Universal Switching Linear Least Squares Prediction

Suleyman S. Kozat
IBM TJ Watson Research Center
Yorktown Heights, NY
Email: kozat@us.ibm.com

Andrew C. Singer
Department of Electrical and Computer Engineering
University of Illinois
Urbana, IL
Email: acsinger@uiuc.edu

***Abstract*— We consider sequential regression of individual sequences under the square error loss. Using a competitive algorithm framework, we construct a sequential algorithm that can achieve the performance of the best piecewise (in time) linear regression algorithm tuned to the underlying individual sequence. The sequential algorithm we construct does not need the data length, number of piecewise linear regions, or the locations of the transition times, however, it can asymptotically achieve the performance of the best piecewise (in time) linear regressor that can choose number of segments, duration of these segments and best regressor in each segment, based on observation of the whole sequence in advance. We use a transition diagram similar to that of [Willems '96] to effectively combine an exponential number of competing algorithms, with a complexity that is only linear in the data length. We demonstrate that the regret of this approach is at most $O(4\ln(n))$ per transition for not knowing the best transition times and at most $O(\ln(n))$ for estimating the best linear regressor in each segment, where n is the total length of the observation process. Lower bounds for any sequential algorithm demonstrate a form of minmax optimality in certain settings. We then extend these results to include a finite collection of competing algorithms within each time segment, rather than linear regressors.**

I. Introduction

A common approach to applications in adaptive signal processing is to take the viewpoint of turning the problem at hand, such as equalization, prediction or some other sequential decision problem, and turn it into an associated parametric modeling or estimation problem. By forcing the problem into this form, we then have to live with the associated performance of the resulting parameter estimation problem, which in general is worse, often significantly worse, than that which could have been obtained by directly addressing the problem at hand. Moreover, if the assumptions in the model do not match reality, then the performance of the algorithm tuned to the assumed statistical model may deteriorate considerably.

In this paper, we approach the problem of prediction from a competitive algorithm point of view. By defining a competitive framework, we try to achieve the performance of the best algorithm from a large class of candidate algorithms, rather than attempting to fit a given model to the data at hand. The performance measure of interest is then defined with respect to the best from this class, instead of the usual parametric modeling error between the output of the modeling algorithm and the desired signal directly. We will show that by not forcing the algorithms to make hard decisions about a set of parameters at each step, but rather permitting a competition among many candidate models, we can obtain algorithms that compete well with respect to all candidate algorithms from a given class such that they sequentially achieve the performance of the best candidate that could have been chosen, given all of the data in advance.

As an example, in [2] we investigated linear regression of real-valued data under the square error loss. We presented a regression algorithm whose accumulated error is asymptotically as small as the best fixed linear regressor for that sequence, taken from the class of all linear regressors of a given order. The algorithm is constructed by observing a performance-weighted (Bayesian) combination of all linear regressors with fixed parameters, as if working in parallel, and selecting a prediction based on this "mixture predictor". The mixture approach is different than the classical "plug-in" approach, which typically involves estimating the parameters of a linear regressor based on data observed so far and plugging them in to a given model. These results are closely connected with methods from data compression and universal probability assignment. A similar approach is also introduced in [4] to derive an algorithm that asymptotically achieves the performance of the best linear predictor from a finite set of linear predictors, for real-valued bounded data. The results in [4] are also in the same spirit and closely tied to universal sequential probability assignment. We call these algorithms that asymptotically achieve the performance of the best algorithm from a given class of algorithms (for any observation sequence) *static* universal algorithms, since the competition class contains a fixed set of predictors, and performance is compared with the single best, fixed element of the class for the duration of the sequence.

In this paper, we extend the results for static algorithms to a framework where the underlying competition class has the ability to switch among the various static elements. Here, each competing algorithm can divide the observation sequence into arbitrary segments, say $k+1$ of them, and fit each contiguous segment with the best algorithm from a given class of static algorithms, such as a fixed linear regressor for that segment. For k transitions, there exist $k+1$ segments. The loss incurred by a class member for any such partition is the sum of the losses of each fixed static algorithm associated with each segment. The best partition is the one which gives the minimum total loss. We can also let the competing algorithm choose the number of possible switches, k. A natural restriction for the number of possible transitions (switches) is $k << n$ where n is the length of the observation sequence, as for $k = n$ a loss of 0 is

readily obtained, and clearly an unrealistic goal of a sequential algorithm.

Unlike [2], [4], here we try to exploit the time varying nature of the best choice of algorithm for any given realization, since the choice of best algorithm from a class of static algorithms can change over time. Nevertheless, instead of trying to find the best partition (possible best switching points) or best number of transitions, our objective is simply to achieve the performance of the best partition directly. The algorithms we provide are strongly sequential such that they do not need the number of transitions, time of these transitions or length of the data in advance. However, they can asymptotically achieve the performance of the best algorithm in the competition class which can select any fixed number of transitions, locations of these transitions and the best static algorithm for each segment, based on observing the whole sequence in advance. In this sense, we call these algorithms "universal switching" algorithms.

One of the main contributions of this paper is making the connection between certain results from sequential probability assignment methods from universal source coding and sequential regression (or adaptive filtering) of individual sequences. In this paper, we show that the performance weighting approach used in [1], [5] can be generalized to yield algorithms that can efficiently compete with the best partition over all partitions for more general loss functions. We demonstrate that the universal probability assignment introduced in [1], [5] can be effectively merged into the prediction framework by using the methodology introduced in [2]. Although we investigate the continuous class of linear regressors or that of a finite number of adaptive filters as our competition class and use the square error loss, the methodology introduced in this paper can be extended to arbitrary competition classes, such as that of certain nonlinear regressors considered in [6], [7] or to more general loss functions as in [8]. Our methodology of combining the resulting exponential number of algorithms with linear complexity is in this sense, generic.

When the algorithms in the competing class can select a different fixed linear regressor for each of the $k+1$ segments in a given partition of the data, from all linear regressors, the sequential algorithm introduced here (without knowledge of k apriori) has a regret (excess loss) of $A^2(k+1)\ln(n/k) + 4A^2k\ln(n/k) + O(k+1)$ over the performance of the best partition, for any bounded data sequence, that is bounded by $\pm A$. For a partition with $k+1$ segments, we also provide a corresponding lower bound on the excess loss of $A^2(k+1)\ln(n/k) + O(k+1)$ for the performance of any sequential algorithm without prior knowledge of k or n. When, $k=0$, i.e., no transition, these upper and lower bounds match. We identify the terms in the regret to include a "parameter-regret" of $A^2\ln(n/k)$ per segment for not knowing the best regressor in each segment in advance, and a switching-regret of $4A^2\ln(n/k)$ per transition for not knowing the transition times in advance. For loss functions other than the square error loss, the parameter-regret would be loss-function dependant, while the switching loss would remain the same.

When the class of algorithms to be used within each segment has a finite number of elements (as in [4]), this problem has been investigated by a number of authors in computational learning theory and has been referred to "tracking the best expert" [10], [11]. In [10] the authors generalize Vovk's Aggregating Algorithm (AA) and introduce an algorithm whose additional loss over the performance of the best partition is given by $O(k\ln(M)+k\ln(n/k))$ for a certain classes of fairly general loss functions, where k is the number of partitions, M is the number of "experts" (models in the class) and n is the length of the data. When the loss is bounded (which is the case for this paper for the square error loss with bounded data) a bound of $O(k\ln(M)+k\ln(L/k))$ where L is the total loss of the best partition is given in [10]. In [11], several randomized algorithms are introduced whose expected excess loss over the performance of the best partition are of the same order as those in [10]. Nevertheless, only in certain cases (when the static class has finite size as in [10]) is an explicit description of the algorithm given. In other cases the algorithms are referred to as "infeasible" due to the difficulty of combining a large number of algorithms. In both cases, one needs to optimize certain parameters a priori for each algorithm to yield tight upper bounds on the total excess loss. For a finite competition class, we present algorithms whose excess loss over the best partition are of the same order as those in [10], i.e., $O(k\ln(M)+k\ln(n/k))$ without the need to optimize parameters a priori. We also provide corresponding upper and lower bounds when the number of segments is large, e.g., comparable to n. For linear regression, in [9], the authors investigate tracking the best linear combination of a finite class of predictors including a Gradient Decent (GD) and an Exponentiated Gradient (EG) method to combine the outputs of a finite class of algorithms. For the setup in [9], the authors provide algorithms whose additional loss over the performance of the best partition is $O(L)$, where L is the total loss of the best partition. Our bounds are with respect to the performance of the best partition and our time averaged excess loss over the best partition asymptotically vanishes.

We begin by studying the prediction problem when the competition class is a special case of linear regressors, i.e., all constant predictors, in Section 2. We continue with the class of fixed-order linear regressors and derive corresponding upper bounds. In Section 3, we investigate the case when the static class contains a finite set of algorithms.

II. Piecewise Linear Predictors

In this section, we investigate the linear regression problem with the square error loss in a competitive framework for deterministic unknown data. The real valued sequence $x^n = \{x[t]\}_{t=1}^n$ and the vector sequence $\vec{y}^n = \{\vec{y}[t]\}_{t=1}^n$ are assumed to be bounded but are otherwise arbitrary, in that $|x[t]| < A_x$ for some $A_x < \infty$ and $|y_r[t]| < A_y$, $r = 1,\ldots,p$ for some $A_y < \infty$. For given sequences x^n and $\vec{y}^n$, a competing algorithm with a transition path $\mathcal{T}_{k,n}$ with k transitions, represented by $(t_1,\ldots,t_k)$, divides x^n into $k+1$ segments such that x^n and $\vec{y}^n$ are represented by the

concatenation of

$$\{x[1],\ldots,x[t_1-1]\}\{x[t_1],\ldots,x[t_2-1]\}\ldots\{x[t_k],\ldots,x[n]\},$$

and

$$\{\vec{y}[1],\ldots,\vec{y}[t_1-1]\}\{\vec{y}[t_1],\ldots,\vec{y}[t_2-1]\}\ldots\{\vec{y}[t_k],\ldots,\vec{y}[n]\},$$

respectively. Given the past values of the desired signal $x[t]$, $t=1,\ldots,n-1$ and the sequence of observation vectors $\vec{y}[t]$, $t=1,\ldots,n$, a competing algorithm forms an estimate of the desired signal in each segment as

$$\hat{x}[t]=\vec{w}_i^T\vec{y}[t]$$

where $\vec{w}_i=[w_{i,1},\ldots,w_{i,p}]^T$, $\vec{w}_i\in R^p$, $t_{i-1}\le t<t_i$, $i=1,\ldots,k+1$. For simplicity we assume $t_0=1$ and $t_{k+1}=n+1$. Given n and k, there exist $\binom{n-1}{k}$ such transition paths $\mathcal{T}_{k,n}$.

We begin the discussion for the particular choice of $y[t]=1$ for all t. In this case, the competing class contains all constant functions in each segment, i.e. $\hat{x}[t]=c_i$, $c_i\in R$, for each sample of the sequence $x[t]$ for $t=t_{i-1},\ldots,t_i-1$. Each c_i can be selected independently for each region, $i=1,\ldots,k+1$ where we denote constant predictors using the variable c to avoid confusion with the scalar linear regression problem. In determining the best algorithm in the competing class, we attempt to outperform all such predictors, including the one that has been selected by choosing the transition path $\mathcal{T}_{k,n}$, the number of transitions k and the constants c_i in each segment based on observing the entire sequence x^n in advance. As such we try to minimize the following regret

$$\sup_{x^n}\left\{\sum_{t=1}^{n}(x[t]-\hat{x}_q[t])^2-\inf_{\substack{c_1,\ldots,c_{k+1}\in R\\ t_1,\ldots,t_k\in\{2,\ldots,n\}}}\sum_{i=1}^{k+1}\sum_{t=t_{i-1}}^{t_i-1}(x[t]-c_i)^2\right\}$$

where $\hat{x}_q[t]$ is the prediction at time t of any sequential algorithm. We will construct a sequential algorithm for which this regret is at most $A_x^2(k+1)\ln(n)+4A_x^2k\ln(n)+O(k+1)$ for any of $\mathcal{T}_{k,n}$, k or n with no knowledge of $\mathcal{T}_{k,n}$, k or n a priori. We recognize the term $A^2(k+1)\ln(n)$ as the parameter regret or additional loss due to the prediction problem with $k+1$ parameters in each of the $k+1$ separate regions and the term $4A^2k\ln(n)$ as the transition path redundancy due to not knowing the best transitions times.

For the more general problem of linear regression algorithms in each segment, a competing algorithm with k transitions and a transition path $\mathcal{T}_{k,n}$ forms its estimate or prediction of $x[n]$ in each segment as

$$\hat{x}[n]=\vec{w}_i^T\vec{y}[n]$$

where $\vec{w}_i=[w_{i,1},\ldots,w_{i,p}]^T$, $\vec{w}\in R^p$ and $i=1,\ldots,k+1$, $\vec{y}[t]=[y_1[t],\ldots,y_p[t]]^T$, $\vec{y}[t]\in[-A_y,A_y]^p$, $|x[n]|\le A_x$. We again identify the best competing algorithm as the one optimized by selecting the transition path $\mathcal{T}_{k,n}$, the number of transitions k and the constants $\vec{w}_i$ based on observing the entire sequence x^n and $\vec{y}^n$ in advance. As such we try to minimize the regret

$$\sup_{x^n,\vec{y}^n}\left\{\sum_{t=1}^{n}(x[t]-\hat{x}_q[t])^2-\inf_{\substack{\vec{w}_1,\ldots,\vec{w}_{k+1}\\ t_1,\ldots,t_k}}\sum_{i=1}^{k+1}\sum_{t=t_{i-1}}^{t_i-1}(x[t]-\hat{x}_{\vec{w}_i}[t])^2\right\}$$

where $\hat{x}_q[t]$ is the prediction at time t of any sequential algorithm and $\hat{x}_{\vec{w}_i}[t]=\vec{w}_i^T\vec{y}[t]$. In linear prediction, the observation sequence $\vec{y}[t]$ is formed by the past observations, i.e., $\vec{y}[t]=[x[t-1],\ldots,x[t-p]]^T$. We will construct a sequential algorithm for which this regret is at most $A_x^2p(k+1)\ln(n)+4A_x^2k\ln(n)+O(k+1)$ with no prior knowledge of $\mathcal{T}_{k,n}$, k or n. We recognize the term $A_x^2p(k+1)\ln(n)$ as the parameter regret due to the prediction problem in $k+1$ regions and $4A_x^2k\ln(n)$ as the transition path regret due to not knowing the best transitions times.

A. Upper Bounds

The main results of this section are the upper bounds contained in Theorem 1 and Theorem 2. Proofs of the theorems are outlined here and provided in more detail in [3]. The proofs are constructive, yielding the corresponding universal sequential algorithms.

Theorem 1 : *Let $x[n]$ be a bounded, real-valued arbitrary sequence, such that $|x[n]|<A$, for all n. Then we can construct a sequential algorithm $\tilde{x}_u[n]$ such that for all $\epsilon>0$ and any k*

$$R_c[n]=\left\{\sum_{t=1}^{n}(x[t]-\tilde{x}_u[t])^2-\inf_{\substack{c_1,\ldots,c_{k+1}\in R\\ 1=t_0<t_1<\ldots<t_n<t_{k+1}=n+1\\ t_1,\ldots,t_k\in Z}}\sum_{i=1}^{k+1}\sum_{t=t_{i-1}}^{t_i-1}(x[t]-c_i)^2\right\}$$

satisfies

$$\frac{R_c[n]}{n}\le A^2(k+1)\frac{\ln(n)}{n}+4A^2(k+\epsilon)\frac{\ln(n)}{n}+O\left(\frac{k}{n}\right)\qquad(1)$$

and

$$\frac{R_c[n]}{n}\le A^2(k+1)\frac{\ln(n/k)}{n}+4A^2\left((k+1)\epsilon+k\right)\frac{\ln(n/k)}{n}+O\left(\frac{k}{n}\right)\qquad(2)$$

for any $\mathcal{T}_{k,n}$ representing transition path $(t_1,\ldots,t_k)$ and any k, such that $\tilde{x}_u[t]$ does not depend on $\mathcal{T}_{k,n}$, k or n.

The upper bound in Equation (1) is better (tighter) when the number of transitions is small, i.e., $O(1)$. If the number of transitions is closer to n, then the upper bound in Equation (2) is better. Theorem 1 states that the average squared prediction error of the universal predictor $\tilde{x}_u[t]$ is within $O((k+1)n^{-1}\ln(n))$ of the best batch piecewise constant

prediction algorithm with k transitions (tuned to the underlying sequence), uniformly, for every individual sequence x^n.

For piecewise linear prediction, we have the following result:

Theorem 2 : *Let $x[n]$ and $\vec{y}[n]$, be bounded, real-valued arbitrary scalar and vector sequences, such that $|x[n]| \leq A_x$ and $|y_s[n]| \leq A_y$, $s = 1, \ldots, p$, for all n. Then we can construct a sequential algorithm $\tilde{x}_u[n]$ such that for all $\delta > 0$, $\epsilon > 0$ and $k > 0$*

$$R_w[n] = \left\{ \sum_{t=1}^{n} (x[t] - \tilde{x}_u[t])^2 - \inf_{\substack{\vec{w}_1, \ldots, \vec{w}_{k+1} \in R^p \\ 1 = t_0 < t_1 < \ldots < t_n < t_{k+1} = n+1 \\ t_1, \ldots, t_k \in Z}} \sum_{i=1}^{k+1} \left(\sum_{t=t_{i-1}}^{t_i - 1} (x[t] - \vec{w}_i^T \vec{y}[t])^2 + \delta \|\vec{w}_i\|^2 \right) \right\}$$

satisfies

$$\frac{R_w[n]}{n} \leq \tag{3}$$
$$pA_x^2(k+1)\frac{\ln(A_y^2 n/\delta)}{n} + 4A_x^2(k+\epsilon)\frac{\ln(n)}{n} + O\left(\frac{k}{n}\right)$$

and

$$\frac{R_w[n]}{n} \leq pA_x^2(k+1)\frac{\ln(A_y^2 n/(k\delta))}{n} \tag{4}$$
$$+ 4A_x^2\left((k+1)\epsilon + k\right)\frac{\ln(n/k)}{n} + O\left(\frac{k}{n}\right)$$

for any $\mathcal{T}_{k,n}$ representing transition path $(t_1, \ldots, t_k)$ and any k, such that $\tilde{x}_u[t]$ does not depend on $\mathcal{T}_{k,n}$, k or n.

The upper bound in Equation (3) is better when the number of transitions is small, i.e., $O(1)$. If the number of transitions is closer to $O(n)$, then the upper bound in Equation (4) becomes tighter. Theorem 2 states that the average squared prediction error of the universal linear regressor is within $O(p(k+1)n^{-1}\ln(n))$ of the best batch piecewise linear pth-order linear regression algorithm with k transitions (tuned to the underlying sequence), uniformly, for every individual sequence x^n and vector sequence $\vec{y}^n$.

Outline of the Proofs of Theorems 1 and 2:
The proof uses ideas from sequential probability assignment. For each possible transition path $\mathcal{T}_{k,n}$ representing $(t_1, \ldots, t_k)$ with k transitions and data length n, we consider a family of predictors, each with its own vector $\vec{c} = [c_1, \ldots, c_{k+1}]^T$ where each c_i represents a constant prediction for the ith region. For each pairing of $\mathcal{T}_{k,n}$ and $\vec{c}$, a measure of the sequential prediction performance or loss of the corresponding competition algorithm is constructed, $l_n(x, \hat{x}_{\vec{c}} \mid \vec{c}, \mathcal{T}_{k,n}) \stackrel{\triangle}{=} \sum_{i=1}^{k+1} \sum_{t=t_{i-1}}^{t_i - 1} (x[t] - c_i)^2$. We then define a function of the loss namely the "probability"

$$P(x^n \mid \vec{c}, \mathcal{T}_{k,n}) \stackrel{\triangle}{=} \exp\left(-\frac{1}{2a} l_n(x, \hat{x}_{\vec{c}} \mid \vec{c}, \mathcal{T}_{k,n})\right),$$

which can be viewed as a probability assignment of a predictor with transition path $\mathcal{T}_{k,n}$ and with parameters $\vec{c}$, to the data $x[t]$, for $1 \leq t \leq n$, induced by the performance of the corresponding algorithm with $\mathcal{T}_{k,n}$ and $\vec{c}$ on the sequence x^n, where a is a positive constant. Our goal is to demonstrate a sequential algorithm which achieves this probability given any k and n, and without a priori knowledge of k or n. We will accomplish this result using a double mixture approach. First, we will demonstrate an algorithm achieving the performance of the competing algorithm with the best constant predictors in each region given any $\mathcal{T}_{q,n}$. Then we will show that a proper weighted combination of all such algorithms over all $\mathcal{T}_{q,n}$, $q = 1, \ldots, n$, can be used to find a sequential algorithm that will achieve the performance of the best algorithm that can choose k, $\mathcal{T}_{k,n}$ based on observing the whole sequence x^n.

To do this, given a $\mathcal{T}_{k,n}$, we first construct a sequential algorithm with performance $\tilde{P}(x^n|\mathcal{T}_{k,n})$ that achieves the performance of the best predictor given $\mathcal{T}_{k,n}$. We then construct a universal estimate of the probability of the sequence x^n as a performance weighted mixture of the probabilities assigned by all such sequential predictors over all possible $\mathcal{T}_{k,n}$ and k

$$P_u(x^n) \stackrel{\triangle}{=} \sum_{k=0}^{n-1} \sum_{\mathcal{T}_{k,n}} P(\mathcal{T}_{k,n}) \tilde{P}(x^n \mid \mathcal{T}_{k,n}), \tag{5}$$

with a suitable prior over the partitions $\mathcal{T}_{k,n}$, $P(\mathcal{T}_{k,n})$. For any transition path $\mathcal{T}_{k,n}$, the weighting or the assigned probability $P(\mathcal{T}_{k,n})$ should be nonnegative and should satisfy $\sum_{k=0}^{n-1} \sum_{\mathcal{T}_{k,n}} P(\mathcal{T}_{k,n}) = 1$. The universal probability obtained here is as large as the probability assigned to the sequence by the predictor with the smallest prediction error. We now must relate this universal probability to an actual prediction by combining the exponential number of such sequential predictors. We accomplish this with using a linear transition diagram as used in [1]. The transition diagram effectively represents an exponential number of transition paths with linear complexity by grouping all paths $\mathcal{T}_{k,n}$ together that share the same state at any given time. ■ The proof of Theorem 2 follows a similar course, where the regret within each region is obtained as in [2] for the competition class of all fixed-order linear regressors within each segment. ■

III. Piecewise Adaptive Filtering

In this section, we investigate the case where we compete against a finite set of static algorithms in each region. Given a real valued, bounded sequence $x[t]$, $t = 1, \ldots, n$, we now consider a class of only M algorithms producing estimates $\hat{x}_m[t]$, $m = 1, \ldots, M$, $t = 1, \ldots, n$. For an observation sequence $x^n = \{x[1], \ldots, x[n]\}$ and outputs of these algorithms $\hat{x}_m^n = \{\hat{x}_m[1], \ldots, \hat{x}_m[n]\}$, $m = 1, \ldots, M$ in the static class, a transition path $\mathcal{T}_{k,n}$ with k transitions, represented by $(t_1, \ldots, t_k)$, divides x^n and $\hat{x}_m^n$ into $k+1$ segments such that x^n and each $\hat{x}_m^n$ can be represented as concatenation of

$$\{x[1], \ldots, x[t_1 - 1]\}\{x[t_1], \ldots, x[t_2 - 1]\} \ldots \{x[t_k], \ldots, x[n]\},$$

and

$$\{\hat{x}_m[1],\ldots,\hat{x}_m[t_1-1]\}\ldots\{\hat{x}_m[t_k],\ldots,\hat{x}_m[n]\},$$

respectively. A "switching" class of competing algorithms could then independently select a different algorithm from the class of M algorithms for each segment.

Given any $\mathcal{T}_{k,n}$, the best switching-based algorithm, i.e. the one with minimum total square error, would choose within each segment the algorithm of the M algorithms whose total square error in that segment was minimum. Here, we demonstrate a sequential algorithm with no knowledge of k, $\mathcal{T}_{k,n}$ or n a priori, that achieves the performance of

$$\sum_{i=1}^{k+1}\sum_{t=t_{i-1}}^{t_i-1}(x[t]-\hat{x}^i[t])^2=\sum_{i=1}^{k+1}\min_{j=1,\ldots,M}\sum_{t=t_{i-1}}^{t_i-1}(x[t]-\hat{x}_j[t])^2$$

for any transition path $\mathcal{T}_{k,n}$, where $\hat{x}^i[t]$ is the best algorithm for the ith segment such that $\hat{x}^i[t]=\hat{x}_m[t]$ if

$$\sum_{t=t_{i-1}}^{t_i-1}(x[t]-\hat{x}_m[t])^2\le\sum_{t=t_{i-1}}^{t_i-1}(x[t]-\hat{x}_j[t])^2, \qquad (6)$$

for $j\neq m$, $j=1,\ldots,M$.

Theorem 5: *Let $x[n]$ be a bounded, real-valued arbitrary sequence, such that $|x[n]|<A$, for all n and $\hat{x}_m[n]$, $m=1,\ldots,M$, are predictions of arbitrary algorithms at time n. Then we can construct a sequential algorithm $\tilde{x}_u[n]$ such that for all $\epsilon>0$*

$$\sum_{t=1}^{n}(x[t]-\tilde{x}_u[t])^2-\sum_{i=1}^{k+1}\sum_{t=t_{i-1}}^{t_i-1}(x[t]-\hat{x}^i[t])^2\le$$
$$2A^2(k+1)\ln(M)+4A^2(k+\epsilon)\ln(n)+O(k+1), \qquad (7)$$

and

$$\sum_{t=1}^{n}(x[t]-\tilde{x}_u[t])^2-\sum_{i=1}^{k+1}\sum_{t=t_{i-1}}^{t_i-1}(x[t]-\hat{x}^i[t])^2\le \qquad (8)$$
$$2A^2(k+1)\ln(M)+4A^2\left((k+1)\epsilon+k\right)\ln(n/k)+O(k+1).$$

for any $\mathcal{T}_{k,n}$ represented by $(t_1,\ldots,t_k)$ and k, without any knowledge of $\mathcal{T}_{k,n}$, k or n apriori, where $x^i[t]$ is the prediction of the best algorithm in the sense of Equation (6) *and $\tilde{x}_u[t]$ does not depend on $\mathcal{T}_{k,n}$, k or n.*

The proof follows from application of the mixture over all transition paths, using the linear transition diagram approach applied in Theorem 1 of this paper, and application of Theorem 1 in [8] within each segment. A linear complexity lattice-filter based algorithm [4] can achieve

$$\sum_{t=1}^{n}(x[t]-\tilde{x}_u[t])^2-\sum_{i=1}^{k+1}\sum_{t=t_{i-1}}^{t_i-1}(x[t]-\hat{x}^i[t])^2$$
$$\le 4A^2(k+1)\ln(M)+2A^2(k+\epsilon)\ln(n)+O(k+1),$$

when the M algorithms are linear predictors of order 1 through M.

REFERENCES

[1] 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

[2] 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

[3] S.S. Kozat, "Competitive Signal Processing," Ph.D. Thesis, University of Illinois at Urbana-Champaign, 2004.

[4] A.C. Singer, M. Feder, "Universal linear prediction by model order weighting," *IEEE Transactions on Signal Processing*, vol. 47, no. 10, October 1999.

[5] 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

[6] D. Luengo, S.S. Kozat, A. C. Singer, "Universal Piecewise Linear Least Squares Prediction," *Iternational Symposium on Information Theory*, Chicago 2004

[7] A.C. Singer, S.S. Kozat,"Universal Context Tree Least Squares Prediction, submitted to *Iternational Symposium on Information Theory*, 2006.

[8] V. Vovk,"Aggregating strategies," *COLT*, 1990, pp.371-383.

[9] M. Herbster, M. K. Warmuth, "Tracking the best linear predictor," *Journal of Machine Learning Research 1* , pp. 281-309, Sept. 2001

[10] M. Herbster and M. K. Warmuth,"Tracking the Best Expert," *International Conference on Machine Learning*,pp. 286-294, 1995.

[11] V. Vovk, "Derandomizing stochastic prediction strategies," *Machine Learning*, 35, 247-282, 1999.

[12] N. Merhav, "On the minimum description length principle for sources with piecewise constant parameters," *IEEE Transactions on Information Theory*, vol. 41, pp. 1962-1967, Nov. 1993