A Universal Predictor Based on Pattern Matching*

September 15, 2001

Philippe Jacquet
INRIA
Rocquencourt
78153 Le Chesnay Cedex
France
Philippe.Jacquet@inria.fr

Wojciech Szpankowski†
Dept. of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.
spa@cs.purdue.edu

Izydor Apostol
Amgen Inc.
One Amgen Center Drive
Thousand Oaks, CA 91320
U.S.A.
iapostol@amgen.com

In Memory of Aaron D. Wyner (1939–1997)

Abstract

We consider a universal predictor based on pattern matching: Given a sequence $X_1, \ldots, X_n$ drawn from a stationary mixing source, it predicts the next symbol X_{n+1} based on selecting a context of X_{n+1}. The predictor, called the *Sampled Pattern Matching* (SPM), is a modification of the Ehrenfeucht–Mycielski pseudo random generator algorithm. It predicts the value of the most frequent symbol appearing at the so called *sampled positions*. These positions follow the occurrences of a fraction of the longest suffix of the original sequence that has another copy inside $X_1 X_2 \ldots X_n$. In other words, in SPM the context selection consists of taking certain fraction of the longest match. The study of the longest match for lossless data compression was initiated by Wyner and Ziv in their 1989 seminal paper. Here, we estimate the redundancy of the SPM universal predictor, that is, we prove that the probability the SPM predictor makes worse decisions than the optimal predictor is $O(n^{-\nu})$ for some $0 < \nu < \frac{1}{2}$ as $n \to \infty$. As a matter of fact, we show that we can predict $K = O(1)$ symbols with the same probability of error.

Index Terms: Optimal predictor, universal predictor, context selection, sequential decision, universal source coding, redundancy of universal predictors, pattern matching, suffix trees.

*The material in this paper was presented in part at the *Colloquium on Mathematics and Computer Science : Algorithms, Trees, Combinatorics and Probabilities*, University of Versailles-St Quentin, 2000.

†The work of this author was supported by the NSF Grant C-CR-9804760, and contract 1419991431A from sponsors of CERIAS at Purdue.

1 Introduction

Prediction is important in communication, control, forecasting, investment, molecular biology, security, and other areas. We understand how to do optimal prediction when the data model is known, but there is a need for designing universal prediction algorithms that will perform well no matter what the underlying probabilistic model is. Universal prediction was subject of extensive research over the last 50 years; it dates back to Shannon [23]. We mention here only a few references: [1, 2, 5, 8, 16, 17, 18, 20, 22]. In this paper we propose a universal predictor based on pattern matching which is a modification of an algorithm proposed by Ehrenfeucht and Mycielski [7] for generating a pseudo random sequence. It could also be viewed as a context selection rule for sequential decision [29], and one can see some resembles to the PPM data compression algorithm [4]. The heart of our scheme is an algorithm that finds the longest suffix of a sequence whose copy is located somewhere inside the sequence. Such a longest match was studied by Wyner and Ziv [30] (cf. also [25]) in the context of lossless compression.

Before we describe in details our algorithm, we first briefly discuss the general prediction problem (cf. [1, 2, 12, 17]). Let $x_1, x_2, \ldots, x_n$ over some finite alphabet $\mathcal{A}$ be given to an observer who tries to predict the next outcome x_{n+1}, or more generally, makes a *decision* b_{n+1} based on the observed data. We consider only *nonanticipatory* predictors whose decisions depend on $x_1, \ldots, x_n$ but not on the future outcomes. Once the real outcome x_{n+1} is revealed, the observer incurs the loss $l(b_{n+1}, x_{n+1})$. The objective of the optimal decision is to minimize this loss function. Throughout the paper, we assume that $b_{n+1} = \hat{x}_{n+1}$ (thus we predict x_{n+1}) and the loss function is the Hamming distance between $\hat{x}_{n+1}$ and x_{n+1}.

The predictor can either be *deterministic* or *random*. For deterministic predictors there is a function f_n such that

$$\hat{x}_{n+1} = f_{n+1}(x_1, \ldots, x_n).$$

For random predictors, one defines a conditional probability distribution, say $q(\cdot|x_1, \ldots, x_n)$, and sets

$$\Pr\{\hat{X}_{n+1} = \hat{x}_{n+1} | X_1 = x_1, \ldots, X_n = x_n\} = q(\hat{x}_{n+1}|x_1, \ldots, x_n),$$

where $X_1, \ldots, X_n$ denote random variables. Finally, we can analyze prediction either in the *probabilistic setting* or the *deterministic setting*. In the probabilistic setting the sequence $X_1, X_2, \ldots$ is generated by a random source with the underlying probability measure P (usually unknown to us) while in the deterministic setting we consider individual sequences.

In this paper, we consider *deterministic* predictors in a *probabilistic setting* with the *Hamming distance* as the loss function. More precisely, we assume that $X_1, X_2, \ldots$ is drawn from a stationary mixing source, and $\hat{X}_{n+1}$ is computed deterministically from the already observed data (i.e., context). In short, the value of $\hat{X}_{n+1}$ is decided by a majority rule of symbols observed at sampled positions that are determined by a pattern matching algorithm described in details below. We shall coin the term *Sampled Pattern Matching* (SPM) predictor for such a scheme.

First, we must understand what is the optimal predictor for *known* source distributions. It is not difficult to prove that for stationary ergodic sources the optimal predictor X^*_{n+1} is given by (cf. [2])

$$X^*_{n+1} := \arg\max_{a\in\mathcal{A}} \Pr\{X_{n+1} = a | X_1 = x_1, \ldots, X_n = x_n\} \tag{1}$$

for all n. The so called *predictability* π^*_n, that is, the average prediction error (in the case of the Hamming distance it is simply the the probability of error $\Pr\{X^*_{n+1} \neq X_{n+1}\}$) is defined as

$$\pi^*_n := \Pr\{X^*_{n+1} \neq X_{n+1}\} = \sum_{x_1,\ldots,x_n} P(x_1,\ldots,x_n) \min_{a\in\mathcal{A}} \left[\Pr\{X_{n+1} \neq a | x_1,\ldots,x_n\}\right], \tag{2}$$

where, throughout this paper, we shall write $P(x_1,\ldots,x_n) := \Pr\{X_1 = x_1,\ldots,X_n = x_n\}$. We illustrate these definitions on memoryless and Markov sources.

Example 1: *Memoryless and Markov Binary Sources* (cf. [16])
1. Memoryless Source. Let $\theta = \Pr\{X_n = 1\}$. Then

$$\begin{aligned} X^*_{n+1} &= 1\left(\theta \geq \frac{1}{2}\right), \\ \pi^*_n &= \min[\theta, 1-\theta], \end{aligned}$$

where $1(A) = 1$ if A occurs, and zero otherwise.
2. Markov Source. Assume for simplicity that X_n is the first order Markov chain. Define $\theta_i = \Pr\{X_{n+1} = 1 | X_n = i\}$ where $i \in \{0,1\}$. Then

$$\begin{aligned} X^*_{n+1} &= 1\left(\theta_i \geq \frac{1}{2}\right), \quad i \in \{0,1\}, \\ \pi^*_n &= \Pr\{X_n = 0\}\min[\theta_0, 1-\theta_0] + \Pr\{X_n = 1\}\min[\theta_1, 1-\theta_1] \end{aligned}$$

for all n. Clearly, $\pi^* = \lim_{n\to\infty} \pi^*_n$ exists for irreducible and aperiodic Markov chains. ■

We now consider *universal* predictors for a *class* of sources $\mathcal{M}$ for which the distribution of the underlying process is not known *a priori* and must be learned from experience. We study here the class $\mathcal{M}$ of stationary mixing sources that we define more precisely in the next section. In this case, the predictability $\hat{\pi}_n(\mathcal{M})$ of the predictor $\hat{X}_{n+1}$ is defined as the average prediction error, that is,

$$\hat{\pi}_n(\mathcal{M}) = \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \Pr\{\hat{X}_i \neq X_i\}.$$

As in source coding, the primary goal of universal prediction is to find predictors that minimize (asymptotically) the predictability $\hat{\pi}_n(\mathcal{M})$ (i.e., they match asymptotically the optimal predictability π_n^*). However, among such predictors one looks for those that minimize the *redundancy*, r_n, defined as the difference between the average prediction error and the *optimal* prediction error presented in (2), that is,

$$r_n := \hat{\pi}_n(\mathcal{M}) - \pi_n^*(\mathcal{M}). \tag{3}$$

Observe, however, that to estimate asymptotically the redundancy it suffices to bound the difference

$$\Pr\{\hat{X}_{n+1} \neq X_{n+1}\} - \Pr\{X_{n+1}^* \neq X_{n+1}\}$$

for $n \to \infty$. But

$$\Pr\{\hat{X}_{n+1} \neq X_{n+1}\} - \Pr\{X_{n+1}^* \neq X_{n+1}\} \leq \Pr\{X_{n+1}^* \neq \hat{X}_{n+1}\}. \tag{4}$$

Thus one can estimate the right-hand side of (4) hoping that the bound is tight enough. This is true for almost all cases (but not all) as illustrated in the next example.

Example 2: *Unbiased versus Biased Binary Memoryless Sources*

Let us consider an unbiased binary memoryless source with both symbols generated with equal probability. By $\widetilde{X}_n$ we denote a very naive estimator that flips an unbiased coin to make decisions whether to predict 0 or 1. We prove that this estimator is optimal. Indeed, for $a = \{0,1\}$ by (1) we have $\Pr\{X_{n+1}^* = a\} = 0.5$, as well as $\Pr\{\widetilde{X}_{n+1} = a\} = 0.5$. Moreover,

$$\Pr\{X_{n+1}^* \neq X_{n+1}\} = \frac{1}{2} \qquad \text{and} \qquad \Pr\{\widetilde{X}_{n+1} \neq X_{n+1}\} = \frac{1}{2},$$

thus $\Pr\{\widetilde{X}_{n+1} \neq X_{n+1}\} - \Pr\{X_{n+1}^* \neq X_{n+1}\} = 0$ and $\widetilde{X}_n$ is an optimal estimator. But the right-hand side of (4) is equal to

$$\Pr\{X_{n+1}^* \neq \widetilde{X}_{n+1}\} = \frac{1}{2}.$$

The bound proposed in (4) is not tight in this case and should not be used (cf. also [8]).

Let us now consider a biased binary source with p denoting the probability of generating 0 and $q := 1-p$, where $p > q$. Clearly, the predictor $\widetilde{X}_n$ suggested above is not good since

$$\Pr\{X_{n+1}^* \neq X_{n+1}\} = q \qquad \text{and} \qquad \Pr\{\widetilde{X}_{n+1} \neq X_{n+1}\} = 2pq > q.$$

We now construct another predictor that makes decisions based on counting the number $N_0(n)$ of 0's and the number $N_1(n)$ of 1's in the sequence $X_1, \ldots, X_n$. The predictor $\hat{X}_{n+1}$ outputs 0 if $N_0(n) \geq N_1(n)$, and predicts 1 if $N_0(n) < N_1(n)$. (We should treat the case $N_0(n) = N_1(n)$ separately, but for our illustrative purpose it is not that important.) Observe that again $\Pr\{X^*_{n+1} \neq X_{n+1}\} = q$ but this time (cf. Lemma 8 of Section 3) for some $\beta > 0$

$$\begin{aligned}\Pr\{\hat{X}_{n+1} \neq X_{n+1}\} &= \Pr\{N_0(n) < N_1(n)\}p + \Pr\{N_0(n) \geq N_1(n)\}q \\ &= pO(e^{-\beta n}) + q(1 - O(e^{-\beta n})) = q + O(e^{-\beta n}).\end{aligned}$$

We also have

$$\Pr\{X^*_{n+1} \neq \hat{X}_{n+1}\} = \Pr\{\hat{X}_{n+1} = 1\} = \Pr\{N_1(n) > N_0(n)\} = O(e^{-\beta n}),$$

therefore, we conclude that the right-hand side of (4) is tight. ∎

In this paper, we propose a universal predictor based on pattern matching that we propose to call the *Sampled Pattern Matching* (SPM). The basic idea of our predictor was already anticipated by Ehrenfeucht and Mycielski [7] (cf. also [12]). The algorithm described in [7] is as follows: For a given $x_1, \ldots, x_n$, let $D_n := n - \ell + 1$ be the maximal suffix $x_\ell, x_{\ell+1}, \ldots, x_n$ that occurs earlier in the sequence $x_1, \ldots, x_n$, that is, the smallest ℓ such that $x_\ell, \ldots, x_n = x_{\ell-i}, \ldots, x_{n-i}$ for some $1 \leq i \leq n$. To construct a predictor, Ehrenfeucht and Mycielski took the smallest i (the most recent occurrence), say I, for which the longest match is found, and set $x_{n+1} = x_{n-I+1}$. It was conjectured in [7, 12] that this is a universal predictor. However, Jacquet [10] (cf. also [18]) proved that the above algorithm is a good density estimator but not a universal predictor. More precisely, Jacquet proved that for memoryless sources $\Pr\{X_{n+1} = a\} = \Pr\{X_{n-I+1} = a\}$ for all $a \in \mathcal{A}$.

To build a universal predictor based on the Ehrenfeucht and Mycielski idea, we consider a fractional maximal suffix, say of length $\lceil \alpha D_n \rceil$ for $0 < \alpha < 1$. We shall show that such a shorter matches occur $O(n^{1-\alpha})$ times with high probability (in short: **whp**) in $X_1, \ldots, X_n$ generated by a stationary mixing source. We find all occurrences of such shorter matches, called further *markers*, in $X_1, \ldots, X_n$ and then apply the majority rule to all symbols that occur just after the markers (i.e., we select the most likely symbol of the *sampled sequence*). We shall prove that such a predictor is asymptotically optimal for mixing sources and its redundancy is $O(n^{-\nu})$ for some $0 < \nu < \frac{1}{2}$ (cf. Theorem 1).

As we mentioned above, there is a large body of literature on prediction (cf. [1, 2, 5, 8, 16, 17, 18, 20, 22]), however, most are either restricted to individual sequences or Markovian models. In particular, in [16] Merhav, Feder, and Gutman proved that a standard majority predictor (as described in the second part of Example 2) is asymptotically optimal for Markov chains of *known* order with the redundancy $O(1/n)$. A more general sources were considered by Weinberger, Rissanen and Feder [29] who proved that for the so called *tree sources* (of finite memory) the majority rule predictor is asymptotically optimal with the redundancy bounded from the above by $\sum_{s\in S} C_s/n = O(1/n)$ where s is the set of context and C_s a constant. In [29] the authors select a context over its parent only if it yields a shorter code length for the past occurrences of symbols in that context. Our SPM predictor is asymptotically optimal for mixing sources that includes Markov sources of unknown order as well as tree sources. However, redundancy of such a predictor is $O(n^{-\nu})$ for some $0 < \nu < \frac{1}{2}$.[1] Also, the SPM predictor seems to have an algorithmic edge since we can provide an efficient implementation based on suffix trees (see Section 2.1).

In passing we mention that the SPM predictor somewhat resembles the PPM (Prediction by Pattern Matching) data compression algorithm of Cleary and Witten [4]. In fact, our context selection rule can be used for a data compression scheme. In PPM the "decision rule" depends on the number of times a (long) match occurs in the text. To be more precise, let the *longest* suffix that occurs at least twice be of the length $1/h(\log n - \ell(n))$ where $\ell(n) = O(\log n)$ and h is the entropy rate of the source. It is not difficult to prove (see Lemma 4) that such a suffix occurs $O(2^{\ell(n)})$ times in the original string of length n. For the Lempel-Ziv scheme we have $\ell(n) = O(1)$ and therefore the *longest* suffix appears $O(1)$ times, while in our SPM algorithm we set $\ell(n) = (1-\alpha)\log n$, and then the α–fractional match occurs $O(n^{1-\alpha})$ times. In PPM $\ell(n)$ seems to be $o(\log n)$.

The paper is organized as follows. In the next section we describe the *Sampled Pattern Matching* predictor, and argue its asymptotic optimality for a class of mixing sources (cf. Theorem 1). The proof of the main result is delayed till the last section. In passing we should mention that we did apply SPM to the prediction of molecular sequences showing its suitability to proteins and DNA predictions (cf. [11]).

2 Main Results

We start this section with a precise description of the Sampled Pattern Matching (SPM) predictor, and show how to implement it efficiently using suffix trees. Then we formulate our main theoretical results.

[1] It is an interesting open problem to determine the best possible redundancy for mixing sources.

2.1 Sampled Pattern Matching Predictor

It is assumed that a sequence $x_1^n := x_1, \ldots, x_n$ is given. Each symbol x_i belongs to a finite alphabet $\mathcal{A}$ of size $V := |\mathcal{A}|$. For a fixed integer $K \geq 1$, the algorithm will predict the next K symbols,[2] that is, $(\hat{x}_{n+1}, \ldots, \hat{x}_{n+K})$. However, throughout the paper we carry out the analysis of the algorithm only for $K = 1$.

Let us fix $0 < \alpha < 1$. The SPM prediction algorithm works as follows:

1. Find the largest suffix of x_1^n whose copy appears somewhere in the string x_1^n. We call this suffix the *maximal suffix* and denote its length by D_n. More precisely, $D_n := l$ where l is the largest integer such that

$$(x_{n-l+1}, \ldots, x_n) = (x_{n-i-l+1}, \ldots, x_{n-i})$$

   for some $1 \leq i \leq n$.

2. Take an α fraction of the maximal suffix of length $k_n := \lceil \alpha D_n \rceil$, that is, the suffix $x_{n-k_n+1}, \ldots, x_n$. Such a fractional suffix occurs more than twice in the original string. Let $L_n \geq 2$ be the actual number of times $x_{n-k_n+1}, \ldots, x_n$ appears in the string x_1^n. Each such a occurrence defines a **marker** (substring), and the K positions after a marker will be called the **marked positions**. Finally, by a **sampled sequence** we mean the sequence composed of all symbols from the K-tuple marked positions. We shall use these notations throughout the paper.

3. Let now $N(x_1, \ldots, x_K)$ be the number of non-overlapping K-tuple $(x_1, \ldots, x_K)$ occurrences in the sampled sequence. The SPM predictor assigns

$$(\hat{x}_{n+1}, \ldots, \hat{x}_{n+K}) = \arg\max N(x_1, \ldots, x_K) \tag{5}$$

   with a tie broken in an arbitrary manner (e.g., by a random selection). In words, $(\hat{x}_{n+1}, \ldots, \hat{x}_{n+K})$ is assigned to the most frequent K-tuple occurring in the sampled sequence.

We illustrate the SPM algorithm in the following example.

Example 3. *SPM Predictor for $K = 1$*

Below is presented a text with the largest suffix and its copy framed (defined in Step 1 of the algorithm):

[2] In some applications (e.g., molecular biology) one may need to predict simultaneously more than one symbol.

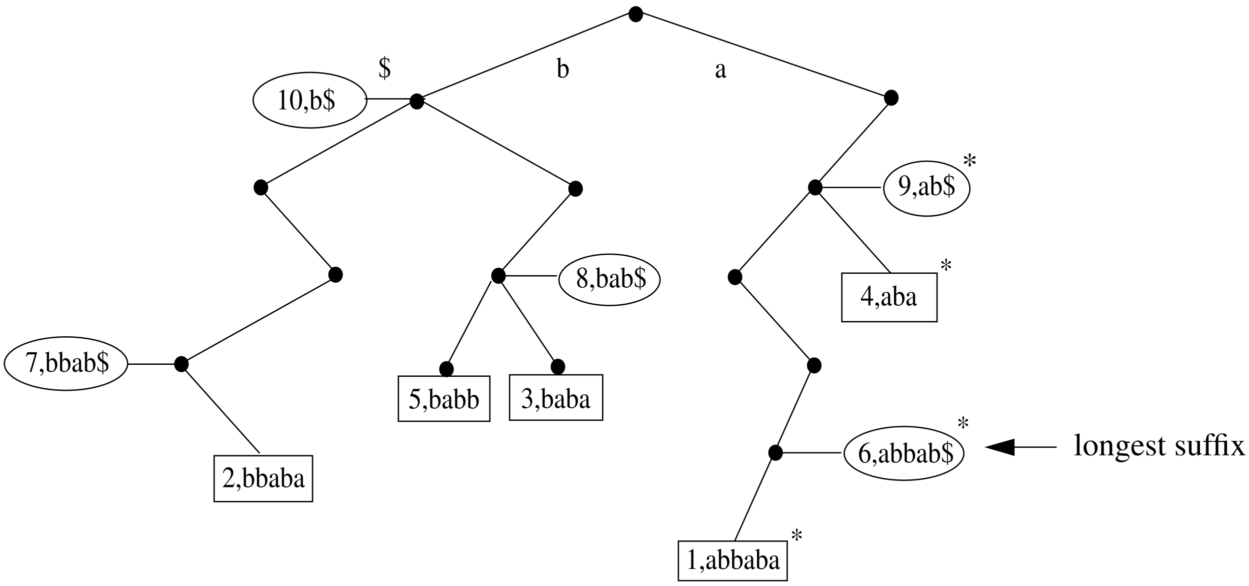

Figure 1: The suffix tree of `abbababbab$` with its longest suffix and markers shown (denoted by asterisks).

$\text{SLJZGGDL}\boxed{\text{YGSJSLJZ}}\underline{\text{K}}\text{GSSLJZIDSLJZJGZ}\boxed{\text{YGSJSLJZ}}$

In fact, $D_{40} = 8$. Let $\alpha = 0.5$. Then the fractional suffix SLJZ is used to find all markers. They are shown below:

$\boxed{\text{SLJZ}}\underline{\text{G}}\text{GDLYGSJ}\boxed{\text{SLJZ}}\underline{\text{K}}\text{GS}\boxed{\text{SLJZ}}\ \underline{\text{K}}\text{LJZJGZYGSJ}\boxed{\text{SLJZ}}$

The sampled sequence is GKK, thus the SPM predicts $\hat{x}_{41} = K$. ■

The next question is how to compute efficiently the longest suffix, markers, and the predicted symbol $\hat{x}_{n+1}$. We propose to use the *suffix tree* construction (cf. [9, 26]). The suffix tree of $x_1, \ldots, x_n$ is a *trie* (i.e., a digital tree) built from all suffixes of $x_1, \ldots, x_n\$$ where $ is a special symbol that does not belong to the alphabet $\mathcal{A}$. External nodes of such a suffix tree contain information about the the suffix positions in the original string and the substring itself that leads to this node (cf. Figure 1). In addition, we keep pointers to those external nodes that contain suffixes ending with the special symbol $ (since one of them will be the longest suffix that we are looking for; in the fact the one with the longest path). Figure 1 shows the suffix tree constructed for $x_1^{10}\$ = $ `abbababbab$`. The external nodes containing suffixes ending with $ are denoted by ovals. Observe that in Figure 1 the node containing (6, `abbab`) leads to the longest suffix $x_6^{10} =$ `abbab` of length $D_{10} = 5$ occurring also at $x_1^5 =$ `abbab`. It is very easy to find all markers once the suffix tree is built. Indeed, they are located in the subtree that can be reached following the last $\lceil \alpha D_n \rceil$ symbols of the longest suffix. In Figure 1 for $\alpha = 0.5$ we chosen the fractional suffix to be `ab` which occurs at position $1, 6, 4$ and 9 as can be read directly from the subtree reached by following the path `ab` (see the nodes denoted by an asterisk). Reading the most frequent symbol (say for $K = 1$) is also simple: We only need to consider strings contained in these nodes (marked by asterisks in Figure 1).

It is well known that a suffix tree of x_1^n can be built in $O(n)$ in the worst case (cf. [9]). This algorithm, due to Weiner (cf. [9]), is quite complicated. One may want to use a simple brute-force algorithm that runs on average in $O(n \log n)$ (cf. [25]). Moreover, it is easy to update the suffix tree when the new symbol x_{n+1} is added. The only nodes that we must look at are the ones with \$ to which we keep pointers. In the worst case, we need to inspect $O(n)$ nodes, but on average only $O(n^{1-\alpha})$ (cf. Lemma 4). This is another advantage over the majority predictor proposed in [16].

2.2 Average Redundancy of the SPM

The prime goal of this work is to derive the redundancy of the SPM algorithm for a class of mixing models $\mathcal{M}$ that we describe next (cf. [3, 24]):

(MX) (Strongly) ψ-Mixing Source

Let $\mathbb{F}_m^n$ be a σ-field generated by $X_{k=m}^n$ for $m \leq n$. The source is called *mixing*, if there exists a bounded function $\psi(g)$ such that for all $m, g \geq 1$ and any two events $A \in \mathbb{F}_1^m$ and $B \in \mathbb{F}_{m+g}^\infty$ the following holds

$$(1 - \psi(g))\Pr\{A\}\Pr\{B\} \leq \Pr\{AB\} \leq (1 + \psi(g))\Pr\{A\}\Pr\{B\}. \qquad (6)$$

If, in addition, $\lim_{g\to\infty} \psi(g) = 0$, then the source is called *strongly* mixing. Hereafter, we consider only strongly ψ-mixing sources and we shall call them *mixing sources.*

It is known that memoryless sources are mixing with $\psi(g) = 0$, and Markov sources over a finite alphabet are strongly mixing with $\psi(g) = O(\rho^g)$ for some $\rho < 1$ (cf. [3, 26]).

Our main result is summarized next. It asserts that the SPM predictor is asymptotically optimal and its average redundancy is $O(n^{-\nu})$ for some $\nu > 0$. We recall the optimal predictability (i.e., the average prediction error) $\pi_n^*(\mathcal{M})$ is computed for the best predictor for known source statistics. In our setting the optimal predictor is defined as

$$(X_{n+1}^*, \ldots, X_{n+K}^*) := \arg \max_{(a_1,\ldots,a_K)\in\mathcal{A}^K} \Pr\{X_{n+1} = a_1, \ldots, X_{n+K} = a_K | x_1, \ldots, x_n\}$$

for all $(x_1, \ldots, x_n) \in \mathcal{A}^n$. The proof of the main result for $K = 1$ is presented in the next section.

Theorem 1 *Let $\alpha > \frac{1}{2}$ and K be fixed. Consider the Sampled Pattern Matching algorithm that predicts the next K outcomes of a sequence $X_1, \ldots, X_n$ drawn from a ψ-mixing source $\mathcal{M}$. Then there exists $0 < \nu < \frac{1}{2}$ such that for $n \to \infty$*

$$r_n = \hat{\pi}_n(\mathcal{M}) - \pi_n^*(\mathcal{M}) = O(n^{-\nu}) \tag{7}$$

provided the ψ mixing coefficient satisfies

$$\lim_{n\to\infty} n^{1-\alpha}\psi(n^{\varepsilon}) = 0 \tag{8}$$

for any arbitrary small $\varepsilon > 0$.

Remark. The restriction $\alpha > \frac{1}{2}$ is necessary to assure that the crucial *marker separation* property (cf. next section) holds. This property says that **whp** two markers are not too close to each others. The SPM may still work for $\alpha < \frac{1}{2}$ but then its average redundancy will decay to zero in a slower pace. However, the proof presented in the next section does not cover such an extension.

3 Proof of the Main Result

We shall prove Theorem 1 using a combination of probabilistic and combinatorial methods. The reader is referred to the recent book [26] for in-depth discussion of these tools. We start with some definitions following by a series of technical lemmas that will lead us to the main result.

In the sequel, we shall need Rényi's entropy, rate of convergence to Shannon entropy, the *Asymptotic Equipartition Property* (AEP), and the Azuma inequality that we briefly review below (cf. [6, 15, 26]).

For $-\infty \le b \le \infty$, the bth *order Rényi entropy* is defined as

$$h_b = \lim_{n\to\infty} \frac{-\log \mathbf{E}[P^b(X_1^n)]}{bn} = \lim_{n\to\infty} \frac{-\log\left(\sum_{w\in\mathcal{A}^n} P^{b+1}(w)\right)^{1/b}}{n}, \tag{9}$$

provided the above limit exists. In the above, we write $P(w) = \Pr\{X_1^n = w\}$ for $w \in \mathcal{A}^n$. It is known (e.g., see [24, 26]) that for mixing processes the Rényi entropies exist. Observe that Shannon entropy $h = \lim_{b\to 0} h_b$. Moreover, by the Shannon-McMillan-Breiman theorem the convergence to Shannon entropy is also in the almost sure sense. Then the AEP states: *For* *a stationary and ergodic sequence X_1^n, for given $\varepsilon > 0$ the state space $\mathcal{A}^n$ can be partitioned into two subsets, $\mathcal{B}_n^\varepsilon$ ("bad set") and $\mathcal{G}_n^\varepsilon$ ("good set"), such that there is N_ε so that for $n \geq N_\varepsilon$ we have*

$$
\begin{aligned}
2^{-nh(1+\varepsilon)} &\leq P(x_1^n) \leq 2^{-nh(1-\varepsilon)} \quad \text{for} \quad x_1^n \in \mathcal{G}_n^\varepsilon, && (10)\\
\lim_{n\to\infty} P(\mathcal{B}_n^\varepsilon) &= 0. && (11)
\end{aligned}
$$

In general, there is no universal rate of convergence to the entropy h, however, for sources satisfying the so called *Blowing-up Property* Marton and Shields [14] proved that the convergence rate in the AEP is exponential, that is, $P(\mathcal{B}_n^\varepsilon)$ converges exponentially fast to zero for such processes. In particular, Shields [24] showed that for mixing processes there exists $\omega > 0$ such that

$$P(\mathcal{B}_n^\varepsilon) = O(2^{-\omega n}) \tag{12}$$

for large n.

3.1 A Road-map to the Proof

Before we proceed with a formal proof we present here a "guided tour" through the main thrust of our approach. As mentioned before, we only consider the case $K = 1$. In order to establish a bound for the prediction redundancy, we shall show that (4), that is,

$$\Pr\{\hat{X}_{n+1} \neq X_{n+1}\} - \Pr\{X_{n+1}^* \neq X_{n+1}\} \leq \Pr\{X_{n+1}^* \neq \hat{X}_{n+1}\}$$

is small for $n \to \infty$. As pointed out in Example 2, the right-hand side of the above might not be tight for some cases (e.g., when probabilities of generating symbols are indistinguishable), and we must handle them separately. However, the core of the proof is common to both cases.

The main theorem will follow from the fact that the sampled sequence is mixing. In Lemma 7 we establish this fact which we call the *mixing property.*

Property 1 (Mixing of the sampled sequence) *The sampled sequence is mixing with probability $P(X_{n+1}|X_1^n)$ provided (8) holds for $n \to \infty$.*

Knowing this, it is easy to prove our main result. Indeed, the majority rule for an (almost) i.i.d. sampled sequence suggests to predict symbol $a \in \mathcal{A}$ such that $\arg\max_a\{P(X_{n+1} = a|X_1^n)\}$ provided that the number of markers tends to infinity. This result is qualitatively equivalent to the main theorem. Details can be found in the forthcoming subsections.

The mixing property of the sampled sequence is a consequence of two crucial properties, namely:

- the *marker separation property*;

- the *marker stability property*.

The marker separation property is formulated next. It is used in establishing the mixing property. We will prove this property in Lemma 3 where the condition $\alpha > 1/2$ is required.

Property 2 (Marker separation property) *There exists $\varepsilon > 0$ such that for $\alpha > \frac{1}{2}$ with high probability two consecutive markers in the string X_1^n cannot be closer than n^ε as $n \to \infty$.*

The separation property together with the mixing condition of the original sequence show that a pair of consecutive markers tend to be independent as $n \to \infty$. This should lead to the proof of the mixing property of the sampled sequence, however, we must first take care of another detail. Observe that a modification of one part of the string may change the positions of the markers in other parts of the string. Fortunately, this happens very rarely as the next marker stability property asserts.

Property 3 (Marker stability property) *There exists $\varepsilon > 0$ such that with high probability there exists no modification of any of the $\lceil n^\varepsilon \rceil$ symbols after a marker that changes string X_1^n into a new string $\tilde{X_1^n}$ with different marker positions.*

In the next subsections we prove in sequel the marker separation property, the marker stability property, and the mixing property of the sampled sequence. Finally, in Section 3.5 we complete the proof of Theorem 1.

3.2 Marker Separation Property

We establish here the marker separation property. We first show in Lemma 1 that the largest suffix D_n is of length $\frac{1}{h}\log n$ **whp** (with high probability). This will lead to Lemma 3 which is a formal statement of the separation property. In addition, we show in Lemma 4 that **whp** the number of markers is $n^{1-\alpha}$ which is also required for the proof of the main result, as discussed above.

Lemma 1 *For a string X_1^n generated by a mixing source, let D_n be the length of largest suffix of X_1^n that has a copy inside X_1^n, that is,*

$$D_n = \max\{l : \exists_{1 \le i \le n-l+1} \;\; X_{n-l+1}^n = X_i^{i+l-1}\}.$$

For any $\varepsilon > 0$

$$\Pr\left\{(1-\varepsilon)\frac{\log n}{h} < D_n < (1+\varepsilon)\frac{\log n}{h}\right\} = 1 - O\left(\frac{\log n}{n^\varepsilon}\right)$$

provided the ψ-mixing coefficient satisfies (8) of Theorem 1.

Proof. This was basically proved in [13, 25, 28] (cf. also [30]) using the first and the second moment methods (cf. [26]). We provide here only a sketch of the proof. Let $w \in \mathcal{G}_k^{\varepsilon/2}$. Then for $k = (1+\varepsilon)h^{-1}\log n$

$$\begin{aligned}
\Pr\{D_n \geq k\} &\leq \sum_{i=1}^{n-k} \sum_{w \in \mathcal{G}_k^{\varepsilon/2}} \Pr\{X_i^{i+k-1} = X_{n-k+1}^n = w\} + P(\mathcal{B}_k^\varepsilon) \\
&\leq \sum_{i=1}^{n-k} (1 + \psi(n-2k-i+2))2^{-kh(1-\varepsilon/2)} + P(\mathcal{B}_k^\varepsilon) \\
&\leq O(\max\{n^{-\varepsilon/2}, P(\mathcal{B}_{\log n}^\varepsilon)\})
\end{aligned}$$

for any $\varepsilon > 0$. By (12) the upper bound is established.

The lower bound is more intricate, but follows the standard approach of "loosing up" the dependency by deleting $n^{\varepsilon/4}$ letters after ever symbol of X_1^n. The derivation from [13] lead us to for $k = (1-\varepsilon)h^{-1}\log n$

$$\Pr\{D_n < k\} \leq 2\psi(n^{\varepsilon/4}) + O(\log n/n^{\varepsilon/4}).$$

This completes the proof since $\psi(n^{\varepsilon/4}) = O(n^{-\varepsilon})$ under (8). ■

Remark. We should point out that (8) is not necessary for Lemma 1 to be true. In general, the rate of convergence is $O(\max\{\psi(n^{-\varepsilon/4}), n^{-\varepsilon}\})$ (cf. [13, 28]). ■

In the sequel, we must study the way markers may overlap. For two strings X and Y we denote $C(X, Y)$ the length of the longest common prefix of both X and Y. The next lemma presents an estimate on the tail of the probability distribution of $C(X_i^\infty, X_j^\infty)$ where X_i^∞ and X_j^∞ are substrings of a string generated by a mixing model.

Lemma 2 *There exists $\xi > 0$ such that for any $1 \leq i \neq j \leq n$*

$$\Pr\{C(X_i^n, X_j^n) \geq k\} \leq c2^{-\xi k} \tag{13}$$

where $c > 0$ is a constant.

Proof. We shall follow the proof of [25]. To simplify the notation let $C_{i,j} = C(X_i^n, X_i^n)$ and $j = i + d - 1$, that is, X_j^∞ is d shifted version of X_i^∞. When $d > k$ the situation is quite simple (there is no overlap), so we concentrate on the case $1 \le d \le k$. Let $w_d \in \mathcal{A}^d$ be a word of length d. Since both strings overlap on $k + d$ positions, there exists w_d such that $X_i^{i+k+d-1} = w_d^{\lfloor \frac{k}{d} \rfloor + 1} \overline{w}_d$ and $X_i + d^{i+k+2d-1} = w_d^{\lfloor \frac{k}{d} \rfloor + 1} \overline{w}_d$ where $\overline{w}_d$ is a prefix of w_d (cf. [25, 26]). Thus we have

$$\begin{aligned}
\Pr\{C_{i,i+d} \ge k\} &= \sum_{\mathcal{A}^d} P(w_d^{\lfloor k/d \rfloor + 1} \overline{w}_d) && (14)\\
&\le c \sum_{\mathcal{A}^d} P(w_d^{\lfloor k/d \rfloor} \overline{w}_d) P(w_d) && (15)\\
&\le c \sqrt{\sum_{\mathcal{A}^d} P^2(w_d^{\lfloor k/d \rfloor} \overline{w}_d) P(w_d)} \le c \sqrt{\sum_{\mathcal{A}^d} P^2(w_d^{\lfloor k/d \rfloor} \overline{w}_d)} && (16)\\
&\le c \sqrt{\sum_{\mathcal{A}^k} P^2(w_k)} = c_1 \sqrt{\mathbf{E}[P(w_k)]} && (17)\\
&\le c 2^{-\frac{1}{2} k h_1 (1-\varepsilon)} && (18)
\end{aligned}$$

where (15) is due to the mixing condition, (16) is a consequence of the *inequality on means* (cf. [26]), (17) follows from $\mathcal{A}^d \subset \mathcal{A}^k$, and (18) is a consequence of the definition (9) of the Rényi's entropy h_1 of order $b = 1$. In the above, the constant c_1 may change from line to line and $\varepsilon > 0$ is any arbitrary small constant. This completes the proof after setting $\xi = \frac{1}{2} h_1 (1 - \varepsilon)$. ■

The next lemma is at the heart of our proof, and it establishes the marker separation property. It says that **whp** markers cannot overlap and in fact cannot be too close to each others. Below $\varepsilon > 0$ stands for a small positive number and c is constant that may change from line to line.

Lemma 3 *For any $\varepsilon > 0$ and $\alpha > \frac{1}{2}$, the probability that for $k \ge \alpha \frac{\log n}{h}$, a string X_1^n contains two consecutive copies of X_{n-k+1}^n that are separated by less than $d = \lceil n^\varepsilon \rceil$ symbols is $O(n^{-\nu})$ with*

$$-\nu = \max\left\{1 - 2\alpha + \varepsilon, -\alpha + \varepsilon, -\omega \frac{\alpha}{h}(1 - \varepsilon), -\frac{h_1}{2h} \alpha \xi \varepsilon\right\},$$

where ω and ξ are defined in (12) and (13), respectively.

Proof. We start by formalizing the statement of the lemma. Define the set $\mathcal{E}_n$ as

$$\mathcal{E}_n := \{X_1^n : \exists_{1 \le i \le n} \exists_{i \le j \le i+d} : X_i^{i+k-1} = X_j^{j+k-1} = X_{n-k+1}^n\}.$$

To prove the lemma it suffices to estimate $P(\mathcal{E}_n)$ and show that it is $O(n^{-\nu})$.

Let us consider two substrings X_i^{i+k-1} and X_j^{j+k-1}. Let the integer $g = \max\{j-i-k+1, 0\}$ be called the *gap* between the substrings. We assume that $g < n^{\varepsilon}$. We define also the distance d between the substrings X_i^{i+k-1} and X_j^{j+k-1} as $d = j - i$ $(j \geq i)$. Clearly $d = j - i \leq k + g$. Observe that strings in $\mathcal{E}_n$ may have markers that may overlap, or may have two markers within distance d without overlapping, or may have a marker within distance d from the suffix X_{n-k+1}^n. To analyze these three case we consider the following three subsets:

- $\mathcal{O}_n$: set of strings X_1^n such that the suffix X_{n-k+1}^n and its copy overlap on more than εk positions;
- $\mathcal{E}_n^1$: set of strings X_1^n such that $X_1^n \notin \mathcal{O}_n$ and X_{n-k-d}^n contains another copy of X_{n-k+1}^n;
- $\mathcal{E}_n^2$: set of strings X_1^n such that $X_1^n \notin \mathcal{O}_n$ and two consecutive copies (i.e., markers) of X_{n-k+1}^n are within distance smaller than d.

Notice that $\mathcal{E}_n \subset \mathcal{O}_n \cup \mathcal{E}_n^1 \cup \mathcal{E}_n^2$. By Lemma 2 we can bound the probability of $\mathcal{O}_n$ as follows

$$P(\mathcal{O}_n) \leq ck2^{-\xi\varepsilon k} = O(n^{-\mu})$$

where $\mu = \frac{h_1}{2h}\xi\alpha\varepsilon$. Thus, now we concentrate on evaluating the probability of the other two sets. Observe that

$$P(\mathcal{E}_n^1) = \sum_{w_k \in \mathcal{A}^k - \mathcal{O}_n} \Pr\{\exists_{0<j\leq k+g} : X_{n-k-i}^{n-i} = X_{n-k}^n = w_k\}.$$

Using Lemma 1, Asymptotic Equipartition Property (AEP), and mixing condition (6), we obtain (to simplify notations we write below $k(1-\varepsilon)$ for $\lfloor k(1-\varepsilon)\rfloor$):

$$\begin{aligned}
P(\mathcal{E}_n^1) &\leq c \sum_{w_{k(1-\varepsilon)} \in \mathcal{A}^{k(1-\varepsilon)}} \Pr\{\exists_{k(1-\varepsilon)\leq i\leq k+g}\ X_{n-k(1-\varepsilon)-i+1}^{n-i} = w_{k(1-\varepsilon)}\}\Pr\{X_{n-k(1-\varepsilon)+1}^n = w_{k(1-\varepsilon)}\} \\
&\leq cP(\mathcal{B}_{k(1-\varepsilon)}^{\varepsilon}) + c(k+g) \sum_{w_{k(1-\varepsilon)} \in \mathcal{G}_{k(1-\varepsilon)}^{\varepsilon}} P^2(w_{k(1-\varepsilon)}) \\
&\leq cP(\mathcal{B}_{k(1-\varepsilon)}^{\varepsilon}) + c(k+g)2^{h(1-\varepsilon)^2 k} \\
&\leq cP(\mathcal{B}_{k(1-\varepsilon)}^{\varepsilon}) + cn^{-\alpha+O(\varepsilon)}.
\end{aligned}$$

The probability of $\mathcal{E}_n^2$, formally satisfies the following identity

$$P(\mathcal{E}_n^2) = \sum_{w_k \in \mathcal{A}^k - \mathcal{O}_n} \Pr\{\exists_{m<n}\exists_{0<j\leq k+d} :\ \ X_{m-k-j}^{m-j} = X_{m-k}^m = X_{n-k}^n = w_k\}. \tag{21}$$

Using the same arguments as above we conclude that

$$
\begin{aligned}
P(\mathcal{E}_n^2) &\leq cP(\mathcal{B}^\varepsilon_{k(1-\varepsilon)}) + cn(k+g) \sum_{w_{k(1-\varepsilon)} \in \mathcal{G}^\varepsilon_{k(1-\varepsilon)}} P^3(w_{k-d}) \\
&\leq cP(\mathcal{B}^\varepsilon_{k(1-\varepsilon)}) + 2cnn^\varepsilon 2^{-2h(1-\varepsilon)^2 k} \\
&\leq cP(\mathcal{B}^\varepsilon_{k(1-\varepsilon)}) + cn^{1-2\alpha+O(\varepsilon)}.
\end{aligned}
$$

Combining the previous estimates we prove the lemma. ∎

Remark: For $\varepsilon \to 0$ we have $0 < \nu < \frac{1}{2}$ for $\alpha > \frac{1}{2}$. The condition $\alpha > \frac{1}{2}$ is required only in the proof of this lemma.

Let now L_n be the number of markers (of length $k = \lfloor \alpha D_n \rfloor$). We shall prove that **whp** $L_n \sim n^{1-\alpha-\varepsilon}$ where $\varepsilon > 0$ is an arbitrary positive number. Actually, we only need a lower bound on the number of markers since we know that $L_n \leq n$ which suits us quite well.

Lemma 4 *For arbitrarily small $\varepsilon > 0$*

$$
\Pr\{L_n < n^{1-\alpha-O(\varepsilon)}\} = O(\max\{n^{\alpha-1+\varepsilon}, \psi(n^\varepsilon)\})
$$

for large n.

Proof. We only consider nonoverlapping markers that are separated by $g = n^\varepsilon$ symbols. Denote this number of markers by L_n^*. Clearly, $L_n \geq L_n^*$ and let Z_i be equal to 1 if a nonoverlapping n^ε-separated marker occurs at position i, where $1 \leq i \leq n/(k+g)$ with $k = \lfloor \alpha D_n \rfloor$. Observe that

$$
\mathbf{E}[L_n^*] = \sum_{i=1}^{n/(k+g)} \mathbf{E}[Z_i] \geq \frac{n}{k+g} \Pr\{Z_i = w_k \in \mathcal{G}_k^\varepsilon,\ k \geq (1-\varepsilon)\alpha h^{-1} \log n\} \geq n^{1-\alpha-\varepsilon}.
$$

Then by Chebyshev's inequality

$$
\Pr\{L_n < n^{1-\alpha+\varepsilon}\} \leq \Pr\{L_n^* < (1-\varepsilon)\mathbf{E}[L_n^*]\} \leq \frac{\mathbf{Var}[L_n^*]}{\varepsilon^2 \mathbf{E}[L_n^*]^2}.
$$

We prove below that

$$
\mathbf{Var}[L_n^*] \leq \mathbf{E}[L_n] + 2\psi(n^\varepsilon)\mathbf{E}[L_n^*]^2.
$$

To estimate the variance $\mathbf{Var}[L_n^*]$ we proceed basically as in [13]. Observe that for $m = n/(k+g)$

$$
\mathbf{Var}[L_n^*] = \sum_{i=1}^{m} \mathbf{Var}[Z_i] + \sum_{|i-j|>n^\varepsilon} \mathbf{Cov}[Z_i Z_j]
$$

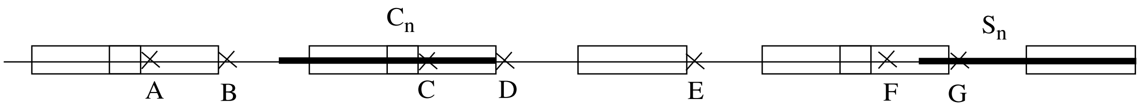

Figure 2: Illustration to Lemma 5: Solid intervals represent the largest suffix and its copy, boxes are markers and sampled positions are marked as crosses.

$$\leq \mathbf{E}[L_n^*] + 2\psi(n^\varepsilon) \sum_{|i-j|>n^\varepsilon} \mathbf{E}[Z_i]\mathbf{E}[Z_j]$$
$$\leq \mathbf{E}[L_n^*] + 2\psi(n^\varepsilon)\mathbf{E}[L_n^*]^2,$$

which, together with our previous estimates, completes the proof. ∎

3.3 Marker Stability Properties

We establish here the marker stability property. Assume now that $m = \lfloor n^\varepsilon \rfloor$ for any arbitrary small $\varepsilon > 0$. In the sequel, we shall work with modified strings $\widetilde{X}_1^n$ in which we change any of the m symbols following a marker. We prove several properties of such modified strings. Among others, in the next lemma we show that **whp** the largest suffix $\widetilde{D}_n$ in the modified strings is equal to the suffix D_n in the original string.

Lemma 5 *Let $\widetilde{X}_1^n$ be a string that differs from the string X_1^n generated by a mixing model on any of $m = \lfloor n^\varepsilon \rfloor$ positions after a marker of X_1^n. Let $\widetilde{D}_n$ be the length of the largest suffix in $\widetilde{X}_1^n$. Then there exists $\varepsilon > 0$ such that*

$$\Pr\{D_n = \widetilde{D}_n\} = 1 - O(n^{-\nu}) \tag{22}$$

for some $0 < \nu < \frac{1}{2}$.

Proof. The thrust of the proof is quite simple. We shall show that the modification defined in the lemma can only concern markers that contains any of these modified symbols. But due to marker separation properties (in particular Lemma 3) such event is quite unlikely as long as $D_n > (1-\varepsilon)\frac{1}{h}\log n$ for ε sufficiently small. Therefore, we assume from now on that $D_n \geq (1-\varepsilon)\frac{1}{h}\log n$, which by Lemma 1 occurs with probability $1 - O(n^{-\varepsilon})$. We consider several cases illustrated in Figure 2 (where $m = 1$ is assumed).

Let S_n be the suffix of length D_n of string X_1^n, that is, $S_n = X_{n-D_n+1}^n$; let C_n be an internal copy of S_n in the original string X_1^n. We assume that C_n starts at position i, i.e., $C_n = X_i^{i+D_n-1}$. We consider two cases:

CASE $D_n < \widetilde{D}_n$.

This can only happen if the modification occurs inside the suffix S_n or the copy C_n (cf. positions C and G in Figure 2). If the change occurs inside S_n, then there must be another marker within distance $O(\log n)$, which happens with probability $O(n^{-\nu})$. If the change is inside C_n (cf. position C in Figure 2), then this will result in producing another marker within distance $O(\log n)$ that by Lemma 3 has probability $O(n^{-\nu})$ to occur.

CASE $D_n > \widetilde{D}_n$.

Again, we must consider a few cases (we refer to positions A, B, E and F in Figure 2). In the first case a change occurs in the new largest suffix of $\widetilde{X}_1^n$, just before S_n. But by Lemma 3 this happens with probability $O(n^{-\nu})$. The second case is more intricate. We assume that the change occurs inside the string which creates a new copy $\widetilde{C}_n$ such that $|\widetilde{C}_n| = \widetilde{D}_n > D_n$ (cf. positions A, B and E in Figure 2). Of course, the new copy $\widetilde{C}_n$ creates a new marker. If this marker does not contain the modified position, then this marker existed before and was within distance $O(n^{\varepsilon})$ from another marker (see A and B) which is unlikely to happen. Finally, we consider the situation as illustrated by position E in Figure 2. We reduce it again to Lemma 3 by considering "new" markers of length $\frac{1}{2} < \alpha' < \alpha$, and see that again these two new markers are close enough so that Lemma 3 can be used. ∎

The last lemma tells us that **whp** strings do not modify the positions of their markers if we alter any of $m = \lfloor n^{\varepsilon} \rfloor$ symbols after a marker. We shall call such strings *favorite* strings. This is made more formal in the next definition.

Definition 1 *A string X_1^n is m-favorite if a modification of any m symbols following a marker does not change locations of nay marker in the new string $\widetilde{X}_1^n$.*

Lemma 5 basically implies that **whp** any string is favorite. This is proved formally in the next lemma.

Lemma 6 *There exists $\varepsilon > 0$ such that the probability that there exists a modification of any $m = O(n^{\varepsilon})$ symbols following a marker in X_1^n changing the position of markers is $O(n^{-\nu})$ for some $0 < \nu < \frac{1}{2}$.*

Proof. By changing a symbol after a marked position we either *create* a new marker that overlap with the previous marker (cf. position E in Figure 2) or *delete* an existing marker that overlapped with the previous marker (cf. position A in Figure 2). Thus by Lemma 3 this can occur with probability $O(n^{-\nu})$. ∎

Before we proceed, we need the following definition.

Definition 2 *Strings* X_1^n *and* $\widetilde{X}_1^n$ *are* m*-paired if:*

- X_1^n *and* $\widetilde{X}_1^n$ *are both* m*-favorite strings;*
- X_1^n *and* $\widetilde{X}_1^n$ *have their markers at the same positions;*
- X_1^n *and* $\widetilde{X}_1^n$ *match on every positions except the marked symbols.*

We define the orbit $\mathcal{F}_n(X_1^n)$ *of* X_1^n *as*

$$\mathcal{F}_n(X_1^n) := \{\widetilde{X}_1^n : \ \widetilde{X}_1^n \text{ is } m-\text{paired with } X_1^n\},$$

and the orbit set (or the set of favorite strings) as

$$\mathcal{F}_n := \bigcup_{X_1^n} \mathcal{F}_n(X_1^n) = \{X_1^n : \ X_1^n \text{ is a favorite string}\}.$$

Given $\mathcal{F} := \mathcal{F}_n(X_1^n)$, let $L_n(\mathcal{F})$ be the number of markers in a string $X_1^n \in \mathcal{F}$. Observe that the favorite strings $\mathcal{F}$ may differ only on m positions following a marker, thus the number of markers is fixed for a given $\mathcal{F}$. Furthermore, the cardinality of $\mathcal{F}$ is $|\mathcal{F}| = V^{mL_n(\mathcal{F})}$. Finally, by Lemma 6 the probability that a string belongs to the set of favorite strings is $1 - O(n^{-\nu})$.

3.4 Mixing Property of Sampled Sequence

The last two facts just proved have far reaching consequences. In particular, in Lemmas 5 and 6 we establish that **whp** markers do not change their positions if we modified any sampled symbol. Strings satisfying this property were called *favorite strings*. They play for our problem the same role as typical sequence for AEP. In Lemma 7 below we shall prove that sampled sequence of favorite strings is mixing. This will allow us to complete the proof of Theorem 1 for strings for which the probabilities of symbol generations are distinguishable (we call them δ-discriminant). When these probabilities are very close (think of an unbiased memoryless source discussed in Example 2) we appeal to the left side of (4) to complete the proof of Theorem 1.

The next lemma summarizes our knowledge about the sampled sequence. It proves that given $\mathcal{F}$ the sampled sequence is mixing. In other words, we shall show that the probability distribution of the marked sequence is within factor $(1 \pm O(\psi(k))^{L_n(\mathcal{F})}$ from an i.i.d. sequence.

Lemma 7 *Let $\mathcal{F} \in \mathcal{F}_n$ be given. Under the condition that $X_1^n \in \mathcal{F}$, the sampled sequence is mixing provided (8) holds. More precisely, let $\ell := L_n(\mathcal{F})$ and let $i_1, i_2, \ldots, i_\ell$ be the marked positions. Then*

$$\left(\frac{1-\psi(n^\varepsilon)}{1+\psi(n^\varepsilon)}\right)^\ell \Pr\{X_{i_1} = x_1 | X_1^n \in \mathcal{F}\} \times \ldots \times \Pr\{X_{i_\ell} = x_\ell | X_1^n \in \mathcal{F}\}$$
$$\leq \Pr\{X_{i_1} = x_1, \ldots, X_{i_\ell} = x_\ell | X_1^n \in \mathcal{F}\} \leq$$
$$\left(\frac{1+\psi(n^\varepsilon)}{1-\psi(n^\varepsilon)}\right)^\ell \Pr\{X_{i_1} = x_1 | X_1^n \in \mathcal{F}\} \times \ldots \times \Pr\{X_{i_\ell} = x_\ell | X_1^n \in \mathcal{F}\}$$

for any arbitrary small $\varepsilon > 0$.

Proof. As in the formulation of the theorem, we let $i_1, i_2, \ldots, i_\ell$ to be the marked positions, where $\ell := L_n(\mathcal{F})$. The sampled sequence is $X_{i_1} X_{i_2} \ldots X_{i_\ell}$. We also define $I_j := \{i_1 + 1, \ldots, i_j + m\}$ for $j = 1, 2, \ldots, \ell$. In words, the sets I_j represent m positions after each marker. Observe that given $\mathcal{F}$ all the other values X_r for $r \notin \bigcup_{j=1}^{\ell} (i_j \cup I_j)$ are fixed. We denote by $X(\mathcal{F})_1^{i_1-1}$ the fixed substring $X_1^{i_1-1}$, $X(\mathcal{F})_{i_k+1}^{i_{k+1}-1}$ the fixed substring $X_{i_k+1}^{i_{k+1}-1}$, and $X(\mathcal{F})_{i_\ell+1}^n$ the fixed substring $X_{i_\ell+1}^n$ when $X_1^n \in \mathcal{F}$. By definitions of the mixing source (MX) and the favorite sequence, we have

$$\begin{aligned}
\Pr\{X_1^n \in \mathcal{F}\} &= \Pr\{X(\mathcal{F})_1^{i_1} X_{i_1}^{i_1+m} \ldots X(\mathcal{F})_{i_{\ell-1}+m+1}^{i_\ell-1} X_{i_\ell}^{i_\ell+m} X(\mathcal{F})_{i_\ell+m+1}^n\} \\
&\geq (1-\psi(m))^\ell \Pr\{X(\mathcal{F})_1^{i_1}\} \times \ldots \times \Pr\{X(\mathcal{F})_{i_{\ell-1}+m}^{i_\ell}\} \Pr\{X(\mathcal{F})_{i_\ell+m}^n\}
\end{aligned}$$

and

$$\begin{aligned}
\Pr\{X_{i_1} &= x_1, \ldots, X_{i_\ell} = x_\ell, X_1^n \in \mathcal{F}\} \leq \\
&\leq (1+\psi(m))^\ell \Pr\{X(\mathcal{F})_1^{i_1} x_1\} \times \ldots \times \Pr\{X(\mathcal{F})_{i_{\ell-1}+m}^{i_\ell} x_\ell\} \Pr\{X(\mathcal{F})_{i_\ell+m}^n\}.
\end{aligned}$$

Combining these two inequalities we obtain the desired upper bound. In a similar manner we obtain the lower bound. This yields the result since $(1+\psi(n^\varepsilon))^{n^{1-\alpha}} \to 1$ as long as (8) holds. ∎

To obtain a complete picture of the probabilistic behavior of the SMP predictor, and to compare it to the optimal predictor X_n^*, we must investigate the distribution of the most frequent symbol in the sampled sequence. We know from Lemma 7 that the sampled sequence is within "distance" $(1+\psi(n^\varepsilon))^{L_n(\mathcal{F})} \to 1$ from an i.i.d. sequence provided (8) holds. However, the distribution of the most frequent symbol depends on how close are the probabilities of the next symbol X_{n+1} given X_1^n. We technically need a different proof of Theorem 1 for these cases, as we have already pointed out in Example 2. Therefore, we introduce the so called δ-discriminant strings.

Definition 3 *A string x_1^n is called δ-discriminant if there exists one the most frequent symbol, say $a_{\max} \in \mathcal{A}$ such that for all $a \in \mathcal{A} - \{a_{\max}\}$*

$$\Pr\{X_{n+1} = a_{\max} | X_1^n = x_1^n\} - \Pr\{X_{n+1} = a | X_1^n = x_1^n\} > \delta \tag{23}$$

for some $\delta > 0$. (Throughout, we assume that $\delta > n^{-\theta}$ for some $\theta > 0$.)

Remark. For memoryless sources all strings are either δ-discriminant or none is δ-discriminant. For sources with memory, some strings might be δ-discriminant while others not, even for the same source.

We need to prove the following simple result before we can complete the proof of Theorem 1.

Lemma 8 *Let Y_1^ℓ be a sequence of length ℓ generated by a δ-discriminant memoryless source over an alphabet $\mathcal{A}$. Let $N_a(Y)$ denote the number of times the symbols "a" occurs in Y. For all $\delta > 0$ there exists $\beta > 0$ such that for all $a \neq a_{\max}$:*

$$\Pr\{N_{a_{\max}}(Y) < N_a(Y)\} \leq \exp(-\beta \ell \delta^2). \tag{24}$$

Proof. We use the Azuma inequality (cf. [15, 26]) applied to $N(Y) := N_{a_{\max}}(Y) - N_a(Y)$ for $a \neq a_{\max}$. Since for any symbol a

$$\mathbf{E}[N(Y)] = \ell(P(a_{\max}) - P(a)) > \ell\delta.$$

Moreover, for any string Y' that differs from Y on a single position we have

$$|N(Y') - N(Y)| \leq 1.$$

Hence, by the Azuma inequality

$$\Pr\{|N(Y) - \mathbf{E}[N(Y)] > \varepsilon \mathbf{E}[N(Y)]\} \leq 2\exp(-\frac{1}{2}\ell\delta^2) \leq \exp(-\beta\ell\delta^2)$$

for some $\beta > 0$. Thus

$$\Pr\{N_{a_{\max}}(Y) - N_a(Y) > 0\} \geq \Pr\{N_{a_{\max}}(Y) - N_a(Y) > (1-\varepsilon)l\delta\} \geq 1 - \exp(-\beta\ell\delta^2),$$

which proves the lemma. ∎

Lemma 9 *For a δ-discriminant string generated by a mixing source and belonging to an orbit $\mathcal{F}$ with $\delta = n^{-\theta}$, we have*

$$\Pr\{\hat{X}_{n+1} \neq a_{\max} | X_1^n \in \mathcal{F}\} = O\left(((1 + \psi(n^\varepsilon))\omega)^{L_n(\mathcal{F})}\right) \tag{25}$$

for some $0 < \omega < 1$ provided $2\theta < 1 - \alpha$.

Proof. We use the previous lemma together with Lemma 7. ∎

3.5 Finishing the Proof of Theorem 1

Now we are in a position to prove Theorem 1 for δ-discriminant strings with $\delta > n^{-\theta}$ for $2\theta < 1-\alpha$. As discussed in Example 2, for this case we show that the right-hand side of (4), namely, $\Pr\{\hat{X}_{n+1} \neq X^*_{n+1}\} = \Pr\{\hat{X}_n \neq a_{\max}\} = O(n^{-\nu})$ for some $0 < \nu < \frac{1}{2}$. Using Lemmas 3–9 we have for $m = \lfloor n^{\varepsilon} \rfloor$ and any $\varepsilon > 0$ (below ν is a positive constant not bigger than $\frac{1}{2}$ that can change from line to line):

$$\begin{aligned}
\Pr\{\hat{X}_{n+1} \neq a_{\max}\} &\leq \Pr\{X_1^n \text{ is not } m\text{-stable }\} \\
&+ \Pr\{X_1^n \text{ is } m\text{-paired and } \hat{X_1^n} \neq a_{\max} \} \\
&\leq O(n^{-\nu}) + \sum_{\mathcal{F}} P(\mathcal{F}) O((1+\psi(n^{\varepsilon}))\omega)^{L_n(\mathcal{F})}) \\
&\leq O(n^{-\nu}).
\end{aligned}$$

This completes the proof for the δ-discriminant strings.

Finally, we consider the remaining non δ-discriminant strings and assume that

$$\Pr\{X_{n+1} = a_{\max} | X_1^n = x_1^n\} - \Pr\{X_{n+1} = a | X_1^n = x_1^n\} \leq \delta = n^{-\theta} \tag{26}$$

for $2\theta < 1-\alpha$ and all $a \in \mathcal{A}$. To simplify the presentation, we now assume that the alphabet $\mathcal{A}$ is binary. Extending to a finite alphabet is straightforward by restricting symbol a to the subset satisfying $\Pr\{X_{n+1} = a | X_1^n = x_1^n\} \geq \Pr\{X_{n+1} = a_{\max} | X_1^n = x_1^n\} - \delta$. As discussed in Example 2, we must consider now the left-hand side of (4), that is, we shall prove that

$$\Pr\{X^*_{n+1} \neq X_{n+1}\} \leq \Pr\{\hat{X}_{n+1} \neq X_{n+1}\} \leq \Pr\{X^*_{n+1} \neq X_{n+1}\} + O(n^{-\nu})$$

for some $0 < \nu < \frac{1}{2}$. The left-hand side of the above inequality is obvious, so we only concentrate on the right-hand side. We have

$$\begin{aligned}
\Pr\{\hat{X}_{n+1} \neq X_{n+1}\} &\leq 1 - \sum_{x_1^n} \Pr\{\hat{X}_{n+1} = X_{n+1} | X_1^n = x_1^n\} P(x_1^n) \\
&\leq 1 - \sum_{x_1^n \in \mathcal{F}_n} \Pr\{\hat{X}_{n+1} = X_{n+1} | X_1^n = x_1^n\} P(x_1^n).
\end{aligned}$$

But due to (26)

$$\Pr\{X_{n+1} = \hat{X}_{n+1} | x_1^n\} \geq \max_{a \in \mathcal{A}} \Pr\{X_{n+1} = a | x_1^n\} - n^{-\theta}.$$

Thus we find

$$\begin{aligned}
\Pr\{\hat{X}_{n+1} \neq X_{n+1}\} &\leq 1 - \sum_{x_1^n \in \mathcal{F}_n} \max_{a \in \mathcal{A}} \Pr\{\hat{X}_{n+1} = X_{n+1} | X_1^n = x_1^n\} P(x_1^n) + n^{-\theta} \\
&= 1 - \sum_{x_1^n} \max_{a \in \mathcal{A}} \Pr\{\hat{X}_{n+1} = X_{n+1} | X_1^n = x_1^n\} P(x_1^n) + n^{-\theta} + O(n^{-\nu}) \\
&= \Pr\{X_{n+1} \neq X^*_{n+1}\} + O(n^{-\nu}).
\end{aligned}$$

This completes the proof of Theorem 1.

Acknowledgment

We thank referees for helpful comments that led to much better presentation of the prediction problem. We are also in debt to Dr. Marcelo Weinberger (HP Labs, Palo Alto) for his interest in this work.

References

[1] P. Algoet, Universal Schemes for Prediction, Gambling and Portfolio Selection, *Ann. Prob.*, 20, 901–941, 1992.

[2] P. Algoet, The Strong Law of Large Numbers for Sequential Decisions Under Uncertainty, *IEEE Trans. Information Theory*, 40, 609-633, 1994.

[3] P. Billingsley, *Convergence of Probability Measures*, John Wiley & Sons, New York, 1968.

[4] J. Cleary and I. Witten, Data Compression Using Adaptive Coding and Partial String Matching, *IEEE Trans. Commun.*, 32, 396–402, 1984.

[5] T.M. Cover, Behavior of Sequential Predictors of Binary Sequence, in *Proc. 4th Prague Conf. Information Theory, Statistical Decision Functions, Random Processes*, 263-272, 1965.

[6] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley&Sons, New York, 1991.

[7] A. Ehrenfeucht and J. Mycielski, A Pseudorandom Sequence — How Random Is It? *Amer. Math. Monthly*, 99, 373-375, 1992

[8] M. Feder, N. Merhav, and M. Gutman, Universal Prediction of Individual Sequences, *IEEE Trans. Information Theory*, 38, 1258–1270, 1992.

[9] D. Gusfield, *Algorithms on Strings, Trees, and Sequences*, Cambridge University Press, Cambridge, 1997.

[10] P. Jacquet, Average Case Analysis of Pattern Matching Predictor, INRIA TR, 1999.

[11] P. Jacquet, W. Szpankowski, and I. Apostol, A Universal Predictor Based on Pattern Matching: Preliminary Results, *Mathematics and Computer Science : Algorithms, Trees, Combinatorics and Probabilities*, (eds. D. Gardy and A. Mokkadem), 75-85, Birkäuser, Basel, 2000.

[12] J. Kieffer, Prediction and Information Theory, preprint, 1998 (available at `ftp://oz.ee.umn.edu/users/kieffer/papers/prediction.pdf`).

[13] T. Łuczak, and W. Szpankowski, A Suboptimal Lossy Data Compression Based in Approximate Pattern Matching, *IEEE Trans. Information Theory*, 43, 1439–1451, 1997.

[14] K. Marton and P. Shields, The Positive-Divergence and Blowing-up Properties, *Israel J. Math*, 80, 331-348 (1994).

[15] C. McDiarmid, On the Method of Bounded Differences, in *Surveys in Combinatorics* (Ed. J. Siemons), vol 141, 148–188, London Mathematical Society Lecture Notes Series, Cambridge University Press, Cambridge, 1989.

[16] N. Merhav, M. Feder, and M. Gutman, Some Properties of Sequential Predictors for Binary Markov Sources, *IEEE Trans. Information Theory*, 39, 887–892, 1993.

[17] N. Merhav, M. Feder, Universal Prediction, *IEEE Trans. Information Theory*, 44, 2124-2147, 1998.

[18] G. Moravi, S. Yakowitz, and P. Algoet, Weak Convergent Nonparametric Forecasting of Stationary Time Series, *IEEE Trans. Information Theory*, 43, 483–498, 1993.

[19] J. Rissanen, A Universal Data Compression System, *IEEE Trans. Information Theory*, 29, 656–664, 1983.

[20] J. Rissanen, Universal Coding, Information, Prediction, and Estimation, *IEEE Trans. Information Theory*, 30, 629–636, 1984.

[21] B. Ryabko, Twice-Universal Coding, *Problems of Information Transmission*, 173–177, 1984.

[22] B. Ryabko, Prediction of Random Sequences and Universal Coding, *Problems of Information Transmission*, 24, 3–14, 1988.

[23] C. Shannon, Prediction and Entropy of Printed English, *Bell System Tech. J.*, 30, 50–64, 1951.

[24] P. Shields, *The Ergodic Theory of Discrete Sample Paths*, American Mathematical Society, Providence, 1996.

[25] W. Szpankowski, A Generalized Suffix Tree and Its (Un)Expected Asymptotic Behaviors, *SIAM J. Computing*, 22, 1176–1198, 1993.

[26] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, John Wiley & Sons, New York, 2001.

[27] J. Vitter and P. Krishnan, Optimal Prefetching via Data Compression, *J. ACM*, 43, 771–793, 1996.

[28] E.H. Yang, and J. Kieffer, On the Redundancy of the Fixed–Database Lempel-Ziv Algorithm for Φ-Mixing Sources, *IEEE Trans. Information Theory*, 43, 1101–1111, 1997.

[29] M. Weinberger, J. Rissanen, and M. Feder, A Universal Finite Memory Source, *IEEE Trans. Information Theory*, 41, 643–652, 1995.

[30] A. Wyner, and J. Ziv, Some Asymptotic Properties of the Entropy of a Stationary Ergodic Data Source with Applications to Data Compression, *IEEE Trans. Information Theory*, 35, 1250–1258, 1989.

BIOGRAPHICAL SKETCHES

Wojciech Szpankowski received the M.S. degree and the Ph.D. degree in Electrical and Computer Engineering from Technical University of Gdańsk in 1976 and 1980, respectively. Currently, he is Professor of Computer Science at Purdue University. Before coming to Purdue, he was Assistant Professor at Technical University of Gdańsk, Poland, and in 1984 he held Visiting Assistant Professor position at the McGill University, Canada. During 1992/1993 he was Professeur Invité in the Institut National de Recherche en Informatique et en Automatique, France, in the Fall of 1999 he was Visiting Professor at Stanford University, and in June 2001 he was Professeur Invité at the Université de Versailles Saint Quentin-en-Yvelines, France.

His research interests cover the design and analysis of algorithms, analytic combinatorics and random structures, pattern matching, information theory (including multimedia data compression), discrete mathematics, performance evaluation, stability problems in distributed systems, modeling of computer systems and computer communication networks, queueing theory, and applied probability. His recent work is mostly devoted to probabilistic analysis and design of algorithms on strings and applications of analytic tools to problems of information theory. His new book: *Average Case Analysis of Algorithms on Sequences*, was just published by John Wiley & Sons, 2001.

Dr. Szpankowski was guest editor for several journals. In 2000–2001, together with H. Prodinger, he edited a special issue in RANDOM STRUCTURES & ALGORITHMS, and in 2002 (with M. Drmota) he will edit another special issue for COMBINATORICS, PROBABILITY, & COMPUTING. He is on the editorial boards of THEORETICAL COMPUTER SCIENCE and DISCRETE MATHEMATICS AND THEORETICAL COMPUTER SCIENCE. In 1999 he co-chaired the *Information Theory and Networking Workshop*, Metsovo, Greece, while in 2000 he was the chair of the *Sixth Seminar on Analysis of Algorithms*, Krynica Morska, Poland.

Philippe Jacquet is a research director in INRIA. He graduated from Ecole Polytechnique in 1981 and from Ecole Nationale des Mines in 1984. He received his Ph.D. degree from Paris Sud University in 1989 and his habilitation degree from Versailles University in 1998. He is currently the head of HIPERCOM project that is devoted to high performance communications. As an expert in telecommunications and information technology, he participated in several standardization committees such as ETSI, IEEE and IETF. His research interests cover information theory, probability theory, quantum telecommunication, evaluation of performance and algorithm design for telecommunication, wireless and ad hoc networking.

Philippe Jacquet is author of several papers that have appeared in international journals. In 1999 he co-chaired the *Information Theory and Networking Workshop*, Metsovo, Greece.

Izydor Apostol, Ph.D. Research Scientist at Amgen Inc. Prior to that he worked for Baxter Hemoglobin Therapeutics Inc., which he joined in 1992 after 10 years of academic experience. Current research involves extensive characterization of recombinant proteins intended for therapeutic use and development of new analytical techniques for forensics investigation of proteins. Also, he conducts research in computation biochemistry. Member of the Protein Society and Association of Biomolecular Resource Facilities.