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

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 apparent only via a short training sequence of $N+1$ letters. We allow N to be a polynomial function of K, the order of the Markov source, rather than the classical case where N is allowed to be exponential in K.

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 past $(N+1)$-sequence which serves as a training sequence is no better than the optimal predictor given only the longest suffix that appeared 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.

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.

1 Introduction

We consider finite-alphabet sequences which are emitted by a stationary source with unknown statistics

$$\begin{aligned} \mathbf{X} &= X_1, X_2, \ldots, X_i, \ldots ; \\ X_1^m &= X_1, X_2, \ldots, X_m ; \\ X_i &\in \mathbf{A} ; \; |\mathbf{A}| = A \, . \end{aligned}$$

Given X^0_{-N}, we need to predict X_1 in cases where the actual measure $P(X_1|X^0_{-N})$ 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), some arbitrary prediction function $f(X^0_{-N})$, 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), namely: $P_{\min}(X^0_{-N}) = \min_{f(*)} E_{X^0_{-N}} \delta(X_1, f(X^0_{-N}))$, where

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

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 $X_1 \in A$ which maximizes $P(X_1|X^0_{-N})$.

Hence, $P_{\min}(X^0_{-N}) = E_{X^0_{-n}} \delta\left(X_1, \arg\max_{X_1 \in A} P(X_1|X^0_{-N})\right)$. In the case of universal prediction, the measure $P(X_1|X^0_{-N})$ is not known, and therefore, $P_{\min}(X^0_{-N})$ is not necessarily achievable (unless N is large enough).

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^0_{-N}, rather than the complete X^0_{-N}, (i.e. $P(X_1|X^0_{-K_0(X^0_{-N})}, K_0(X^0_{-N}))$ where $X^0_{-K_0(X^0_{-N})}$ is the longest suffix of X_1 in X^0_{-N} [1]. More precisely, we make the following restriction:

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

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

($K_0 = -1$ *if* X_0 *does not appear in* X^{-1}_{-N} *where* X^0_1 *is the null string). We restrict our attention to the class of predictors*

$$\begin{aligned} G = [g : \mathbf{A}^{N+1} \quad &\rightarrow \quad \mathbf{A} | E\delta(X_1, g(X^0_{-N}) \geq \\ & E\delta(X_1, \arg\max_{X \in \mathbf{A}} P(X_1 | X^0_{-K_0(X^0_{-N})}, K_0(X^0_{-N}))] \end{aligned}$$

Define,

$$P^u_{\min}(X^0_{-N}) = \min_{g(*)\in G} E_{X^0_{-N}} \delta\Big(X_1, g(X^0_{-N})\Big)$$

where the expectation is taken w.r.t. X_1 given the context X^0_{-N}.
Clearly,

$$P^u_{\min}(X^0_{-N}) \geq P_{\min}(X^0_{-N})\,. \tag{3}$$

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 observed in the training data. Assume that the source that emits X^1_{-N} is a stationary ergodic K-th order Markov source (K is unknown to 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 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, given the positive parameters T_1 and N, satisfy:

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

for some positive number H_1 and some positive integer $\ell \leq \frac{5\log N}{H_1}$. 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

$polynomial$ in K. The algorithm finds the longest suffix of X^0_{-N} that 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 $X' \in \mathbf{A}$ 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 algorithms, where it is inherently assumed that N is, 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).

2 Main Results

Let Y_{-N}^{0} be a realization of the source process, which is independent 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 $0 \leq j \leq N - i$. ($K_0(X_{-N}^{0}|Y_{-N}^{0}) = -1$ if X_0 does not appear in Y_{-N}^{0}). Also, consider the following restriction:

Restriction 2 *We restrict our attention to the class of predictors* $G' = [g' : \mathbf{A}^{N+1} \to \mathbf{A} | g'(X^0_{-N}|Y^0_{-N}) = g'(Z^0_{-N}|Y^0_{-N})$ *whenever* $K_0(X^0_{-N}|Y^0_{-N}) = K_0(Z^0_{-N}|Y^0_{-N}) = k$ *and* $X^0_{-k-1} = Z^0_{-k-1}]$*. Thus,*

$$
\begin{aligned}
5P^u_{\min}\Big(X^0_{-N}|Y^0_{-N}\Big) &= \min_{g'(*)\in G'} E_{X^0_{-N},Y^0_{-N}}\delta\Big(X_1, g'(X^0_{-N}|Y^0_{-N})\Big) \\
&= P_{\min}\Big(X^0_{-K_0(X^0_{-N}|Y^0_{-N})-1}\Big) . \qquad (4)
\end{aligned}
$$

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

Then,

Lemma 1 *Let* Y^0_{-N} *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* *integer. Then,*

1. $$EP^u_{\min}(X^0_{-N}|Y^0_{-N}) \leq EP_{\min}(X^0_{-K(X^0_{-N})-1}) + O(\frac{\log N}{N^\delta}) + \frac{2}{T_1}$$

   $$\leq EP^u_{\min}(X^0_{-N}) + O(\frac{\log N}{N^\delta}) + \frac{2}{T_1}$$

   provided that the (unknown) order of the Markovian source satisfies $K \leq O(N^{\frac{1}{3}-3\delta})$ *where* $0 \leq \delta \leq \frac{1}{9}$.

2. *For any predictor* $g(X^0_{-N})$ *that satisfies*

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

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

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

3.

$$\Pr\Big[K_0(X^0_{-N}|Y^0_{-N}) \geq \frac{5}{H_1}\log N\Big] \leq \frac{2T_1}{H_1 N} + \frac{2}{T_1}\,.$$

$$\Pr\Big[K_0(X^0_{-N}) \geq \frac{5}{H_1}\log N\Big] \leq \frac{2T_1}{H_1 N} + \frac{2}{T_1}\,. \tag{5}$$

The proof of Lemma 1 appears in the Appendix.

Discussion:

1. 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 by an independent training vector ${Y^0}_{-N}$ and a short suffix of ${X^0}_{-N}$, namely ${X^0}_{-K_{\max}}$ where $K_{\max} = O(\log N)$ with only a 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 $g^u({X^0}_{-N}; T_2)$.

3 A Universal Prediction Algorithm

Consider the suffix of X_1, X^0_{-N}.

Let $\hat{N} = \frac{N}{(1+{T_1}^2 T_2)}$ where T_1 and T_2 are some positive numbers $(T_1 = {T_2}^2)$ and let j be a positive integer $1 \leq j \leq T_2$.

1. Evaluate $K_0(X^0_{-\hat{N}})$.

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

3. Predict X_1 to be the letter $\hat{X} \in \mathbf{A}$ that minimizes: $\sum_{j=1:t^j>-N-1}^{T_2} \delta(X_{t_j+1}, \hat{X})$.

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:*

$$\begin{aligned} 3aE\delta\Big(X_1; g^u(X^0_{-N}; T_2)\Big) &= EP^u(X^0_{-N}, T_2) \leq EP_{\min}(X^0_{-K(X^0_{-\hat{N}})})) \\ &+O\left(\frac{1}{T_2}\right) + O\left(\frac{T_2 \log N}{\hat{N}^\delta}\right) \qquad (6) \end{aligned}$$

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

Discussion:

1. By Lemma 1, with probability larger than about $1 - \frac{1}{T_1}$, $K_0(X^0_{-N}) \leq 0(\log N)$. Also, $K_0(X^0_{-\hat{N}}) \leq O\left(\log N \left(1 - \frac{1+\log {T_1}^2 T_2}{\log N}\right)\right) = O(\log N)$. Hence, for $\log N \gg \log({T_1}^2 T_2)$, $EP^u(X^0_{-N}, T_1)$ is expected to be roughly equal to $EP^u_{\min}(X^0_{-\hat{N}})$, 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.

2. Despite the fact that we require N to be large, the results are non-asymptotic since we allow the order K to be $K = O(\hat{N}^{\frac{1}{3}-3\delta}) = O\Big((\frac{N}{1+{T_1}^2 T_2})^{\frac{1}{3}-3\delta}\Big)$ 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 $S = [X^0_{-N} : P(X^0_{-i}) \leq 2^{-H_1 i}; i \geq \frac{5 \log N}{H_1}]$. By the Chebytchev ineqality,

$$\Pr\left[P(X^{-j}_{-j-i}) = P(X^0_{-i} | X^0_{-i}, S) \geq N^2 T_1 P(X^0_{-i} | S)\right] \leq \frac{1}{N^2 T_1} \, .$$

Also,

$$\Pr\left(K_0(X^0_{-N}) = i | S\right) \leq \sum_{j=1}^{N} P\Big(X^{-j}_{-j-i} = X^0_{-i} | S\Big) \, .$$

But, for some $i \leq \frac{5 \log N}{H_1}$,

$$\Pr\left[P(X^0_{-i}|S) \leq 2^{-H_1 i}\right].$$

Hence, it follows that,

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

which proves Lemma 1 (part 3). $\square$

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* $M = mK + t$ *and* $m \geq 2$*; Then,*

$$\Pr\left[P(X^1_{-t}, X^{-M}_{-N}) \leq P(X^1_{-t})P(X^{-M}_{-N})(1-\epsilon)\right] \leq (1-\epsilon)^{\frac{M-t}{K}-1}$$

where ϵ *is an arbitrarily small positive number.*

Proof of Lemma A1: The fact that $M > 2K + t$ makes X^0_{-t} and X^{-M}_{-N} essentially independent of each other. This is established by following the (improved and corrected) derivation in the Appendix of [1] below.

By the Markovity of the source,

$$\begin{aligned} P\Big(X_{-N}^{1}\Big) &= P\Big(X_{-N}^{-M}\Big)P\Big(X_{-M+1}^{-M+K}\Big|X_{-M-K+1}^{-M}\Big) \\ &\cdots P\Big(X_{-M+iK+1}^{-M+(i+1)K}\Big|X_{-M+(i-1)K+1}^{-M+iK}\Big) \\ &\cdots P\Big(X_{-t)}^{1}\Big|X_{-t-K+1}^{-t-1}\Big) \end{aligned}$$

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

Now, if for some i,
$P\Big(X_{-M+iK+1}^{-M+(i+1)K}|X_{-M+(i-1)K+1}^{-M+iK}\Big) \geq (1-\epsilon)P\Big(X_{-M+iK+1}^{-M+(i+1)K}\Big)$ then it follows that,

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

The probability that no such i exists is the probability of the event that for **each** i, given $X_{-M+(i-1)K+1}^{-M+iK}$, $X_{-M+iK+1}^{-M+(i+K)}$ satisfies:

$$P\Big(X_{-M+iK+1}^{-M+(i+1)K}\Big|X_{-M+(i-1)K+1}^{-M+iK}\Big) < (1-\epsilon)P\Big(X_{-M+iK+1}^{-M+(i+1)K}\Big)\,.$$

Now, for each i the probability of this event, given $X_{-M+(i-1)K+1}^{-M+iK}$, is upper-bounded by

$$\sum_{X_{-M+iK+1}^{-M+(i+1)K}\in\mathbf{A}^K:P\left(X_{-M+iK+1}^{-M+(i+1)K}\middle|X_{-M+(i-1)K+1}^{-M+iK}\right)<(1-\epsilon)P\left(X_{-M+iK+1}^{-M+(i+1)K}\right)} P\Big(X_{-M+iK+1}^{-M+(i+1)K}\Big|X_{-M+(i-1)K+1}^{-M+iK}\Big)$$

$$< (1-\epsilon)\sum_{X_{-M+iK+1}^{-M+(i+1)K}\in\mathbf{A}^K} P\Big(X_{-M+iK+1}^{-M+(i+1)K}\Big) \leq (1-\epsilon)\,.$$

Lemma A1 then follows. $\square$

Clearly, Lemma 1 (part 3) also guarantees that, with high probability, $K_0(X^0_{-N}) \leq t < K < M$. Thus, if the training data is confined to X_{-N}^{-M} it can be treated as being essentially independent of X^0_{-t} . Thus, X_{-N}^{-M} may be replaced by Y_{-N}^{-M} where Y^0_{-N} 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 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_{-N}^{-M}.

Now, for any positive number Z and any integer $i : P(X^0_{-i}; K_0 \leq t) \geq \frac{Z}{N}$, we have, by Kac's Lemma [4] and by the

Chebytchev inequality that,

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

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

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

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

$$\Pr\left[\Pr(X^0_{-K_0-1}|K_0 \le t) \ge \frac{T}{N}\right] \le \frac{t}{T\Pr(K_0 \le t)}\,.$$

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

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

Also,

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

therefore,

$$\Pr\left[P(X^0_{-K_0}|K_0 \le t) \ge \frac{AT^2}{N}\right] \le \frac{t}{T}\,. \tag{7}$$

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

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

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:

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

By adding up the probabilities of the r.h.s. of Eq. (8), by Lemma 1 (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 part of Lemma 1 follows from the r.h.s. of Lemma A1, and the third part of Lemma 1 follows from Lemma A1. $\square$

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^0_{-N}, by observing that the proposed algorithm is a function of strings of 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 more than $O(\frac{\log N)}{N^\delta}) + \frac{2T_2}{T_1}$ to the prediction error).

By the union bound by the proof of Lemma 1 part 1,

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

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 in any of the T_2 vectors of length ${T_1}^2\hat{N}$ is upper-bounded by:

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

(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, for $A = 2$. Similar results may be obtained for $A > 2$).

Concluding Remarks:

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 $P(X_1|X^0_{-t})$ (in order to predict X_1 given X^0_{-N}, or compress X_1 given X^0_{-N} etc.), in cases where the actual measure $P(X_1|X^0_{-N})$ is not available to us.

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

K-L divergence:

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

Similar to Restriction 1 above, we now restrict our attention to the class of predictors

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

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

$$E \log \frac{P(X_1|X^0_{-N})}{Q(X_1|X^0_{-N})} \geq H(X_1|X^0_{-K_0(X^0_{-N})}, K_0(X^0_{-N})) - H(X_1|X^0_{-N})$$

where $H(*)$ denotes entropy.

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 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 it approximates $H(X_1|{X^0}_{-K_0({X^0}_{-\hat{N}})}, K_0({X^0}_{-N}))$.

This may be achieved by the following algorithm:

1. Set $\hat{N} = \frac{N}{(1+A{T_1}^3)}$.
2. Evaluate $K_0({X^0}_{-\hat{N}})$.

   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 $X^0_{K_0({X^0}_{-\hat{N}})}, X_1$

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 t in $X^{-\hat{N}-1}_{-(AT_1{}^3+1)\hat{N}}$ for which $X_t^{t+K_0(X^0_{-\hat{N}})} = X_1, {X^0}_{-K_0(X^0_{-\hat{N}})}$. If no such instant exists, set $t(X_1, {X^0}_{-K_0(X^0_{-\hat{N}})}) = -N - 1$. Given ${X^0}_{-K_0(X^0_{-\hat{N}})}$, order the A letters lexicographically according to the values of their corresponding $t(X_1, {X^0}_{-K_0(X^0_{-\hat{N}})})$. For each $X_1 \in \mathbf{A}$, let $j(X_1 | {X^0}_{-K_0(X^0_{-\hat{N}})}))$ denote the place of X_1 in this lexicographic list.

4. By Kac's Lemma and by Lemma 1 $E(\log j(X_1) | X^0_{K_0(X^0_{-\hat{N}})}) \leq H(X_1 | X^0_{K_0(X^0_{-\hat{N}})}) + O(\frac{1}{AT_1})$. Furthermore, it is easy to show that

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

Thus, setting

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

yields:

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