Prediction of Binary Sequences using Finite Memory

Meir Feder and Eitan Federovski
Department of EE-Systems
Tel-Aviv University
Tel-Aviv, 69978, Israel

Introduction

We consider universal prediction of binary sequences, where the criterion is the minimum number of prediction errors. The simplest case of this problem in the probabilistic setting is when the data is generated by an unknown Bernoulli source with, say, P(1)=p and so the goal is to make the expected number of errors close to min{Np,N(1-p)} where N is the sequence length. In the deterministic, individual sequence, setting, the goal is to approach min{N_0,N_1) where N_0 = N-N_1 is the number of zeros in the sequence. In the probabilistic setting this goal is essentially achieved within a constant number (that depends on p, but independent of N) by a majority predictor that counts at each time the number of zeros and ones observed so far, and then predicts the value that has the highest count. In the deterministic setting, the optimal number of errors is attained within a $O(N^{1/2})$ term by a majority predictor whose decision is randomized when the difference between the counts is less than $O(N^{1/2})$. In any case, even in this simple problem the optimal universal predictor requires an infinite state predictor that uses an unbounded counter. See, e.g., [1,2] for further discussion on universal prediction of binary sequences.

It is interesting to explore the performance of the universal predictor under finite memory constraints. The finite memory constraints may come up due to implementation efficiency requirements. In addition, when the sequence behavior is non-stationary, a finite memory predictor which can forget the past and adapt faster to the new statistics is more suitable and has a better prediction performance than the infinite state predictor.

We present and analyze specifically in this work two finite memory predictors:

1. A predictor that uses a saturated counter.
2. A predictor that uses a finite past time window.

The analysis is performed for the Bernoulli case, and for individual sequences. We explicitly provide non-asymptotic results and we try to point out the effect of the memory size M and the interplay between M and the sequence length N. We distinguish in the analysis between the case where the initial counter state is at the origin (i.e. balanced between zeros and ones) and the case where the counter starts at another state. The latter case refers to the non-stationary scenario where the predictor initially is tuned to the past behavior. As will be seen, in this case increasing M may result in a poorer predictor, and there will be an optimal value of M as a function of N.

Main Results:

1. Asymptotic Performance in the Probabilistic Bernoulli Case

Consider a Bernoulli source with P(1)=p. Suppose we use a deterministic M-state machine where in each state we predict either one or zero with probability 1. We partition the states into two sets, one for predicting 1 and the other for 0 . Let the probabilities of the sets be denoted:

$$\Pr(\Omega_0) \ , \ \Pr(\Omega_1) \ , \ \Pr(\Omega_0) + \Pr(\Omega_1) = 1$$

Thus, the mean of the fraction of correct predictions is given by:

$$\Pi = p * \Pr(\Omega_1) + (1-p) * \Pr(\Omega_0) = p * \Pr(\Omega_1) + (1-p) * (1 - \Pr(\Omega_1))$$
$$= 1 - p + (2p-1) * \Pr(\Omega_1) = p + (1-2p) * \Pr(\Omega_0)$$

We use this relation to find the asymptotic mean fraction of correct predictions for the analyzed predictors. In the expression below it is assumed that p>1/2 (symmetric expressions exist for p<1/2, while for p=1/2 any predictor makes errors half of the time). First, recall that results in [3] determine the optimal performance that can be achieved by any M state predictor:

$$\Pr^{Theory}(\Omega_0) = \frac{1}{1 + \left(\frac{p}{1-p}\right)^{M-1}}$$

$$\Pi^{Theory} = p - \frac{2p-1}{1 + \left(\frac{p}{1-p}\right)^{M-1}} \approx p - (2p-1)\left(\frac{1-p}{p}\right)^{M-1}$$

Now, our analysis of the saturated counter predictor, with M states leads to:

$$\Pr^{Sat}(\Omega_0) = \frac{1}{1 + \left(\frac{p}{1-p}\right)^{\frac{M}{2}}}$$

$$\Pi^{Sat} = p - \frac{2p-1}{1 + \left(\frac{p}{1-p}\right)^{\frac{M}{2}}} \approx p - (2p-1)\left(\sqrt{\frac{1-p}{p}}\right)^{M}$$

The asymptotic mean fraction of correct predictions for the finite past-time window predictor, with a window size k is given by:

$$\Pr^{Win}(\Omega_0) = \sum_{i=\frac{k+1}{2}}^{k} \binom{k}{i} p^i (1-p)^{k-i}$$

$$\Pi^{Win} = p + (1-2p) * \sum_{i=\frac{k+1}{2}}^{k} \binom{k}{i} p^i (1-p)^{k-i} \approx p - \frac{1}{\sqrt{2\pi k}} \left(2\sqrt{p(1-p)}\right)^{k+1}$$

Note that the number of states for the finite window predictor is $M=2^K$. Thus, it performs much worse than the saturated counter with the same number of states. The finite window predictor can be approximated, however, by a predictor with M=-k+1 states, in which each state corresponds to, say, the number of ones in the finite window. As the window is shifted the predictor state can be updated probabilistically, as in [4]. Even with this smaller number of states, the finite window predictor is worse than the saturated counter predictor with the same number of states.

2. Non-asymptotic Performance in the Probabilistic Bernoulli Case

We have developed general expressions for finite time analysis of a predictor with finite number of states, in the probabilistic setting. This analysis follows by realizing that the predictor's states progress is governed by a Markov chain, whose transition probabilities depends on p. In a deterministic finite state predictor there are the two sets of states, defined above, corresponding to the two possible predictions. This Markov chain analysis can determine the probability of these two sets, as a function of time and the initial state probabilities.

The Saturated Counter Predictor:

Consider, first, the saturated counter predictor. In the non-stationary scenario we partition the data into segments, of size N, where we assume that p is constant along the segment. The initial predictor state, however, is determined by the previous segment. Thus for the saturated counter, the initial state probability should be given by

$$\pi_0 = \left[1/2, 0, \ldots., 0, \ldots., 0, 1/2\right]$$

which mean that we have equal probability for the polarity of p in the previous segment, implying that the predictor is at one of the extreme states with equal probability at the beginning of the new segment.

Under this assumption we get that the mean fraction of correct predictions, for a segment of size N is given by (up to lower order term)

$$\Pi_N = q + (p-q)\left(1 - \left(\frac{q}{p}\right)^{\frac{M}{2}} - \frac{M}{4N(p-q)}\right) + o\left(\left[\frac{p}{q}\right]^M\right)$$

where p>1/2, and q=1-p. In other words:

$$\Pi_N \approx p - (p-q)\left(\frac{q}{p}\right)^{\frac{M}{2}} - \frac{M}{4N}$$

The expression above show the loss over the optimal fraction p, as a function of M and N. We see that by increasing M the loss over the optimal fraction may not decrease necessarily. We can find the value of M that minimize the loss in the expression above:

$$M^* = 2\frac{\ln N}{\ln\frac{q}{p}} + 2\frac{\ln\left(2(p-q)\ln(p/q)\right)}{\ln\frac{q}{p}}$$

Substituting this value, we get that the approximated mean fraction of prediction errors for this optimal memory size is given by:

$$\Pi_N{}^* = p - \frac{1}{2\ln\frac{q}{p}}\frac{\ln N}{N} - \frac{\ln\left(2e(p-q)\ln(p/q)\right)}{2\ln\frac{q}{p}}\frac{1}{N}$$

Note that the standard analysis of the saturated counter corresponds to the a scenario where the predictor is initialized at the balanced point. This may correspond to a case where if the data is non-stationary the predictor knows that the statistics has been changed and it resets its counts. In this case it can be shown for large M and p >1/2> q that the mean fraction of prediction errors is given by :

$$\Pi_N \approx p - (p-q)\left(\frac{q}{p}\right)^{\frac{M}{2}} - \frac{1}{2(p-1)}\frac{1}{N}$$

We see that the optimal memory in this case is infinite, i.e., increasing the memory improves performance. There are two terms in the loss over the optimal fraction: The term that is exponential in M follows since we use a finite memory. The 1/N follows from the finite segment time.

The Finite Past-time Window Predictor:

In this analysis the window size is k. The interesting case for analyzing the non-stationary scenario is when the initial statistics are at some extreme point. For that we assumed that with probability half the previous k-segment was all zeros, and with probability half it was all ones .

Here we assume again that p>1/2>q=1-p. In the analysis we distinguish between the probability of correct prediction at time i which is less or equal k the window size, and time i>k. In the latter case, the probability of correct prediction is the same for all i>k, and is given by

$$\Pi(i) = 1 - p + (2p-1)\sum_{m=\frac{k+1}{2}}^{k}\binom{k}{m}p^m(1-p)^{k-m} = p - (2p-1)\sum_{m=0}^{\frac{k-1}{2}}\binom{k}{m}p^m(1-p)^{k-m}$$

We can also come up with an expression that sums all the mean fractions of correct prediction for i=0,...,k, given by:

$$\sum_{i=0}^{k}\Pi(i) = \frac{(k+1)}{2} + \frac{2p-1}{2}\frac{2p-1}{2p}(k+1) = (k+1)\left(\frac{1}{2} + \frac{(p-q)^2}{4p}\right)$$

Thus, for a sequence of finite length N , we get that the average mean fraction of correct predictions is given by

$$\Pi_N = \frac{1}{N}\sum_{i=0}^{N-1}\Pi(i) = \frac{1}{N}\left(\sum_{i=0}^{k}\Pi(i) + \sum_{i=k+1}^{N-1}\Pi(i)\right)$$

$$= \frac{1}{N}\left((k+1)\left(\frac{1}{2} + \frac{(p-q)^2}{4p}\right) + (N-k-1)\left(p - (2p-1)\sum_{m=0}^{\frac{k-1}{2}}\binom{k}{m}p^m(1-p)^{k-m}\right)\right)$$

$$= p - (p-q)\sum_{m=0}^{\frac{k-1}{2}}\binom{k}{m}p^m(1-p)^{k-m} - \frac{k+1}{N}(p-q)\left(\frac{1}{4p} - \sum_{m=0}^{\frac{k-1}{2}}\binom{k}{m}p^m(1-p)^{k-m}\right)$$

For large k we get :

$$\Pi \approx p - (p-q)\sqrt{\frac{2}{\pi k}}\frac{\sqrt{pq}}{p-q}\left(2\sqrt{pq}\right)^k - \frac{k+1}{N}(p-q)\left(\frac{1}{4p} - \sqrt{\frac{2}{\pi k}}\frac{\sqrt{pq}}{p-q}\left(2\sqrt{pq}\right)^k\right)$$

or :

$$\Pi \approx p - \sqrt{\frac{2pq}{\pi k}}\left(2\sqrt{pq}\right)^k - \frac{k}{N}\frac{(p-q)}{4p}$$

We can now find, as before, the optimal window size, that minimize the loss over over the optimal fraction p. We get :

$$k^* = \frac{\ln N}{-\ln 2\sqrt{pq}} + \frac{\ln\left(-\frac{p-q}{4p}\sqrt{k^*}\right)}{\ln 2\sqrt{pq}} - \frac{\ln\left(\sqrt{\frac{2pq}{\pi}}\left(\ln\left(2\sqrt{pq}\right) - \frac{1}{2k^*}\right)\right)}{\ln 2\sqrt{pq}}$$

that is, for large N :

$$k^* \approx \frac{\ln N}{-\ln 2\sqrt{pq}}$$

The optimal fraction of correct prediction is then:

$$\Pi^* \approx p - \frac{(p-q)}{-4p\ln 2\sqrt{pq}}\frac{\ln N}{N}$$

We see again that in the non-stationary scenario, where the predictor has to "forget" the effects of the past, there is an optimal memory, or window size of O(ln N). This leads to a loss of O(lnN/N) of prediction errors, over the optimal performance. Note that when the predictor is initialized properly, i.e., we can reset the predictor whenever the statistics changes, increasing the memory, or the window size, improves the prediction performance, and as was shown in many previous results (see e.g. [2]) the loss over the optimal performance is O(1/N).

3. Saturated Counter with Finite Memory for Individual Sequences

We first recall that for individual sequences, where the universal predictor should perform well for any possible sequence, the predictor must be randomized. Otherwise, for each predictor there exists a sequence, the inverse of the prediction sequence, for which the predictor makes only errors. Also, the optimal rate in which the goal of attaining a fraction of correct prediction $\max\{N_0,N_1\}$ can be achieved is given by

$$\frac{1}{2^N}\binom{N-1}{\frac{N-1}{2}} \approx \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{N}}$$

for large N.

In this section we analyze the saturated counter, and we denote by k the number of states, assuming that k is odd. The states will be labeled as -(k-1)/2 ... 0 ... (k+1)/2 . The predictor state moves up when 1 is received and moves down for 0. The predictor is randomized and we denote by p_I the probability that we predict 1 and state i. We assume that $p_0 = \frac{1}{2}$ and $p_{\frac{k-1}{2}} \geq \cdots \geq p_1 \geq p_0 = \frac{1}{2} \geq p_{-1} \geq \cdots \geq p_{-\frac{k-1}{2}}$ and for symmetry we set $p_i = 1 - p_{-i}$.

Saturated Counter with Zero Initial Condition:

We begin by analyzing the predictor whose initial state is 0, corresponding to the case where we can reset the counter before prediction when the data behavior changes. We also assume that $\frac{p_{i-1} + p_{i+1}}{2} \leq p_i$. With this condition, as in [1] it is easy to see that out of all the sequences of size N whose number of ones is N_1, the worst sequence for prediction starts from the origin with a zigzag pattern. The expected (over the predictor randomization) fraction of correct prediction in this case is given by:

$$\Pi = (N - N_1)p_0 + (N - N_1)(1 - p_1) + \sum_{i=0}^{\frac{k-1}{2}-1} p_i + \left(2N_1 - N - \frac{k-1}{2}\right)p_{\frac{k-1}{2}}$$

where the sequence is partitioned into three segments : zigzag, raised counts and saturated count. We can now find the optimal p_I that maximizes the correct predictions. It turns out that the maximal value is achieved when $p_{i+1} = 2 * p_i - p_{i-1}$, i.e.,

$$p_i = \frac{1}{2} + \left(\frac{1}{2} - p_1\right) * i \quad \text{for} \quad 0 \leq i \leq \frac{k-1}{2} \quad \text{if} \quad p_i \leq 1$$

Note that the value of i , in which p_I is 1 is:

$$1 = \frac{1}{2} + \left(\frac{1}{2} - p_1\right) * i_{\max} \quad \Rightarrow \quad i_{\max} = \frac{1}{2p_1 - 1}$$

In the analysis we have to distinguish between the two cases :

1. $i_{\max} = \frac{1}{2p_1 - 1} < \frac{k-1}{2}$ saturation of the predictor occurs in the raising segment
2. $i_{\max} = \frac{1}{2p_1 - 1} \geq \frac{k-1}{2}$ saturation does not occur along the sequence

Consider first the case where saturation occurs. We get after some analysis that the fraction of correct predictions (which will depend on N_1 but also on p_1

$$\Pi = \frac{N + N_1}{2} - (N - N_1)p_1 - \frac{p_1}{2(2p_1 - 1)}$$

In the second case we get after some analysis:

$$\Pi = N - \frac{N_1}{2} + \left(p_1 - \frac{1}{2}\right)\left(\frac{k-1}{2}\right)\left(\frac{1}{2}\left(\frac{k-1}{2} - 1\right) + 2N_1 - N - \frac{k-1}{2}\right) - (N - N_1)p_1$$

We can now find the optimal p_1 and calculate the resulting fraction of correct prediction. It turns out that optimality is achieved in the first case, with the value

$$p_1^* = \frac{1}{2} + \frac{1}{2\sqrt{2(N - N_1)}}$$

This requires that the number of states must satisfy:

$$k \geq 2\sqrt{2(N - N_1)} + 1$$

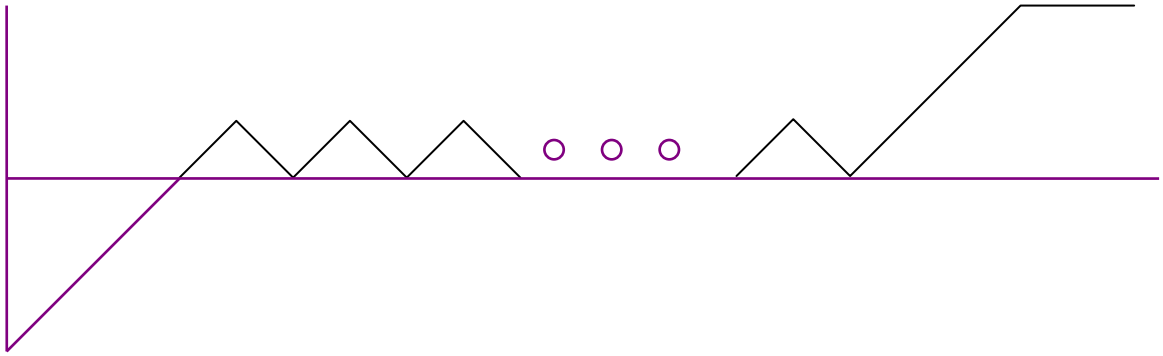
There is no need to increase the memory beyond what is required for saturation, so the optimal choice is $k \geq 2\sqrt{2(N - N_1)} + 1$ and the optimal fraction of correct prediction is given by:

$$\Pi = N_1 - \sqrt{\frac{N - N_1}{2}} - \frac{1}{4}$$

Saturated Counter with Worst Case Initial Condition

The worst case occurs when the predictor starts with some extreme case, and that is not tuned to the sequence empirical counts. In this case the predictor should adapt to the changing behavior. As will be seen, large memory, i.e., a large number of states, can make this task difficult, resulting in poorer prediction performance.

Schematically we have:



and we assume N_1 >N-N_1 . We get after some calculations that :

$$\Pi = \frac{k-1}{2} - 1 + p_{\frac{k-1}{2}} + p_0 + (N - N_1)p_0 + (N - N_1)(1 - p_1) + \left(2N_1 - N - (k-1)\right)p_{\frac{k-1}{2}}$$

As above, the optimal randomization is linear in the counts:

$$p_i = \frac{1}{2} + \left(\frac{1}{2} - p_1\right) * i$$

Again, the analysis can proceed by considering the two cases where saturation either occurs or does not occur along the sequence. In the first case where $i_{\max} = \frac{1}{2p_1 - 1} < \frac{k-1}{2}$, we get

$$\Pi = \frac{N + N_1}{2} - (N - N_1)p_1 - \frac{k}{2}$$

Clearly the optimal value of p_1 is the minimum value $p_1 = \frac{1}{2} + \frac{1}{k-1}$ which leads to

$$\Pi = N_1 - \frac{N - N_1}{k-1} - \frac{k}{2}$$

Now the optimal value of k is : $k^* = \sqrt{2(N - N_1)} + 1$ and so the mean fraction of prediction errors is given by:

$$\Pi^* = N_1 - \sqrt{2(N - N_1)} - \frac{1}{2}$$

Interestingly, an analysis of the second case leads to the same results, meaning that the number of states should be chosen so saturation is exactly achieved at the point where the predictor becomes non-random.

Summary:

We consider prediction with finite memory. Such predictors are useful for practical reasons. In addition, in case the sequence behavior changes, the finite, or even small memory makes the predictor more flexible to adapt to the changing statistics. The analysis was performed in the stochastic setting and in the individual sequence setting.

The main conclusions are:

The optimal memory size is O(ln N) where N is the segment length, in the stochastic setting, in non-stationary scenario.

The optimal memory size is $O(N^{1/2})$ in the individual setting.

The saturated counter is preferable over the finite window predictor.

Bibliography:

[1] M. Feder, N. Merhav, and M. Gutman, "Universal Prediction of Individual Sequences," *IEEE Transactions on Information Theory*, pp. 1258-1270, July 1992.

[2] N. Merhav, M. Feder and M. Gutman, "Some properties of Sequential Predictors for Binary Markov Sources," *IEEE Transactions on Information Theory*, pp. 887--892, May 1993.

[3] T.M. Cover, , "hypothesis testing with finite statistics," *The Annals of Mathematical Statistics, 40, (3), 828-835, 1969.*

[4] B. Ya. Ryabko, "The imaginary window for universal coding," *Private Correspondence.*