An Efficient Universal Prediction Algorithm for **Unknown Sources with Limited Training Data**

Jacob Ziv

Department of Electrical Engineering

Technion–Israel Institute of Technology

Haifa 32000, Israel

April 27, 2002

Abstract

 \bigwedge

 $\sqrt{2}$

In of cc fir th the trace of contract of the trace of the trace of the trace of the trace of the class of \bigcup
 \bigcup Inspired by C.E Shannon's celebrated paper: "Prediction and entropy of printed English" (BSTJ 30:50–64, 1951), we consider, in this correspondence, the optimal prediction error for unknown finite-alphabet ergodic Markov sources, for prediction algorithms that make inference about the most probable incoming letter, where the distribution of the unknown source is apparen^t only via ^a short finite-alphabet ergodic Markov sources, for prediction algorithms
that make inference about the most probable incoming letter, where
the distribution of the unknown source is apparent only via a short
training sequence of function of K , the order of the Markov source, rather than the the distribution of the unknown source is apparent only vi
training sequence of $N + 1$ letters. We allow N to be a por
function of K, the order of the Markov source, rather than
classical case where N is allowed to be exp

 \sqrt{A} A ur ot as

as (*l* th so

F(so Pr ...) \bigwedge $\begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$
is e A lower bound on the prediction error is formulated for such universal prediction algorithms, that are based on suffixes that were A lower bound on the prediction error is formulated for such
universal prediction algorithms, that are based on suffixes that were
observed somewhere in the past "training-sequence" X_{-N}^{-1} (i.e. it is assumed that the universal predictor, given the pas^t universal prediction algorithms, that are based on suffixes that wer
observed somewhere in the past "training-sequence" X_{-N}^{-1} (i.e. it is
assumed that the universal predictor, given the past
 $(N + 1)$ -sequence which se than the optimal predictor given only the longest suffix that appeared $(N + 1)$ -sequence which serves as a training sequence is no better
than the optimal predictor given only the longest suffix that appear
somewhere in the past X_{-N}^{-1} vector).

For ^a class of stationary Markov sources (which includes all Markov sources with positive transition probabilities), ^a particular universal predictor is introduced, and it is demonstrated that its performance is "optimal" in the sense that it yields ^a prediction-error which is close

to the lower-bound on the universal prediction-error, with limited training data.

 \bigwedge

 $\sqrt{2}$

to tra Tl

Tl

ef pr

ba

Fl

f3

ne re % The results are non-asymptotic in the sense that they express the effect of limited training data on the efficiency of universal predictors. An asymptotically optimal universal predictor which is based on pattern matching appears elsewhere in the literature (e.g [3], [5]). However, the prediction error of these algorithms does not necessarily come close to the lower bound in the non-asymptotic region.

$\sqrt{1}$ **1 Introduction**

We consider finite-alphabet sequences which are emitted by ^a stationary source with unknown statistics Iphabet sequences which are e
th unknown statistics
 $X = X_1, X_2, ..., X_i, ...;$

Example: **equences which are**

\nwith unknown statistics

\n
$$
\mathbf{X} = X_1, X_2, \dots, X_i, \dots;
$$
\n
$$
X_1^m = X_1, X_2, \dots, X_m;
$$
\n
$$
X_i \in \mathbf{A}; |\mathbf{A}| = A.
$$

 \bigwedge

 $\begin{pmatrix} 1 \\ w \end{pmatrix}$ % $X_i \in \mathbf{A}$; $|\mathbf{A}| = A$.
Given X_{-N}^0 , we need to predict X_1 in cases where the actual measure $X_i \in \mathbf{A}$; $|\mathbf{A}| = A$.
Given X_{-N}^0 , we need to predict X_1 in cases where the actual meas $P(X_1|X_{-N}^0)$ is not available to us. In order to predict X_1 one may $P(X_1|X_{-N}^0)$ is not available to us. In order to predict X_1 one may assign, for any suffix X_{-N}^{-0} (which serves as a training sequence),

 $\sqrt{2}$ so pr de an de la pr de \bigcap $\begin{pmatrix} 1 \\ 1 \\ 2 \\ 3 \end{pmatrix}$
 A some arbitrary prediction function $f(X_{-N}^0)$, hoping that this assigned predictor will yield ^a small prediction error which will be as close as possible to the minimal prediction error (for known statistics), some arbitrary prediction function $f(X_{-N}^0)$, hoping that this
predictor will yield a small prediction error which will be as
possible to the minimal prediction error (for known statistics
namely: $P_{\min}(X_{-N}^0) = \min_{f(*)} E$ the minimal prediction error (for known statistics),
 $\lim_{n \to \infty} (X^0_{-N}) = \min_{f(*)} E_{X^0_{-N}} \delta(X_1, f(X^0_{-N}))$, where
 $\delta(a, b) = 0$ if $a = b$; else $\delta(a, b) = 1$, (1)

$$
\delta(a, b) = 0 \text{ if } a = b; \qquad \text{else} \qquad \delta(a, b) = 1, \tag{1}
$$

 $\delta(a, b) = 0$ if $a = b$; else $\delta(a, b) = 1$,
and where $E_{X_{-N}^0}(*)$ denotes conditional expectation. The optimal $\delta(a, b) = 0$ if $a = b$; else $\delta(a, b) = 1$, (1)
and where $E_{X^0_{-N}}(*)$ denotes conditional expectation. The optimal
prediction of X_1 , given X^0_{-N} , is achieved by picking the one $X_1 \in A$ and where $E_{X_{-N}^0}(*)$ denotes co
prediction of X_1 , given X_{-N}^0 , is
which maximizes $P(X_1|X_{-N}^0)$. prediction of X_1 , given X_{-N}^0 , is achieved by picking the one $X_1 \in$
which maximizes $P(X_1|X_{-N}^0)$.
Hence, $P_{\min}(X_{-N}^0) = E_{X_{-n}^0} \delta(X_1, \arg \max_{X_1 \in A} P(X_1|X_{-N}^0))$. In

 $\sqrt{\text{th}}$ the case of universal prediction, the measure $P(X_1|X_{-N}^0)$ is not the case of universal prediction, the measure $P(X_1|X_{-N}^0)$ is no
known, and therefore, $P_{\min}(X_{-N}^0)$ is not necessarily achievable the case of universal predic
known, and therefore, P_{min}
(unless N is large enough).

 \bigwedge

th kr (u W in du m in di. $\begin{array}{c}\n\text{at} \\
\hline\n\text{at} \\
\hline\n\end{array}$ We consider the class of universal predictors that satisfy the highly intuitive restriction that each universal predictor in the class may not outperform the predictor which is based on the true probability measure, conditioned on the longest suffix that appeared somewhere in the past $(N + 1)$ –sequence X_{-N}^0 , rather than the complete X_{-N}^0 , measure, conditioned on the longest suffix that appeared somewher
in the past $(N + 1)$ –sequence X_{-N}^0 , rather than the complete X_{-N}^0 ,
(i.e. $P(X_1|X_{-K_0(X_{-N}^0)}^0, K_0(X_{-N}^0))$ where $X_{-K_0(X_{-N}^0)}^0$ is the longe in the past $(N + 1)$ -sequence X_{-N}^S , rather than the complete X_{-N}^S
(i.e. $P(X_1 | X_{-K_0(X_{-N}^0)}^0, K_0(X_{-N}^0))$ where $X_{-K_0(X_{-N}^0)}^0$ is the long
suffix of X_1 in X_{-N}^0 [1]. More precisely, we make the follo restriction:

Restriction 1 Let $K_0(X_{-N}^0)$ be the largest integer $i \leq N-1$ such that $X_{-i}^0 = X_{-i-i}^{-j}$ for some

$$
1 \le j \le N - i \tag{2}
$$

 $(K_0 = -1$ if X_0 does not appear in X_{-N}^{-1} where X_1^0 is the null string). We restrict our attention to the class of predictors

$$
G = [g : \mathbf{A}^{N+1} \to \mathbf{A} | E \delta(X_1, g(X_{-N}^0) \ge
$$

$$
E \delta(X_1, \arg \max_{X \in \mathbf{A}} P(X_1 | X_{-K_0(X_{-N}^0)}^0, K_0(X_{-N}^0))]
$$

 \sqrt{D} Define,

Define,
\n
$$
P_{\min}^u(X_{-N}^0) = \min_{g(*) \in G} E_{X_{-N}^0} \delta\left(X_1, g(X_{-N}^0)\right)
$$
\nwhere the expectation is taken w.r.t. X_1 given the context X_{-N}^0 .

Clearly, $\begin{aligned} \n\binom{0}{-N} &= \min_{g(*) \in G} E_{X_{-N}^0} \delta\left(X_1, g(X_{-N}^0)\right) \n\end{aligned}$ n is taken w.r.t. X_1 given the context X_{-N}^0 .
 $P_{\min}^u(X_{-N}^0) \ge P_{\min}(X_{-N}^0)$. (3)

$$
P_{\min}^u(X_{-N}^0) \ge P_{\min}(X_{-N}^0). \tag{3}
$$

 \bigwedge

D

W

C

T

T

P

T

T

P

T

T

P

T

F

d

d

is $\begin{pmatrix} 3 \\ N \\ N \end{pmatrix}$ The l.h.s. of Eq. (3) therefore serves as ^a lower bound on the prediction error that may be achieved by any predictor in the restricted class. Roughly speaking, we are treating this problem by deriving performance bounds for ^a restricted class of prediction algorithms that only make inferences about the "optimal" predictor for the (unknown) random process based only on what has been algorithms that only make inferences about the "optimal" predictor
for the (unknown) random process based only on what has been
observed in the training data. Assume that the source that emits X_{-N}^1 deriving performance bounds for a restricted class of prediction
algorithms that only make inferences about the "optimal" predictor
for the (unknown) random process based only on what has been
observed in the training dat

 $\sqrt{\text{th}}$ th ur [5 ap a 1 m ln gi fo W pr contained to the set of t \bigwedge or

or

ug

at,

(the predictor). If N is large enough (say, exponential in K) the universal prediction error may approach the minimal prediction error [5]. Intuitively, this exponential growth is needed since the prediction the predictor). If *N* is large enough (say, exponential in *K*) the universal prediction error may approach the minimal prediction error [5]. Intuitively, this exponential growth is needed since the prediction approach i [5]. Intuitively, this exponential growth is needed since the prediction
approach is based on estimating K by some order \hat{K} and performing
a majority vote conditioned on the context of length \hat{K} . However, in many cases, such ^a large number of samples is not available.

In this correspondence, we consider K -th order Markov sources that, In this correspondence, we consider K -th order given the positive parameters T_1 and N , satisfy: large number of samples is no
nce, we consider *K*-th order N
arameters *T*₁ and *N*, satisfy:
 $Pr\left[P(X_{-\ell}^0) \ge 2^{-H_{\min}\ell}\right] \le \frac{1}{T_1}$

$$
\Pr\left[P(X_{-\ell}^0) \ge 2^{-H_{\min}\ell}\right] \le \frac{1}{T_1}
$$

given the positive parameters T_1 and N , satisfy:
 $\Pr\left[P(X_{-\ell}^0) \ge 2^{-H_{\min}\ell}\right] \le \frac{1}{T_1}$

for some positive number H_1 and some positive integer $\ell \le \frac{5 \log N}{H_1}$. We derive an efficient universal prediction algorithm that yields a prediction error close to $P_{\min}^u(X_{-N}^0)$ for values of N which are

pc

re

cc

X

m

(e

pr

ar

er

lo at \bigcup polynomial in K. The algorithm finds the longest suffix of X_{-N}^0 that polynomial in K. The algorithm finds the longest suffix of X_{-N}^0 recurs at least T_2 times, where T_2 is some predetermined positive recurs at least T_2 times, where T_2 is some predetermined positive constant. The predicted X_1 is taken to be the most frequent letter polynomial in K. The algorithm finds the longest suffix of X_{-N}^0 recurs at least T_2 times, where T_2 is some predetermined positiv constant. The predicted X_1 is taken to be the most frequent lette $X' \in \mathbf{A}$ majority rule). This in contrast with other results in the literature (e.g. [3], [5]) that describe asymptotically optimal universal $X' \in$ **A** among the letters that followed these recurrences (i.e. a majority rule). This in contrast with other results in the literature (e.g. [3], [5]) that describe asymptotically optimal universal prediction algorith apparently, exponential in K , and therefore the associated prediction error of these algorithms does not necessarily come close to the lower bound of Eq. (3).

 \bigwedge

 $\sqrt{2}$

2 Main Results

 $\sqrt{2}$

& $\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$ **2 Main Results**
Let Y_{-N}^0 be a realization of the source process, which is independent Let Y _{of X^0} X_{-N}^0 be a realization of the source process, which is independed
 X_{-N}^0 . Define $K_0(X_{-N}^0|Y_{-N}^0)$ to be the largest integer *i* such that Let
of X^0 **i** Y_{-N}^0 be a realization of the source process, which is independen
 X_{-N}^0 . Define $K_0(X_{-N}^0|Y_{-N}^0)$ to be the largest integer *i* such that
 $X_{-i}^0 = Y_{-i-j}^{-j}$ for some $0 \le j \le N - i$. $(K_0(X_{-N}^0|Y_{-N}^0) = -1$ Let Y_{-N}^s be a realization of the source process, which is independ
of X_{-N}^0 . Define $K_0(X_{-N}^0|Y_{-N}^0)$ to be the largest integer *i* such that
 $X_{-i}^0 = Y_{-i-j}^{-j}$ for some $0 \le j \le N - i$. $(K_0(X_{-N}^0|Y_{-N}^0) = -1$

 \bigwedge

Restriction 2 We restrict our attention to the class of predictors
\n
$$
G' = [g' : \mathbf{A}^{N+1} \to \mathbf{A}] g'(X^0_{-N} | Y^0_{-N}) = g'(Z^0_{-N} | Y^0_{-N})
$$
\nwherever\n
$$
K_0(X^0_{-N} | Y^0_{-N}) = K_0(Z^0_{-N} | Y^0_{-N}) = k \text{ and } X^0_{-k-1} = Z^0_{-k-1}].
$$
\nThus,
\n
$$
5P^u_{\min}(X^0_{-N} | Y^0_{-N}) = \min_{g'(\ast) \in G'} E_{X^0_{-N}, Y^0_{-N}} \delta(X_1, g'(X^0_{-N} | Y^0_{-N}))
$$
\n
$$
= P_{\min}(X^0_{-K_0(X^0_{-N} | Y^0_{-N})-1}). \tag{4}
$$

Observe that, given the suffix $X^0_{-K_0(X^0_{N} | Y^0_{N})-1}$, X_1 is independent of $K_0(X_{-N}^0|Y_{-N}^0)$.

Then,

Lemma 1 Let Y_{-N}^0 be a realization of an admissible K-th order stationary ergodic Markov process, that is independent of X^0_{-N} which is emitted from the same source, and let T_1 be a positive

 $\sqrt{2}$ $\left(\begin{array}{c} \begin{array}{c} \begin{array}{c} \end{array}\\ \begin{array}{c} \end{array}\\$ *integer. Then,*

1.

For example, the following equations:

\n
$$
EP_{\min}^{u}(X_{-N}^{0}|Y_{-N}^{0}) \leq EP_{\min}(X_{-K(X_{-N}^{0})-1}^{0}) + O(\frac{\log N}{N^{\delta}}) + \frac{2}{T_{1}}
$$
\n
$$
\leq EP_{\min}^{u}(X_{-N}^{0}) + O(\frac{\log N}{N^{\delta}}) + \frac{2}{T_{1}}
$$
\nprovided that the (unknown) order of the Markovian source satisfies $K \leq O(N^{\frac{1}{3}-3\delta})$ where $0 \leq \delta \leq \frac{1}{9}$.

\nFor any predictor $g(X_{-N}^{0})$ that satisfies

\n
$$
g(X_{-N}^{0}) = g(X_{-N}^{-M}, X_{-K}^{0})
$$

 \bigwedge

$$
\leq EP^{u}_{\min}(X^{0}_{-N})+O(\frac{\log N}{N^{\delta}})+\frac{2}{T_{1}}
$$

provided that the (unknown) order of the Markovian source $\leq EP_{\min}^u(X_{-N}^0) + O(\frac{\log N}{N^{\delta}})$
provided that the (unknown) order of the M
satisfies $K \leq O(N^{\frac{1}{3}-3\delta})$ where $0 \leq \delta \leq \frac{1}{9}$. *provided that the (unknown) order of the Mar*
 satisfies $K \leq O(N^{\frac{1}{3}-3\delta})$ *where* $0 \leq \delta \leq \frac{1}{9}$.

2. For any predictor $g(X_{-N}^0)$ that satisfies
 $g(X_{-N}^0) = g(X_{-N}^{-M}, X_{-K}^0)$

$$
g(X_{-N}^0) = g(X_{-N}^{-M}, X_{-K}^0)
$$

$$
for\ every\ X_{-N}^{0} \in \mathbf{A}^{N+1} \ where\ M \geq N^{\frac{1}{3}-\delta} \ we\ have\ that,
$$

\n
$$
E\delta\Big(X_1, g(Y_{-N}^{-M}, X_{-K}^{0})\Big) \leq E\delta\Big(X_1, g(X_{-N}^{0})\Big) + O\left(\frac{1}{N^{\delta}}\right).
$$

\n3.
\n
$$
\Pr\Big[K_0(X_{-N}^{0}|Y_{-N}^{0}) \geq \frac{5}{H_1}\log N\Big] \leq \frac{2T_1}{H_1N} + \frac{2}{T_1}.
$$

\n
$$
\Pr\Big[K_0(X_{-N}^{0}) \geq \frac{5}{H_1}\log N\Big] \leq \frac{2T_1}{H_1N} + \frac{2}{T_1}.
$$

\nThe proof of Lemma 1 appears in the Appendix.

 \bigwedge

The proof of Lemma 1 appears in the Appendix.

 $\sqrt{2}$

$\sqrt{2}$ **Discussion:**

 $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ 1. It should be noted that any given stationary ergodic source is **ussion:**
It should be noted that any given stationary ergodic source is
admissible as N tends to infinity. However, we are dealing with **Example 18**
It should be noted that any given stationary ergodic source is
admissible as N tends to infinity. However, we are dealing with
a class of Markov sources that are characterized by an order K that is allowed to grow with N. 2. Lemma 1 indicates that one may replace X^0_{-N} as a training data 2. Lemma 1 indicates that one may replace X^0_{-N} as a training data

 \bigwedge

 $\left(\begin{array}{c} \mathbf{D} \vdots \\ \mathbf{D} \end{array}\right)$ that is allowed to grow with N.
Lemma 1 indicates that one may replace X_{-N}^0 as a training
by an independent training vector Y_{-N}^0 and a short suffix of Lei $_{\rm{by}}^{\rm{be}}$ at is allowed to grow with N.

Emma 1 indicates that one may replace X_{-N}^0 as a training or
 X_{-N}^0 an independent training vector Y_{-N}^0 and a short suffix of
 X_{-N}^0 , namely $X_{-K_{\text{max}}}^0$ where $K_{\text{max}} = O(\$ negligible deterioration in the prediction error.

Lemma 1 will be used as an analysis tool for the prediction algorithm that is proposed below, which is denoted by $g^u(X^0_{-N}; T_2)$.

$\sqrt{3}$ **3 A Universal Prediction Algorithm** 3 A Universal Predication of X_1, X_{-N}^0 . **3 A Universal Prediction Algorithm**
Consider the suffix of X_1, X_{-N}^0 .
Let $\hat{N} = \frac{N}{(1+T_1^2T_2)}$ where T_1 and T_2 are some positive numbers

 \bigwedge

 $\begin{pmatrix} 3 \\ 1 \\ 7 \end{pmatrix}$ Consider the suffix of X_1, X_{-N}^0 .

Let $\hat{N} = \frac{N}{(1+T_1^2T_2)}$ where T_1 and T_2 are some positive $(T_1 = T_2^2)$ and let *j* be a positive integer $1 \le j \le T_2$. $(T_1 = T_2^2)$ and let *j* be a positive integer $1 \le j \le T_2$.
1. Evaluate $K_0(X^0 \circ_{\hat{N}})$.

-
- % 1. Evaluate $K_0(X^0_{-\hat{N}})$.
2. For each j, denote by t^j the first instant in $X^{-(jT_1^2+1)\hat{N}-1}_{((j+1)T_1^2+1)}$ $-(jT_1^2+1)\hat{N}-1$
 $-((j+1)T_1^2+1)\hat{N}$ for Evaluate $K_0(X_{-\hat{N}}^0)$.

Evaluate $K_0(X_{-\hat{N}}^0)$.

For each j, denote by t^j the first instant in $X_{-(j+1)T_1^2+1)\hat{N}}^{-(j+1)}$ for which $X_{t^j-K_0(X_{-\hat{N}}^0)}^{t^j} = X_{-K_0(X_{-\hat{N}}^0)}^0$. If no such instant exists, set Evaluate $K_0($

For each j, de

which $X_{tj-K_0}^{tj}$
 $t^j = -N - 1$. 2. For each *j*, denote by t^j the first instant in $X_{-(t+1)}^{-(j+1)}$
which $X_{t^j - K_0(X^0_{-\hat{N}})}^{t^j} = X_{-K_0(X^0_{-\hat{N}})}^0$. If no such inst
 $t^j = -N - 1$.
3. Predict X_1 to be the letter $\hat{X} \in \mathbf{A}$ that minimizes:
- $t^j = -N 1.$
3. Predict X_1 to be the letter $\hat{X} \in \mathbf{A}$ that minimizes:
 $\sum_{i=1 \cdot t^j > -N-1}^{T_2} \delta(X_{t_j+1}, \hat{X}).$

 $\begin{pmatrix} 1 \ 1 \ 1 \end{pmatrix}$ **Theorem 1** *Let us assume that the source that emits* X_{-N}^1 *is a stationary ergodic K-th order Markov source that satisfies the condition in Lemma 1 (part 3). (K is unknown to the predictor.) Then, the prediction error that is associated with the universal prediction algorithm above is upper-bounded by:* condition in Lemma 1 (part 3). (K is unknown to the predictor.)
Then, the prediction error that is associated with the universal
prediction algorithm above is upper-bounded by:
 $3aE\delta\left(X_1;g^u(X_{-N}^0;T_2)\right) = E P^u(X_{-N}^0,T_2$ \bigwedge

Theorem 1 Let us assume that the source that emits
$$
X^1_{-N}
$$
 is a
stationary ergodic K-th order Markov source that satisfies the
condition in Lemma 1 (part 3). (K is unknown to the predictor.)
Then, the prediction error that is associated with the universal
prediction algorithm above is upper-bounded by:
 $3aE\delta\left(X_1; g^u(X^0_{-N}; T_2)\right) = E P^u(X^0_{-N}, T_2) \leq E P_{\min}(X^0_{-K(X^0_{-\hat{N}})}))$
 $+ O\left(\frac{1}{T_2}\right) + O\left(\frac{T_2 \log N}{\hat{N}^{\delta}}\right)$ (6)
provided that $K \leq O(\hat{N}^{\frac{1}{3}-3\delta}).$

provided that $K \leq O(\hat{N}^{\frac{1}{3} - 3\delta})$ *.*

 $\sqrt{2}$

Discussion:

 $\sqrt{2}$

& % **iscussion:**
1. By Lemma 1, with probability larger than about $1 - \frac{1}{T_1}$, ussion:
By Lemma 1, with probabilit
 $K_0(X_{-N}^0) \le 0(\log N)$. Also, **ussion:**
By Lemma 1, with probability larger than about $1 - \frac{1}{T_1}$,
 $K_0(X_{-N}^0) \le 0(\log N)$. Also,
 $K_0(X_{-\hat{N}}^0) \le O\left(\log N \left(1 - \frac{1 + \log T_1^2 T_2}{\log N}\right)\right) = O(\log N)$. Hence,
for $\log N \gg \log(T_1^2 T_2)$, $EP^u(X_{-N}^0, T_1)$ is expected $K_0(X_{-N}^0) \le 0(\log N)$. Also,
 $K_0(X_{-\hat{N}}^0) \le O\left(\log N \left(1 - \frac{1 + \log T_1^2 T_2}{\log N}\right)\right) = O(\log N)$.

for $\log N \gg \log(T_1^2 T_2)$, $EP^u(X_{-N}^0, T_1)$ is expected to

roughly equal to $EP^u_{\min}(X_{-\hat{N}}^0)$, which demonstrates the efficiency of the proposed prediction algorithm. This particular algorithm was introduced here because it lends itself to ^a simple analysis. Other similar algorithms might be, perhaps, more efficient.

 \bigwedge

2. Despite the fact that we require N to be large, the results are non-asymptotic since we allow the order K to be $K = O(\hat{N}^{\frac{1}{3}-3\delta}) = O((\frac{N}{1+T_1^2T_2})^{\frac{1}{3}-3\delta})$ and not $K = O(\log N)$ as is customary to assume.

The proof of Theorem 1 appears in the Appendix.

Appendix

Proof of Lemma 1 (part 3): Define $S = [X_{-N}^0 : P(X_{-i}^0) \le 2^{-H_1 i}; i \ge \frac{5 \log N}{H_1}]$. By the Chebytchev inequality,

$$
\Pr\left[P(X_{-j-i}^{-j}) = P(X_{-i}^0 | X_{-i}^0, S) \ge N^2 T_1 P(X_{-i}^0 | S)\right] \le \frac{1}{N^2 T_1}
$$

Also,

$$
\Pr\left(K_0(X_{-N}^0) = i|S\right) \le \sum_{j=1}^N P\left(X_{-j-i}^{-j} = X_{-i}^0|S\right)
$$

But, for some $i \leq \frac{5 \log N}{H_1}$, $\Pr\left[P(X_{-i}^0|S) \leq 2^{-H_1i}\right].$

Hence, it follows that,

$$
\Pr\left[K_0(X_{-N}^0) \ge \frac{5}{H_1}\log N\right] \le \frac{1}{T_1} + \sum_{i=\frac{5}{H_1}\log N}^{N} \sum_{j=1}^{N} \frac{T_1 N^2}{N^5} \le \frac{T_1}{N} + \frac{2}{T_1}
$$

which proves Lemma 1 (part 3).

 $\sqrt{1}$ $\sum_{i=1}^{n}$ **Lemma 1** Lemma A1 is a simple version of strong-mixing $[2]$ and is *an improved and generalized version of Lemma A1 in [1].* **Lemma 1** *Lemma A1 is a simple version of strong-mixing* [2] *and is an improved and generalized version of Lemma A1 in [1].*
Let $t = \frac{5}{H_1} \log N$ *and let* M *and* m *be two positive integers such that* **Lemma 1** Lemma A1 is a simple
an improved and generalized ver
Let $t = \frac{5}{H_1} \log N$ and let M and
 $M = mK + t$ and $m \geq 2$; Then,

improved and generalized version of Lemma A1 in [1].
 $t t = \frac{5}{H_1} \log N$ and let M and m be two positive integers such th
 $= mK + t$ and $m \ge 2$; Then,
 $Pr [P(X_{-t}^1, X_{-N}^{-M}) \le P(X_{-t}^1) P(X_{-N}^{-M})(1 - \epsilon)] \le (1 - \epsilon)^{\frac{M-t}{K} - 1}$

$$
\Pr\left[P(X_{-t}^1, X_{-N}^{-M}) \le P(X_{-t}^1)P(X_{-N}^{-M})(1 - \epsilon)\right] \le (1 - \epsilon)^{\frac{M - t}{K} - 1}
$$
\nwhere ϵ is an arbitrarily small positive number.

\n**Proof of Lemma A1:** The fact that $M > 2K + t$ makes X_{-t}^0 and

where ϵ *is an arbitrarily small positive number.*

 $\begin{pmatrix} \mathbf{L}^d & \mathbf{L}^$ $\begin{pmatrix} i_s \\ i_s \end{pmatrix}$ where
Prood
 X^{-M}_{N} X_{-N}^{-M} essentially independent of each other. This is established by following the (improved and corrected) derivation in the Appendix of [1] below.

 $\sqrt{2}$ By the Markovity of the source,

$$
\begin{array}{ll}\n\text{Markovity of the source,} \\
P\left(X_{-N}^1\right) & = & P\left(X_{-N}^{-M}\right)P\left(X_{-M+1}^{-M+K}\middle|X_{-M-K+1}^{-M}\right) \\
&\cdots P\left(X_{-M+iK+1}^{-M+(i+1)K}\middle|X_{-M+(i-1)K+1}^{-M+iK}\right) \\
&\cdots P\left(X_{-t}^1\middle|X_{-t-K+1}^{-t-1}\right)\n\end{array}
$$

 \bigwedge

where $1 \leq i \leq m - 1$.

 $\begin{pmatrix}\nB \\
B\n\end{pmatrix}$ % Now, if for some i , $\cdots P(X_{-t}^1 | X_{-t-k+1}^{-t-1})$

where $1 \le i \le m-1$.

Now, if for some *i*,
 $P(X_{-M+iK+1}^{-M+iK}| X_{-M+(i-1)K+1}^{-M+iK}) \ge (1 - \epsilon) P(X_{-M+iK+1}^{-M+(i+1)K})$ then it follows that, some i ,
 ${}_{+1}^{(n+1)K}|X-M+iK \atop +1}|X-M+iK \atop +1}(X-M+iK+1}) \geq (1-\epsilon)P(X-M+iK+1)$,
 $P(X-M, X_{-t}^{1}) \geq (1-\epsilon)P(X-M)P(X_{-t}^{1}).$

$$
P\left(X_{-N}^{-M}, X_{-t}^{1}\right) \geq (1 - \epsilon)P\left(X_{-N}^{-M}\right)P\left(X_{-t}^{1}\right).
$$

 \sqrt{T} The Mood of Contracts of Co The probability that no such i exists is the probability of the event The probability that no such *i* exists is the probability of th
that for **each** *i*, given $X_{-M+(i-1)K+1}^{-M+(i+K)}$, $X_{-M+iK+1}^{-M+(i+K)}$ satisfies: probability that no such *i* exists is the probability of the even
for **each** *i*, given X_{-M+i}^{-M+iK} , $X_{-M+iK+1}^{-M+(i+K)}$ satisfies:
 $P(X_{-M+iK+1}^{-M+(i+1)K} | X_{-M+(i-1)K+1}^{-M+iK}) < (1 - \epsilon) P(X_{-M+iK+1}^{-M+(i+1)K})$. $\begin{array}{c} \hline \begin{array}{c} \text{s is}\end{array} \ \hline \begin{array}{c} \text{is}\end{array} \ \hline \begin{array}{c} \text{this}\end{array} \end{array}$

that for each *i*, given
$$
X_{-M+iK}^{-M+iK}
$$
, $X_{-M+iK+1}^{-M+(i+K)}$ satisfies:
\n
$$
P(X_{-M+iK+1}^{-M+(i+1)K}|X_{-M+(i-1)K+1}^{-M+iK}) < (1 - \epsilon)P(X_{-M+iK+1}^{-M+(i+1)K})
$$
\nNow, for each *i* the probability of this event, given $X_{-M+(i-1)K+1}^{-M+iK}$, is

 \bigwedge

 $\begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}$ upper-bounded by

, for each *i* the probability of this event, given
$$
X_{-M+i(-1)K+1}^{-M+iK}
$$
,
\nr-bounded by
\n
$$
\sum_{X_{-M+iK+1}^{-M+(i+1)K} \in \mathbf{A}^K : P\left(X_{-M+iK+1}^{-M+(i+1)K}|X_{-M+(i-1)K+1}^{-M+iK}\right) < (1-\epsilon)P\left(X_{-M+iK+1}^{-M+(i+1)K}|X_{-M+(i-1)K+1}\right)
$$
\n
$$
P\left(X_{-M+iK+1}^{-M+(i+1)K}|X_{-M+(i-1)K+1}^{-M+(i+1)K}\right) \leq (1-\epsilon).
$$
\n
$$
X_{-M+iK+1}^{-M+(i+1)K} \in \mathbf{A}^K
$$

Lemma A1 then follows.

 $\left(\begin{array}{c} \mathbf{C} \ \mathbf{C} \ \mathbf{K} \ \mathbf{X} \ \mathbf{T} \ \mathbf{R} \ \mathbf{C} \ \mathbf{L} \ \mathbf{C} \ \mathbf{I} \ \mathbf{n} \ \mathbf{p} \ \mathbf{I} \ \mathbf{n} \ \mathbf{p} \ \mathbf{I} \ \mathbf$ $\begin{pmatrix} x \\ y \\ z \\ z \\ z \end{pmatrix}$ Clearly, Lemma 1 (part 3) also guarantees that, with high probability, Clearly, Lemma 1 (part 3) also guarantees that, with high probabil $K_0(X_{-N}^0) \le t < K < M$. Thus, if the training data is confined to Clearly, Lemma 1 (part 3) also guarantees that, with high proba $K_0(X_{-N}^0) \le t < K < M$. Thus, if the training data is confined X_{-N}^{-M} it can be treated as being essentially independent of X_{-t}^0 . X_{-N}^{-M} it can be treated as being essentially independent of X_{-t}^{0} .
Thus, X_{-N}^{-M} may be replaced by Y_{-N}^{-M} where Y_{-N}^{0} is an independent realization of the source process. In the following, we will establish the fact that this is essentially the case when the whole training Thus, X_{-N} may be replaced by Y_{-N}
realization of the source process. In the fact that this is essentially the case
vector X_{-N}^0 is being replaced by Y_{-N}^0 .

 \bigwedge

 $\sqrt{2}$

In order to establish that fact, it is enough to show that, with high vector X^0_{-N} is being replaced by Y^0_{-N} .
In order to establish that fact, it is enough to show that, with high
probability, the first recurrence of $X^0_{-K_0}$ in X^0_{-N} occurs within X^{-M}_{-N} . Vector A_{-N}^* is being replaced by Y_{-N}^* .
In order to establish that fact, it is enough to sho
probability, the first recurrence of $X_{-K_0}^0$ in X_{-N}^0
Now, for any positive number Z and any integer Now, for any positive number Z and any integer
 $i : P(X_{-i}^0; K_0 \le t) \ge \frac{Z}{N}$, we have, by Kac's Lemma [4] and by the

Chebytchev inequality that,

$$
Pr\left[n_{X_{-\infty}^{-1}}(X_{-i}^0; K_0 \le t) \ge N\right] \le \frac{1}{Z}.
$$

Also, by definition of K_0 , $X_{-K_0-1}^0$ does not recur in X_{-N}^{-1} and therefore $n_{X_{-\infty}}(X_{-K_0-1}^0; K_0 \le t) \ge N$. Thus, for any $Z \ge 0$,

$$
\Pr\left[P_r(X_{-K_0-1}^0; K_0 \le t) \ge \frac{Z}{N}\right] \le \frac{t}{Z}
$$

Hence, by setting $Z = T Pr(K_0 \le t)$,

$$
\Pr\left[\Pr\left(X_{-K_0-1}^0|K_0\leq t\right)\geq \frac{T}{N}\right]\leq \frac{t}{T\Pr\left(K_0\leq t\right)}
$$

Now, for any positive integer $i \leq t$,

$$
P(X_{-i-1}^0|K_0 \le t) = P(X_{-i}^0|K_0 \le t)P(X_{-i-1}|X_{-i}^0;K_0 \le t).
$$

 \sqrt{A} Also,

$$
\Pr\left[P(X_{-i-1}^0 | X_{-i}^0; K_0 \le t) \le \frac{1}{AT}\right] \le \frac{1}{T},
$$

therefore,

Pr
$$
\left[P(X_{-i-1}^0 | X_{-i}^0; K_0 \le t) \le \frac{1}{AT} \right] \le \frac{1}{T}
$$
,
\npre,
\nPr $\left[P(X_{-K_0}^0 | K_0 \le t) \ge \frac{AT^2}{N} \right] \le \frac{t}{T}$. (7)
\nobserve that for any positive integers *i*, *j*,
\nPr $\left[P(X_{-j-i}^{-j} | X_{-i}^0, K_0 \le t) \ge TP(X_{-i}^0 | K_0 \le t) \right] \le \frac{1}{T}$. (8)
\nby Lemma 1 (part 3) and by the union-bound and by Eq. (8),
\nobability of X_{-K_0} recurring in X_{-M+1} is upper bounded by:
\n
$$
\frac{t}{Pr(K_0 \le t)} M \left[AT^2 \frac{1}{N} + \frac{2}{T} \right] + \frac{2}{T_1} + \frac{T_1}{N}
$$
.
\nling up the probabilities of the r. (8), by Lemma 1

 \bigwedge

Also, observe that for any positive integers i, j ,

$$
\Pr\left[P(X_{-j-i}^{-j}|X_{-i}^{0}, K_0 \le t) \ge TP(X_{-i}^{0}|K_0 \le t)\right] \le \frac{1}{T}.
$$
 (8)

A

th

A

H

H

B

B Then, by Lemma 1 (part 3) and by the union-bound and by Eq. (8), Then, by Lemma 1 (part 3) and by the union-bound and by Eq. (8) the probability of X_{-K_0} recurring in X_{-M+1} is upper bounded by: $\chi \over t$ $(X_{-j-i}^{-j}|X_{-i}^{0}, K_0 \le t) \ge TP(X_{-i}^{0}|K_0 \le t)]$
nma 1 (part 3) and by the union-bound and
y of X_{-K_0} recurring in X_{-M+1} is upper bo
 $\frac{t}{Pr(K_0 \le t)} M \left[AT^2 \frac{1}{N} + \frac{2}{T} \right] + \frac{2}{T_1} + \frac{T_1}{N}$.

$$
\frac{t}{Pr(K_0 \le t)} M \left[AT^2 \frac{1}{N} + \frac{2}{T} \right] + \frac{2}{T_1} + \frac{T_1}{N}
$$

By adding up the probabilities of the r.h.s. of Eq. (8), by Lemma 1

 $\sqrt{\frac{1}{2}}$ \bigwedge (part 3), and by setting $\log T = \frac{\log N}{3}$, $\log M = (\frac{1}{3} - 2\delta) \log N$ and $\epsilon = N^{-\delta}$ one gets the first part of Lemma 1, assuming $K \leq \frac{M\epsilon}{\log N}$. The second par^t of Lemma 1 follows from the r.h.s. of Lemma A1, and the third par^t of Lemma 1 follows from Lemma A1.

We now proceed to outline the proof of Theorem 1.

 $\left(\begin{array}{ccc} \text{(p } & \text{if } &$ $\begin{CD} T_2 \end{CD}$
 \blacksquare Theorem 1 then follows from a variant of Lemma 1 (part 2) where T_2 We now proceed to outline the proof of Theorem 1.
Theorem 1 then follows from a variant of Lemma 1 (part 2) where
independent vectors are being used rather than the single Y_{-N}^0 , by observing that the proposed algorithm is ^a function of strings of We now proceed to outline the proof of Theorem 1.
Theorem 1 then follows from a variant of Lemma 1 (part 2) where T_2
independent vectors are being used rather than the single Y_{-N}^0 , by
observing that the proposed a each other and therefore may be treated as being mutually independent, without essentially affecting the prediction error (thus length $K_0(X^0_{-\hat{N}}) + 1$ which are, with high probability, far apart from each other and therefore may be treated as being mutually independent, without essentially affecting the prediction error (thus contributing no mo

By the union bound by the proof of Lemma 1 par^t 1,

$$
\Pr\left[Pr(K_0(X_{-\hat{N}}^0)) \le \frac{1}{\hat{N}T_1}\right] \le \hat{N}\frac{1}{\hat{N}T_1} + \frac{2}{T_1} + O\left(\frac{\log N}{N^{\delta}}\right).
$$

 \bigwedge

 $\sqrt{2}$

Thus, by applying arguments similar to those that led to par^t 1 of Thus, by applying arguments similar to those that led to part 1 of
Lemma 1, it follows that the probability of $X^0_{-K_0(X^0_{-\hat{N}})}$ not recurring Thus, by applying arguments similar to those that led to part
Lemma 1, it follows that the probability of $X_{-K_0(X_{-\tilde{N}}^0)}^0$ not re
in any of the T_2 vectors of length $T_1^2 \hat{N}$ is upper-bounded by:
 $O\left(\frac{1}{T_2}\right$

$$
O\left(\frac{1}{T_2}\right) + O\left(\frac{\log N}{N^{\delta}}\right)
$$

T

T

L

in

(E

er

e

fo % (By the Chernoff bound, the probability that the empirical prediction (By the Chernoff bound, the probability that the empirical prediction
error which is based on T_2 independent recurrences, will not not be equal to the optimal prediction error, decays exponentially with T_2 , error which is based on T_2 independent recurrences, will not not be equal to the optimal prediction error, decays exponentially with T_2 , for $A = 2$. Similar results may be obtained for $A > 2$).

Concluding Remarks:

 $\sqrt{2}$

It thes X is In CC as se $\begin{array}{c}\n\text{ss} \\
\text{y} \\
\text{y} \\
\text{z}\n\end{array}$ It should be noted that prediction also has ^a connotation other than **CONCRUCING NETALAS.**
It should be noted that prediction also has a connotation other
the one that was introduced here. Given Y_{-N}^{-1} , we may need to It should be noted that prediction also has a connotation other than
the one that was introduced here. Given Y_{-N}^{-1} , we may need to
estimate $P(X_1|X_{-t}^0)$ (in order to predict X_1 given X_{-N}^0 , or compress It should be noted that prediction also has a connotation other than
the one that was introduced here. Given Y_{-N}^{-1} , we may need to
estimate $P(X_1|X_{-t}^0)$ (in order to predict X_1 given X_{-N}^0 , or compres
 X_1 is not available to us. estimate $P(X_1|X_{-t}^{\circ})$ (in order to predict X_1 given X_{-N}° , or X_1 given X_{-N}^0 etc.), in cases where the actual measure $P(X_1|X_{-N}^0)$ is not available to us.
In order to estimate $P(X_1|X_{-N}^0)$ one as

 \bigwedge

In order to estimate $P(X_1|X_{-N}^0)$ one assigns some arbitrary
conditional probability measure $Q(X_1|X_{-N}^0)$ of X_1 hoping that this assigned conditional probability measure will be "close" in some sense to the true $P(X_1|X_{-N}^0)$. Assume that we want to minimize the \sqrt{K} K-L divergence:

$$
E \log \frac{P(X_1|X_{-N}^0)}{Q(X_1|X_{-N}^0)}.
$$

 \bigwedge

 $\begin{pmatrix}\nK \\
S_1 \\
c_2\n\end{pmatrix}$ Similar to Restriction 1 above, we now restrict our attention to the class of predictors L divergence:
 $E \log \frac{P(X_1|X_{-N}^0)}{Q(X_1|X_{-N}^0)}$

milar to Restriction 1 above, we now restrict of

uss of predictors
 $Q = [q : A^{N+1} \rightarrow A] - E \log q(X_1|X_{-N}^0)$

milar to Restriction 1 above, we now restrict our attention to the

\n
$$
\begin{aligned}\n\mathbf{Q} &= [\mathbf{q} : \mathbf{A}^{N+1} \to \mathbf{A}] - E \log q(X_1 | X_{-N}^0) \\
&\geq -E \log P(X_1 | X_{-K_0(X_{-N}^0)}^0, K_0(X_{-N}^0))]\n\end{aligned}
$$

Then, for any predictor in the restricted class and for any finite-alphabet stationary ergodic source,

K-L divergence:
\n
$$
E \log \frac{P(X_1|X_{-N}^0)}{Q(X_1|X_{-N}^0)}.
$$
\nSimilar to Restriction 1 above, we now restrict our attention to the
\nclass of predictors
\n
$$
Q = [q : A^{N+1} \rightarrow A] - E \log q(X_1|X_{-N}^0)
$$
\n
$$
\geq -E \log P(X_1|X_{-K_0(X_{-N}^0)}^0, K_0(X_{-N}^0))]
$$
\nThen, for any predictor in the restricted class and for any
\nfinite-alphabet stationary ergodic source,
\n
$$
E \log \frac{P(X_1|X_{-N}^0)}{Q(X_1|X_{-N}^0)} \geq H(X_1|X_{-K_0(X_{-N}^0)}^0, K_0(X_{-N}^0)) - H(X_1|X_{-N}^0)
$$
\nwhere $H(*)$ denotes entropy.

where $H(*)$ denotes entropy.

 $\sqrt{F_{\rm R}}$ $\begin{pmatrix} Fu & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\$ \bigcap $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ Furthermore, for the class of Markov sources described here, and by using similar arguments as above, ^a variant of the HZ-universal data-compression algorithm [1] (i.e. generating ^a length-function data-compression algorithm [1] (i.e. generating a length-f
that is based on a conditional recurrence time of X_1 given that is based on a conditional recurrence time of X_1 given $X^0_{-K_0(X^0_{N})}$) can be shown to be an efficient predictor in the sense that is based on a conditional recurrence time of X_1
 $X_{-K_0(X_{-N}^0)}^0$ can be shown to be an efficient predict

that it approximates $H(X_1|X_{-K_0(X_{-\hat{N}}^0)}^0, K_0(X_{-N}^0))$. $\int_{-K_0(X_{-N}^0)}^{0}$ can be shot can be shot it approximates H
his may be achieved
1. Set $\hat{N} = \frac{N}{(1+AT_1^3)}$.

This may be achieved by the following algorithm:

1. Set
$$
\hat{N} = \frac{N}{(1 + AT_1^3)}
$$
.

2. Evaluate $K_0(X^0_{\hat{M}})$.

Note that \hat{N} here is set so as to enable, with probability higher Set $\hat{N} = \frac{N}{(1 + AT_1^3)}$.
Evaluate $K_0(X_{-\hat{N}}^0)$.
Note that \hat{N} here is set so as to enable, with probability higher
than $1 - \frac{1}{T_1}$ the recurrence of the concatenation of $X_{K_0(X_{-\hat{N}}^0)}^0$, X_1 $\sqrt{2}$ & \bigwedge $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ for every $X_1 \in \mathbf{A}$ such that $P(X_1 | X^0_{K_0(X^0_{\hat{\sigma}})}) \geq \frac{1}{AT_1}$ $\frac{1}{X}$ $\frac{1}{1}$ $\frac{1}{2}$ $\frac{1}{\sqrt{2}}$ \overline{a} $\frac{1}{X}$ $\frac{0}{K}$ $\begin{array}{c}\n0 \\
0\n\end{array}$ \hat{N} $\frac{1}{\sqrt{2}}$ $\frac{1}{2}$ $\frac{1}{1}$. for every $X_1 \in \mathbf{A}$ such that $P(X_1 | X^0_{K_0(X^0_{-\hat{N}})}) \geq \frac{1}{AT_1}$.
3. For each $X_1 \in \mathbf{A}$, denote by $t(X_1, X^0_{-K_0(X^0_{\hat{N}})})$ the first instant \overline{X} |
|
|
| $\begin{array}{c}\n\vdots \\
\in\n\end{array}$ \int_0^1 $\binom{1}{1}$ $\frac{1}{X}$ $\frac{1}{X_{K_0(X_{-N}^0)}^0}$ $\begin{array}{c}\n0 \\
0 \\
0\n\end{array}$ $\frac{1}{K}$ \hat{v} \hat{N} t in $\frac{1}{x}$ e V^{\prime} $\frac{c}{\hat{N}}$ $\frac{y}{1}$ $\frac{1}{1}$ very $X_1 \in \mathbf{A}$ such that $P(X_1 | X^0_{K_0(X^0_{-\hat{N}})}) \ge \frac{1}{AT_1}$.

each $X_1 \in \mathbf{A}$, denote by $t(X_1, X^0_{-K_0(X^0_{-\hat{N}})})$ the first instant
 $-\hat{N}-1$
 $-(AT_1^3+1)\hat{N}$ for which $X_t^{t+K_0(X^0_{-\hat{N}})} = X_1, X^0_{-K_0(X^0_{-\hat{N}})}$ $\frac{1}{\log X}$ t
t $\begin{array}{c} \n\frac{1}{t} \\
+ \n\end{array}$ $\begin{pmatrix} 1 \\ 2 \\ K \end{pmatrix}$ ζ_1 , ζ_0 (X $\begin{bmatrix} 1 \\ 2 \\ 0 \\ -1 \end{bmatrix}$ $\frac{\hat{\mathbf{N}}^{\prime}}{\hat{N}}$)
)
) $\frac{t}{t}$ $(X$
 $\overline{X_0}$ $\overline{X^0}$
 $\overline{X^0}$ $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ $\overline{A^{\prime}}$ ne $\frac{1}{2}$
0 \hat{N} such instant exists, set $t(X_1, X_{\lceil K_{\varepsilon}(X_1) \rceil}) = -N - 1$. Given i
t $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ ha
e t
X t $P(X_1|X)$
by $t(X_1, X)$
 $X_t^{t+K_0(X^0_{-t})}$ $\angle K$
0 ζ_1 o(\hat{V}
0 \hat{N} $\frac{1}{\sqrt{2}}$ X^0
 X^0 \equiv)
 \hat{N})
 \cdot , \cdot \geq) t
 X_-^0 $\frac{1}{A'}$
he Fc
 in
 su
 X $\frac{1}{2}$
0
= $\frac{K}{K}$ ach $X_1 \in \mathbf{A}$, denote by $t(X_1, X_{-K_0(X_{-\hat{N}}^0)}^0)$ the first instart
 $-\hat{N}-1$
 $-(AT_1^3+1)\hat{N}$ for which $X_t^{t+K_0(X_{-\hat{N}}^0)} = X_1, X_{-K_0(X_{-\hat{N}}^0)}^0$. If r

instant exists, set $t(X_1, X_{-K_0(X_{-\hat{N}}^0)}^0) = -N - 1$. Given $\begin{bmatrix} 1 \ 1 \end{bmatrix}$ \hat{N} $_{\rm)}$, order the the values of their corresponding $t(X_1, X_{-K_0(X^0_{\delta)})}^0$. For each $\frac{1}{t}$ $\begin{pmatrix} 1 \end{pmatrix}$ X^0
 X^0
 X) $=$
 $\frac{1}{2}$
 $\frac{1}{2}$
 $\frac{1}{2}$ $\begin{array}{c}\n\frac{1}{2} \\
\frac{1}{2} \\
\frac{1$ $=$
 $\frac{1}{K}$ $\begin{aligned} -I \\ \text{call} \\ \frac{0}{X} \end{aligned}$ $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ \hat{N} $\frac{1}{X}$ su X the X $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $\begin{array}{c} \mathbf{n} \ \mathbf{w}_0 \ \mathbf{v}_1 \in \mathbb{R} \end{array}$ A, let j $\frac{1}{2}$ $\begin{align} N &\text{is} \ \text{de} \ \text{nei} \ X \end{align}$ $\begin{align*} &\text{ts, s} \ &\text{sr, th} \ &\text{r, cc} \ &\text{r, t} \ &\text{r, c} \end{align*}$ $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ \angle re 0(X $\begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}$ $\frac{1}{\hat{N}}$ $\lim_{t \to \infty} t(X_1, X^0_{-K_0(X^0_{-\hat{N}})})$
(denote the place of $\begin{array}{c} (X \ 1. \end{array}$ co $\begin{array}{c} \text{F} \ X \end{array}$ For each $\frac{1}{1}$ in this lexicographic list. $\begin{array}{c} \text{as} \ \text{ap} \ \text{is} \ \text{I} \ \text{S} \ \text{I} \ \text{X} \end{array}$ $\begin{aligned} &\frac{1}{2}(X_1|X^0_{-K_0(X^0_{-\hat{N}})})) \text{ deno} \ &\text{is } \text{list.} \\\\text{erman and by Lemma 1} \ &\text{if } \text{Hom}(X^0_{K_0(X^0_{-\hat{N}})}) \leq H(X_1|X) \end{aligned}$ $\begin{array}{c} \n\Gamma \ X' \end{array}$ $\frac{\hat{N}}{\hat{N}}$)
em \mathfrak{g}) (\mathfrak{m} а X $\frac{1}{10}$ \hat{N}
: O

4. By Kac's Lemma and by Lemma 1 lexicographic list.

By Kac's Lemma and by Lemma 1
 $E(\log j(X_1) | X^0_{K_0(X^0 \underset{\hat{Y}}{\delta})}) \leq H(X_1 | X^0_{K_0(X^0 \underset{\hat{Y}}{\delta})}) + O(\frac{1}{AT_1}).$ $\frac{1}{j}$ $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ |
|
| ist
na
0
K $\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ $\frac{1}{\hat{N}}$ $\frac{b_1}{b_2}$ \mathbf{u} $\frac{0}{K}$ $\frac{0}{1}$ \hat{N} $)$ \overline{a} |
|
| Furthermore, it is easy to show that

$$
\log j(X_1)|X_{-K_0(X_{-\hat{N}}^0)}^0 + \log \log(A+1)
$$
 is a proper
length-function.

Thus, setting

$$
Q(X_1|X_{-N}^0) = \frac{2}{\sum_{X_1 \in \mathbf{A}} 2^{-\log j(X_1)|X_{-K_0(X_{-N}^0)}^0}}
$$

yields:

$$
-E \log Q(X_1|X_{-N}^0) \le H\left(X_1|X_{-K_0(X_{-\hat{N}}^0)+1}^0, K_0(X_{-\hat{N}}^0)\right) + \log \log(A+1) + O\left(\frac{1}{AT_1}\right).
$$