

Universal prediction of Stationary Processes

Benjamin Weiss

Joint work with
Gustav Morvai

Universal Prediction of Stationary Processes

1. Introduction

minimax prediction

T. Cover

2. Markov Chains

3. Learning from the Past

4. Random Markov Chains

5. Universal Prediction for Random Markov Chains

PERSONAE DRAMATIS

D. Blackwell

T. Cover

D. Ornstein

D. Bailey

G. Morvai

L. Györfi

S. Yakowitz

P. Shields

I. Csiszar

B. Ya. Ryabko

⋮

1. Introduction

Stationary processes

$\{X_n\}_{-\infty}^{\infty}$, $A = \text{finite alphabet}$

NOTATION: $X_i^j = X_i X_{i+1} \dots X_j$

$P(X | Y) = \text{conditional distribution of } X - \text{ not a number - but rather an element of } \mathcal{P}(A), \text{ if } X \text{ takes values in } A.$

Predicting values ^(DB) or distribution

Weather Forecast - - RAIN

- NOT RAIN

vs distribution on

- RAIN

- NOT RAIN

"FORWARD" Prediction

Given $x_0, x_1, x_2, \dots, x_n$

GUESS the conditional distribution of x_{n+1} .

A scheme for guessing is

$$g_n(\xi_0, \xi_1, \dots, \xi_n) \in \text{Dist.}(A)$$

A is the alphabet of the process.

- GOAL -

$$\lim_{n \rightarrow \infty} \|P(x_{n+1} | x_0^n) - g_n(x_0^n)\|_1 = 0$$

almost surely

LEARNING from the PAST ("Backward" Prediction)

GOAL: A guessing scheme $g_n(x_{-n}^o)$ such that

$$\lim_{n \rightarrow \infty} \|P(x_1 | x_{-n}^o) - g_n(x_{-n}^o)\|_1 = 0$$

almost surely.

Remark: Using martingale convergence
thm can replace $P(x_1 | x_{-n}^o)$
by $P(x_1 | x_{-\infty}^o)$.

2. Markov CHAINS

d-step Markov Chain

$$P(X_0 | X_{-\infty}^-) = P(X_0 | X_{-d}^-)$$

g_n^d - a guessing scheme
for all d-step M.C.

$g_n^d(\xi_0^n)(a) =$ relative frequency
in ξ_0^n of the appearance
of $\xi_{n-d+1}^n a$ compared
to the total # of appearances
of ξ_{n-d+1}^n .

When $d=0$ - just the empirical
distribution of a itself.

ALL THAT MATTERS is X_{n+1-d}^{n+1}

If d is not known - but we know it is a M.C. then estimate d - and use preceding.

If we don't know that it is a M.C. and we are penalized for guessing wrongly - but are allowed to pass - then we want a scheme to test for (M.C.) \sim NOT (M.C.).

D. Bailey showed - that this is not possible -

Moreover

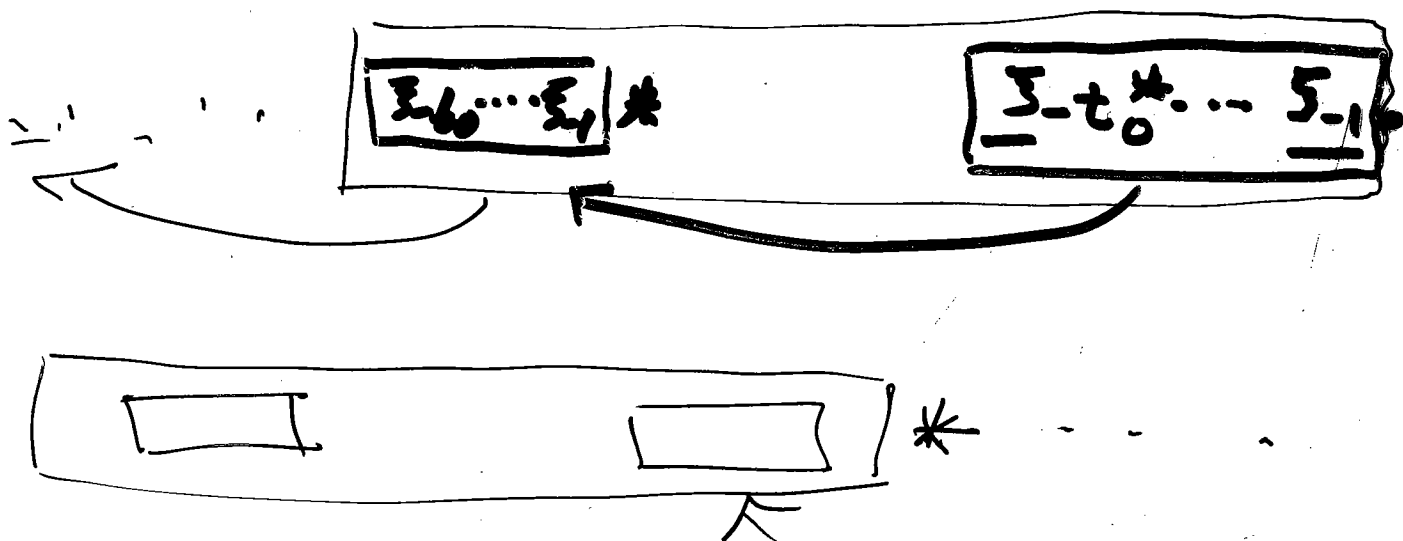
Bailey showed that there is
no way of achieving our
goal in "Forward Prediction" for
ALL - STATIONARY processes.

However - in the problem of
learning from the past - the
situation is different - .

3. Learning from the Past

D. Ornstein -

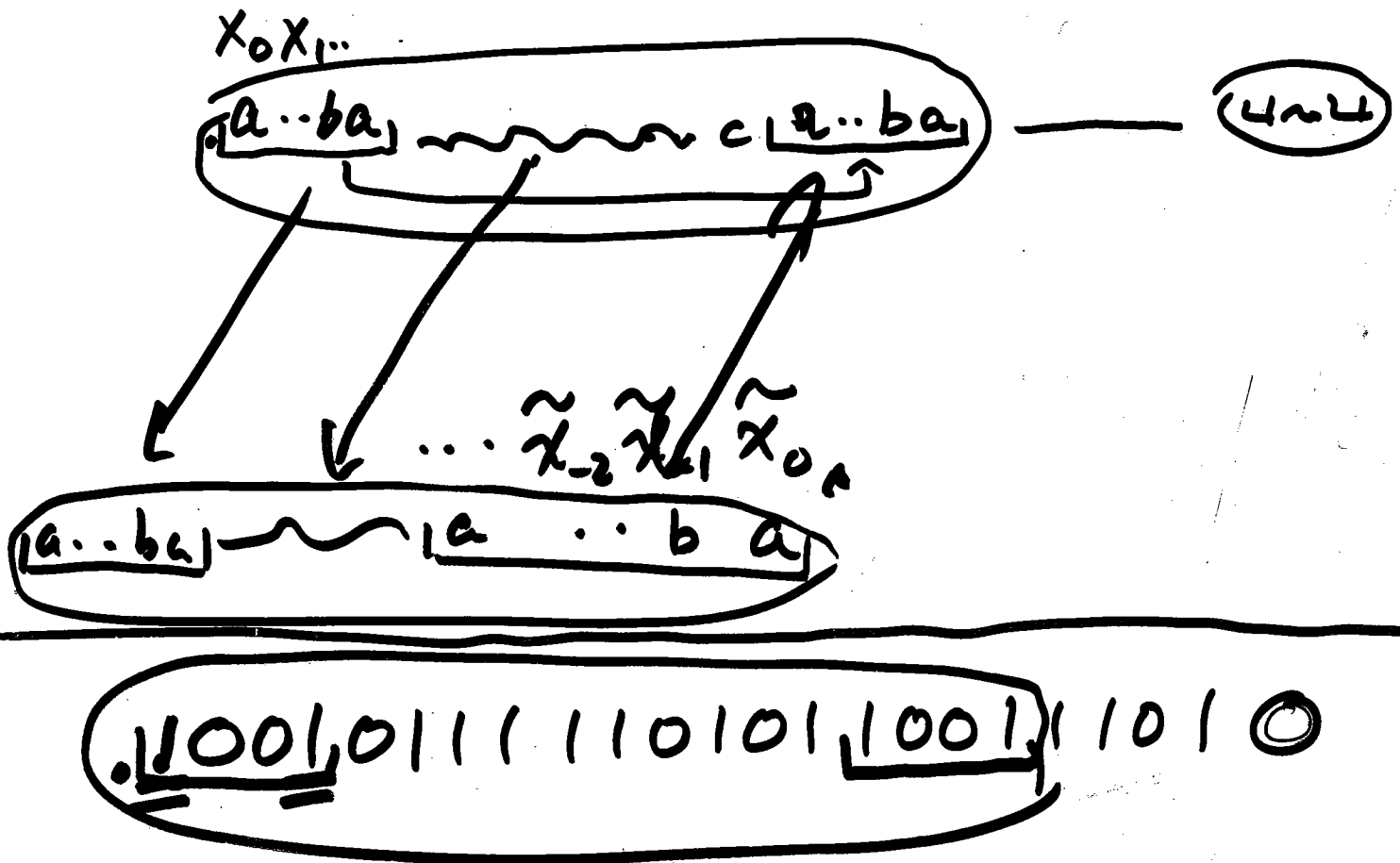
GMY



not very efficient
uses only small part of
the information -

but shows that it is possible -
still open - to find more
efficient schemes!

Adapting to the future -



$\leftarrow 100101111101011001.$

This maps: past \leftarrow future
 and yields a process
 which is distributed just like
 the original one!

This yields times $\tau_1 < \tau_2 < \dots$
such that we put forward
at time τ_k a guess for
 $\mathbb{P}(X_{\tau_k+1} | X_0^{\tau_k})$

and almost surely this
is converging to the truth -
(i.e. the difference
tends to zero).

Can we guess more often?

Yes - but not for all processes
need some condition -

a.s. - continuity of $P(x_0 | x_{-1}^n)$
in \mathcal{X}^n metric that makes

\mathcal{A}^n a compact metric

space.

Scheme - uses similar idea

but after $x_0 = a$ repeats

at time t look for next

occurrence of $x_{t+1} = a$

a...ba...ba
..cba

4. Random M.Chains

for each d - a table of conditional probabilities, T_d
 $P(X_1 | X_{-\infty}^0)$ is $P(X_1 | X_{-\tau}^0)$
where for $\tau = d$ use the table T_d . and τ is chosen ind. of everything with a fixed distribution.

S. Kalikow - continuous conditional Probabilities.

- wide class - all zero entropy processes can be coded from such random m.c.

(Kalikow - Katznelson - W.)

45. UNIVERSAL PREDICTION for RM's.

To define \mathcal{L}_n -

for each k (not too big,
say $< .1 \log n$) check to
see if ξ_{-k+n+1}^n occurs in
 ξ_0^n at least \sqrt{n} times.

If it does - then look at
its occurrences and the
distribution it yields. Call
that $D_{n,k}(\xi_0^n)$, Now
average these up and
that is \mathcal{L}_n .

IN FORMULAS

$$\tau_i^k(n) = \min \{ t > \tau_{i-1}^k(n) : X_{n-k+1-t}^{n-t} = X_{n-k+1}^n \}$$

$$\tau_0^k(n) = 0$$

$$P_n^k(a) = \frac{1}{[\sqrt{n}]} \sum_{j=1}^{[\sqrt{n}]} \mathbb{1}_{\{X_{n-\tau_j^k(n)+1}^n = a\}}$$

$$K_n = \max \{ 1 \leq k \leq [1 \log n] : \tau_{[\sqrt{n}]}^k(n) \leq n-k+1 \}$$

if there is such a k - and 0 otherwise.

then

$$g_n(a) = \frac{1}{K_n} \sum_{k=1}^{K_n} P_n^k(a)$$

the sequence X_0^n has been fixed throughout

Theorem: If $P(X_0 = a | X_{-\infty}^{-1})$
is a continuous fcn of
 $X_{-\infty}^{-1}$ for all $a \in A$ then
with \mathcal{G}_n defined as in the
previous page, almost surely

$$\lim_{n \rightarrow \infty} |\mathcal{G}_n(a) - P(X_{n+1} = a | X_0^n)| = 0$$

for each $a \in A$.

$$\mathcal{G}_n(a) = \frac{1}{K_n} \sum P_n^k(a)$$