

ON UNIVERSAL LINEAR PREDICTION OF GAUSSIAN DATA

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

In this paper, we derive some of the stochastic properties of a *universal linear predictor*, through analyses similar to those generally made in the adaptive signal processing literature. In [1], a predictor was introduced whose sequentially accumulated mean squared error for any bounded individual sequence was shown to be as small as that for any linear predictor of order less than some maximum order m . For stationary Gaussian time series, we generalize these results, and remove the boundedness restriction. In this paper we show that the learning curve of this universal linear predictor is dominated by the learning curve of the best order predictor used in the algorithm.

1. INTRODUCTION

Autoregressive (AR) modeling by predictive least squares is a widely studied method for time series analysis, with many applications including channel equalization, speech modeling, coding and parametric spectral estimation. For an m th order linear predictor, the observed data at time n is modeled by a linear combination of the previous m samples, i.e.,

$$\hat{x}_m(n) = \sum_{i=1}^m a_i x(n-i).$$

For a given order predictor, a common way to select the unknown coefficients is to minimize the total squared prediction error, i.e., minimize,

$$E_n(x, \hat{x}_m) = \sum_{t=1}^n \left(x(t) - \sum_{i=1}^m a_i x(t-i) \right)^2.$$

When the best order is not known, a wide variety of methods have been proposed for model order selection. In [1], instead of fixing a specific model order, an algorithm was developed to dynamically adjust a weighted-combination, or mixture, of all model orders up to some

m . This results in a sequential predictor whose sequentially accumulated mean square prediction error for any bounded individual sequence is as small as that attainable by any linear predictor of order less than m . In this sense, the predictor is “universal” with respect to both model parameters and model orders. Nevertheless, for a probabilistic setting, say for Gaussian data, the probability that an individual sequence is bounded $|x(n)| < A$ goes to zero as n goes to infinity, for any $A \in R^+$. In this paper, we explore new results in a stochastic context, without the boundedness restriction.

The organization of this paper will be as follows. In Section 2, we define several terms related to the universal linear predictor. In the third section, we explore the convergence of time-varying weights of the mixture in the mean value. The limiting values and rate of convergence of these weights are important for calculation of the mean-squared error of the universal linear predictor. In Section 4, the mean-squared error of the universal linear predictor is derived, and some simulations are provided.

2. DEFINITIONS

Let $\hat{x}_k(n)$ be the output of a sequential linear predictor as obtained by the recursive least squares (RLS) algorithm with model order k . Define a universal linear predictor as a weighted sum over linear predictors of order less than or equal to m ,

$$\hat{x}_u(n) = \sum_{k=1}^m \mu_k(n) \hat{x}_k(n), \quad (1)$$
$$\mu_k(n) = \frac{\exp(-c l_{t-1}(x, \hat{x}_k))}{\sum_{k=1}^m \exp(-c l_{t-1}(x, \hat{x}_k))},$$

where c is a positive constant and $\mu_k(n)$, the weights in the mixture, are proportional to the performance of the k th order predictor on the data observed so far. The performance, $l_{t-1}(x, \hat{x}_k)$ is the accumulated squared prediction error that results from sequential

application of the time varying set of predictor coefficients, a_1^t, \dots, a_p^t , i.e., by using $\hat{x}_k(t)$. For each new sample at time n , these coefficients are obtained such that $E_{n-1}(x, \hat{x}_k)$ is minimized over these coefficients. Then,

$$l_n(x, \hat{x}_k) = \sum_{t=1}^n \left(x(t) - \sum_{j=1}^k a_j^{t-1} x(t-j) \right)^2.$$

Because these linear prediction coefficients are optimized only over the data available (up to but not including the value to be predicted), the sequential prediction error is a fair measure of performance of each predictor.

3. CONVERGENCE OF WEIGHT COEFFICIENTS IN THE MEAN VALUE

Suppose the underlying process to be estimated is a stationary Gaussian random process with unknown covariance but zero mean. In this probabilistic setting, Davisson [3] has shown that, the expected squared sequential prediction error of an RLS linear predictor of order p for any n satisfies,

$$\sigma^2[p; n] \stackrel{def}{=} E[(x(n) - \hat{x}_p(n))^2] = \sigma^2[p; \infty] \left(1 + \frac{p}{n} + o(n^{-1}) \right), \quad (2)$$

where $\sigma^2[p; \infty] = \lim_{n \rightarrow \infty} \sigma^2[p; n]$ exists and is the optimal expected square error without the sequentiality constraint on the linear predictor and $no(n^{-1}) \rightarrow 0$. The quantity $\sigma^2[p; \infty]$ is a non-increasing function of p such that the p th order linear predictor asymptotically outperforms (or at least gives the same minimum error of) any predictor with order less than p . The accumulated additional mean-squared prediction error of an RLS algorithm will therefore be the harmonic sum of terms of the form p/n which is approximately $p \ln(n)$. Hence,

$$E[l_n(x, \hat{x}_p)] = \sigma^2[p; \infty](n + p \ln(n)) + o(\ln(n)). \quad (3)$$

For calculation of the mean values of the mixture coefficients, $\mu_p(n)$, we make the assumption that,

$$E[\mu_p(n)] = \frac{\exp(-c E[l_{t-1}(x, \hat{x}_p)])}{\sum_{k=1}^m \exp(-c E[l_{t-1}(x, \hat{x}_k)])}.$$

Then by (3),

$$E[\mu_p(n)] = \frac{1}{1 + \sum_{k=1, k \neq p}^m \exp(-c A_k)}, \quad (4)$$

where,

$$A_k = \sigma^2[k; \infty](n + k \ln(n)) - \sigma^2[p; \infty](n + p \ln(n)) + o(\ln(n)).$$

Since each exponent in (4) is approximately $n(\sigma^2[p; \infty] - \sigma^2[k; \infty])$ for large n , the sign of the difference, $(\sigma^2[p; \infty] - \sigma^2[k; \infty])$ is important for the limiting value and the rate of the convergence. Suppose the underlying process is a general Gaussian random process, such that $\sigma^2[k; \infty]$ is monotonic and strictly decreasing in k . Then, for the maximum order predictor, m , the sign of the difference is always negative. Thus every exponential in the denominator vanishes as n goes to infinity, such that,

$$\lim_{n \rightarrow \infty} E[\mu_m(n)] = 1.$$

For any other $\mu_p(n)$, with $p < m$, at least one of the exponentials in the denominator will diverge, yielding,

$$\lim_{n \rightarrow \infty} E[\mu_p(n)] = 0, \quad p = 1, \dots, m-1.$$

Nevertheless, suppose $x(n)$ is a w th order Gaussian AR process,

$$x(n) = \sum_{k=1}^w c_k x(n-k) + \varepsilon(n), \quad (5)$$

where $\varepsilon(n)$ is a sequence of iid Gaussian random variables with zero mean and variance σ_ε^2 . When $(m > w)$, the term $\sigma^2[p; \infty]$ is not a monotonically decreasing, but rather a monotonically non-increasing function of p . For sufficient order predictors ($p \geq w$), $(\sigma^2[p; \infty] = \sigma^2[w; \infty] = \sigma_\varepsilon^2)$. Thus the previous analysis is true for only non-sufficient order predictors, i.e.,

$$\lim_{n \rightarrow \infty} E[\mu_p(n)] = 0, \quad p = 1, \dots, w-1.$$

For a sufficient order predictor p , the exponents in (4) with $(k > w)$ are approximately $(p-k) \ln(n)$ for large n . Thus at least one of the exponentials will diverge for $(p > w)$,

$$\lim_{n \rightarrow \infty} E[\mu_p(n)] = 0, \quad p = w+1, \dots, m.$$

Since by definition, $\sum_{k=1}^m E[\mu_k(n)] = 1$, we conclude that,

$$\lim_{n \rightarrow \infty} E[\mu_p(n)] = 1, \quad p = w.$$

As an example, suppose a second order Gaussian AR process, $x(n) + 0.5x(n-1) + 0.25x(n-2) = \varepsilon(n)$, with $\sigma_\varepsilon^2 = 1$, is estimated by a fourth-order universal linear predictor with $c = 0.5$. As seen from Figure-1, all $E[\mu_p(n)]$'s except $E[\mu_2(n)]$, goes to zero as n goes to infinity, which is in accordance with the results derived in this section.

4. MEAN-SQUARED ERROR

Deriving the true learning curve of this predictor is cumbersome due to the time-dependent weight coefficients, $\mu_k(n)$, which depend upon the data $x(n)$ in a

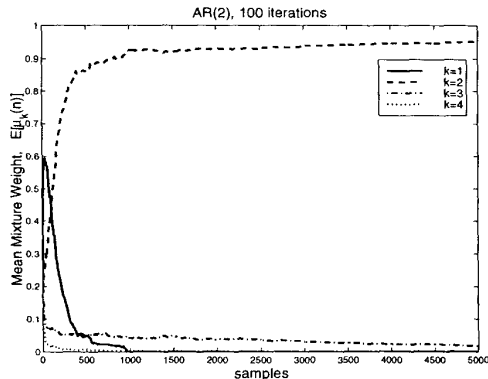


Figure 1: Mean Mixture weight, $E[\mu_k(n)]$, $k=1,2,3,4$.

non-linear manner. To make these calculations tractable, a few assumptions are made (A1 through A4).

A1) The weight coefficients $\mu_k(n)$, are statistically independent from predictions $\hat{x}_k(n)$, and data $x(n)$, but dependent upon one-another otherwise.

A2) The signal $x(n)$ and the predictions $\hat{x}_k(n)$ are jointly wide-sense stationary (WSS) Gaussian random processes, with zero mean.

This yields,

$$\begin{aligned} J_u(n) &= E[(x(n) - \sum_{k=1}^m \mu_k(n) \hat{x}_k(n))^2], \\ &= \sigma_x^2 - 2 \sum_{k=1}^m E[\mu_k(n)] E[x(n) \hat{x}_k(n)] \quad (6) \\ &\quad + \sum_{k=1}^m \sum_{l=1}^m E[\mu_k(n) \mu_l(n)] E[\hat{x}_k(n) \hat{x}_l(n)], \end{aligned}$$

where $E[x^2(n)] = \sigma_x^2$. By the law of iterated expectations,

$$\begin{aligned} E[x(n) \hat{x}_k(n)] &= E[E[x(n) \hat{x}_k(n) | x(n)]], \\ &= E[x(n) E[\hat{x}_k(n) | x(n)]]. \quad (7) \end{aligned}$$

Note that $E[\hat{x}_k(n) | x(n)]$ is the minimum mean square error (MMSE) estimate of $\hat{x}_k(n)$, given $x(n)$. The MMSE estimate of a Gaussian random variable based on another jointly Gaussian random variable is a linear function of that variable, [5], hence,

$$\begin{aligned} E[\hat{x}_k(n) | x(n)] &= a x(n), \\ a &= E[x(n) \hat{x}_k(n)] / E[x^2(n)]. \quad (8) \end{aligned}$$

When the prediction $\hat{x}_k(n)$ is the output of the MMSE-optimal predictor of order k , the error of prediction, $e_k(n) \stackrel{\text{def}}{=} x(n) - \hat{x}_k(n)$, and the prediction $\hat{x}_k(n)$ are orthogonal in probabilistic sense, [3],

$$E[e_k(n) \hat{x}_k(n)] = 0. \quad (9)$$

Suppose we make the third assumption such that,

A3) The error of prediction $e_k(n)$ and the prediction $\hat{x}_k(n)$ are orthogonal in probabilistic sense, so that (9) holds for all n , (even if $\hat{x}_k(n)$ is not the output of the MMSE-optimal predictor).

Then we can express 'a' as,

$$\begin{aligned} a &= (E[x(n) \hat{x}_k(n)] / E[x^2(n)]), \\ &= (E[\hat{x}_k^2(n)] / E[x^2(n)]) = \frac{\sigma_x^2 - J_k(n)}{\sigma_x^2}, \quad (10) \end{aligned}$$

where $J_k(n) = E[e_k^2(n)]$ is given by (2). Thus by (6),

$$\begin{aligned} J_u(n) &= \sigma_x^2 - 2 \sum_{k=1}^m E[\mu_k(n)] (\sigma_x^2 - J_k(n)) \\ &\quad + \text{cross terms}. \quad (11) \end{aligned}$$

An explicit analysis of the cross terms is cumbersome. Suppose the underlying process is a w th order AR Gaussian random process where ($w < m$). For all pairs of predictors of order k and l , make the last assumption that,

$$\mathbf{A4)} \quad E[\hat{x}_k(n) \hat{x}_l(n)] \approx E[\hat{x}_w^2(n)]$$

This equation is true for the output of the sufficient order predictors, when each predictor converges to the MMSE-optimal predictor, [3]. From the results of section 3, for the insufficient order predictors, the terms with $E[\mu_k(n) \mu_l(n)]$ will converge to zero at an exponential rate. Thus we can argue that the contribution of the insufficient order terms will vanish from the cross terms exponentially, making the assumption plausible and more accurate as n increases.

If the underlying random process is a general Gaussian process such that $\sigma^2[k, \infty]$ is monotonic and decreasing in k , then the assumption is

$$E[\hat{x}_k(n) \hat{x}_l(n)] \approx E[\hat{x}_m^2(n)]. \quad (12)$$

The same argument can be made for this case also, so that the contributions of all the lower order terms will vanish from the cross terms exponentially, as n goes to infinity.

Then for a Gaussian AR process of order ($w < m$), we can simplify the cross terms as,

$$\begin{aligned} \text{cross terms} &= \sum_{k=1}^m \sum_{l=1}^m E[\mu_k(n) \mu_l(n)] E[\hat{x}_k(n) \hat{x}_l(n)], \\ &\approx \sum_{k=1}^m \sum_{l=1}^m E[\mu_k(n) \mu_l(n)] E[\hat{x}_w^2(n)]. \quad (13) \end{aligned}$$

Since by definition $\sum_{k=1}^m E[\mu_k(n)] = 1$, and $\sum_{k=1}^m \sum_{l=1}^m E[\mu_k(n) \mu_l(n)] = 1$, we can simplify the

$J_u(n)$ considerably, using (13),

$$\begin{aligned}
 J_u(n) &= \sigma_x^2 - 2 \sum_{k=1}^m E[\mu_k(n)](\sigma_x^2 - J_k(n)) \\
 &\quad + \sigma_x^2 - J_w(n), \\
 &= E[\mu_w(n)]J_w(n) \\
 &\quad + \sum_{k=1, k \neq w}^m E[\mu_k(n)](2J_k(n) - J_w(n)),
 \end{aligned} \tag{14}$$

where $E[\mu_k(n)]$ is given by (4) and $J_k(n)$ is given by (2). The terms $J_w(n)$ and $\mu_w(n)$ will be replaced by $J_m(n)$ and $\mu_m(n)$, if the process is a general Gaussian random process such that the term $\sigma^2[k, \infty]$ is monotonic and decreasing in k . Then by (14), we conclude that,

$$E[(x(n) - \hat{x}_u(n))^2] \rightarrow \min_{k=1, \dots, m} E[(x(n) - \hat{x}_k(n))^2], \tag{15}$$

i.e., the universal linear predictor is universal in a stochastic sense as well.

This approximation of $J_u(n)$ for the mean-squared error (MSE) of the universal linear predictor will improve as n increases, as each of the approximations improve. For a general Gaussian random process, $J_u(n)$ converges to $\sigma^2[m, \infty]$ (which is $\sigma^2[m, \infty] = \sigma^2[w, \infty] = \sigma_\varepsilon^2$ for an AR process of order ($w < m$)). Then $J_u(n)$ is (asymptotically) unbiased.

As a second example, suppose a third order Gaussian AR process, $x(n) - 2.4x(n-1) - 1.91x(n-2) - 0.50x(n-3) = \varepsilon(n)$, with $\sigma_\varepsilon^2 = 0.1$ is predicted by a third order universal predictor, with $c = 1$. As seen from Figure-2, the plot of $J_u(n)$ curve matches the decay and convergence characteristics of the MSE curve of universal linear predictor. The simulations for weight coefficients, Figure-3, also agrees with the results of section 3 (as in the first example).

5. CONCLUSION

In this paper, we investigated the MSE for the universal linear predictor presented in [1], by making a few plausible assumptions whose affects diminish as n increases. It is shown that the learning curve of this universal linear predictor can be approximated as a weighted sum over all predictors' learning curves used in algorithm. As n goes to infinity, the MSE of universal linear predictor converges to the MSE of the best order linear predictor used in algorithm. Thus we can conclude that the universal linear predictor is also universal in this stochastic context.

6. REFERENCES

[1] A. C. Singer, M. Feder, "Universal Linear Prediction by Model Order Weighting," *IEEE Trans. on Signal Proc.*, vol. 47, no. 10, pp. 2685-2700, Oct. 1999.

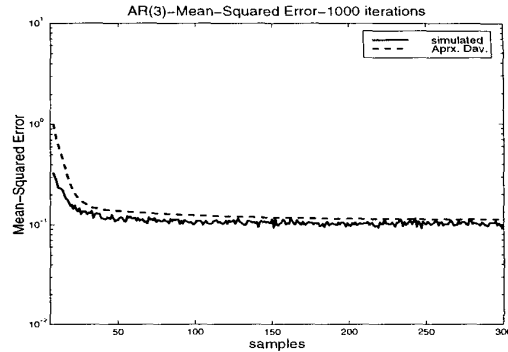


Figure 2: MSE of the Universal Predictor and Davison Approximation.

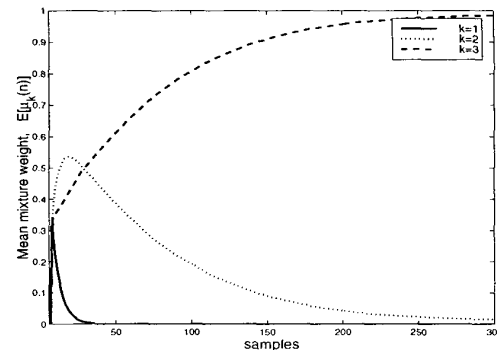


Figure 3: Mean Mixture weight, $E[\mu_k(n)]$, $k=1,2,3$.

[2] M. Feder, A. C. Singer, "Universal Data Compression and Linear Prediction," *Proc. 1998 IEEE Data Compression Conference, 1998*.

[3] L. D. Davison, "The Prediction Error of Stationary Gaussian Time Series of Unknown Covariance," *IEEE Trans. on Information Theory*, vol. 11, no. 4, pp. 527-532, Oct. 1965.

[4] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, Upper Saddle River, NJ 07458, 1996.

[5] H. Stark, J.W. Woods, *Probability, Random Processes, and Estimation Theory for Engineers*, Prentice-Hall, Upper Saddle River, NJ 07458, 1994.

[6] T. L. Lai, C. C. Wei, "Asymptotic Properties of Projections with Applications to Stochastic Regression Problems," *J. Mult. Anal.*, vol. 12, pp. 346-370, 1982.

[7] R.J. Bhansali, "Effects of Not Knowing the Order of an Autoregressive Process on the Mean Squared Error of Prediction," *Jour. of the American Statistical Association*, vol. 76, no. 375, pp. 588-597, September 1981.