

Universal Linear Least-Squares Prediction

Andrew C. Singer

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
e-mail: acsinger@uiuc.edu

Meir Feder

Department of Electrical Engineering-Systems
Tel Aviv University
e-mail: meir@eng.tau.ac.il

Abstract — An approach to the problem of linear prediction is discussed that is based on recent developments in the universal coding and computational learning theory literature. This development provides a novel perspective on the adaptive filtering problem, and represents a significant departure from traditional adaptive filtering methodologies. In this context, we demonstrate a sequential algorithm for linear prediction whose accumulated squared prediction error, for every possible sequence, is asymptotically as small as the best fixed linear predictor for that sequence.

I. LINEAR PREDICTION

In this work, we consider the problems of adaptive filtering and linear prediction in a competitive algorithm framework. Given a data sequence $x^n = \{x[t]\}_{t=1}^n$, the optimal set of p coefficients, w_k , $k = 1, \dots, p$, that minimizes the total prediction error

$$E_w[N] = \sum_{n=1}^N (x[n] - \sum_{k=1}^p w_k x[n-k])^2,$$

is uniquely determined and certainly depends on the input sequence. Recently, a linear prediction algorithm was presented that asymptotically achieves the minimum average *sequentially accumulated* prediction error over all linear predictors of order p , i.e. $\min_w E_w[N]$, for every individual sequence [1]. In this work, we somewhat modify the algorithm, and as a result improve both the algorithm performance, in terms of the bound on the redundancy, and provide a more intuitive proof of this bound.

II. p -TH-ORDER LINEAR PREDICTION

We consider the problem of linear prediction with a filter of fixed-order p , parameterized by the vector $\vec{w} = [w_1, \dots, w_p]^T$, with predicted value $\hat{x}_{\vec{w}}[n] = \vec{w}^T \vec{x}[n]$, where $\vec{x}[n] = [x[n-1], \dots, x[n-p]]^T$. Let $x[n]$, $n = 1, \dots, N$, be a bounded, but otherwise arbitrary, sequence such that $|x[n]| < A$, where A need not be known in advance. Let $l_n(x, \hat{x}_{\vec{w}})$ be the running total squared prediction error, i.e. $l_n(x, \hat{x}_{\vec{w}}) = \sum_{t=1}^n (x[t] - \hat{x}_{\vec{w}}[t])^2$. Define a universal predictor $\hat{x}_u[n]$, as $\hat{x}_u[n] = \vec{w}_u[n-1]^T \vec{x}[n]$, where, $\vec{w}_u[n] = [R_{xx}^{n+1} + \delta I]^{-1} r_{xx}^n$, $R_{xx}^n = \sum_{k=1}^n \vec{x}[k] \vec{x}[k]^T$, $r_{xx}^n = \sum_{k=1}^n x[k] \vec{x}[k]$, and $\delta > 0$ is a positive constant.

Theorem 1 *The total squared prediction error of the p -th-order universal predictor, $l_n(x, \hat{x}_u) = \sum_{t=1}^n (x[t] - \hat{x}_u[t])^2$, satisfies*

$$l_n(x, \hat{x}_u) \leq \min_{\vec{w}} \{l_n(x, \hat{x}_{\vec{w}}) + \delta \|\vec{w}\|^2\} + A^2 \ln |I + R_{xx}^n \delta^{-1}|,$$

$$\frac{1}{n} l_n(x, \hat{x}_u) \leq \min_{\vec{w}} \frac{1}{n} \{l_n(x, \hat{x}_{\vec{w}}) + \delta \|\vec{w}\|^2\} + \frac{A^2 p}{n} \ln \left(1 + \frac{A^2 n}{\delta}\right).$$

Theorem 1 states that the average squared prediction error of the p -th-order universal predictor is within $O(A^2 p \ln(n)/n)$ of the best batch p -th-order linear prediction algorithm, for every individual sequence x^n . The idea behind the universal predictor and the proof of the Theorem is as follows. We define a “probability” assignment of each of the continuum of predictors $\vec{w} \in R^p$ to the data sequence x^n such that the probability will be an exponentially decreasing function of the total squared-error for that predictor. Over the continuum of predictors with coefficients \vec{w} , we assign a Gaussian prior over these probabilities, and define the universal probability to be the Bayesian mixture of these probabilities. With the Gaussian prior, we can obtain the universal probability in closed form. Since the probabilities assigned by every predictor can also be found in closed form, we can compare the universal probability to that of the best batch predictor for each sequence.

We note that the conditional universal probability is Gaussian distributed about some Bayesian (time-varying) mixture of predictor outputs as that applied to the individual predictor probabilities, however it is not in the form of an exponentially decreasing function of the prediction error of a particular predictor. In [1], the conditional mean of this distribution was used as a predictor and was shown to be universal using a convexity argument to bound its excess prediction error. However, the convexity argument required construction of a new Gaussian, centered about the same mean, which was both larger than the universal probability over the range of the data and also in the form of an exponentially decreasing function of the accumulated prediction error. This led to a redundancy proportional to $O(4A^2 p \ln(n)/n)$, four times larger than that achieved here. In this work, we search for a new Gaussian in the proper form, with a different mean and variance, that is larger than the universal probability over the range of the data. By symmetry arguments, we obtain the new mean and variance that minimize the resulting redundancy of the universal predictor. The resulting predictor $\hat{x}_u[n]$ can be viewed as the least-squares batch solution over the past, where we assume that $x[n] = 0$ and update $r_{xx}^n[0]$ and $R_{xx}^n[0]$ accordingly before predicting $x[n]$.

REFERENCES

- [1] A. Singer and M. Feder, “Twice universal linear prediction of individual sequences,” in *1998 IEEE Int. Symp. on Info. Theory*, 1998.