Finite-Memory Least Squares Universal Prediction of Individual Continuous Sequences

Ronen Dar
Dept. of EE-Systems, TAU
Tel Aviv, Israel
Email: ronendar@post.tau.ac.il

Meir Feder
Dept. of EE-Systems, TAU
Tel Aviv, Israel
Email: meir@eng.tau.ac.il

Abstract—In this paper we consider the problem of universal prediction of individual continuous sequences with square-error loss, using a deterministic finite-state machine (FSM). The goal is to attain universally the performance of the best constant predictor tuned to the sequence, which predicts the empirical mean and incurs the empirical variance as the loss. The paper analyzes the tradeoff between the number of states of the universal FSM and the excess loss (regret). We first present a machine, termed Exponential Decaying Memory (EDM) machine, used in the past for predicting binary sequences, and show bounds on its performance. Then we consider a new class of machines, Degenerated Tracking Memory (DTM) machines, find the optimal DTM machine and show that it outperforms the EDM machine for a small number of states. Incidentally, we prove a lower bound indicating that even with large number of states the regret of the DTM machine does not vanish. Finally, we show a lower bound on the achievable regret of any FSM, and suggest a new machine, the Enhanced Exponential Decaying Memory, which attains the bound and outperforms the EDM for any number of states.

Index Terms—Universal prediction, individual continuous sequences, finite-memory, least-squares.

I. INTRODUCTION

Consider a continuous-valued individual sequence \(x_1, x_2, \ldots, x_t, \ldots\), where each sample is assumed to be bounded in the interval \([a, b]\) but otherwise arbitrary with no underlying statistics. At each time \(t\), after observing \(x_1\), a predictor guesses the next outcome \(\hat{x}_{t+1}\), and incurs a square error prediction loss \((x_{t+1} - \hat{x}_{t+1})^2\). Suppose one can tune a (non-universal) predictor to the sequence, from a given class of predictors. For example, the best constant predictor for a given sequence, i.e., a predictor that uses a constant prediction for all the sequence outcomes, is the empirical mean \(\bar{x} = \frac{1}{n} \sum_{t=1}^{n} x_t\). The square error loss incurred by this predictor is the sequence’s empirical variance \(\frac{1}{n} \sum_{t=1}^{n} (x_t - \bar{x})^2\). Thus, for a given sequence \(x_1^n\), the excess loss of a universal predictor \(U\) (that predicts \(\hat{x}_{u,1}, \ldots, \hat{x}_{u,n}\) over the best constant predictor is termed the regret of the sequence w.r.t \(U\):

\[
R(U, x_1^n) = \frac{1}{n} \sum_{t=1}^{n} (x_t - \hat{x}_{u,t})^2 - \frac{1}{n} \sum_{t=1}^{n} (x_t - \bar{x})^2. \tag{1}
\]

In the individual setting, we analyze the performance of a universal predictor by the maximal excess loss, that is, the incurred regret of the worst sequence. An extensive survey on universal prediction is given in [1]. Some aspects of the universal prediction problem for individual continuous sequences with square error loss were already explored by Merhav and Feder in [2]. That work actually considered a more general case and showed that the Recursive Least Squares (RLS) algorithm [3], [4] generates a universal predictor that attains the performance of the best (non-universal) \(L\)-order linear predictor [5] tuned to the sequence. When specialized to the case of “zero-order”, i.e. the case where the non-universal predictor is the constant empirical mean predictor, the resulting universal predictor is the Cumulative Moving Average (CMA):

\[
\hat{x}_{t+1} = (1 - \frac{1}{t+1}) \hat{x}_t + \frac{1}{t+1} x_t, \tag{2}
\]

where \(\hat{x}_t\) is the prediction at time \(t\). The regret of this predictor tends to zero with the sequence length \(n\).

Note that while the reference non-universal constant predictor needs a single state, the universal predictor (2) requires an ever growing amount of memory. What happens when the universal predictor is constrained to be a finite \(k\)-state machine? Universal estimation and prediction problems where the estimator/predictor is a \(k\)-state machine have been explored extensively in the past years. Cover [6] studied hypothesis testing problem where the tester has a finite memory. Hellman [7] studied the problem of estimating the mean of a Gaussian (or more generally stochastic) sequence using a finite state machine. This problem is closely related to our problem and may be considered as a stochastic version of it: if one assumes that the data is Gaussian than predicting it with a minimal mean square error essentially boils to estimating its mean. More recently, the finite-memory universal prediction problem for individual binary sequences with various loss functions was explored thoroughly in [8]–[13]. The finite-memory universal portfolio selection problem (that dealt with continuous-valued sequences but considered a very unique loss function) was also explored recently [14]. Yet, the basic problem of finite-memory universal prediction of continuous-valued, individual sequences with square error loss was left unexplored so far. This paper provides a solution for this problem, presenting such universal predictors attaining a vanishing regret when a large memory is allowed, but also maintaining an optimal tradeoff between the regret and the number of states used by the universal predictor.
The outline of the paper is as follows. We start with general definitions and guidelines in section II. In section III we propose the EDM machine, proving bounds on the worst regret. Section IV is devoted to universal prediction with a small number of states, where we present the class of DTM machines. Sections V and VI are devoted to universal prediction using a large number of states - in section V we present an asymptotic lower bound on the achievable regret of any deterministic k-states machine, where in section VI we present the E-EDM machine for any vanishing desired regret. For further information and more detailed proofs, see [15].

II. DEFINITIONS

Finite-state machine (FSM) is a commonly used model for sequential machines with a limited amount of storage. In our work we focus on time-invariant FSM.

Definition 1: A deterministic FSM is defined by an array of k states where \( \{S_1, \ldots, S_k\} \) denote the value assigned to each state. The prediction of the machine at time \( t \), denoted \( \hat{x}_t \), is the value assigned to the current state. The transition of the machine between states is defined in the continuous case by the maximum up and down steps from each state \( i \), denoted \( m_{u,i} \) and \( m_{d,i} \), correspondingly, and by a thresholds set \( \{T_{i,-m_{d,i}-1}, T_{i,-m_{d,i}}, \ldots, T_{i,m_{u,i}-1}, T_{i,m_{u,i}}\} \) for each state \( i \). Thus, if at time \( t \) the machine is at state \( i \), it jumps \( j \) states \( (-m_{d,i} \leq j \leq m_{u,i}) \) if the input sample \( x_t \) satisfies \( T_{i,j-1} \leq x_t < T_{i,j} \). Note that the thresholds are non-intersecting, where the union of them covers the interval \([a, b]\) (each input sample is assumed to be bounded in \([a, b]\)).

Throughout this paper we discuss predictors designed for input samples that are bounded in \([0, 1]\). It is easily notable that any FSM designed to achieve a regret smaller than \( R \) for any sequence bounded in \([0, 1]\), can be transformed into a FSM that achieves a regret smaller than \((b - a)^2 R\) for any sequence bounded in \([a, b]\), where \( a, b \in \mathbb{R} \), by applying a simple transformation - each state value \( S_i \) is transformed into \( a + (b - a)S_i \) and each thresholds set \( \{T_{i,j}\} \) is transformed to \( a + (b - a)T_{i,j} \). Since the transformation is bijective, all results presented in this paper can be expanded to the more general case, where each individual sequence is assumed to be bounded in \([a, b]\).

We further present a theorem that will use throughout this paper.

Definition 2: A circle is a cyclic closed set of L states/predictions \( \{\hat{x}_t\}_{t=1}^L \), if there are input samples \( \{x_t\}_{t=1}^L \) that rotate the machine between these states. A minimal circle is a circle that does not contain the same state more than once.

Theorem 1: The worst sequence for a given FSM takes the machine to a minimal circle and rotates in it endlessly.

Proof: The proof for the binary case w.r.t the log-loss function is given in details in [16]. The proof in the continuous case w.r.t the square-error loss is exactly the same.

III. THE EXPONENTIAL DECAYING MEMORY MACHINE

In [17] the Exponential Decaying Memory (EDM) machine has been presented as a universal predictor for individual binary sequences. It was further shown that the EDM machine with \( k \) states achieves an asymptotic regret of \( O(k^{-2/3}) \) compared to the constant predictors class w.r.t the log-loss (code length) and square-error functions.

We start by describing and adjusting the EDM machine for our case, predicting individual continuous sequences:

Definition 3: The Exponential Decaying Memory (EDM) machine is defined by \( k \) states \( \{S_1, \ldots, S_k\} \) distributed uniformly over \([k^{-1/3}, 1 - k^{-1/3}]\) axis. Hence, the spacing gap, denoted \( \Delta \), satisfies:

\[
\Delta = \frac{1 - 2k^{-1/3}}{k^{-1}} \sim k^{-1}
\]

Let \( \hat{x}_t \) be the prediction at time \( t \) (recalling that if \( \hat{x}_t = S_i \), then the machine is at state \( i \) at time \( t \)).

The transition function between states is defined by:

\[
\hat{x}_{t+1} = Q(\hat{x}_t(1 - k^{-2/3}) + x_t k^{-2/3})
\]

where \( Q \) is the quantization function to the nearest state, i.e. \( Q(y) = \hat{x}_{t+1}, \) if \( y \) satisfies \( \hat{x}_{t+1} - \frac{1}{2}\Delta \leq y < \hat{x}_{t+1} + \frac{1}{2}\Delta \).

We now present asymptotic bounds on the regret achieved by the EDM machine when used to predict individual continuous sequences.

Theorem 2: The maximal regret of the \( k \)-states EDM machine (denoted \( U_{EDM_k} \)), attained by the worst continuous sequence, is asymptotically bounded by:

\[
\frac{1}{2} k^{-2/3} + O(k^{-1}) \leq \max R(U_{EDM_k}, x^n_t) \leq \frac{17}{4} k^{-2/3}
\]

Proof: The input sample at each time \( t \) can be written as follows:

\[
x_t = \hat{x}_t + (P_t\Delta + \delta_t)k^{2/3}
\]

where \( P_t \in \mathbb{Z} \) denotes the number of states crossed by the machine at time \( t \), \( \delta_t \) is a quantization addition that satisfies \( |\delta_t| < \frac{1}{2}\Delta \) and has no impact on the jump at time \( t \), i.e. has no impact on the prediction at time \( t+1 \). We prove that the regret of any sequence that endlessly rotates the EDM machine in a minimal circle of \( L \) states is upper bounded by:

\[
R(U_{EDM_k}, x^n_t) \leq \text{quantization loss + spacing loss },
\]

where the quantization loss depends only on the quantization of the input samples, \( \delta_t \), and the spacing loss depends only on the spacing gap between states, \( \Delta \). Hence, the regret of the sequence is upper bounded by a loss incurred by the quantization of the input samples and a loss incurred by the quantization of the states’ values, i.e. the prediction values. It can be shown that the quantization loss is no more than \( \frac{1}{2} k^{-2/3} \) and the spacing loss is no more than \( 4k^{-2/3} \). Thus, by using Theorem 1, the upper bound is proven. To prove the lower bound we find a sequence that endlessly rotates the EDM machine in a minimal circle and achieves regret \( \frac{1}{2} k^{-2/3} + O(k^{-1}) \).

IV. DESIGNING AN OPTIMAL FSM WITH A SMALL NUMBER OF STATES

In this section we want to design a finite-state universal predictor using a small number of states. We start by defining a new class of machines.
**Definition 4:** The class of all k-states Degenerated Tracking Memory (DTM) machines is of the form:

- An array of k states.
- The maximum down-step in the lower half, i.e. from all states satisfying \( S_i \leq \frac{1}{2} \), is no more than a single state jump. The maximum up-step in the upper half, i.e. from all states satisfying \( S_i \geq \frac{1}{2} \), is no more than a single state jump.
- A transition between the lower and upper halves is allowed only from and to the nearest state to \( \frac{1}{2} \) in each half.

**A. Building the optimal DTM machine**

Given a desired regret, \( R_d \), the task of finding the optimal DTM machine can be viewed as a covering problem, meaning assigning the smallest number of states over the \([0, 1]\) axis achieving a regret lower than \( R_d \) for all sequences. At [15] we present and prove an algorithm for finding the optimal DTM machine. Here we present an outline of the algorithm. Recursively, the algorithm finds the optimal states allocation for the lower half, \([0, \frac{1}{2}]\). By the symmetry property of the optimal DTM machine, the optimal upper half is the mirror image of the lower half. By definition, a transition from the lower half to the upper half is allowed only from one state, denoted \( S_1 \), which is the nearest state to \( \frac{1}{2} \). We show that \( S_1 = \frac{1}{2} \) in the optimal DTM machine with odd number of states, while for even number of states

\[
S_1 = \max \left\{ 1 - \sqrt{R_d + \frac{1}{4}}, 2 + \sqrt{R_d} - 2\sqrt{R_d + \sqrt{R_d + \frac{1}{2}}} \right\}. \tag{7}
\]

Given states \( \{S_{i-1}, ..., S_1\} \) in the lower half (in descending order) and their thresholds set \( \{T_{i-1}, ..., T_1\} \), satisfying a regret smaller than \( R_d \) for all minimal circles between them, the algorithm finds the minimum \( S_i \) and a thresholds set \( T_i \), satisfying a regret smaller than \( R_d \) for all minimal circles starting at that state. Since a transition from the lower half to the upper half is allowed only from \( S_1 \), all minimal circles starting at state \( i \) are within the lower half, for any \( i > 1 \). The algorithm finishes when \( S_i \leq \sqrt{R_d} \). Thus, given a desired regret, one should run the algorithm twice - for odd and even number of states with the corresponded starting state, \( S_1 \). The optimal DTM machine is the one with the least states among the two (differ by a single state).

**B. Lower bound - The Limitation of the DTM Machines**

**Theorem 3:** The achievable regret of any DTM machine is lower bounded by

\[
R \geq \left( \frac{1}{6} \right)^2 = 0.0278 .
\]

**Proof:** See [15].

**C. Numerical results**

Figure 1 shows numerical results (number of states vs. regret) of the optimal DTM machine and the asymptotic EDM machine (regret of \( \frac{1}{4}k^{-2/3} \)). While the EDM machine can achieve any vanishing regret with large enough number of states, the lower bound for the DTM machine is depicted - as the number of states grows, the achievable regret goes to 0.0278.

In [15] we further show that the optimal DTM machine with a single, two and three states is the optimal solution for these machines, concluding that up to a certain number of states, our algorithm generates the optimal solution in sense of achieving the lowest maximal regret for a given number of states. Yet, it is unresolved up to which number of states.

![Fig. 1. The performance of the optimal DTM machine and the EDM machine.](image)

**V. LOWER BOUND ON THE ACHIEVABLE REGRET OF ANY k-STATES MACHINE**

In this section we present a lower bound on the achievable regret of any \( k \)-states universal predictor for continuous-valued sequences. We start with the following Lemma which lower bounds the farthest up and down jumps from any state in a FSM.

**Lemma 1:** Consider a FSM that achieves a maximal regret \( R \). The maximum number of states crossed in an up-step and in a down-step from state \( S_i \), for any \( i \), must satisfy

\[
m_{u,i} \geq 1 - \frac{(S_i + \sqrt{R})}{2\sqrt{R}} ,
\]

\[
m_{d,i} \geq \frac{S_i - \sqrt{R}}{2\sqrt{R}} . \tag{9}
\]

**Proof:** See [15].

Note that Lemma 1 implies the same lower bound on the achievable regret of any DTM machine, \( R \geq \left( \frac{1}{6} \right)^2 \) (as presented in section IV). Any DTM machine allows only a single state down-jump from all states below \( \frac{1}{2} \). Thus, a DTM machine can achieve regret \( R \) if all states below \( \frac{1}{2} \) satisfy Equation (9) with \( m_{d,i} = 1 \), hence:

\[
\frac{\frac{1}{2} - \sqrt{R}}{2\sqrt{R}} \leq 1 . \tag{10}
\]

Furthermore, Lemma 1 provides a lower bound on the maximal regret of any machine that allocates a state \( S_i \) with maximum up and down jumps of \( m_{u,i} \) and \( m_{d,i} \) states.
Theorem 4: The number of states in any deterministic FSM that achieves a regret smaller than \( R \) for any continuous sequence, is lower bounded by

\[
\frac{1}{2\pi} R^{-3/2} + O(R^{-1}) .
\]

Proof: See [15].

VI. ENHANCED EXPONENTIAL DECAYING MEMORY MACHINE

In this section we present a new FSM named the Enhanced Exponential Decaying Memory (E-EDM) machine, targeting to achieve any vanishing desired regret.

A. Designing the E-EDM machine

The algorithm for constructing the E-EDM machine for a desired regret, denoted \( R_d \), is as follows:

Set \( R = \frac{R_d}{2} \) and divide the \([0,1] \) axis into segments, where a segment \( A(m_u, m_d) \) is defined as the set of all \( x \)'s satisfying:

\[
m_u = \left[ \frac{1-x-\sqrt{x}}{2\sqrt{R}} \right] \quad \forall \ x \in A(m_u, m_d) ,
\]

\[
m_d = \left[ \frac{x-\sqrt{x}}{2\sqrt{R}} \right] \quad \forall \ x \in A(m_u, m_d) . \quad (11)
\]

All states in segment \( A(m_u, m_d) \) are assigned with a maximum up and down jump of \( m_u, m_d \) states, accordingly. Linearly spread states in each segment \( A(m_u, m_d) \) with a \( \Delta(m_u, m_d) \) spacing gap between them where

\[
\Delta(m_u, m_d) = \frac{\sqrt{R}}{2m_u - m_d} . \quad (12)
\]

We further need to guarantee the desired regret when the machine traverses between segments. Consider two adjacent segments \( A(m_{u,1}, m_{d,1}) \) and \( A(m_{u,2}, m_{d,2}) \) and suppose the spacing gap in the second segment is smaller. Add states to the first segment such that the closest \( m_{u,1} + m_{d,1} \) states to the second segment have a spacing gap of \( \Delta(m_{u,2}, m_{d,2}) \). It can be shown that at most two states need to be added to each segment. Finally, Assign transition thresholds for each state \( i \) as follows:

\[
T_{i,j} = S_i + (2j + 1)\sqrt{R} \quad \forall \quad m_{d,i} \leq j \leq m_{u,i} , \quad (13)
\]

that is, if the machine at time \( t \) is at state \( i \), it jumps \( j \) states if the current outcome, \( x_t \), satisfies:

\[
S_i + (2j - 1)\sqrt{R} \leq x_t < S_i + (2j + 1)\sqrt{R} . \quad (14)
\]

Note that as required, the transition thresholds cover the \([0,1] \) axis (arises from the chosen maximum up and down jumps).

Theorem 2 implies that the maximal regret of the \( k \)-states EDM machine is at least \( \frac{1}{2} R^{-2/3} \). Note that if equality holds, the definitions of the EDM machine, excluding the part of allocating states, are identical to the definitions of the E-EDM machine. Thus, the new machine presented here can be regarded as an improvement of the EDM machine by better allocating the states - the states of the EDM are uniformly distributed over the interval \([0,1] \) while in the E-EDM machine the interval \([0,1] \) is divided into segments and states are uniformly distributed with a different spacing in each segment.

Theorem 5: The regret of the E-EDM machine is smaller than \( R_d \) for any input sequence.

Proof: With similar technics to those presented in the proof of Theorem 2, we can show that also for the E-EDM machine (denoted \( U_e-EDM \)), the regret of any minimal circle of \( L \) states satisfies:

\[
R(U_e-EDM, x(t)) \leq \text{quantization loss + spacing loss} . \quad (15)
\]

Each loss can be upper bounded by \( R \), finalizing that the regret of any minimal circle is upper bounded by

\[
R(U_e-EDM, x(t)) \leq 2R = R_d . \quad (16)
\]

Theorem 6: The number of states in an E-EDM machine designed to achieve a regret \( R_d \) is

\[
\frac{1}{12}(\frac{R_d}{2})^{-3/2} + O(R_d^{-1}) .
\]

Proof: See [15].

B. Numerical results

Theorem 2 implies that the asymptotic worst regret of the \( k \)-states EDM machine is at least \( \frac{1}{2} k^{-2/3} \). Thus, the number of states in an EDM machine that achieves a regret \( R_d \), is at least \( (2R_d)^{-3/2} \) states. Theorem 4 implies that the asymptotic number of states of any deterministic FSM that achieves a maximal regret \( R_d \) is at least \( \frac{1}{2\pi} R_d^{-3/2} \). Theorem 6 implies that the asymptotic number of states in an E-EDM machine that achieves a regret \( R_d \) is \( \frac{1}{12}(\frac{R_d}{2})^{-3/2} \). Thus we can conclude that for a given desired regret, the E-EDM machine outperforms the EDM machine in number of states by a factor of:

\[
\frac{2^{3/2} R_d^{-3/2}}{(2\pi)^{-3/2}} = \frac{2}{3} ,
\]

i.e. uses only \( \frac{2}{3} \) of the states needed for the EDM machine to achieve the same maximal regret. For a given desired regret, the E-EDM machine approaches the lower bound with a factor of about:

\[
\frac{2^{3/2} R_d^{-3/2}}{(2\pi)^{-3/2}} = 2^{5/2} = 5.6 .
\]

Simulation results are presented in Figure 2. Note that the theoretical results match the numerical results.

VII. SUMMARY AND CONCLUSIONS

In this paper we studied the problem of universal least squares prediction of individual continuous-alphabet sequences when limited resources are available.

For universal predictors with a small number of states, or equivalently for high desired regret, we presented the optimal Degenerated Tracking Memory (DTM) machine, which performs well with a small number of states yet its achievable regret is lower bounded by \( R = 0.0278 \). Numerical results showed that the optimal DTM machine indeed outperforms
Fig. 2. Comparing the performance of the E-EDM machine, the EDM machine and the lower bound.

any other machine for a small enough number of states. However, it is still unknown up to which number of states it is the best universal predictor. For number of states larger than that, one can try to attain better performance by easing the constrains of the DTM machines and allowing more than a single state down-jump (up-jump) from all states in the lower (upper) half. However, the construction of the optimal machine in this case is much more complex.

For universal predictors with a large number of states, or equivalently for any vanishing desired regret, we proved a lower bound of $O(k^{-2/3})$ on the achievable regret of any $k$-state machine. We proposed the Exponential Decaying Memory (EDM) machine and showed that the worst sequence incurs a bounded regret of $O(k^{-2/3})$, where $k$ is the number of states. We further presented the Enhanced Exponential Decaying Memory (E-EDM) machine which outperforms the EDM machine. The E-EDM machine can be regarded as an improvement of the EDM machine by better allocating the states over the interval $[0, 1]$. Recalling that the EDM machine is a finite-memory approximation of the Cumulative Moving Average predictor which is the best unlimited resources universal predictor (w.r.t the non-universal empirical mean predictor), we can understand why both the EDM and the E-EDM machines approach optimal performance.

Analyzing the performance of the EDM and the E-EDM machines showed that the regret of any sequence can be upper bounded by the sum of two losses - quantization loss, the loss incurred by the quantization of the input samples, and spacing loss, the loss incurred by the quantization of the prediction values. It is worth mentioning that the worst regret of the optimal DTM machine can also be upper bounded by the sum of these losses. As the number of states in the optimal DTM machine increases, the quantization loss goes to the lower bound, $R = 0.0278$, and the spacing loss goes to zero. Thus, understanding the optimal allocation between these two losses may lead to the answer of up to which number of states the optimal DTM machine is the best universal predictor. It is also worth mentioning that the E-EDM machine is constructed with allocating half of the desired regret to the quantization loss and the other half to the spacing loss. A further optimization may be obtained by a different allocation.

Throughout this paper we assumed that the sequence’s outcomes are bounded. Note that this constraint is mandatory since the performance of a universal predictor is analyzed by the regret of the worst sequence. In the unbounded case, for any finite-memory predictor one can find a sequence that incurs an infinite regret. However, an optional further study is to expand the results we achieved to a more relaxed case, e.g. sequences with a bounded difference between consecutive outcomes.

In this study we essentially examined finite-memory universal predictors trying to attain the performance of the (non-universal) “zero-order” predictor, i.e. trying to attain the empirical variance of any individual continuous sequence. We believe that our work is the first step in the search for the best finite-memory universal predictor trying to attain the performance of the best (non-universal) $L$-order predictor, for any $L$.

REFERENCES