

# Universal Context Tree Least Squares Prediction

Andrew C. Singer

Department of Electrical and Computer Engineering  
University of Illinois at Urbana Champaign  
Urbana, IL 61801, USA  
Email: acsinger@uiuc.edu

Suleyman S. Kozat

International Business Machines, Inc.  
Hawthorn, NY, USA  
Email: kozat@us.ibm.com

**Abstract**— We investigate the problem of sequential prediction of individual sequences using a competitive algorithm approach. We have previously developed prediction algorithms that are universal with respect to the class of all linear predictors, such that the prediction algorithm competes against a continuous class of prediction algorithms, under the square error loss. In this paper, we introduce the use of a “context tree,” to compete against a doubly exponential number of piecewise linear models. We use the context tree to achieve the performance of the best piecewise linear model that can choose its partition of the real line and real-valued prediction parameters, based on observing the entire sequence in advance, for the square error loss, uniformly, for any individual sequence. This performance is achieved with a prediction algorithm whose complexity is only linear in the depth of the context tree.

## I. INTRODUCTION

A number of results in machine learning [4], [5], adaptive signal processing [6], [7], and information theory [8] describe prediction algorithms that are universal in that they sequentially achieve the performance of the best model in a particular class, for every bounded sequence and for a variety of loss functions. While some of this work considers parametrically continuous linear classes, the structural constraint on linearity considerably limits the modeling power of the class. In this paper, we increase the richness of the competition class by including a doubly-exponential number of piecewise-linear structures, that can approximate any smooth nonlinear model arbitrarily well.

For piecewise linear modeling, the space spanned by past observations is partitioned into a union of disjoint regions. For each region, an estimate of the desired signal is given as the output of a fixed linear regressor. As an example, suppose for a scalar piecewise linear predictor, that the past observation space,  $x[t-1] \in [-A, A]$ , is parsed into  $J$  disjoint regions  $R_j$  where  $\bigcup_{j=1}^J R_j = [-A, A]$ . At each time  $t$ , the underlying predictor forms its prediction of  $x[t]$  as  $\hat{x}[t] = w_j x[t-1]$ ,  $w_j \in \mathcal{R}$ , when  $x[t-1] \in R_j$ .

In this paper, we first present results for the piecewise linear regression problem when the boundaries of each region are fixed and known. We will demonstrate an algorithm that achieves the performance of the best piecewise linear regressor for a given partition of a real line. We then extend these results to when the boundaries of each region are also design parameters in this class. In this case, we try to achieve the performance of the best piecewise linear regressor when the re-

gressor has the additional freedom of choosing the boundaries of each region from a large class of possible partitions. These partitions will be compactly represented using a “context-tree” [1]. Here, we have neither a priori knowledge of the selected partition nor the best model parameters given that partition. We focus on scalar piecewise linear regression, such that each prediction algorithm in the competition class is a function of only the latest observation, i.e.,  $x[t-1]$ . These results can be extended to higher-order regression models by considering context tree partitionings of multiple past observations.

We start our discussion when the boundaries of each region are fixed and known. Given such a partition  $\bigcup_{j=1}^J R_j = [-A, A]$ , real valued sequences  $x^n = \{x[t]\}_{t=1}^n$  and  $y^n = \{y[t]\}_{t=1}^n$  are assumed to be bounded but are otherwise arbitrary, in that  $|x[t]| < A_x$  for some  $A_x < \infty$  and  $|y[t]| < A_y$  for some  $A_y < \infty$ , respectively. Given past values of the desired signal  $x[t]$ ,  $t = 1, \dots, n-1$ , and a sequence of observations  $y[t]$ ,  $t = 1, \dots, n$ , we define a competing algorithm from the class of all piecewise scalar regressors as  $\hat{x}[t] = w_{s[t-1]} y[t]$ , where  $s[t-1] = j$  when  $x[t-1] \in R_j$ , and  $w_j \in \mathcal{R}$ ,  $j = 1, \dots, J$ . For each region,  $w_j \in \mathcal{R}$ ,  $j = 1, \dots, J$ , can be selected independently. For the linear prediction problem, we have  $y[t] = x[t-1]$ .

Here we try to minimize the following regret:

$$\sup_{x^n} \left\{ \sum_{t=1}^n (x[t] - \hat{x}_q[t])^2 - \inf_{\substack{w_j \in \mathcal{R} \\ j \in \{1, \dots, J\}}} \sum_{t=1}^n (x[t] - \hat{x}_w[t])^2 \right\}, \quad (1)$$

where,  $\hat{x}_w[t] = w_{s[t-1]} y[t]$ , and  $\hat{x}_q[t]$  is the prediction of a sequential algorithm; i.e., we try to achieve the performance of the best piecewise linear regressor tuned to the underlying sequences  $x^n$  and  $y^n$ .

We will demonstrate an algorithm whose prediction error over that of the best piecewise linear predictor is upper bounded by  $O(JA^2 \ln(n/J))$ . We then extend these results to when the partition itself is varied within the competition class. We define a depth- $K$  context tree structure for a partition of size  $2^K$  of the real line as in Figure 1, where, for this tree,  $K = 2$ . For a depth- $K$  context tree, the  $2^K$  finest partition bins are leaves of the tree. On this tree, each of the bins are equal in size and assigned to  $[A_x, A_x/2]$ ,  $[A_x/2, 0]$ ,  $[0, -A_x/2]$ ,  $[-A_x/2, -A_x]$ . In general, this need not be true.

- Use a *context-tree* to represent partitions of  $\mathcal{R}$
- Depth- $K$  full tree embeds  $N(K)$  different context-tree partitions in the set

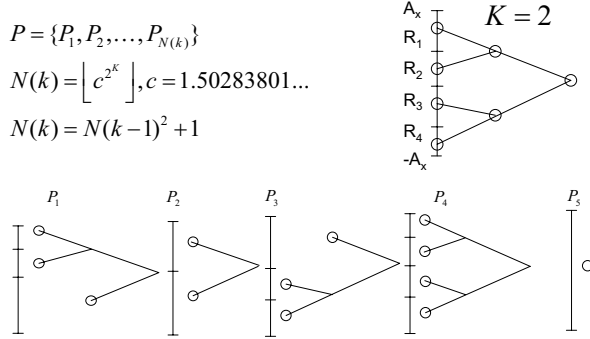


Fig. 1. A full tree of depth 2 that represents all context-tree partitions of the real line  $[-A_x, A_x]$  into at most four possible regions.

For a tree of depth- $K$ , there exist  $2^{K+1} - 1$  nodes, including leaf nodes and internal nodes. Each node  $\eta$  on this tree represents a portion of the real line,  $R_\eta$ . The region corresponding to each node  $\eta$ ,  $R_\eta$ , (if it is not a leaf) is constructed by the union of regions represented by the nodes of its children; the upper node  $R_{\eta_u}$  and the lower node  $R_{\eta_l}$ ,  $R_\eta = R_{\eta_u} \cup R_{\eta_l}$ . By this definition, any inner node is the root of a subtree and represents the union of its corresponding leaves (or bins).

We define a “partition” of the real line as a specific partitioning  $\mathcal{P}_i = \{R_{i,1}, \dots, R_{i,J_i}\}$  with  $\bigcup_{j=1}^{J_i} R_{i,j} = [-A, A]$ , where each  $R_{i,j}$  is represented by a node on the tree in Figure 1 and  $R_{i,j}$  are disjoint. The competitive framework defined to minimize the regret in (1) corresponds to a single, fixed partition,  $\mathcal{P}_i$ , where we compete against the best piecewise linear regressor given the partition  $\mathcal{P}_i$ . There exist  $N_K \approx (1.5)^{2^K}$  different such partitions,  $\mathcal{P}_i$ ,  $i = 1, \dots, N_K$ , embedded within a depth- $K$  tree. This is equivalent to the number of “proper binary trees” of depth  $K$ , and is given by Sloane’s sequence A003095[2], [3].

To achieve the performance of the best partition, we try to minimize the following regret

$$\sup_{x^n, y^n} \left\{ \sum_{t=1}^n (x[t] - \hat{x}_q[t])^2 - \inf_{\mathcal{P}_i} \inf_{w_{i,j} \in \mathcal{R}} \sum_{t=1}^n (x[t] - \hat{x}_w[t])^2 \right\}, \quad (2)$$

where  $\hat{x}_w[t] = w_{i,s_i[t-1]}y[t]$  and  $s_i[t-1]$  identifies the region within the partition  $\mathcal{P}_i$ . The algorithms introduced here are “twice-universal” such that the competition class can both select the regression parameters of the piecewise linear model and also the partition of the model itself based on observing the whole sequence in advance. Instead of trying to find, or estimate, the best partition from the class of all partitions and best parameters given the partition, we will use a performance weighted combination of all partitions with the corresponding parameters to construct a strictly sequential universal algo-

rithm that asymptotically achieves the performance of the best partition with the best parameters. This approach is based on sequential probability assignment from universal source coding [1], [11], [10]. We use the notion of a context tree from [1] to compactly represent the  $N(K)$  partitions of the real line. Here, instead of making hard decisions in every step of our algorithm, we use a soft combination of all possible models and parameters to achieve the performance of the best model, with complexity that remains linear in the data length. In contrast to the individual sequence results presented here, in [9], a regression tree is developed for statistical estimation of an unknown nonlinear state space model, nearly attaining the minimum mean square error, for that unknown model.

We begin with the case when the partition of the piecewise linear model is fixed and known in Section II. We then extend these results using context trees in Section III to unknown partitions. In each section, we provide theorems upper-bounding the regret with respect to the best competing algorithm. The derivations of the theorems are constructive, yielding the corresponding algorithms.

## II. KNOWN REGIONS

We begin with the class of piecewise linear predictors in each region for a known partitions and seek to minimize the following regret:

$$\sup_{x^n, y^n} \left\{ \sum_{t=1}^n (x[t] - \hat{x}_q[t])^2 - \inf_{w_j \in \mathcal{R}} \sum_{t=1}^n (x[t] - \hat{x}_w[t])^2 \right\}, \quad (3)$$

where  $\hat{x}_q[t]$  is the prediction at time  $t$  of any sequential algorithm,  $\hat{x}_w[t] = w_{s[t-1]}y[t]$  and the variable  $s[t-1] = j$  when  $x[t-1] \in R_j$ ,  $j = 1, \dots, J$ . That is, we wish to obtain a sequential algorithm that can predict every sequence  $x^n$  as well as the best fixed piecewise linear algorithm for that sequence with the given partition of the real line as  $\bigcup_{j=1}^J R_j = [-A_x, A_x]$ , where  $|x[n]| < A_x$  and  $|y[n]| < A_y$ , for all  $n$ .

A competing algorithm from the class of all piecewise linear predictors given this partition would be represented by the parameter vector  $\vec{w} = [w_1, \dots, w_J]$  and accumulate the total loss

$$l_n(x, \hat{x}_{\vec{w}}) = \sum_{t=1}^n (x[t] - w_{s[t-1]}y[t])^2. \quad (4)$$

Defining  $J$  time vectors (or index sequences) of length  $n_j$ ,  $t_j^{n_j} = \{t : s[t-1] = j\}$ , with  $j = 1, \dots, J$ , and sequences  $x_j^{n_j} = \{x[t_j[k]]\}_{k=1}^{n_j}$  and  $y_j^{n_j} = \{y[t_j[k]]\}_{k=1}^{n_j}$ , the loss (4) can be rewritten as  $l_n(x, \hat{x}_{\vec{w}}) = \sum_{j=1}^J \sum_{t=1}^{n_j} (x_j[t] - w_j y_j[t])^2$ . Since the number and boundaries of the regions are known, we have  $J$  independent least-squares problems in  $J$  regions. Applying the regression algorithm introduced in [6] for each region independently yields our candidate algorithm:  $\tilde{x}_w[n] = \tilde{w}_{s[n-1]}[n-1]y[n]$ , with

$$\tilde{w}_j[n] = \frac{R_{x_j y_j}^{n_j}}{R_{y_j y_j}^{n_j+1} + \delta_j},$$

where  $n_j$  is the number of points of  $x^{n-1}$  that belong to  $R_j$ ,  $\delta_j > 0$  is a positive constant, and  $R_{xy}^n = \sum_{t=1}^n x[t]y[t]$ . The following theorem relates the performance of the universal predictor,  $l_n(x, \tilde{x}_w) = \sum_{t=1}^n (x[t] - \tilde{x}_w[t])^2$ , to that of the best batch piecewise linear predictor.

**Theorem 1:** *Let  $x^n$  and  $y^n$  be bounded, real-valued sequences, such that  $|x[t]| < A_x$ , and  $|y[t]| < A_y$  for all  $t$ . Then for  $\Delta = \text{diag}(\delta_1, \dots, \delta_J)$ ,  $l_n(x, \tilde{x}_w)$  satisfies*

$$l_n(x, \tilde{x}_w) \leq \min_{\tilde{w}} \{l_n(x, \hat{x}_{\tilde{w}}) + \tilde{w}^T \Delta \tilde{w}\} + \sum_{j=1}^J h_j \ln \left( 1 + \frac{n_j A_y^2}{\delta_j} \right)$$

with  $h_j = \frac{1}{n_j} \sum_{k=1}^J n_{jk} A_{x,k}^2$ , where  $n_{jk}$  is the number of elements of region  $k$  that result from a transition from region  $j$  and  $|x[t]| \leq A_{x,k}$  when  $x[t] \in R_k$ . Here,  $l_n(x, \hat{x}_{\tilde{w}}) = \sum_{t=1}^n (x[t] - w_{s[t-1]})^2$  and  $s[t-1]$  is the state indicator variable.

The proof of Theorem 1 is based on sequential probability assignment and follows directly from [12].

Maximizing the upper bound with respect to  $n_j$ , replacing  $A_{x,k}$  with  $A_x$  and  $\delta_j$  with  $\delta$  yields

$$l_n(x, \tilde{x}_w) - \min_{\tilde{w}} \{l_n(x, \hat{x}_{\tilde{w}}) + \delta \|\tilde{w}\|^2\} \leq J A_x^2 \ln(n/J) + O(1).$$

### III. CONTEXT TREES

Given a context tree as in Figure 1 of depth  $K$ , we try to minimize the regret in (2), where  $s_i[t-1] = j$  if  $x[t-1] \in R_{i,j}$ , and  $\mathcal{P}_i = \{R_{i,1}, \dots, R_{i,J_i}\}$  with  $\bigcup_{j=1}^{J_i} R_{i,j} = [-A_x, A_x]$  for each  $i$ . Each  $R_{i,j}$  is represented by a node on the full tree and  $R_{i,j}$  are disjoint. The partition  $\mathcal{P}_i$  can be viewed as a subtree or ‘‘context tree’’ of the depth  $K$  full tree with the  $R_{i,j}$  corresponding to the leaves of the tree.

From [11], we define  $C(\mathcal{P}_i)$  as the total number of bits to represent each partition  $\mathcal{P}_i$  on the tree  $C(\mathcal{P}_i) = J_i + n_{\mathcal{P}_i} - 1$ , where  $n_{\mathcal{P}_i}$  is the total number of internal nodes in  $\mathcal{P}_i$ . Since  $n_{\mathcal{P}_i} \leq J_i$ ,  $C(\mathcal{P}_i) \leq 2J_i - 1$ . Given the tree, we construct a sequential algorithm with linear complexity in the data length  $n$  that asymptotically achieves the performance of the any partition with the best piecewise linear predictors as stated in the following theorem.

**Theorem 2:** *Let  $x^n$  and  $y^n$  be bounded scalar real-valued sequences, with  $|x[t]| < A_x$ , and  $|y[t]| < A_y$ , for all  $t$ . Then we can construct a sequential predictor  $\tilde{x}_w[t]$  with complexity linear in the data length  $n$  such that*

$$\begin{aligned} & \sum_{t=1}^n (x[t] - \tilde{x}_w[t])^2 \leq \\ & \inf_{\mathcal{P}_i} \inf_{w_{i,j} \in \mathcal{R}} \left\{ \sum_{t=1}^n (x[t] - w_{i,s_i[t-1]})^2 + \delta \|\tilde{w}_i\|^2 \right\} \\ & + 4A_x^2 C(\mathcal{P}_i) + J_i A_x^2 \ln(A_y^2 n / J_i) + O(1) \end{aligned}$$

where  $\delta > 0$  and  $C(\mathcal{P}_i)$  is a constant that is less than or equal to  $2J_i - 1$ .

#### A. Proof of Theorem 2

Given a partition  $\mathcal{P}_i = \bigcup_{j=1}^{J_i} R_{i,j}$ , we consider a family of predictors, each with its own prediction vector  $\tilde{w}_i = [w_{i,1}, \dots, w_{i,J_i}]^T$ . Here, each  $w_{i,j}$  represents a linear prediction for the  $j$ th region of partition  $\mathcal{P}_i$ , i.e., when  $x[t-1] \in R_{i,j}$ ,  $\hat{x}_{\tilde{w}_i}[t] = w_{i,j}y[t]$ . For each pairing of  $\mathcal{P}_i$  and  $\tilde{w}_i$ , we also consider a measure of the sequential prediction performance, or loss, of the corresponding algorithm,  $l_n(x, \hat{x}_{\tilde{w}_i} | \tilde{w}_i, \mathcal{P}_i) \triangleq \sum_{t=1}^n (x[t] - w_{i,s_i[t-1]})^2$ , where  $s_i[t-1]$  is the state indicator variable for partition  $\mathcal{P}_i$ , i.e.,  $s_i[t-1] = j$  if  $x[t-1] \in R_{i,j}$ . We define a function of the loss, namely, the ‘‘probability’’

$$P(x^n | \tilde{w}_i, \mathcal{P}_i) \triangleq \exp \left( -\frac{1}{2a} l_n(x, \hat{x}_{\tilde{w}_i} | \tilde{w}_i, \mathcal{P}_i) \right),$$

which can be viewed as a probability assignment of  $\mathcal{P}_i$ , with parameters  $\tilde{w}_i$ , to  $x^n$  induced by the performance of the corresponding predictor with  $\mathcal{P}_i$  and  $\tilde{w}_i$  on the sequence  $x^n$ , where  $a$  is a positive constant. Given  $\mathcal{P}_i$ , the algorithm in the family with the best predictors in each region assigns to  $x^n$  the probability

$$P^*(x^n | \mathcal{P}_i) \triangleq \exp \left( -\frac{1}{2a} \inf_{\tilde{w}_i} l_n(x, \hat{x}_{\tilde{w}_i} | \tilde{w}_i, \mathcal{P}_i) \right).$$

Maximizing  $P^*(x^n | \mathcal{P}_i)$  over all  $\mathcal{P}_i$  (on the tree) yields  $P^*(x^n | \mathcal{P}_i^*) \triangleq \sup_{\mathcal{P}_i} P^*(x^n | \mathcal{P}_i)$ . Here,  $P^*(x^n | \mathcal{P}_i^*)$  corresponds to the best piecewise constant predictor in the class on the tree of depth  $K$ . Our goal is to demonstrate a sequential algorithm which achieves  $P^*(x^n | \mathcal{P}_i)$ , for any  $\mathcal{P}_i$ . We accomplish this with a double mixture approach. First, we derive a universal probability assignment  $\tilde{P}_u(x^n)$  to  $x^n$  as a weighted combination of probabilities on the context tree. We will then demonstrate that this universal probability indeed achieves  $P^*(x^n | \mathcal{P}_i)$ , for any  $\mathcal{P}_i$ . As the final step we construct a sequential prediction algorithm, of linear complexity, whose associated probability assignment to  $x^n$  is as large as  $\tilde{P}_u(x^n)$  and hence the desired result.

Given any  $\mathcal{P}_i$ , using the sequential algorithm for  $\tilde{x}_w[n]$  for the partition  $\mathcal{P}_i$  yields  $\tilde{P}(x^n | \mathcal{P}_i) \triangleq \exp \left( -\frac{1}{2a} \sum_{t=1}^n (x[t] - \tilde{w}_{s_i[t-1]}[t-1]y[t])^2 \right)$ . As the next step, we assign to each node  $\eta$  on the context tree a sequential predictor working on the data observed by this particular node. For a node  $\eta$  representing the region  $R_\eta$ , we first assign a time vector (or index sequence) of length  $n_\eta$ ,  $t_\eta^{n_\eta} = \{t : x[t-1] \in R_\eta\}$  and a sequence  $d_\eta^{n_\eta} = \{x[t_\eta^{n_\eta}[k]]\}_{k=1}^{n_\eta}$ . Clearly, for each node  $\eta$ , there corresponds a portion of the observation sequence of length  $n_\eta$  and for a parent node in the tree with upper and lower children we have  $n_\eta = n_{\eta_u} + n_{\eta_l}$ , where  $n_{\eta_u}$  is the length of the subsequence that is shared with the upper child and  $n_{\eta_l}$  is the partition shared with the lower tree. For each node, we assign a predictor  $\tilde{c}_\eta[n] = w_\eta[n-1]y[n]$ , where

$$w_\eta[n] = \frac{R_{x_\eta y_\eta}^{n_\eta}}{R_{y_\eta y_\eta}^{n_\eta+1} + \delta}$$

and  $R_{xy}^n = \sum_{t=1}^n x[t]y[t]$  and  $\delta > 0$ . Here  $x_\eta[t]$  and  $y_\eta[t]$  are the samples that belong to node  $\eta$ .

We then define a weighted probability of a leaf node as  $\tilde{P}_\eta(x^n) = \exp\left(-\frac{1}{2a} \sum_{t=1}^n (d_\eta[t] - \tilde{c}_\eta[t-1])^2\right)$ , which is a function of the performance of the node predictor on the sequence  $d_\eta^{n\eta}$ . The probability of an inner node is defined as  $\tilde{P}_\eta(x^n) = \frac{1}{2} \tilde{P}_{\eta_u}(x^n) \tilde{P}_{\eta_l}(x^n) + \frac{1}{2} \exp\left(-\frac{1}{2a} \sum_{t=1}^n (d_\eta[t] - \tilde{c}_\eta[t-1])^2\right)$ , which is a weighted combination of the probabilities assigned to the data by each of the child nodes,  $\tilde{P}_{\eta_u}(x^n)$  and  $\tilde{P}_{\eta_l}(x^n)$ , and the probability assigned to  $d_\eta^{n\eta}$  by the sequential predictor of  $R_\eta$ . We then define the universal probability  $\tilde{P}_u(x^n)$  of  $x^n$  as the probability of the root node  $\tilde{P}_u(x^n) = \tilde{P}_r(x^n)$ , where we represent the root node with  $\eta = r$ . Using the recursion for the probability of an inner node, it can be shown [11] that the root probability  $\tilde{P}_r(x^n)$  is given by the sum of weighted probabilities of partitions  $\mathcal{P}_i$   $\tilde{P}_u(x^n) = \sum_{\mathcal{P}_i} 2^{-C(\mathcal{P}_i)} \tilde{P}(x^n|\mathcal{P}_i)$ , where  $C(\mathcal{P}_i) = J_i + n\mathcal{P}_i - 1$  is defined as the ‘‘cost’’ of partition  $\mathcal{P}_i$  and  $P(\mathcal{P}_i) \triangleq 2^{-C(\mathcal{P}_i)}$  can be viewed as a prior weighting, or prior probability, of the partition  $\mathcal{P}_i$ . It can also be shown that  $\sum_{\mathcal{P}_i} 2^{-C(\mathcal{P}_i)} = 1$  [1].

Hence, for any  $\mathcal{P}_i$ ,  $\tilde{P}_u(x^n) \geq 2^{-C(\mathcal{P}_i)} \tilde{P}(x^n|\mathcal{P}_i)$  yielding,  $-2a \ln(\tilde{P}_u(x^n)) \leq 2aC(\mathcal{P}_i) - 2a \ln(\tilde{P}(x^n|\mathcal{P}_i))$ . Applying Theorem 1 to  $\tilde{P}(x^n|\mathcal{P}_i)$ , we have

$$-2a \ln(\tilde{P}_u(x^n)) \leq \quad (5)$$

$$2aC(\mathcal{P}_i) + \inf_{w_{i,j} \in \mathcal{R}} \left\{ \sum_{t=1}^n (x[t] - w_{i,s_i[t-1]} y[t])^2 + \delta \|\tilde{w}_i\|^2 \right\} + J_i A_x^2 \ln(A_y^2 n / J_i) + O(1).$$

Hence we have a probability assignment  $\tilde{P}_u(x^n)$  which is as large as the probability assignment of the best partition  $P^*(x^n|\mathcal{P}_i^*)$  to  $x^n$ , to first order in the exponent. Nevertheless,  $\tilde{P}_u(x^n)$  is not in the form of the assigned probability from a valid sequential predictor. We now demonstrate a sequential prediction algorithm whose probability assignment to  $x^n$  is as large as  $\tilde{P}_u(x^n)$  and which is also in the proper prediction form.

The universal probability  $\tilde{P}_u(x^n)$  can be calculated recursively by defining a conditional probability, from the induced probability, i.e.,

$$\tilde{P}_u(x[n]|x^{n-1}) \triangleq \frac{\tilde{P}_u(x^n)}{\tilde{P}_u(x^{n-1})},$$

where  $\tilde{P}_u(x^n) = \prod_{t=1}^n \tilde{P}_u(x[t]|x^{t-1})$ . To achieve  $\tilde{P}_u(x^n)$ , we will demonstrate a sequential algorithm whose probability assignment is as large or larger than  $\tilde{P}_u(x[t]|x^{t-1})$  for all  $t$ .

Given  $x^{n-1}$  and  $\tilde{P}_u(x^{n-1})$ , node probabilities  $P_\eta(x^{n-1})$  should be adjusted after observing  $x[n]$  to form  $\tilde{P}_u(x^n)$ . However, owing to the tree structure, only probabilities of nodes that include  $x[n-1]$  need to be updated. Only  $K+1$  nodes contain  $x[n-1]$ : the leaf node that contains  $x[n-1]$  and all the nodes that contain that leaf. Hence, at each time

$n$ , only  $K+1$  node probabilities need be adjusted to form  $\tilde{P}_u(x^n)$ . This enables us to update  $\tilde{P}_u(x^{n-1})$ , a mixture of  $N(k)$ , doubly exponential in  $K$ , predictors with only  $K+1$  updates.

Tracing this path through the context tree, along ‘‘dark nodes’’,  $\tilde{P}_u(x^{n-1})$  can be compactly represented as sum of  $K+1$  terms, collecting all terms that will not be affected by  $x[n]$ , i.e.,

$$\tilde{P}_u(x^{n-1}) = \sum_{k=0}^K \sigma_k[n-1] e^{\left(-\frac{1}{2a} \sum_{t=1}^{n\eta_k} (d_{\eta_k}[t] - \tilde{c}_{\eta_k}[t-1])^2\right)}. \quad (6)$$

We enumerate dark nodes as  $\eta_k$ , where  $k = 0, \dots, K$ . For each dark node  $\eta_k$ ,  $\sigma_k[n-1]$  are products of node probabilities  $\tilde{P}_\eta(x^n)$  that share the same parent nodes with  $\eta_k$  but will be unchanged by  $x[n]$ . Hence, at each time  $n-1$ ,  $\sigma_k[n-1]$  can be calculated recursively with only  $K$  updates. In the calculation of  $\sigma_k[n-1]$ , we use the nodes that will be unchanged by  $x[n]$ , i.e.,  $\tilde{P}_{r_u}(x^n) = \tilde{P}_{r_u}(x^{n-1})$ ,  $\tilde{P}_{r_{lu}}(x^n) = \tilde{P}_{r_{lu}}(x^{n-1})$ . Thus, to obtain  $\tilde{P}_u(x^n)$ , we need to update only the exponential terms in (6). Since also  $d_r[n_r] = d_{\eta_1}[n\eta_1] = d_{\eta_2}[n\eta_2] = \dots = x[n]$ ,

$$\tilde{P}_u(x^n) = \sum_{k=0}^K \sigma_k[n-1] e^{\left(-\frac{1}{2a} \sum_{t=1}^{n\eta_k} (d_{\eta_k}[t] - \tilde{c}_{\eta_k}[t-1])^2\right)} \times \exp\left(-\frac{1}{2a} (x[n] - \tilde{c}_{\eta_k}[n-1])^2\right), \quad (7)$$

yielding the sequential update for  $\tilde{P}_u(x^n)$ .

Then,  $\tilde{P}_u(x[n]|x^{n-1})$  can be written  $\tilde{P}_u(x[n]|x^{n-1}) = \sum_{k=0}^K \mu_k[n-1] \exp\left(-\frac{1}{2a} (x[n] - \tilde{c}_{\eta_k}[n-1])^2\right)$ , where the weights  $\mu_k[n-1]$  are defined as

$$\mu_k[n-1] \triangleq \frac{\sigma_k[n-1] e^{\left(-\frac{1}{2a} \sum_{t=1}^{n\eta_k} (d_{\eta_k}[t] - \tilde{c}_{\eta_k}[t-1])^2\right)}}{\tilde{P}_u(x^{n-1})}.$$

We now construct sequential prediction algorithms whose associated probability assignments asymptotically achieve  $\tilde{P}_u(x^n)$  by upper bounding  $\tilde{P}_u(x[n]|x^{n-1})$  at each time  $n$ .

If we can find a prediction algorithm such that

$$\exp\left\{-\frac{1}{2a} (x[n] - \tilde{x}_c[n])^2\right\} \geq \tilde{P}_u(x[n]|x^{n-1}), \quad (8)$$

the result is obtained. One method to obtain the desired result relies on the concavity of the exponentiated loss function for  $a > 4A_x^2$ . By Jensen’s inequality,

$$\tilde{P}_u(x[n]|x^{n-1}) \leq e^{\left\{-\frac{1}{2a} (x[n] - \sum_{k=0}^K \mu_k[n-1] \tilde{c}_{\eta_k}[n-1])^2\right\}}$$

which gives the universal predictor as

$$\tilde{x}_c[n] = \sum_{k=0}^K \mu_k[n-1] \tilde{c}_{\eta_k}[n-1], \quad (9)$$

where  $\eta_k$  are the nodes such that  $x[n-1] \in R_{\eta_k}$ , i.e., dark nodes. From (5) we conclude that

$$\begin{aligned} & \sum_{t=1}^n (x[t] - \tilde{x}_w[t])^2 \\ & \leq 8A_x^2 C(\mathcal{P}_i) + \inf_{w_{i,j} \in \mathcal{R}} \left\{ \sum_{t=1}^n (x[t] - w_{i,s_i[t-1]})^2 + \delta \|\tilde{c}_i\|^2 \right\} + \\ & J_i A_x^2 \ln(A_y^2 n / J_i) + O(1). \end{aligned}$$

Another approach arises by noting that  $P_u(x[n] | x^{n-1})$  is in the form of the Aposteriori Prediction Algorithm (APA) of [4]. For values of  $a \geq 2A_x^2$ , there exists an interval of the real line that satisfies Equation (8) for  $\tilde{x}_c[n]$  within the interval and a value within this interval can be found in polynomial time [4]. Using this constraint on  $a$  yields

$$\begin{aligned} & \sum_{t=1}^n (x[t] - \tilde{x}_w[t])^2 \leq \\ & 4A_x^2 C(\mathcal{P}_i) + \inf_{w_{i,j} \in \mathcal{R}} \left\{ \sum_{t=1}^n (x[t] - w_{i,s_i[t-1]})^2 + \delta \|\tilde{w}_i\|^2 \right\} + \\ & J_i A_x^2 \ln(A_y^2 n / J_i) + O(1) \end{aligned}$$

completing the proof of Theorem 2 ■.

#### ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation under grant CCR-0092598 (CAREER).

#### REFERENCES

- [1] F.M.J. Willems, Y.M. Shtarkov, T.J. Tjalkens, "The context-tree weighting method: basic properties," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 653-664, May 1995.
- [2] Sloane, N. J. A. Sequence A003095/M1544 in "The On-Line Encyclopedia of Integer Sequences."
- [3] A. V. Aho, N. J. A. Sloane, "Some Doubly Exponential Sequences" *Fibonacci Quarterly*, vol. 11, pp. 429-437, 1970.
- [4] V.Vovk, "Aggregating strategies" *COLT*, pp. 371-383, 1990.
- [5] N. Cesa-Bianchi, P. M. Long, M.K. Warmuth, "Worst-case quadratic loss bounds for prediction using linear functions and gradient descent," *IEEE Transactions on Neural Networks*, vol. 7, no. 3, pp. 604 - 619, May 1996.
- [6] A. C. Singer, S. S. Kozat, M. Feder, "Universal linear least squares prediction: upper and lower bounds," *IEEE Transactions on Information Theory*, vol. 48, no. 8, pp. 2354-2362, Aug. 2002.
- [7] A. C. Singer, M. Feder, "Universal linear prediction by model order weighting," *IEEE Transactions on Signal Processing*, vol. 47, no. 10, pp. 2685-2699, October 1999.
- [8] N. Merhav, M. Feder, "Universal schemes for sequential decision from individual data sequences," *IEEE Transactions on Information Theory*, vol. 39, no. 4, pp. 1280-1292, July 1993.
- [9] O.J.J. Michel, A.O. Hero, A.E. Badel, "Tree-structured nonlinear signal modeling and prediction," *IEEE Transactions on Signal Processing*, vol. 47, no. 11, pp. 3027-3041, Nov. 1999.

#### Variables:

- $\eta = 1, \dots, 2^{K+1} - 1$ :  
 $P_\eta[n-1] \triangleq \tilde{P}_\eta(x^{n-1})$   
 $E_\eta[n-1] \triangleq \exp\left(-\frac{1}{2a} \sum_{t=1}^{n-1} (d_\eta[t] - \tilde{c}_\eta[t-1])^2\right)$   
 $C_\eta[n-1] \triangleq \tilde{c}_\eta[n-1]$  : Prediction of node  $\eta$  for  $x[n]$ .  
 $\delta, \delta_1, \delta_2$  : small real positive constants.  
 $A$  : Upper bound for the absolute value of the underlying process  $x[n]$ .  
 $\vec{d}[k]$  : the  $k$ th component of vector  $\vec{d}$ .

#### Initialization:

- For  $\eta = 1, \dots, 2^{K+1} - 1$ :  $P_\eta[0] = \delta_1^{-1}$ ,  $E_\eta[0] = \delta_2^{-1}$ ,  $C_\eta[0] = 0$ .  
 For  $k = 1, \dots, K+1$ :  $\mu_k[0] = 0$ ,  $\sigma_k[0] = 0$

#### Algorithm:

- For  $n = 1, \dots, N$ ,  
 $\vec{d} = []$  (vector containing indices of dark nodes).  
 For  $\eta = 1, \dots, 2^{K+1} - 1$ , (find dark nodes)  
   if  $x[n-1] \in R_\eta$ ,  
      $\vec{d} = [\vec{d}; \eta]$   
 $\sigma_0[n-1] = \frac{1}{2}$  (find weight for each node)  
 For  $\eta = \vec{d}[2], \dots, \vec{d}[K+1]$ ,  
    $\sigma_k[n-1] = \frac{1}{2} P_s[n-1] \sigma_{k-1}[n-1]$   
   where  $R_{\vec{d}[k]} \cup R_s = R_{\vec{d}[k-1]}$   
   (i.e.,  $s$  is the sibling node of  $\vec{d}[k]$ )  
    $\mu_k[n-1] = \frac{\sigma_k[n-1] E_{\vec{d}[k]}[n-1]}{P_{\vec{d}[1]}[n-1]}$   
 $\tilde{x}_c[n] = \sum_{k=0}^K \mu_k[n-1] C_{\vec{d}[k]}[n-1]$  (prediction)

For  $k = K+1, \dots, 1$ , (update node probabilities)

$$\begin{aligned} E_{\vec{d}[k]}[n] &= E_{\vec{d}[k]}[n-1] \exp\left(-\frac{1}{2a} (x[n] - C_{\vec{d}[k]}[n-1])^2\right) \\ \text{if } k &= K+1, P_{\vec{d}[k]}[n] = P_{\vec{d}[k]}[n] \text{ (leaf node).} \\ \text{elseif } k &\neq K+1, \\ P_{\vec{d}[k]}[n] &= \frac{1}{2} P_{\vec{d}[k]_u}[n-1] P_{\vec{d}[k]_l}[n-1] + \frac{1}{2} E_{\vec{d}[k]}[n]. \\ C_{\vec{d}[k]}[n] &= x[n] \frac{n_{\vec{d}[k]} - 1 + \delta}{n_{\vec{d}[k]} + \delta} + \frac{C_{\vec{d}[k]}[n-1]}{n_{\vec{d}[k]} + \delta} \end{aligned}$$

TABLE I

COMPLETE DESCRIPTION OF THE CONTEXT TREE ALGORITHM.

- [10] G. I. Shamir, N. Merhav, "Low-Complexity Sequential Lossless Coding for Piecewise-Stationary Memoryless Sources," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1498-1519, July 1999.
- [11] F. M. J. Willems, "Coding for a Binary Independent Piecewise-Identically-Distributed Source," *IEEE Transactions on Information Theory*, vol. 42, pp. 2210-2217, Nov. 1996.
- [12] David Luengo, Suleyman S. Kozat, Andrew C. Singer, "Universal Piecewise Linear Least Squares Prediction: Upper and Lower Bounds," *International Symposium on Information Theory*, p. 198, Chicago, 2004.
- [13] R. E. Krichevsky and V. K. Trofimov, "The Performance of Universal Encoding," *IEEE Transactions on Information Theory*, vol. 27, pp. 190-207, March 1981.