

A Lower bound on the Performance of Sequential Prediction

Suleyman S. Kozat and Andrew C. Singer

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][2] 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 paper, we 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 $A^2 p \ln(n)/n$.

I. SUMMARY

We investigate the problem of predicting a real-valued 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 sequences we consider are assumed to be bounded, in that $|x[t]| < A$ for some $0 < A < \infty$, but are otherwise arbitrary.

Rather than assuming a statistical ensemble of sequences, and attempting to achieve good expected performance, the goal of this game is to sequentially predict a sequence as well as the best linear predictor out of a large class of predictors. Hence, to assess the performance of any sequential predictor, we lower bound the following worst-case regret,

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

where $\hat{x}_a[t]$ is the prediction of the sequential predictor at time t , and $\hat{x}_c[t]$ is the prediction of the predictor from the class C of predictors. By lower bounding this regret, we demonstrate that no sequential predictor can outperform the best linear predictor for each sequence, and that for any predictor, there will always be a sequence for which the best linear predictor is better by at least this amount.

We first consider the class of first order linear predictors such that the competing class of predictors form their predictions as $\hat{x}_c[t] = wx[t-1]$ for each sample of the sequence x^n , where $w \in R$ is a scalar. We lower bound the regret (1) by defining a suitable distribution on x^n and taking expectation over this distribution. This distribution on

Andrew Singer and Suleyman Kozat are with the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign. email:acsinger@uiuc.edu, kozat@ifp.uiuc.edu. This research was supported in part by the National Science Foundation, under grants number CCR-0092598 (CAREER), CCR 99-79381, and ITR 00-85929, and the Office Of Naval Research under Award No.: N000140110117.

x^n is generated by a two-state Markov chain, with a beta prior on the state transition probability. This enables us to demonstrate the following result:

Theorem 1: Let $x[t]$ be a bounded, real-valued, but otherwise arbitrary sequence such that $|x[t]| \leq A$ for some $A > 0$. Then for any $\epsilon > 0$ there exists a constant G such that the total prediction error of any sequential predictor satisfies,

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

where a is the class of all sequential predictors.

We use this result to then consider the problem of linear prediction with linear predictors of fixed-order p . Now, at each time instant t , the prediction is given by $\hat{x}_c[t] = \underline{w}^T \underline{x}[t-1]$, where $\underline{w} = [w_1, \dots, w_p]$ and $\underline{x}[t-1] = [x[t-1], \dots, x[t-p]]^T$. The distribution on the sequence x^n is constructed as follows. We create a new sequence by interleaving p independent sequences of the type described previously. That is, we interleave p independent two-state Markov chains, and then take expectations over this distribution. This choice of distribution enables derivation of the following:

Theorem 2: Let $x[t]$ be a bounded, real-valued, but otherwise arbitrary sequence such that $|x[t]| \leq A$ for some $A > 0$. Then for any $\epsilon > 0$ there exists a constant G such that the total prediction error of any sequential predictor satisfies,

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

where a is the class of all sequential predictors.

The lower bounds we derived here are asymptotically tight such that in [2][3] a prediction algorithm is demonstrated whose accumulated excess prediction error over the best p th-order linear predictor is upper bounded by $A^2 p \ln(n)$ for every bounded sequence. Thus, this universal prediction algorithm is optimal in a min-max sense, such that no sequential predictor can ever do better.

REFERENCES

- [1] V. Vovk, "Competitive on-line statistics," *International Statistical Review*, vol. 69, pp. 213–248, 2001.
- [2] A. Singer and M. Feder, "Universal linear least-squares prediction," *2000 IEEE Int. Symp. on Info Theory, June 25-30, 2000, Sorrento, Italy*.
- [3] A. C. Singer, S. S. Kozat and M. Feder, "Universal Linear Least-Squares Prediction: Upper and Lower Bounds," to appear on *IEEE Trans. Inf. Theo*.