

Universal Data Compression and Linear Prediction

Meir Feder* and Andrew C. Singer†

January 2, 1998

The relationship between prediction and data compression can be extended to universal prediction schemes and universal data compression. Recent work shows that minimizing the sequential squared prediction error for individual sequences can be achieved using the same strategies which minimize the sequential codelength for data compression of individual sequences. Defining a “probability” as an exponential function of sequential loss, results from universal data compression can be used to develop universal linear prediction algorithms. Specifically, we present an algorithm for linear prediction of individual sequences which is twice-universal, over parameters and model orders.

1 Introduction

We describe a sequential linear prediction algorithm which is “twice universal,” over parameters and model orders, for individual sequences under the square-error loss function; the sequentially accumulated mean-square prediction error is as good as any linear predictor of order up to some M , where the parameters may be tuned to the data. The linear prediction problem is transformed into one of sequential probability assignment, equivalent to lossless compression, which is accomplished through a double mixture; first over all linear predictors of a given model order using a Gaussian prior, and then over all model orders up to some maximum order M . For square error loss functions, the Gaussian prior enables the mixture probability over the continuum of models to be found in closed form. With respect to model orders, a finite mixture is used with an arbitrary prior. Using lattice filters, the coding distributions of all possible linear predictors with model orders up to M can be weighted in an efficient recursive procedure whose complexity is not larger than that for a conventional linear predictor of the largest model order. We derive an upper bound on the excess prediction error which can be identified with the excess coding redundancy in the assigned

*Meir Feder is with the Department of Electrical Engineering - Systems, Tel-Aviv University, Tel-Aviv, 69978, ISRAEL, E-mail: meir@eng.tau.ac.il

†Andrew Singer is with the Advanced Systems Directorate at Sanders, A Lockheed Martin Company, Nashua, NH 03061-0868, Tel: (603) 645-5647, Fax: (603) 645-5731, E-mail: acs@alum.mit.edu

mixture probabilities. The bound holds for all individual sequences of all lengths, not only for asymptotically long sequences. The two terms in the bound correspond to a parameter redundancy term, which is proportional to $p \ln(n)/n$, and a model order redundancy term which is proportional to $\ln(p)/n$, where n is the data length, and p is the best model order.

2 Statement of the Problem and Main Result

Consider the problem of designing a causal predictor which observes a sequence $x_1^{n-1} \triangleq x[0], x[1], \dots, x[n-1]$, and then computes a prediction of the value of $x[n]$ given the past. We assume that the sequence $x[n]$ is bounded such that $|x[t]| < A < \infty$ for all t , but is otherwise an arbitrary, real-valued sequence. We would like to design a predictor whose performance is at least as good as the best batch linear predictor of any order less than some $M < \infty$. This goal will be accomplished in two steps. First, we will demonstrate a fixed-order sequential prediction algorithm which performs as well as the best batch linear predictor of that order. We will then construct a predictor which performs as well as the best fixed-order predictor of order less than M .

Theorem 1 *Let x_1^n be a bounded, real-valued arbitrary sequence, such that $|x[t]| < A$, $1 \leq t \leq n$. Let R_{xx}^n and r_x^n be the p -th order deterministic autocorrelation matrix and vector defined as $R_{xx}^n = \sum_{t=1}^n \underline{x}[t] \underline{x}[t]^T$, and $r_x^n = \sum_{t=1}^n x[t] \underline{x}[t]$, where $\underline{x}[t] = [x[t-1], \dots, x[t-p]]^T$. Also assume that $\frac{1}{t} R_{xx}^t$ has a unique minimum eigenvalue bounded away from zero, $\lambda_0 \geq \lambda_\infty > 0$, $1 \leq t \leq n$. Let $\hat{x}_{\underline{a}}[n] = \underline{a}^T \underline{x}[n]$ be the fixed linear predictor with parameters \underline{a} . Define a universal p -th order linear predictor as $\hat{x}_p[n] = \hat{\underline{a}}_{\underline{a}}[n]^T \underline{x}[n]$, where $\hat{\underline{a}}_{\underline{a}}[n] = [R_{xx}^{n-1} + \frac{c}{\sigma^2} I]^{-1} r_x^{n-1}$, and σ and c are positive constants. Let $l(x_1^n, \hat{x}_{p,1}^n)$ be the running total squared prediction error for the p -th order universal linear predictor, i.e. $l(x_1^n, \hat{x}_{p,1}^n) = \sum_{t=1}^n (x[t] - \hat{x}_p[t])^2$. Define a twice-universal predictor $\hat{x}_{tu}[n]$, as $\hat{x}_{tu}[n] = \sum_{i=1}^M \mu_i[n] \hat{x}_i[n]$, where $\mu_i[n]$ is defined as*

$$\mu_i[n] = \frac{\exp(-\frac{1}{2c} l(x_1^{n-1}, \hat{x}_{i,1}^{n-1}))}{\sum_{k=1}^M \exp(-\frac{1}{2c} l(x_1^{n-1}, \hat{x}_{k,1}^{n-1}))}.$$

Then the total squared prediction error of the twice-universal predictor, $l(x_1^n, \hat{x}_{tu,1}^n) = \sum_{t=1}^n (x[t] - \hat{x}_{tu}[t])^2$, satisfies

$$\frac{1}{n} l(x_1^n, \hat{x}_{tu,1}^n) \leq \min_{p, \underline{a}} \frac{1}{n} l(x_1^n, \hat{x}_{p,1}^n) + \frac{4A^2 p}{n} \left(\ln \left(\frac{A^4 (p+1)n}{8 \lambda_\infty^2} \right) + 1 \right) + \frac{8A^2}{n} \ln(M) + O(n^{-2}).$$

Theorem 1 tells us that the average squared prediction error of the universal prediction algorithm is within $O(p \ln(n)/n)$ of the best batch linear prediction algorithm, uniformly, for every individual sequence x_1^n . As we shall see, the cost terms can be identified as a parameter redundancy term, proportional to $p \ln(n)/n$ and a model order redundancy term, proportional to $\ln(M)/n$. The proof of Theorem 1 is completed in two steps. First we demonstrate that a predictor generated by a mixture over all

p -th-order linear predictors is universal with respect to the class of all p -th-order linear predictors. We then show that a second mixture over all model orders provides a predictor which is universal with respect to both model orders and parameters. Each of these steps are contained in the proofs of Theorems 2, and 1, in Sections 3, and 4, respectively. The result is a twice-universal [1] [2] linear predictor which implements a double-mixture over model orders and parameters. This resembles the context tree weighting procedure in [3] which implements a double-mixture over the parameters and model orders of context-trees used in data compression. Key to the development of such universal algorithms is that the mixture be implementable by an efficient algorithm. We will show that the computational complexity of this twice-universal predictor is no larger than that for a conventional linear predictor of the order M .

3 Fixed-Order Linear Prediction

In this section, we consider the problem of linear prediction with a predictor of fixed-order p . The predictor is parameterized by the vector $\underline{a} = [a_1, \dots, a_p]^T$, and the predicted value can be written $\hat{x}_{\underline{a}}[t] = \underline{a}^T \underline{x}[t]$, where $\underline{x}[t] = [x[t-1], \dots, x[t-p]]^T$. If the parameter vector \underline{a} 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,

$$\underline{a}[n] = \arg \min_{\underline{a}} \sum_{t=1}^n (x[t] - \underline{a}^T \underline{x}[t])^2.$$

The well-known least-squares solution to this problem is given by $\underline{a}[n] = (R_{xx}^n)^{-1} r_x^n$, where $R_{xx}^n = \sum_{t=1}^n \underline{x}[t] \underline{x}[t]^T$, and $r_x^n = \sum_{t=1}^n x[t] \underline{x}[t]$.

The parameters $\underline{a}[n]$ can be computed recursively with the recursive least squares (RLS) algorithm. A common approach to sequential prediction is to use the parameters $\underline{a}[t-1]$ to predict $\hat{x}[t] = \underline{a}[t-1]^T \underline{x}[t]$. This is the so-called “plug-in” approach, since the best estimate of the parameters based on the data x_1^{t-1} are “plugged-in” to the predictor model for $x[t]$. It can be shown [4] [5] that the least-squares optimal batch prediction error can be achieved sequentially by the plug-in approach of the RLS algorithm to within $O(p^2 \ln(n)/n)$. This indicates that the rate at which RLS achieves the batch performance is slower than the $(p/2) \ln(n)/n$ which might be expected from universal coding results [6] [7], and is in agreement with the result in [7] which demonstrates that although the plug-in approach to sequential probability assignment can be optimal for certain model classes in the stochastic context, it is not optimal for individual sequences.

For this reason, rather than selecting a single set of parameters to use for prediction, we use the mixture approach of universal coding to obtain the universal predictor coefficients. This idea has already been applied in [2] for prediction in a probabilistic context. By transforming the problem into one of probability assignment, we can sequentially assign a probability to the sequence which is almost as good as that assigned by the best linear predictor. As such, we consider a means of estimating the parameters of the p -th order linear predictor $\underline{a}[t]$ through an a priori mixture over

the continuum of all possible parameters according to some prior. We now show that the prediction error of this universal predictor is as good as the best linear predictor determined from all of the data.

Theorem 2 *Let x_1^n be a bounded, real-valued arbitrary sequence, such that $|x[t]| < A$ for all t and $\frac{1}{t}R_{xx}^t$ has a unique minimum eigenvalue bounded away from zero, $\lambda_0 \geq \lambda_\infty > 0$. Let $\hat{x}_a[t]$ be the output of a p -th-order linear predictor with parameter vector \underline{a} , and $l(x_1^n, \hat{x}_{\underline{a},1}^n)$ be the running total squared prediction error, i.e. $l(x_1^n, \hat{x}_{\underline{a},1}^n) = \sum_{t=1}^n (x[t] - \hat{x}_{\underline{a},1}[t])^2$, where, $\hat{x}_a[n] = \underline{a}^T \underline{x}[n]$. Define a universal predictor $\hat{x}_u[n]$, as $\hat{x}_u[n] = \underline{a}_u[n-1]^T \underline{x}[n]$, where,*

$$\underline{a}_u[n] = \left[R_{xx}^n + \frac{c}{\sigma^2} I \right]^{-1} r_x^n,$$

$R_{xx}^{n-1} = \sum_{k=1}^{n-1} \underline{x}[k] \underline{x}[k]^T$, $r_x^n = \sum_{k=1}^n x[k] \underline{x}[k]$, and σ and c are positive constants. Then the total squared prediction error of the p -th-order universal predictor, $l(x_1^n, \hat{x}_{\underline{a}_u,1}^n) = \sum_{t=1}^n (x[t] - \hat{x}_u[t])^2$, satisfies

$$\frac{1}{n} l(x_1^n, \hat{x}_{\underline{a}_u,1}^n) \leq \min_{\underline{a}} \frac{1}{n} l(x_1^n, \hat{x}_{\underline{a},1}^n) + \frac{4A^2 p}{n} \ln \left(\frac{A^4(p+1)n}{8\lambda^2} \right) + \frac{4A^2 p}{n} + O(n^{-2}).$$

Theorem 2 tells us that the average squared prediction error of the p -th-order universal predictor is within $O(p \ln(n)/n)$ of the best batch p -th-order linear prediction algorithm, uniformly, for every individual sequence x_1^n . The basic idea behind the proof for Theorem 2 will be the following. We define a ‘‘probability’’ assignment of each of the continuum of predictors to the data sequence x_1^n such that the probability will be an exponentially decreasing function of the total squared-error for that predictor. This use of prediction error as a probability or likelihood was also used by Rissanen [6] and Vovk[8]. By defining a universal probability as an a priori average of the assigned probabilities, then to first order in the exponent, the universal probability will be dominated by the largest exponential, i.e., the probability assignment of the model order with the smallest total squared error. For a finite collection of predictors, the redundancy of the mixture can be bounded by the negative logarithm of the weight assigned to the best model. However, for a mixture over a continuum of models, we must seek an alternate bound on the redundancy. Specifically, we obtain the conjugate prior such that the mixture over the parameters can be obtained in closed form. We then relate the universal probability assignment to the accumulated squared error of the universal predictor, giving the desired result.

Proof of Theorem 2: For each set of parameters \underline{a} , we define the probability $P_{\underline{a}}(x_1^n) = B \exp(-\frac{1}{2c} l(x_1^n, \hat{x}_{\underline{a},1}^n))$ as an exponential function of the sequential loss on the data. Over the continuum of predictors with coefficients \underline{a} , we assign the a priori Gaussian mixture

$$p(\underline{a}) = (\sqrt{2\pi}\sigma)^{-p} \exp \left\{ \frac{1}{2\sigma^2} \underline{a}^T \underline{a} \right\},$$

and define the universal probability

$$P_u(x_1^n) = \int_{\underline{a}} p(\underline{a}) P_{\underline{a}}(x_1^n) d\underline{a}.$$

We can then obtain this universal probability in closed form,

$$P_u(x_1^n) = B\sigma^{-p} \left| \left[\frac{1}{c}R_{xx}^n + \frac{1}{\sigma^2}I \right] \right|^{-1/2} \exp \left\{ -\frac{1}{2c} \left(R_x^n[0] - r_x^{nT} \left[R_{xx}^n + \frac{c}{\sigma^2}I \right]^{-1} r_x^n \right) \right\},$$

where $R_x^n[0] = \sum_{k=1}^n x^2[k]$.

To compare the universal probability with the maximum probability over all parameters \underline{a} , observe that

$$\max_{\underline{a}} P_{\underline{a}}(x_1^n) = P_{\underline{a}}(x_1^n)|_{\underline{a}=\hat{\underline{a}}_{ML}} = B \exp \left\{ -\frac{1}{2c} l(x_1^n, \hat{\underline{a}}_{ML,1}^n) \right\},$$

where, $\hat{\underline{a}}_{ML} = (R_{xx}^n)^{-1}r_x^n$. Since,

$$\begin{aligned} l(x_1^n, \hat{\underline{a}}_{ML,1}^n) &= \sum_{k=1}^n \left(x[k] - ((R_{xx}^n)^{-1}r_x^n)^T \underline{x}[k] \right)^2 \\ &= R_x^n[0] - r_x^{nT} (R_{xx}^n)^{-1} r_x^n, \end{aligned}$$

we obtain

$$P_{\hat{\underline{a}}_{ML}}(x_1^n) = B \exp \left\{ -\frac{1}{2c} (R_x^n[0] - r_x^{nT} (R_{xx}^n)^{-1} r_x^n) \right\}.$$

Taking their ratio, and after some algebra, we obtain,

$$\frac{P_u(x_1^n)}{\max_{\underline{a}} P_{\underline{a}}(x_1^n)} = \sigma^{-p} \left| \left[\frac{1}{c}R_{xx}^n + \frac{1}{\sigma^2}I \right] \right|^{-1/2} \exp \left\{ -\frac{1}{2} \left(r_x^{nT} [R_{xx}^n \sigma^2 R_{xx}^n + cR_{xx}^n]^{-1} r_x^n \right) \right\}. \quad (1)$$

Taking the logarithm, and substituting $R_{xx}^n = n\bar{R}_{xx}^n$, yields

$$\begin{aligned} -\ln \left(\frac{P_u(x_1^n)}{\max_{\underline{a}} P_{\underline{a}}(x_1^n)} \right) &= \frac{1}{2} \ln \left| \frac{1}{c}\sigma^2 R_{xx}^n + I \right| + \frac{1}{2} r_x^{nT} [R_{xx}^n \sigma^2 R_{xx}^n + cR_{xx}^n]^{-1} r_x^n \\ &= \frac{1}{2} \ln \left| \frac{1}{c}\sigma^2 n\bar{R}_{xx}^n + I \right| + \frac{1}{2} n\bar{r}_x^{nT} [n\bar{R}_{xx}^n \sigma^2 \bar{R}_{xx}^n n + cn\bar{R}_{xx}^n]^{-1} n\bar{r}_x^n \\ &= \frac{p}{2} \ln(n) + \frac{1}{2} \ln \left| \frac{1}{c}\sigma^2 \bar{R}_{xx}^n + \frac{1}{n}I \right| + \frac{1}{2} \bar{r}_x^{nT} [\bar{R}_{xx}^n \sigma^2 \bar{R}_{xx}^n + \frac{c}{n}\bar{R}_{xx}^n]^{-1} \bar{r}_x^n \\ &\leq \frac{p}{2} \ln(n) + \frac{1}{2} \ln(c^{-p}\sigma^{2p}) + \frac{1}{2} \ln \left| \bar{R}_{xx}^n + \frac{c}{n\sigma^2}I \right| + \frac{1}{2} pA^2 \lambda_{\infty}^{-2} \sigma^{-2} A^2. \end{aligned} \quad (2)$$

To continue, we need the following lemma bounding the logarithm of the determinant of a positive definite matrix, which is proved in the appendix.

Lemma 1 *For a $p \times p$ positive definite matrix M whose elements are each bounded by C , i.e., $|M_{i,j}| < C$, and positive constant δ , the logarithm of the determinant of the matrix $M + \delta I$ satisfies*

$$\ln |M + \delta I| \leq p \ln \left(\frac{p+1}{2} \right) + p \ln(C) + p \ln \left(1 + \frac{\delta}{\lambda_0} \right),$$

where λ_0 is the smallest eigenvalue of M .

Applying Lemma 1 to (2), we obtain

$$-2c \ln P_u(x_1^n) \leq \min_{\underline{a}} l(x_1^n, \hat{x}_{\underline{a},1}^n) + cp \ln \left(\frac{n\sigma^2(p+1)}{c} \frac{A^2}{2} \left(1 + \frac{c}{n\sigma^2\lambda_\infty} \right) \right) + \frac{cpA^4}{\lambda_\infty^2\sigma^2}. \quad (3)$$

We expand the definition of the universal conditional probability as

$$P_u(x_n|x_1^{n-1}) = \frac{\int_{\underline{a}} p(\underline{a}) P_{\underline{a}}(x_1^n) d\underline{a}}{\int_{\underline{a}} p(\underline{a}') P_{\underline{a}'}(x_1^{n-1}) d\underline{a}'} = \int_{\underline{a}} \mu_n(\underline{a}) P_{\underline{a}}(x_n|x_1^{n-1}) d\underline{a},$$

i.e.,

$$\mu_n(\underline{a}) = \frac{p(\underline{a}) P_{\underline{a}}(x_1^{n-1})}{\int_{\underline{a}} p(\underline{a}') P_{\underline{a}'}(x_1^{n-1}) d\underline{a}'}.$$

Note that $\mu_n(\underline{a})$ is proportional to the performance of the model \underline{a} on the data up to time $n-1$, $P_{\underline{a}}(x_1^{n-1})$. That is, while the universal probability is an a priori Gaussian mixture over the probabilities assigned to the sequence by each of the parameters \underline{a} , in order to maintain this a priori probability, the conditional probabilities, $P_u(x_n|x_1^{n-1})$ must be weighted according to their performance on the data so far, $\mu_n(\underline{a})$.

We define the universal predictor as a mixture over the parameters \underline{a} using the same conditional weights as the conditional probabilities $\mu_n(\underline{a})$. A straightforward but tedious calculation verifies that the universal predictor defined by this mixture uses the parameter vector $\underline{a}_u[t-1]$ at each time t for prediction of the sample $x[t]$, where

$$\underline{a}_u[t] = \int_{\underline{a}} \mu_t(\underline{a}) \underline{a} d\underline{a} = \left[R_{xx}^t + \frac{c}{\sigma^2} I \right]^{-1} r_x^t.$$

Defining $\tilde{P}_u(x_1^n)$ as the probability from the predictor which is a mixture over the parameters \underline{a} using the same weights as the mixture over the probabilities $P_{\underline{a}}(x_1^n)$, we have

$$\tilde{P}_u(x_n) = B \exp \left\{ -\frac{1}{2c} \sum_{k=1}^n \left(x[k] - \left(\int_{\underline{a}} \underline{a} \mu_k(\underline{a}) d\underline{a} \right) \underline{x}[k] \right)^2 \right\}. \quad (4)$$

Comparing $P_u(x_n|x_1^{n-1})$ and $\tilde{P}_u(x_n|x_1^{n-1})$,

$$\tilde{P}_u(x_n|x_1^{n-1}) = B \exp \left\{ -\frac{1}{2c} \left(x[n] - \int_{\underline{a}} \mu_n(\underline{a}) \underline{a}^T \underline{x}[n] d\underline{a} \right)^2 \right\},$$

and,

$$P_u(x_n|x_1^{n-1}) = \int_{\underline{a}} \mu_n(\underline{a}) B \exp \left\{ -\frac{1}{2c} \left(x[n] - \underline{a}^T \underline{x}[n] \right)^2 \right\},$$

we observe that $\tilde{P}_u(x_n|x_1^{n-1})$ is a function of a convex combination of the predicted values $\hat{x}_{\underline{a}}[n]$, while $P_u(x_n|x_1^{n-1})$ is the same convex combination of the function evaluated at the same values. By Jensen's inequality,

$$\tilde{P}_u(x_n|x_1^{n-1}) \geq P_u(x_n|x_1^{n-1}), \quad (5)$$

provided that the function $f(z) = B \exp(-(x[t] - z)^2/2c)$ is concave over the domain of z , which leads to

$$-\sqrt{c} \leq (x[k] - \hat{x}_{\underline{a}}[k]) \leq \sqrt{c}.$$

Since $|a_i| < A/\lambda_\infty$, (see [4]), the inequality (5) holds for $c \geq \left(A + \frac{A^2}{\lambda_\infty}\right)^2$. However, since $x[n]$ is bounded, we can always decrease the prediction error by enforcing $|\hat{x}_{\underline{a}}[n]| < A$, which leads to the selection $c \geq 4A^2$.

Using $c = 4A^2$ and (4) in (3), we obtain

$$l(x_1^n, \hat{x}_{\underline{a},1}^n) \leq \min_{\underline{a}} l(x_1^n, \hat{x}_{\underline{a},1}^n) + 4A^2 p \ln \left(\frac{n\sigma^2(p+1)}{8} \left(1 + \frac{4A^2}{n\sigma^2\lambda_\infty} \right) \right) + \frac{4A^6 p}{\lambda_\infty^2 \sigma^2}. \quad (6)$$

Our ‘‘probability’’ assignment algorithm had two free constants to be set. Now that we have selected a range for the constant c , we can investigate the constant σ^2 . Minimizing the expression in (6) with respect to σ^2 yields,

$$\frac{1}{n} l(x_1^n, \hat{x}_{\underline{a},1}^n) \leq \min_{\underline{a}} \frac{1}{n} l(x_1^n, \hat{x}_{\underline{a},1}^n) + \frac{4A^2 p}{n} \left(\ln \left(\frac{A^4(p+1)n}{8\lambda^2} \right) + 1 \right) + O(n^{-2})$$

where, $\sigma^2 = (A^4/\lambda^2) + O(n^{-1})$.

We note in particular that the parameter redundancy term in (6) is proportional to $p \ln(n)/n$ rather than the $p^2 \ln(n)/n$ redundancy shown for the plug-in method of RLS. The redundancy is actually of the form $(p/2) \ln(n)/n$, scaled by the factor $2c$ which accounts for the effect of range A of the sequence $x[n]$. Comparing this result with a finite number M models, where the parameter redundancy term would be bounded by $O(\ln(M)/n)$, we see that the ‘‘effective’’ number of models for the Gaussian mixture, grows linearly with n . This completes the proof of Theorem 2. ■

4 Proof of the Main Result

The proof of the main result of the paper, Theorem 1, uses the results from Section 3 which bound the parameter redundancy of the mixture model and a result from [4] bounding the model order redundancy from a second mixture over the model orders.

Proof of Theorem 1: Suppose a set of linear predictors of order k , $1 \leq k \leq M$, are given, such that at each time sample, the k -th linear predictor produces the estimate $\hat{x}_k[n]$. For the ‘‘loss’’ of the k -th order predictor defined as its running total squared prediction error, define the probability

$$P_k(x_1^n) \triangleq B \exp \left(-\frac{1}{2c} l(x_1^n, \hat{x}_{k,1}^n) \right),$$

and the universal probability $P_u(x_1^n)$

$$P_u(x_1^n) = \frac{1}{M} \sum_{i=1}^M P_i(x_1^n).$$

When $\hat{x}_u[n]$ is defined as a universal predictor obtained by the same sequential mixture over the individual predictors as over the probabilities, Theorem 1 in [4] shows that

$$\frac{1}{n} l(x_1^n, \hat{x}_{u,1}^n) \leq \min_i \frac{1}{n} l(x_1^n, \hat{x}_{i,1}^n) + \frac{8A^2}{n} \ln(M).$$

When each of the fixed-order predictors are k -th-order universal linear predictors as defined in Section 3, then the overall predictor is formed by a double-mixture; first over parameters, and then over model orders. The resulting prediction error of this twice-universal predictor, $\hat{x}_{tu}[n]$, satisfies,

$$\frac{1}{n} l(x_1^n, \hat{x}_{tu,1}^n) \leq \min_{p, \underline{a}} \frac{1}{n} l(x_1^n, \hat{x}_{\underline{a}^p, 1}^n) + \frac{4A^2 p}{n} \left(\ln \left(\frac{A^4(p+1)n}{8\lambda^2} \right) + 1 \right) + \frac{8A^2}{n} \ln(M) + O(n^{-2}). \quad (7)$$

This completes the proof of Theorem 1. ■

Theorem 1, the main result of this paper, demonstrates that a prediction algorithm based on a double-mixture over model orders and parameters, is indeed twice-universal. One observation from this result, is that the predictor parameters are very similar to those which arise from the recursive least squares procedure. In fact, if the covariance matrix of the RLS algorithm is initialized with the value of $R_{xx}^0 = (c/\sigma^2)I \approx 4(\lambda_\infty^2/A^2)I$, then the remaining RLS procedure is unchanged. For $c \geq 4A^2$, we see that c is greater than the largest instantaneous square prediction error. We also have that $\sigma \approx A^2/\lambda$ is a ratio of the maximum possible square value to the minimum average square value, or a measure of the “spread” of the sequence. To be universal, the a priori mixture over the parameters should have a large enough “variance” to cover this range.

The first term of the redundancy in (7) can be identified as a parameter redundancy term, since this is the excess prediction error induced above the batch error for a given model order due to the lack of knowledge of the best batch parameters for that model order a priori. Note that the parameter redundancy term here is of the form $O(p \ln(n)/n)$, which is in agreement with the stochastic case, as implied both by Davisson in [9] and the more general MDL [6]. We also note that the model order redundancy term, $8A^2 \ln(M)/n$, can be slightly improved upon. Rather than using a priori weights, $w_i = 1/M$, we could have weighted each of the models inversely proportional to their model order, i.e.,

$$w_i = \frac{i^{-1}}{\sum_{j=1}^M j^{-1}}.$$

The proof in [4] remains intact with the model order redundancy being $-\ln(w_p)/n$ rather than $-\ln(1/M)/n$, where p is the order of the model with the smallest prediction error. The resulting model order redundancy term becomes $\ln(p)/n + \ln \ln(M)/n$.

5 Algorithmic Issues

An issue that remains is the computational complexity of the universal approach which incorporates the $\{1, \dots, M\} \times R^p$ predicted values from each of the M model orders and the each of the continuum of predictors within a given model order along with their sequential prediction errors to compute each predicted value. At first glance, it might appear that the cost of universality is rather high, requiring the solution of an infinite number of linear prediction problems in parallel. However, since the mixture over the parameters can be accomplished through a properly initialized RLS algorithm, it only remains to solve for each of the RLS predictors for $i = 1, \dots, M$. The linear prediction problems for each model order have a great deal in common with one-another, and this structure can be exploited. Indeed, just as the RLS algorithm for a given model order can be written as a time-recursion, there exist time- and order-recursive solutions to the least squares prediction problem, in which at each time step, the M -th order prediction problem can be constructed by recursively solving for each of the predictors of lower order. The resulting complexity of these algorithms can be made to have $O(M)$ operations per time sample which results in a total complexity of $O(Mn)$. An example lattice prediction algorithm is given in [4].

6 Concluding Remarks

The main result of this paper, stated in Theorem 1, is an algorithm which is “twice universal” [1] [2] for linear prediction with respect to model orders and parameters. The universal predictor presented in this paper will perform as well as the best linear predictor of any order up to some maximum order, uniformly, for every individual sequence. With this algorithm, the problems of model order selection and parameter estimation for linear prediction have been mitigated in favor of a performance-weighted average among all model orders and all parameters. Efficient lattice algorithms which recursively generate all of the linear predictors at the computational price of only the largest model order and closed-form mixture parameters yield an algorithm that is computationally very efficient. Since the mixture parameters of the universal predictor can be identified as the RLS parameters with a properly initialized covariance, this paper also gives a concrete rationale for initializing an RLS or Kalman filter algorithm with an a priori covariance; it makes the algorithm universal with respect to parameters for individual sequences.

A Proof of Lemma 1:

To prove this bound, we note that $|M| < p!C^p$. Therefore,

$$\begin{aligned}\ln |M| &\leq \ln(p!) + p \ln(C) = p \sum_{k=1}^p \frac{1}{p} \ln(k) + p \ln(C) \\ &\leq p \ln \left(\sum_{k=1}^p \frac{1}{p} k \right) + p \ln(C) = p \ln \left(\frac{p+1}{2} \right) + p \ln(C).\end{aligned}$$

Therefore, for eigenvalues of M , $\lambda_k \geq \lambda_0$,

$$\begin{aligned}\ln |M + \delta I| &\leq p \ln \left(\frac{p+1}{2} \right) + p \ln(C) + \sum_{k=1}^p \ln \left(1 + \frac{\delta}{\lambda_k} \right) \\ &\leq p \ln \left(\frac{p+1}{2} \right) + p \ln(C) + p \ln \left(1 + \frac{\delta}{\lambda_0} \right).\end{aligned}$$

■

References

- [1] B. Y. Ryabko, “Twice-universal coding,” *Prob. Inf. Trans.*, vol. 20, pp. 173–7, Jul-Sep 1984.
- [2] B. Y. Ryabko, “Prediction of random sequences and universal coding,” *Prob. Inf. Transmission*, vol. 24, pp. 87–96, Apr-June 1988.
- [3] F. Willems, Y. Shtarkov, and T. Tjalkens, “The context-tree weighting method: Basic properties,” *IEEE Trans. Info. Theory*, vol. IT-41, pp. 653–664, May 1995.
- [4] A. Singer and M. Feder, “Universal linear prediction over parameters and model orders,” *submitted to IEEE Transactions on Signal Processing*.
- [5] N. Merhav and M. Feder, “Universal schemes for sequential decision from individual sequences,” *IEEE Trans. Info. Theory*, vol. 39, pp. 1280–1292, July 1993.
- [6] J. Rissanen, “Universal coding, information, prediction, and estimation,” *IEEE Trans. Info. Theory*, vol. IT-30, pp. 629–636, 1984.
- [7] M. J. Weinberger, N. Merhav, and M. Feder, “Optimal sequential probability assignment for individual sequences,” *IEEE Trans. Info. Theory*, vol. 40, pp. 384–396, March 1994.
- [8] V. Vovk, “Aggregating strategies (learning),” in *Proceedings of the Third Annual Workshop on Computational Learning Theory* (M. Fulk and J. Case, eds.), (San Mateo, CA), pp. 371–383, Morgan Kaufmann, 1990.
- [9] L. D. Davisson, “The prediction error of stationary Gaussian time series of unknown covariance,” *IEEE Trans. Info. Theory*, vol. IT-11, pp. 527–532, Oct. 1965.