On Context–Tree Prediction of Individual Sequences

Jacob Ziv and Neri Merhav

Department of Electrical Engineering

Technion—Israel Institute of Technology

Haifa 32000, Israel

Mathematical Foundations of Learning Theory – II:

Ecole Normale Supérieure, Paris, May–June, 2006

General Motivation

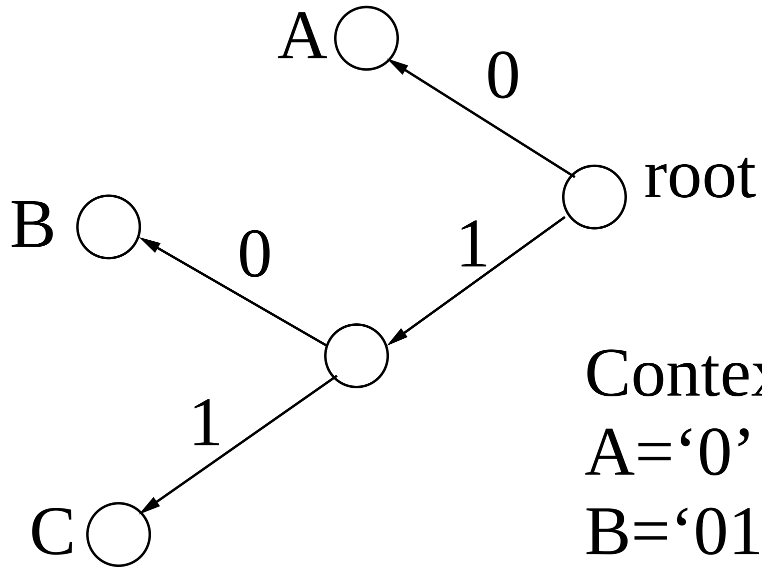
Motivated by the success of context–tree methods in compression, we wish to study them in the scenario of prediction of individual sequences.

Letting $x_1, x_2, \ldots$, be a binary individual sequence, a context–tree predictor is one of the form

$$\hat{x}_{t+1} = f(s_t),$$

where the 'state' s_t is a suffix of $(\ldots, x_{t-1}, x_t)$ derived by some rule, in particular, by a tree.

An Example



Context set:
A=‘0’
B=‘01’
C=‘11’

0 1 1 1 0 1 0 0 0 0 1 0 1 1 0 1

A B C C A B A A A A B A B C A B

Earlier Work

In [FederMerhavGutman92], universal prediction relative to general finite–state (FS) predictors, was investigated, where

$$s_{t+1} = g(s_t, x_t) \quad t = 1, 2, \ldots$$

for an arbitrary next–state function g:

Given an infinite sequence $\boldsymbol{x} = (x_1, x_2, \ldots)$, the finite–state predictability was defined as

$$\pi(\boldsymbol{x}) = \lim_{S\to\infty} \limsup_{N\to\infty} \pi_S(x_1, \ldots, x_N),$$

where $\pi_S(x_1, \ldots, x_N)$ is the minimum fraction of errors that is attained by the best FS predictor with $\leq S$ states on $(x_1, \ldots, x_N)$.

It was shown in [FederMerhavGutman92] that $\pi(\boldsymbol{x})$ is achievable by a universal predictor based on the LZ algorithm, or a Markov (finite–memory) predictor of growing order.

The asymptotic regime is such that $N >> S$.

Earlier Work (Cont'd)

Context–based methods are extensively used in data compression

- Weinberger and Seroussi, 1994
- Weinberger, Seroussi, and Sapiro, 1996
- Shtar'kov, Tjalkens, and Willems, 1997
- Willems, Shtar'kov, and Tjalkens, 1998
- Willems, 2004
- Martin, Seroussi, and Weinberger, 2004.

In predicton, studied by:

- Jacquet, Szpankowski, and Apostol, 2002
- Ziv, 2002, 2004

for random processes under certain regularity conditions.

Objectives

- Study context–tree prediction in the individual sequence regime.

Objectives

- Study context–tree prediction in the individual sequence regime.
- Study the case where $S = S_N$ grows concurrently with N.

Objectives

- Study context–tree prediction in the individual sequence regime.
- Study the case where $S = S_N$ grows concurrently with N.
- Propose a context–based prediction algorithm.

Motivation

- Context–tree predictors are more powerful than ‘Markov’ predictors. It is expected that their relative advantage would be emphasized in this asymptotic regime.

Motivation

- Context–tree predictors are more powerful than ‘Markov’ predictors. It is expected that their relative advantage would be emphasized in this asymptotic regime.
- We wish to understand fundamental limits of universality: How fast can S_N grow without sacrificing universal achievability of optimum performance?

Motivation

- Context–tree predictors are more powerful than 'Markov' predictors. It is expected that their relative advantage would be emphasized in this asymptotic regime.
- We wish to understand fundamental limits of universality: How fast can S_N grow without sacrificing universal achievability of optimum performance?
- Explore the regime where N is not necessarily very large relative to S.

Summary of Main Results

- We show that this critical growth rate of S_N is linear with N: If $S_N/N \to$ const., then the context-predictability cannot be universally approached.

Summary of Main Results

- We show that this critical growth rate of S_N is linear with N: If $S_N/N \to$ const., then the context-predictability cannot be universally approached.
- For a sublinear growth rate of S_N, we propose a universal predictor that achieves it uniformly.

Summary of Main Results

- We show that this critical growth rate of S_N is linear with N: If $S_N/N \to$ const., then the context-predictability cannot be universally approached.
- For a sublinear growth rate of S_N, we propose a universal predictor that achieves it uniformly.
- It is possible to control the growth rate of the number of contexts generated by the algorithm. For the best choice, the regret decays like $(S_N/N)^{1/3}$ in the horizon–dependent case.

Summary of Main Results

- We show that this critical growth rate of S_N is linear with N: If $S_N/N \to$ const., then the context-predictability cannot be universally approached.
- For a sublinear growth rate of S_N, we propose a universal predictor that achieves it uniformly.
- It is possible to control the growth rate of the number of contexts generated by the algorithm. For the best choice, the regret decays like $(S_N/N)^{1/3}$ in the horizon–dependent case.
- An horizon–independent algorithm is proposed too.

Problem Formulation

A context–tree predictor with S contexts is given by

$$\hat{x}_{t+1} = f(s_t),$$

where s_t takes values in a finite set $\mathcal{S}$ of $|\mathcal{S}| = S$ contexts defined by the leaves of a complete binary tree.

$f : \mathcal{S} \rightarrow \{0, 1\}$ may be randomized.

In the earlier example, $\mathcal{S} = \{0, 01, 11\}$, thus $S = 3$, and a predictor is defined by three probability distributions, $P(\cdot|0)$, $P(\cdot|01)$, and $P(\cdot|11)$.

Problem Forumlation (Cont'd) – Extension

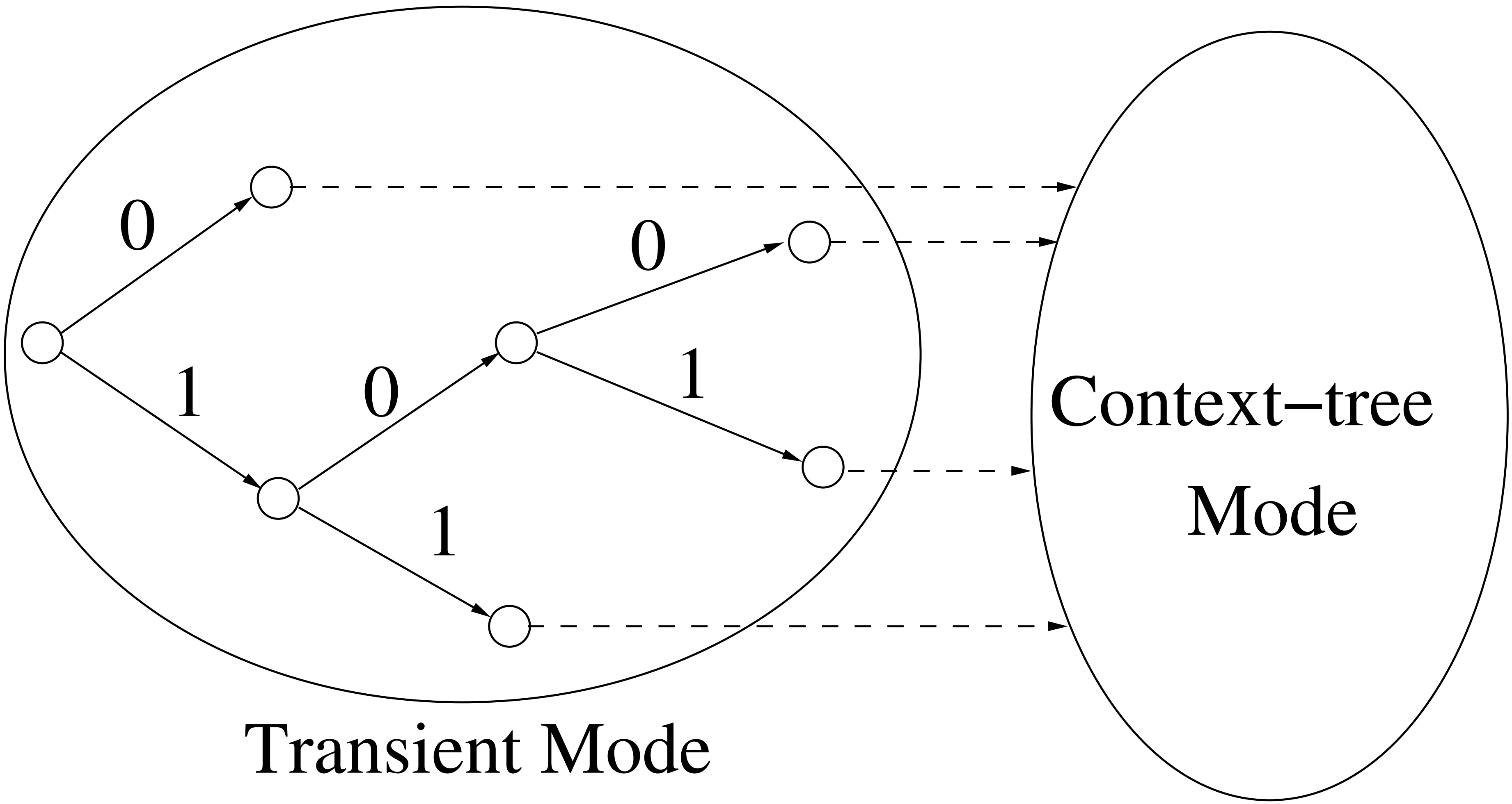
Given a total budget of S states, let us split it between:

- $S^C \leq S$ context states as before, plus
- $S^T \leq S - S^C$ transient states used to store the first few (training) samples $x_1, x_2, \ldots, x_\ell$, where ℓ may be context–dependent.

The set of transient states is defined by the internal nodes of a tree, whose root serves as the initial state.

The system begins at the transient mode, but at a certain stage, switches to the context–tree mode.

In the transient mode, the transient mode tree is traversed according to the incoming symbols. Once a leaf is reached, the system passes to the context–tree mode.


0
1
0
1
0
1
Context–tree
Mode
Transient Mode

Problem formulation (Cont'd)

$\mathcal{P}_S$ – the class of all predictors with $S^T + S^C \leq S$ states.
The Sth order context–predictability, $\kappa(x^N, S)$, is the minimum fraction of prediction errors attained over x^N by the best member of $\mathcal{P}_S$.

Given $\{S_N\}_{N\geq 1}$, the context predictability is universally achievable w.r.t. $\{S_N\}_{N\geq 1}$, if $\exists$ predictor such that for every $\boldsymbol{x} = (x_1, x_2, \ldots)$:

$$\limsup_{N\to\infty} \left[\frac{1}{N} \sum_{t=1}^{N} \Pr\{\hat{X}_t \neq x_t\} - \kappa(x^N, S_N) \right] \leq 0.$$

A predictor is said to achieve the context predictability w.r.t. $\{S_N\}_{N\geq 1}$ uniformly if

$$\limsup_{N\to\infty} \max_{x^N} \left[\frac{1}{N} \sum_{t=1}^{N} \Pr\{\hat{X}_t \neq x_t\} - \kappa(x^N, S_N) \right] \leq 0.$$

Main Result

Theorem: The context predictability w.r.t. $\{S_N\}_{N\geq 1}$ is uniformly universally achievable iff $\lim_{N\to\infty} S_N/N = 0$.

Sufficiency: We propose a universal (context–based) prediction algorithm which achieves the context predictability whenever $\lim_{N\to\infty} S_N/N = 0$.

Necessity: We show that for $a \in (0,1]$, there is a set $\mathcal{B}$ of sequences for each of which $\kappa(x^N, aN+1) = 0$, but $\forall$ predictor $\exists\, x^N \in \mathcal{B}$ such that

$$\frac{1}{N}\sum_{t=1}^{N} \text{Pr}\{\hat{X}_t \neq x_t\} - \kappa(x^N, aN+1) \geq \frac{a}{2}.$$

The question of universal achievability which is not uniforml, in this case, remains open.

The Algorithm

Horizon–dependent version

For a given N, choose a positive integer M_N. Let $k_0 = k_0(x_1, \ldots, x_t)$ denote the largest positive integer k such that the following two conditions hold at the same time:

- $(x_{t-k+1}, \ldots, x_t)$ appears at least M_N times along $(x_1, \ldots, x_t)$, and
- $(x_{t-k+2}, \ldots, x_t)$ has already served as prediction context $\geq M_N$ times previously.

If no such k exists, set $k_0 = 0$. $(x_{t-k_0+1}, \ldots, x_t)$ is the prediction context at time t. For $k_0 = 0$, the context s_t is “null.”

Having selected $s_t = (x_{t-k_0+1}, \ldots, x_t)$ according to these rules, randomly draw $\hat{x}_{t+1}$ according to Pr$\{\hat{x}_{t+1} = 1 | s_t\} = \phi(\hat{p}_t(1|s_t), N(s_t))$, where ϕ is defined as follows:

$$\phi(\alpha, n) = \begin{cases} 0 & \alpha < \frac{1}{2} - \epsilon_n \\ \frac{1}{2\epsilon_n}(\alpha - \frac{1}{2}) + \frac{1}{2} & \frac{1}{2} - \epsilon_n \leq \alpha \leq \frac{1}{2} + \epsilon_n \\ 1 & \alpha > \frac{1}{2} + \epsilon_n \end{cases}$$

Performance

We next show that the excess fraction of prediction errors, beyond $\kappa(x^N, S_N)$, is upper bounded by

$$\left(2\sqrt{\frac{2}{M_N}+\frac{1}{M_N^2}}+\frac{1}{M_N}\right)\cdot\left(1+\frac{M_N}{2N}\right)+\frac{(2M_N+1)S_N}{N},$$

which $\to 0$ iff $M_N \to \infty$ and $M_N S_N/N \to 0$.

These two conditions can be met at the same time whenever $S_N/N \to 0$.

Comments:

- Optimum M_N is prop. to $(N/S_N)^{2/3}$ yielding redundancy prop. to $(S_N/N)^{1/3}$.
- Horizon–independent version of the algorithm: can be obtained by defining M as function of k (length of examined context) rather than function of N.

Analysis

An upper bound on the redundancy,

$$\frac{1}{N}\sum_{t=1}^{N}[\mathrm{Pr}\{\hat{x}_t \neq x_t\} - \kappa(x^N, S_N)]$$

will be obtained by bounding $(1/N)\sum_{t=1}^{N} \mathrm{Pr}\{\hat{x}_t \neq x_t\}$ from above, and bounding $\kappa(x^N, S_N)$ from below.

As for the latter, we have:

$$\begin{aligned}\kappa(x^N, S_N) &\geq \frac{1}{N}\left[\sum_{s\in\mathcal{S}_N^C} \min\{N(s,0), N(s,1)\} - S_N^T\right] \\ &\geq \frac{1}{N}\left[\sum_{s\in\mathcal{S}_N^C} \min\{N(s,0), N(s,1)\} - S_N\right],\end{aligned}$$

where $N(s,x)$ is the count of $(s_t = s, x_{t+1} = x)$.

Analysis (Cont'd)

As was shown in [FederMerhavGutman92], when the proposed predictor is applied, the contribtution of each state s to the expected number of prediction errors,

$$EN_e(s) = \sum_{t:s_t=s} \mathrm{Pr}\{\hat{x}_t \neq x_t\},$$

is upper bounded by

$$EN_e(s) \leq \min\{N(s,0), N(s,1)\} + \sqrt{N(s)+1} + \frac{1}{2}, \tag{1}$$

where $N(s) = N(s,0) + N(s,1)$ is the number of occurrences of s.

Consider the above prediction scheme applied to x^N, and denote sequence of contexts, generated by this algorithm, as $\hat{s}^N = (\hat{s}_1, \ldots, \hat{s}_N)$.

Analysis (Cont'd)

By the construction of the algorithm, every one of $S_N^C - 1$ internal nodes of the reference predictor in $\mathcal{P}_{S_N}$ is used as a prediction context $\leq 2M_N$ times. The reason is that in the $(2M_N + 1)$–st time, it was either preceded by '0' or by '1'at least M_N times, and so, the conditions for extending the context are met.

Thus, except for $2M_N(S_N^C - 1) < 2M_N S_N$ time instants, $\hat{s}$ is a refinement of the reference state, s.

Let $\mathcal{T}_s$ denote the sub–tree of prediction contexts rooted at s. Then,

$$\begin{aligned}\frac{1}{N}\sum_{t=1}^{N}\text{Pr}\{\hat{x}_t \neq x_t\} &\leq 2M_N S_N + \sum_{s\in\mathcal{S}_N^C}\sum_{\hat{s}\in\mathcal{T}_s}\min\{N(\hat{s},0),N(\hat{s},1)\} + \\ &\quad + \sum_{\hat{s}\in\mathcal{T}_s}\left[\sqrt{N(\hat{s})+1}+\frac{1}{2}\right] \\ &\stackrel{\Delta}{=} 2M_N S_N + A + B.\end{aligned}$$

Analysis (Cont'd)

Now,

$$
\begin{aligned}
A &= \sum_{s \in \mathcal{S}_N^C} \sum_{\hat{s} \in \mathcal{T}_s} \min\{N(\hat{s},0), N(\hat{s},1)\} \\
&\leq \sum_{s \in \mathcal{S}_N^C} \min\left\{ \sum_{\hat{s} \in \mathcal{T}_s} N(\hat{s},0), \sum_{\hat{s} \in \mathcal{T}_s} N(\hat{s},1) \right\} \\
&\leq \sum_{s \in \mathcal{S}_N^C} \min\{N(s,0), N(s,1)\} \\
&\leq N \cdot \kappa(x^N, S_N) + S_N .
\end{aligned}
$$

Analysis (Cont'd)

As for

$$B = \sum_{s \in \mathcal{S}_N^C} \sum_{\hat{s} \in \mathcal{T}_s} \left[\sqrt{N(\hat{s}) + 1} + \frac{1}{2} \right],$$

we again use the fact that $N(\hat{s}) \leq 2M_N$, and so,

$$\begin{aligned} B &\leq \sum_{s \in \mathcal{S}_N^C} \sum_{\hat{s} \in \mathcal{T}_s} \left(\sqrt{2M_N + 1} + \frac{1}{2} \right) \\ &= \left(\sqrt{2M_N + 1} + \frac{1}{2} \right) \cdot \sum_{s \in \mathcal{S}_N^C} |\mathcal{T}_s| \end{aligned}$$

and $\sum_{s \in \mathcal{S}_N^C} |\mathcal{T}_s|$ is in turn upper bounded by the total number of contexts generated by the algorithm, which is $\leq (2N/M_N + 1)$ because each context pertaining to an internal node is used as a prediction context at least M_N times.

Horizon–Independent Algorithm

Defining the H–I algorithm in terms of a sequence $\{M(k)\}_{k\geq 1}$, let

$$\psi(N) = 2\min_k \left[\frac{2^k}{N} + \frac{1}{M(k)}\right],$$

then the redundancy is upper bounded by

$$\frac{2S_N(M(S_N)+1)}{N} + \sqrt{\psi(N)[1+\psi(N)]} + \frac{\psi(N)}{2}.$$

The choice of $M(k)$ controls the trade–off between the allowed growth rate of S_N and the redundancy rate.

Faster convergence than the LZ–based algorithm in [FederMerhavGutman92].

Necessity

For each one of the 2^{aN} sequences

$$x_1, x_2, \ldots, x_{aN}, 0, \ldots, 0$$

there exists a member in $\mathcal{P}_{aN+1}$ which gives error–free prediction, thus

$$\kappa(x^N, aN+1) = 0.$$

This is easily seen by using aN transient states and only one context–tree state.

On the other hand, $\forall$ predictor

$$\max_{(x^{aN},0,\ldots,0)} \frac{1}{N} \sum_{t=1}^{N} \mathsf{Pr}\{\hat{x}_t \neq x_t\} \geq \frac{1}{N} \sum_{t=1}^{aN} E\mathsf{Pr}\{\hat{x}_t \neq X_t\} \geq \frac{a}{2}.$$

Conclusion

- Context–tree predictability of individual sequences was investigated.

Conclusion

- Context–tree predictability of individual sequences was investigated.
- Asymptotic regime allows $S = S_N$.

Conclusion

- Context–tree predictability of individual sequences was investigated.
- Asymptotic regime allows $S = S_N$.
- The critical growth rate is linear.

Conclusion

- Context–tree predictability of individual sequences was investigated.
- Asymptotic regime allows $S = S_N$.
- The critical growth rate is linear.
- A context–based algorithm.

Conclusion

- Context–tree predictability of individual sequences was investigated.
- Asymptotic regime allows $S = S_N$.
- The critical growth rate is linear.
- A context–based algorithm.
- Open question no. 1: what about non–uniform achievability?

Conclusion

- Context–tree predictability of individual sequences was investigated.
- Asymptotic regime allows $S = S_N$.
- The critical growth rate is linear.
- A context–based algorithm.
- Open question no. 1: what about non–uniform achievability?
- Open question no. 2: can we get rid of the transient states?

Conclusion

- Context–tree predictability of individual sequences was investigated.
- Asymptotic regime allows $S = S_N$.
- The critical growth rate is linear.
- A context–based algorithm.
- Open question no. 1: what about non–uniform achievability?
- Open question no. 2: can we get rid of the transient states?
- Open question no. 3: sharper upper and lower bounds on the regret.