1]x[n-1]\right)^2\right\}\end{aligned}$$

for the sequences of interest, i.e., for $|x[n]| \leq A$, then we would have

$$l(x^n, \tilde{x}_u^n) \leq -2h\,\ln P_u(x^n)$$

completing the proof of the theorem.

Comparing $\tilde{P}_u(x_n|x^{n-1})$ and $P_u(x_n|x^{n-1})$, we obtain

$$\begin{aligned}P_u\left(x_n\middle|x^{n-1}\right) &= \alpha\exp\left\{-\frac{1}{2h}\,\alpha^2\left(x[n] - \hat{x}_u[n]\right)^2\right\}\\ &\leq \exp\left\{-\frac{1}{2h}\left(x[n] - \tilde{x}_u[n]\right)^2\right\}\end{aligned}$$

for $\hat{x}_u[n] = w^*[n-1]x[n-1]$, and for some $\tilde{x}_u[n]$. Note that these are two Gaussians, with different means and different variances. We would like to select an appropriate mean for $\tilde{P}_u(x_n|x^{n-1})$, i.e., $\tilde{x}_u[n]$, such that over the range $x[n] \in [-A, A]$, $\tilde{P}_u(x_n|x^{n-1})$ is *larger* than $P_u(x_n|x^{n-1})$. This would ensure that the loss of the predictor $\tilde{x}_u[n]$ satisfies Theorem 1.

The locations of the crossover points for the two Gaussians, x_l and x_r, where $\tilde{P}_u = P_u$, are obtained by setting $\tilde{x}_u[n] = \gamma\hat{x}_u[n]$ and solving, yielding the equation shown at the bottom of the page. Note that the size of the region $[x_l, x_r]$ over which $\tilde{P}_u \geq P_u$ grows with increasing h. We would like to select h as small as possible, since it appears as a constant multiplier of the redundancy, or excess prediction error of the universal predictor. Since we require that $\tilde{P}_u \geq P_u$ for all $x[n] \in [-A, A]$, the smallest value of h can be selected only when the region $[x_l, x_r]$ is centered about $x[n] = 0$. This can be achieved only by the choice $\gamma = \alpha^2$. For this choice of γ, we have

$$\tilde{P}_u\left(x_n\middle|x^{n-1}\right) = \exp\left\{\frac{-1}{2h}\left(x[n] - \tilde{x}_u[n]\right)^2\right\}$$

where the prediction is given by

$$\begin{aligned}\tilde{x}_u[n] &= \alpha^2 w^*[n-1]x[n-1]\\ &= \frac{R_{xx}^{n-1}[-1]}{R_{xx}^{n-1}[0]+\delta}\,x[n-1].\end{aligned}$$

Note that $\tilde{x}_u[n]$ can be viewed as using $w^*[n]x[n-1]$ where we assume that $x[n] = 0$ to update $R_{xx}^n[-1]$ (which remains at $R_{xx}^{n-1}[-1]$) and $R_{xx}^{n-1}[0]$ accordingly before computing $w^*[n]$.

Now we can select the *smallest* value of h so that the region $[-A, A] \subseteq [x_l, x_r]$, i.e.,

$$\begin{aligned}A &\leq \frac{\sqrt{2h\,\ln(\alpha)(\alpha^2-1) + \alpha^2\hat{x}_u[n]^2(1-\alpha^2)}}{(1-\alpha^2)}\\ h &\geq \frac{A^2(1-\alpha^2) - \alpha^2\hat{x}_u[n]^2}{-2\,\ln(\alpha)}\end{aligned}$$

which must hold for all values of $\hat{x}_u[n] \in [-A, A]$. Therefore,

$$h \geq A^2\,\frac{(1-\alpha^2)}{-2\,\ln(\alpha)}$$

where $\alpha < 1$. Note that for $0 < \alpha < 1$ the function

$$0 < \frac{(1-\alpha^2)}{-2\,\ln\alpha} < 1$$

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 α and $\hat{x}_u[n]$, and is only tight for $\alpha \to 1$, 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, and Theorem 1 will remain valid.

Our "probability" assignment algorithm had two free constants to be set, h and σ^2. Now that we have selected a range for the constant h, the constant σ^2 can be chosen such that $\delta = h/\sigma^2$ is an arbitrary positive constant which does not depend on knowing A in advance. It is worth noting that while the parameter h appears in the upper bound on the excess prediction error (via (12)), it is independent of the algorithm. As a result, while the excess prediction error will depend on the smallest

$$x[n] = \frac{\hat{x}_u[n](\alpha^2-\gamma)}{(\alpha^2-1)} \pm \frac{\sqrt{-2\hat{x}_u[n]^2\gamma\alpha^2 + 2\alpha^2\,\ln(\alpha)h + \alpha^2\hat{x}_u[n]^2\gamma^2 - 2\,\ln(\alpha)h + \alpha^2\hat{x}_u[n]^2}}{(\alpha^2-1)}.$$

value of h for which the theorem holds, this value of h, and therefore A, need not be known in advance.

This leads to the following result:

$$l(x^n, \tilde{x}_u^n) \le \min_w \left\{ l(x^n, \hat{x}_w^n) + \delta |w|^2 \right\} + A^2 \ln\left(1 + R_{xx}^{n-1}[0]\delta^{-1}\right)$$

or

$$\frac{1}{n} l(x^n, \tilde{x}_u^n) \le \frac{1}{n} \min_w \left\{ l(x^n, \hat{x}_w^n) + \delta |w|^2 \right\} + \frac{A^2}{n} \ln\left(1 + \frac{nA^2}{\delta}\right)$$

implying that the universal predictor performs as well as the best parameter w to within a parameter redundancy term of $O(\ln(n)/n)$.

B. Proof of Theorem 2

This can be shown in the scalar case using the Schwarz inequality

$$\tilde{x}_u[n] = \frac{\sum_{k=1}^{n-1} x[k]x[k-1]}{\sum_{k=1}^{n-1} x[k]x[k] + \delta} x[n-1]$$

$$|\tilde{x}_u[n]| \le \left| \frac{\sum_{k=1}^{n-1} x[k]x[k-1]}{\sum_{k=1}^{n-1} x[k]x[k]} x[n-1] \right|$$

$$\le A \left| \frac{\sum_{k=1}^{n-1} x[k]x[k-1]}{\sum_{k=1}^{n-1} x[k]x[k]} \right|.$$

Using the vector notation, $x_1^n = [x[1]x[2]\cdots x[n]]^T$, we have

$$|\tilde{x}_u[n]| \le A \frac{\left| \left[\left(x_2^{n-1}\right)^T 0 \right] \left(x_1^{n-1}\right) \right|}{\left| \left(x_1^{n-1}\right)^T \left(x_1^{n-1}\right) \right|}$$

$$\le A \frac{\left\| \begin{bmatrix} x_2^{n-1} \\ 0 \end{bmatrix} \right\| \left\| x_1^{n-1} \right\|}{\left\| x_1^{n-1} \right\|^2}$$

$$\le A$$

completing the proof.

For the regression problem, the corresponding universal regression $\hat{y}_u[n]$ would be given by

$$\tilde{y}_u[n] = \frac{\sum_{k=1}^{n-1} x[k]y[k]}{\sum_{k=1}^{n} x[k]x[k] + \delta} x[n].$$

For a sequences

$$x_1^{n-1} = [\sqrt{\delta}, \ldots, \sqrt{\delta}, A]^T \quad \text{and} \quad y_1^n = [A, \ldots, A]^T$$

this yields,

$$\tilde{y}_u[n] = \frac{(n-1)\sqrt{\delta}A^2}{(n-1)\delta + A^2 + \delta}$$

which, for $n > (A^2 + \sqrt{\delta}A)/(\sqrt{\delta}A - \delta)$, yields, $|\tilde{y}_u[n]| > A$, i.e., a prediction outside the range $[-A, A]$. Note that for this sequence, $\tilde{y}_u[n] \to A^2/\sqrt{\delta}$, which can be made arbitrarily large by adjusting δ.

C. Proof of Lemma 1

We calculate each term of the quadratic sum separately. First, the cross term

$$E\left[x[t]\left(\frac{t-2-2F_{t-2}}{t-2+2C}\right)x[t-1]\right]$$
$$= \frac{t-2}{t-2+2C} E[x[t]x[t-1]] - \frac{2}{t-2+2C} E[F_{t-2}x[t]x[t-1]]$$
$$= -\frac{2}{t-2+2C} E[F_{t-2}x[t]x[t-1]]$$

where the second line follows from

$$E[x[t]x[t-1]] = E[E[\theta A^2 + (1-\theta)(-A^2)|\theta]]$$
$$= E[(2\theta - 1)A^2] = 0. \tag{13}$$

Again, using conditional expectations, we obtain

$$E[F_{t-2}x[t]x[t-1]]$$
$$= E[E[E[F_{t-2}x[t]x[t-1]|\theta, x[t]x[t-1]]|x[t], x[t-1]]].$$

Given θ, $x[t]$, and $x[t-1]$, F_{t-2} is still a binomial random variable with parameters $(1-\theta)$ and size $(t-2)$. Thus,

$$E[F_{t-2}|\theta, x[t], x[t-1]] = (t-2)(1-\theta).$$

By this

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

where the third line follows from

$$E[\theta x[t]x[t-1]] = E[E[\theta x[t]x[t-1]|\theta]]$$
$$= E[\theta E[x[t]x[t-1]|\theta]]$$
$$= E[\theta(2\theta-1)A^2] = \frac{A^2}{2(2C+1)}$$

and $E[x[t]x[t-1]] = 0$.

For evaluating the square term, we expand the term

$$E\left[\left(\frac{t-2-2F_{t-2}}{t-2+2C} x[t-1]\right)^2\right]$$
$$= \frac{A^2}{(t-2+2C)^2} E\left[(t-2)^2 - 4(t-2)F_{t-2} + 4F_{t-2}^2\right].$$

Since, $E[F_{t-2}] = E[E[F_{t-2}|\theta]] = E[(1-\theta)(t-2)] = (t-2)/2$, only $E[F_{t-2}^2]$ is unknown. Since

$$\mathrm{Var}(F_{t-2}|\theta) = E[F_{t-2}^2|\theta] - (E[F_{t-2}|\theta])^2$$

and noting the variance of a binomial-distributed random variable, we have

$$(t-2)\left((1-\theta) - (1-\theta)^2\right) = E[F_{t-2}^2|\theta] - (t-2)^2(1-\theta)^2.$$

Taking the expectation of both sides, and rearranging the terms

$$E[F_{t-2}^2] = E\left[(t-2)\left((1-\theta)-(1-\theta)^2\right)+(t-2)^2(1-\theta)^2\right] = \frac{C(t-2)}{2(2C+1)}+\frac{(C+1)(t-2)^2}{2(2C+1)}.$$

This yields

$$E\left[\left(\frac{t-2-2F_{t-2}}{t-2+2C}\,x[t-1]\right)^2\right] = \frac{A^2}{(t-2+2C)^2}\left((t-2)^2-4\,\frac{(t-2)^2}{2}+4\,\frac{(t-2)C}{2(2C+1)}+4\,\frac{(C+1)(t-2)^2}{2(2C+1)}\right), = \frac{A^2}{(t-2+2C)^2}\left(\frac{(t-2)^2}{2C+1}+\frac{2C}{2C+1}\,(t-2)\right).$$

Thus, the first sum in (5) of Lemma 1 becomes

$$\sum_{t=1}^{n}(x[t]-\hat{x}_{\mathcal{A}}[t])^2 = \sum_{t=1}^{n}\left(A^2-2\,\frac{t-2}{(2C+1)(t-2+2C)}\,A^2+\frac{A^2}{(t-2+2C)^2}\left(\frac{(t-2)^2}{2C+1}+\frac{2C}{2C+1}\,(t-2)\right)\right).$$

D. Proof of Lemma 2

After expanding the sum, we obtain

$$E\left[\left(x[t]-\frac{n-2F_n}{n-1}\,x[t-1]\right)^2\right] = E\left[x^2[t]-2x[t]\,\frac{n-2F_n}{n-1}\,x[t-1]+\left(\frac{n-2F_n}{n-1}\,x[t-1]\right)^2\right].$$

For calculation of the cross term, we obtain

$$E\left[\frac{n-2F_n}{n-1}\,x[t-1]x[t]\right] = -\frac{2}{n-1}\,E[F_n x[t]x[t-1]]] = -\frac{2}{n-1}\,E[E[E[F_n x[t]x[t-1]|\theta, x[t]x[t-1]]|x[t], x[t-1]]] = -\frac{2}{n-1}\,E[E[x[t]x[t-1]E[F_n|\theta, x[t]x[t-1]]|x[t], x[t-1]]].$$

The inner expectation is given by

$$E[F_n|\theta,\, x[t]x[t-1]] = (n-1)(1-\theta)+\frac{|x[t]-x[t-1]|}{2A}.$$

This equality follows, since, given $x[t]$ and $x[t-1]$, there remains only $n-1$ possible transitions between each sample and there is also an additional transition if $x[t]\neq x[t-1]$. The cross term yields

$$E\left[\frac{n-2F_n}{n-1}\,x[t-1]x[t]\right] = -\frac{2}{n-1}\,E\left[x[t]x[t-1]\left((n-1)(1-\theta)+\frac{|x[t]-x[t-1]|}{2A}\right)\right] = -\frac{2}{n-1}\left[E[x[t]x[t-1](1-\theta)(n-1)]+E\left[x[t]x[t-1]\,\frac{|x[t]-x[t-1]|}{2A}\right]\right].$$

Since these equations are analogous with the ones carried out for the calculation of $\inf_{a\in\mathcal{A}} E[l(x^n,\,\hat{x}_a^n)]$, the derivations will follow along the same lines. From (13)

$$E[x[t]x[t-1](1-\theta)] = -\frac{A^2}{2(2C+1)}$$

and

$$E\left[x[t]x[t-1]\,\frac{|x[t]-x[t-1]|}{2A}\right] = E[(1-\theta)(-A^2)] = -A^2/2.$$

Thus, we obtain

$$E\left[\frac{n-2F_n}{n-1}\,x[t-1]x[t]\right] = -\frac{2}{n-1}\left(-\frac{A^2}{2(2C+1)}\,(n-1)-\frac{A^2}{2}\right) = \frac{A^2}{2C+1}+\frac{A^2}{n-1}.$$

The square term is given by

$$E\left[\left(\frac{n-2F_n}{n-1}\,x[t-1]\right)^2\right] = \frac{A^2}{(n-1)^2}\,E[n^2-4nF_n+4F_n^2].$$

As before, $E[F_n] = n/2$ and $E[F_n^2] = \frac{Cn}{2(2C+1)}+\frac{(C+1)n^2}{2(2C+1)}$. The square term then becomes

$$E\left[\left(\frac{n-2F_n}{n-1}\,x[t-1]\right)^2\right] = \frac{A^2}{(n-1)^2}\left(\frac{2Cn}{2C+1}+\frac{n^2}{2C+1}\right).$$

By this, the second term in the (5) can be evaluated as

$$\sum_{t=1}^{n}E[(x[t]-\hat{x}_a[t])^2] = \sum_{t=1}^{n}\left(A^2-2\left(\frac{A^2}{2C+1}+\frac{A^2}{n-1}\right)+\frac{A^2}{(n-1)^2}\left(\frac{n^2}{2C+1}+\frac{2Cn}{2C+1}\right)\right).$$

E. Proof of Lemma 3

After combining results of Lemmas 1 and 2, the overall lower bound $L(n)$ is given by

$$\begin{aligned}
&= \sum_{t=1}^{n}\left\{\left(A^2-2\,\frac{t-2}{(2C+1)(t-2+2C)}\,A^2+\frac{A^2}{(t-2+2C)^2}\left(\frac{(t-2)^2}{(2C+1)}+\frac{2C}{2C+1}\,(t-2)\right)\right)\right.\\
&\quad -\left(A^2-2\left(\frac{A^2}{2C+1}+\frac{A^2}{n-1}\right)+\frac{A^2}{(n-1)^2}\cdot\left(\frac{n^2}{2C+1}+\frac{2Cn}{2C+1}\right)\right)\Bigg\}\\
&= A^2\sum_{t=1}^{n}\left\{\frac{-2}{2C+1}+\frac{4C}{(2C+1)(t-2+2C)}+\frac{(t-2)^2}{(2C+1)(t-2+2C)^2}+\frac{2C(t-2)}{(2C+1)(t-2+2C)^2}\right.\\
&\quad +\frac{2}{2C+1}+\frac{2}{n-1}-\frac{n^2}{(2C+1)(n-1)^2}-\left.\frac{2Cn}{(2C+1)(n-1)^2}\right\}\\
&= A^2\sum_{t=1}^{n}\left\{\frac{4C}{(2C+1)(t-2+2C)}+\frac{(t-2+2C)^2}{(2C+1)(t-2+2C)^2}-\frac{4C(t-2+2C)}{(2C+1)(t-2+2C)^2}+\frac{4C^2}{(2C+1)(t-2+2C)^2}\right.\\
&\quad +\frac{2C(t-2+2C)}{(2C+1)(t-2+2C)^2}-\frac{4C^2}{(2C+1)(t-2+2C)^2}+\frac{2}{n-1}-\frac{(n-1)^2}{(2C+1)(n-1)^2}-\frac{2(n-1)}{(2C+1)(n-1)^2}\\
&\quad -\left.\frac{1}{(2C+1)(n-1)^2}-\frac{2Cn}{(2C+1)(n-1)^2}\right\}\\
&= A^2\sum_{t=1}^{n}\left\{\frac{2C}{2C+1}\,\frac{1}{t-2+2C}\right\}+O(1).
\end{aligned}$$

ACKNOWLEDGMENT

The authors would like to thank the Associate Editor Gábor Lugosi and the anonymous reviewers for their insightful comments and suggestions.

REFERENCES

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

[2] K. S. Azoury and M. K. Warmuth, "Relative loss bounds for on-line density estimation with the exponential family of distributions," *J. Machine Learning*, vol. 43, no. 3, pp. 211–246, June 2001.

[3] A. Singer and M. Feder, "Universal linear least-squares prediction," in *Proc. 2000 IEEE Int. Symp. Information Theory*, Sorrento, Italy, June 25–30, 2000.

[4] V. Vovk, "Competitive on-line linear regression," in *Advances in Neural Information Processing Systems*, M. I. Jordan, M. J. Kearns, and S. A. Solla, Eds. Cambridge, MA: MIT Press, 1998, pp. 364–370.

[5] M. J. Weinberger, N. Merhav, and M. Feder, "Optimal sequential probability assignment for individual sequences," *IEEE Trans. Inform. Theory*, vol. 40, pp. 384–396, Mar. 1994.

[6] G. Shamir and N. Merhav, "Low complexity sequential lossless coding for piecewise stationary memoryless sources," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1498–1519, July 1999.

[7] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 530–536, Sept. 1978.

[8] P. A. J. Volf and F. M. Willems, "Switching between two universal source coding algorithms," in *Proc. 1998 Data Compression Conf.*, Snowbird, UT, 1998, pp. 491–500.

[9] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629–636, July 1984.

[10] F. Willems, Y. Shtarkov, and T. Tjalkens, "The context-tree weighting method: Basic properties," *IEEE Trans. Inform. Theory*, vol. 41, pp. 653–664, May 1995.

[11] T. Shamoon and C. Heegard, "Adaptive update algorithms for fixed dictionary lossless data compressors," in *Proc. 1994 IEEE Int. Symp. Information Theory*, 1994, p. 14.

[12] A. H. Sayed and T. Kailath, "A state-space approach to adaptive RLS filtering," *IEEE Signal Processing Mag.*, pp. 18–60, July 1994.

[13] H. Stark and J. W. Woods, "Probability, random processes, and estimation theory for engineers," manuscript, 1994.

[14] D. Foster, "Prediction in the worst case," *Ann. Statist.*, vol. 19, no. 2, pp. 1084–1090, 1991.

Analysis of a Complexity-Based Pruning Scheme for Classification Trees

Andrew B. Nobel, *Member, IEEE*

***Abstract*—A complexity-based pruning procedure for classification trees is described, and bounds on its finite sample performance are established. The procedure selects a subtree of a (possibly random) initial tree in order to minimize a complexity penalized measure of empirical risk. The complexity assigned to a subtree is proportional to the square root of its size. Two cases are considered. In the first, the growing and pruning data sets are identical, and in the second, they are independent. Using the performance bound, the Bayes risk consistency of pruned trees obtained via the procedure is established when the sequence of initial trees satisfies suitable geometric and structural constraints. The pruning method and its analysis are motivated by work on adaptive model selection using complexity regularization.**

***Index Terms*—Bayes risk consistency, classification trees, complexity regularization, pruning, tree structured partitions.**

I. INTRODUCTION

Let $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$ be a jointly distributed pair of random variables, where the covariate vector X contains the outcomes of a sequence of experiments, and the binary response variable Y is an associated class label of interest. For example, X may contain the results of d diagnostic tests performed on a patient, and Y might indicate whether or not the patient has a particular disease. A classification rule is a deterministic map $\phi\colon \mathbb{R}^d \to \{0, 1\}$ that assigns a class label to each possible value of X. The performance of ϕ is measured by its probability of error, or risk

$$R(\phi) = \mathbb{P}\{\phi(X) \neq Y\}.$$

(We assume throughout this correspondence that classes zero and one have equal prior probabilities and identical misclassification costs.) The best achievable risk of any prediction rule is given by the Bayes probability of error

$$R^* = \inf_{\phi} R(\phi)$$

where the infimum is taken over all measurable functions $\phi\colon \mathbb{R}^d \to \{0, 1\}$. The infimum is achieved by the Bayes rule

$$\phi^*(x) = I\{E(Y|X = x) > 1/2\}$$

which can be deduced from the joint distribution of (X, Y). A comprehensive treatment of probabilistic pattern recognition can be found in [6], [13].

Histogram classification rules are defined by partitioning the space $\mathbb{R}^d$ of the covariates into disjoint regions, and then assigning a class label to each region. Binary classification trees, also known as decision trees, are a widely used family of histogram rules. A binary classification tree is described by a labeled binary tree, each of whose leaves corresponds to a unique cell of a partition of $\mathbb{R}^d$. The tree structure makes computation of the corresponding classification rule fast, and

Manuscript received October 10, 2000; revised December 11, 2001. This work was supported in part by the National Science Foundation under Grants DMS-9501926 and DMS-9971964.

The author is with the Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260 USA (e-mail: nobel@stat.unc.edu).

Communicated by G. Lugosi, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Publisher Item Identifier 10.1109/TIT.2002.800482.