Electronic Journal of Probability

Electron. J. Probab. **17** (2012), no. 25, 1–9.
ISSN: 1083-6489 DOI: 10.1214/EJP.v17-1878

First occurrence of a word among the elements of a finite dictionary in random sequences of letters[*]

Emilio De Santis[†] Fabio Spizzichino[‡]

Abstract

In this paper we study a classical model concerning occurrence of words in a random sequence of letters from an alphabet. The problem can be studied as a game among $(m+1)$ words: the winning word in this game is the one that occurs first. We prove that the knowledge of the first m words results in an advantage in the construction of the last word, as it has been shown in the literature for the cases $m=1$ and $m=2$ [1, 2]. The last word can in fact be constructed so that its probability of winning is strictly larger than $1/(m+1)$. For the latter probability we will give an explicit lower bound. Our method is based on rather general probabilistic arguments that allow us to consider an arbitrary cardinality for the alphabet, an arbitrary value for m and different mechanisms generating the random sequence of letters.

Keywords: Competing words; Sub-words; Renewal Theorem; Ergodic Theorem.
AMS MSC 2010: 60C05; 65C50.
Submitted to EJP on September 12, 2011, final version accepted on March 12, 2012.

1 Introduction

The theme of the occurrence of *words* in random sequences of *letters* from an *alphabet* is a rather classical one in discrete probability. The related literature has a long tradition and papers with new insights and deep results continue to appear from time to time.

This topic has, among others, the following interesting aspects: it has a number of important applications and it is characterized by surprising results which, at a first glance, can sometimes appear even contradictory. Feller's book is a starting point for the study of occurrence of words in a Bernoulli scheme [3]. Different types of interesting problems arise in this field and many important papers appeared in the related literature; see in particular [4, 1, 2, 5, 6, 7] and references cited therein.

One interesting problem considers a finite set of given words, a *dictionary*, and concerns the probability that a fixed word occurs as the first. This problem can be seen

[*] Supported by PRIN 2009Y3PX3X_004 Applicazioni fisiche di campi aleatori; 2009 Research Project "Modelli e Algoritmi Stocastici: Convergenza e Ottimizzazione" University "La Sapienza", Rome.

[†] La Sapienza Università di Roma, Italy. E-mail: desantis@mat.uniroma1.it

[‡] La Sapienza Università di Roma, Italy. E-mail: fabio.spizzichino@uniroma1.it

as related to a game among different words, where the winner is the word which occurs first.

In this respect, given m words $\mathbf{w}_1, \dots, \mathbf{w}_m$, we construct a word $\mathbf{w}_{m+1}$ such that its probability of winning is larger than $1/(m+1)$.

In [1] the case of two competing words (i.e. $m = 1$) on a binary alphabet has been considered. In [2], the analysis has been extended in a thorough way to the case of three words (i.e. $m = 2$).

Provided that the length of the words is sufficiently large, and by introducing suitable probabilistic arguments, we solve the problem for an arbitrary value of m and for an alphabet of arbitrary size. In particular we provide an explicit lower bound for the probability of first occurrence for the constructed word $\mathbf{w}_{m+1}$.

In the next section, we introduce some useful notation to formalize our result in Theorem 2.3. Then we give our constructive proof after presenting the preliminary Lemmas 1-3. Section 3 is devoted to a short discussion containing some comments and concluding remarks.

2 Construction of efficient words and probability of winning

Let $\mathcal{A}_N = \{a_1, \dots, a_N\}$ be an alphabet composed of N distinct letters. We consider $(m+1)$ words $\mathbf{w}_1, \dots, \mathbf{w}_{m+1}$ of a fixed length k, i.e. $(m+1)$ elements of $\mathcal{A}_N^k$ and let $\mathcal{W}_{m+1} = \{\mathbf{w}_1, \dots, \mathbf{w}_{m+1}\}$. We write $w_{i,l}$ for the i-th letter of the word $\mathbf{w}_l$ and we say that the word $(w_{i,l}, \dots, w_{j,l})$, for $1 \leq i \leq j \leq k$, is a *sub-word* of $\mathbf{w}_l$.

At any instant $n = 1, 2, \dots$ a letter is drawn from the alphabet $\mathcal{A}_N$. Drawings are supposed to be independent and uniformly distributed over $\mathcal{A}_N$. We define the space $\Omega = \mathcal{A}_N^{\mathbb{N}}$; for $\omega = (\omega_1, \omega_2, \dots) \in \Omega$, we refer to ω_n as *the letter at time* $n \in \mathbb{N}$. The probability measure on Ω is then the product measure that, at any drawing, assigns probability $1/N$ to each letter of $\mathcal{A}_N$:

$$P(\omega_n = a) = \frac{1}{N}, \quad a \in \mathcal{A}_N, \quad n \in \mathbb{N}.$$

We now consider a *game* that ends at the random time R_1, where

$$R_1 = \inf\{n \geq k : (\omega_{n-k+1}, \dots, \omega_n) \in \mathcal{W}_{m+1}\}, \tag{2.1}$$

and the *winner* is the word $\mathbf{w}_l$ such that

$$(\omega_{R_1-k+1}, \dots, \omega_{R_1}) = \mathbf{w}_l. \tag{2.2}$$

Next we define the events

$$E_l = \{\omega \in \Omega : (\omega_{R_1+1-k}, \dots, \omega_{R_1}) = \mathbf{w}_l\} \text{ for } l = 1, \dots, m+1. \tag{2.3}$$

Hence, the event E_l means that $\mathbf{w}_l$ occurs first within $\mathcal{W}_{m+1}$.

We assume that $\mathbf{w}_{m+1}$ can be chosen as a function of the other words $\mathbf{w}_1, \dots, \mathbf{w}_m$ and we show that this can be done in such a way that the winning probability of $\mathbf{w}_{m+1}$ is greater than $\frac{1}{m+1}$.

For this purpose it is convenient to assume that drawings of letters go on indefinitely also beyond time R_1, so that, a.s., we will have an infinite number of games.

Let us introduce the random variables $V_{l,n}$, for each $l = 1, \dots, m+1$ and $n \in \mathbb{N}$, as the number of wins of word $\mathbf{w}_l$ until time n. Obviously the probability law of $V_{l,n}$ also depends on the ordered sequence $(\mathbf{w}_1, \dots, \mathbf{w}_{m+1})$, however these words are fixed once forever and for shortness sake we will omit to indicate this dependence. Recursively define

$$R_{h+1} = \inf\{n \geq R_h + k : (\omega_{n-k+1}, \dots, \omega_n) \in \mathcal{W}_{m+1}\}, \tag{2.4}$$

for $h = 1, 2, \dots$ where R_1 is the random variable defined in (2.1). Thus, by using this notation, the random variables $V_{1,n}, \dots, V_{m+1,n}$ can be more formally defined as

$$V_{l,n} = \sum_{s=1}^{\infty} \mathbf{1}_{\{R_s \le n\}} \mathbf{1}_{\{(\omega_{R_s - k+1}, \dots, \omega_{R_s}) = \mathbf{w}_l\}}, \tag{2.5}$$

for $l = 1, \dots, m+1$ and $n = 1, 2, \dots$. Let moreover $\mathcal{R} = \{R_h : h \in \mathbb{N}\}$ and define the random variable $N_{l,n}$ as the number of times in which the word $\mathbf{w}_l$ occurs inside the interval $[0, n]$, i.e.

$$N_{l,n} = \sum_{s=k}^{n} \mathbf{1}_{\{(\omega_{s-k+1}, \dots, \omega_s) = (w_{1,l}, \dots, w_{k,l})\}}. \tag{2.6}$$

Furthermore we consider the random times $T_h = \inf\{n : \sum_{l=1}^{m+1} N_{l,n} = h\}$ and put $\mathcal{T} = \{T_h : h \in \mathbb{N}\}$; clearly $\mathcal{R} \subset \mathcal{T}$.

We present a remark that will be useful for the proof of Lemmas 2.5-2.6.

Remark 2.1. *Let $n \ge k$ be fixed. An event of the form*

$$\{\omega \in \Omega : (\omega_{n-k+1}, \dots, \omega_n) = (w_{1,l}, \dots, w_{k,l})\} \text{ for } l = 1, \dots, m+1$$

implies the event $\{n \in \mathcal{T}\}$ but does not imply the event $\{n \in \mathcal{R}\}$. In order to guarantee $\{n \in \mathcal{R}\}$ it is sufficient (but not necessary), see definition (2.4), to exclude that, for some $s = n-k+1, \dots, n-1$ and some $j = 1, \dots, m+1$ it happened

$$\{\omega \in \Omega : (\omega_{s-k+1}, \dots, \omega_s) = (w_{1,j}, \dots, w_{k,j})\}. \tag{2.7}$$

Notice on the other hand that, again by (2.4), the event $\{s \in \mathcal{R}\}$ excludes the event $\{n \in \mathcal{R}\}$ for $n = s+1, \dots, s+k-1$. In fact we can not observe two wins at a distance less than k.

By a renewal-theorem, or an ergodic-theorem, argument the limit $\lim_{n\to\infty} V_{l,n}/n$ exists almost surely for $l = 1, \dots, m+ 1$ and it is constant. Thus we define the quantities q_l as follows

$$q_l = \lim_{n\to\infty} \frac{V_{l,n}}{n} \ \ a.s. \tag{2.8}$$

By taking into account the latter equation we see that

$$\lim_{n\to\infty} \frac{V_{l,n}}{\sum_{h=1}^{m+1} V_{h,n}} = P(E_l) \ \ a.s. \tag{2.9}$$

Remark 2.2. *Denote $\mu = \frac{1}{E(R_1)}$. Concerning the probabilities $P(E_l)$, we can also write*

$$P(E_l) = \lim_{n\to\infty} \frac{V_{l,n}}{\mu n} \ \ a.s. \tag{2.10}$$

The above identity is obtained by recalling (2.9) and by noticing that, by the renewal theorem or by the ergodic theorem, one must have

$$\lim_{n\to\infty} \frac{\sum_{h=1}^{m+1} V_{h,n}}{n} = \mu \ \ a.s. \tag{2.11}$$

As the main achievement of our paper we can state the following result. It is convenient first to introduce the notation:

$$L = L(N, m, k) = \lfloor \log_N(mk) \rfloor + 1, \tag{2.12}$$

for given N, m, k.

Theorem 2.3. *Let k be such that*

$$\frac{1}{(m+1)N^{2L}} - \frac{2L}{N^k} > \frac{2L(m+1)}{N^{k-2L}} \tag{2.13}$$

and let $\mathbf{w}_1, \ldots, \mathbf{w}_m \in \mathcal{A}_N^k$ be any m distinct words. Then there exists a word $\mathbf{w}_{m+1} \in \mathcal{A}_N^k$ such that $P(E_{m+1}) > \frac{1}{m+1}$.

The proof of Theorem 2.3 will be presented at the end of this section as a direct consequence of Lemmas 2.4, 2.5, 2.6 below.

The interest of $P(E_{m+1}) > \frac{1}{m+1}$ becomes clear when we consider the following case: each of $(m+1)$ players bets one dollar on a word of length k and the one who has chosen the winning word receives $(m+1)$ dollars. Even if the drawings are independent and the letters are equiprobable, the word $\mathbf{w}_{m+1}$ can be constructed, for any given $\mathbf{w}_1, \ldots, \mathbf{w}_m$, in such a way that the game is unfair, namely it is favorable for $(m+1)$-th player.

It is intuitive that, for fixed N and m, the length k of the words should be large enough. We shall see in Remark 2.7, as a consequence of (2.13), that $\log_N m$ is the appropriate order.

As mentioned, the word $\mathbf{w}_{m+1}$ will be obtained by means of a constructive procedure. We roughly anticipate that the word $\mathbf{w}_{m+1}$ can be constructed according to the following steps:

Step 1. The second part of $\mathbf{w}_{m+1}$, of a suitable length r, must coincide with the initial part of the word $\mathbf{w}_1$.

Step 2. The first $k-r$ letters of $\mathbf{w}_{m+1}$ must give rise to a sub-word which does not coincide with any sub-word drawn from $\mathbf{w}_1, \ldots, \mathbf{w}_m$ (see Lemma 2.4).

We will discover that a suitable value for r is $2L = 2\lfloor \log_N(mk) \rfloor + 2$. Now, we proceed to explain how to explicitly construct the word $\mathbf{w}_{m+1}$.

Let us consider the set $W_{L,m}$ of all the words $\widetilde{\mathbf{w}}_{i,l} = (w_{i+1,l}, w_{i+1,l}, \ldots, w_{i+L,l}) \in \mathcal{A}_N^L$ with $l = 1, \ldots m$, $i = 0, \ldots, k-L$ and with L defined in (2.12). Hence we are considering all the sub-words of length L of the words $\mathbf{w}_1, \ldots, \mathbf{w}_m$.

Clearly $|\mathcal{A}_N^L| = N^L$ and $|W_{L,m}| < mk$, therefore $|\mathcal{A}_N^L| > |W_{L,m}|$; thus, the set $\mathcal{A}_N^L \setminus W_{L,m}$ is not empty, and we can choose a word $(v_1, \ldots, v_L) \notin W_{L,m}$.

Let us arbitrarily take a letter $\widetilde{v} \neq v_1$ and consider

$$\mathbf{w}_{m+1} = (\underbrace{\widetilde{v}, \ldots, \widetilde{v}}_{L}, \underbrace{v_1, \ldots, v_L}_{L}, \underbrace{w_{1,1}, \ldots, w_{k-2L,1}}_{k-2L}). \tag{2.14}$$

Concerning such a choice, the following lemma shows that any possible matching between an *initial* sub-word of $\mathbf{w}_{m+1}$ and a *final* sub-word of any word in $\mathcal{W}_{m+1}$ must be sufficiently short.

Lemma 2.4. *For $l = 1, \ldots, m+1$, if $i < k$ and*

$$(w_{k-i+1,l}, w_{k-i+2,l}, \ldots, w_{k,l}) = (w_{1,m+1}, w_{2,m+1}, \ldots, w_{i,m+1}), \tag{2.15}$$

then $i \leq 2L$.

Proof. First we prove the case $l = m+1$. For $i = k-L, \ldots, k-1$, we compare the $(i+L-k+1)$- th letter of the two sub-words in (2.15). The $(i+L-k+1)$-th letter of the word on the l.h.s. is $w_{L+1,m+1} = v_1$, while the $(i+L-k+1)$-th letter of the word on the r.h.s. is $\widetilde{v}$. Then the validity of (2.15) excludes the possibility that $k-L \leq i \leq k-1$.

As to $i = 2L+1, \ldots, k-L-1$, we notice that the sub-word $(w_{L+1,m+1}, \ldots, w_{2L,m+1}) = (v_1, \ldots, v_L)$ on the r.h.s. is different from the corresponding word on the l.h.s. for the construction presented before (2.14). Then the validity of (2.15) also excludes the possibility that $2L+1 \leq i \leq k-L-1$.

For what concerns $l = 1, \ldots, m$ the latter argument is sufficient to solve directly all the cases $i = 2L+1, \ldots, k-1$ and this completes the proof. □

For the words $\mathbf{w}_{m+1}$ and $\mathbf{w}_1$ we now respectively consider the ratios $\frac{N_{m+1,n}}{V_{m+1,n}}$ and $\frac{N_{1,n}}{V_{1,n}}$. The latter expresses the ratio between the numbers of times where the word $\mathbf{w}_1$ appears within the first n drawings and the corresponding number of wins of the same word. Similarly for $\frac{N_{m+1,n}}{V_{m+1,n}}$ concerning $\mathbf{w}_{m+1}$. The following lemma provides an almost sure lower bound for the limit of the ratio $\frac{V_{m+1,n}}{N_{m+1,n}}$. Notice that the existence of such a limit can be guaranteed by ergodicity arguments. An upper bound for $\lim_{n\to\infty} \frac{V_{1,n}}{N_{1,n}}$ will be provided in Lemma 2.6.

Lemma 2.5.

$$\lim_{n\to\infty} \frac{V_{m+1,n}}{N_{m+1,n}} \geq 1 - \frac{2L(m+1)}{N^{k-2L}} \quad a.s. \tag{2.16}$$

Proof. For $n \geq k$, let us define the events

$$F_n = \{\omega \in \Omega : (\omega_{n-k+1}, \ldots, \omega_n) = (w_{1,m+1}, \ldots, w_{k,m+1})\}, \tag{2.17}$$

$$H_n = \{\omega \in \Omega : n \in \mathcal{R}(\omega)\}, \tag{2.18}$$

$$G_n^{(i,l)} = \{\omega \in \Omega : (\omega_{n-k-i+1}, \ldots, \omega_{n-k}) \neq (w_{1,l}, \ldots, w_{i,l})\}, \tag{2.19}$$

for $i = 1, \ldots, k-1$, for $l = 1, \ldots, m+1$. Let

$$G_n = \bigcap_{l=1}^{m+1} \bigcap_{i=k-2L}^{k-1} G_n^{(i,l)}. \tag{2.20}$$

Clearly, for $n \geq k$, the event $H_n \cap F_n$ means that word $\mathbf{w}_{m+1}$ wins at n. Moreover the probabilities $P(G_n \cap F_n)$ and $P(F_n)$ do not depend on n and the events G_n, F_n are independent.

Now, in view of Lemma 2.4 and the above definition (2.20), we will show that

$$G_n \cap F_n \subset H_n \cap F_n \tag{2.21}$$

whenever $n \geq 2k$. First notice that the event $H_n \cap F_n$ is equivalent to $\{V_{m+1,n} - V_{m+1,n-1} = 1\}$. In order to prove inclusion (2.21), we can argue as follows.

Remark 2.1 says that

$$F_n \cap (\bigcap_{l=1}^{m+1} \bigcap_{i=1}^{k-1} G_n^{(i,l)}) \subset \bigcap_{s=n-k+1}^{n-1} \{s \notin \mathcal{T}\}.$$

Hence

$$F_n \cap (\bigcap_{l=1}^{m+1} \bigcap_{i=1}^{k-1} G_n^{(i,l)}) \subset F_n \cap (\bigcap_{s=n-k+1}^{n-1} \{s \notin \mathcal{T}\}).$$

On the other hand

$$F_n \cap (\bigcap_{s=n-k+1}^{n-1} \{s \notin \mathcal{T}\}) \subset F_n \cap H_n = \{V_{m+1,n} - V_{m+1,n-1} = 1\}.$$

Therefore

$$F_n \cap (\bigcap_{l=1}^{m+1} \bigcap_{i=1}^{k-1} G_n^{(i,l)}) \subset H_n \cap F_n.$$

At this point, we can use Lemma 2.4 to ensure that the above argument is still valid if we replace $\bigcap_{l=1}^{m+1}\bigcap_{i=1}^{k-1} G_n^{(i,l)}$ with G_n.

Now notice that $P(F_{2k+i}) = P(F_{2k})$, $P(G_{2k+i}) = P(G_{2k})$, and $P(F_{2k+i}\cap G_{2k+i}) = P(F_{2k}\cap G_{2k})$ for $i \geq 0$. We set $p_F = P(F_{2k})$, $p_G = P(G_{2k})$, and $p_{F\cap G} = P(F_{2k}\cap G_{2k})$. Independence between F_{2k} and G_{2k} immediately yields $p_{F\cap G} = p_F p_G$. By ergodic theorem the following equalities hold almost surely:

$$\lim_{n\to\infty}\frac{V_{m+1,n}}{N_{m+1,n}} = \lim_{n\to\infty}\frac{n}{N_{m+1,n}}\lim_{n\to\infty}\frac{V_{m+1,n}}{n} = \frac{1}{p_F}\lim_{n\to\infty}P(H_n\cap F_n). \tag{2.22}$$

By (2.21) and (2.22), we conclude

$$\lim_{n\to\infty}\frac{V_{m+1,n}}{N_{m+1,n}} \geq \frac{p_{F\cap G}}{p_F} = p_G \quad a.s. \tag{2.23}$$

Now

$$p_G = 1 - P(\bigcup_{l=1}^{m+1}\bigcup_{i=k-2L}^{k-1}\overline{G_{2k}^{(i,l)}}) \geq 1 - \sum_{l=1}^{m+1}\sum_{i=k-2L}^{k-1}\frac{1}{N^i} \geq 1 - \frac{2L(m+1)}{N^{k-2L}}.$$

□

By following a same type of argument as in the previous proof we can now obtain an asymptotic upper bound for the ratio $V_{1,n}/N_{1,n}$.

Lemma 2.6.

$$\lim_{n\to\infty}\frac{V_{1,n}}{N_{1,n}} \leq 1 - \frac{1}{N^{2L}} + \frac{2L(m+1)}{N^k} \quad a.s. \tag{2.24}$$

Proof. We consider $n \geq 3k$ and define the events

$$\widehat{F}_n = \{\omega\in\Omega : (\omega_{n-k+1},\ldots,\omega_n) = (w_{1,1},\ldots,w_{k,1})\}, \tag{2.25}$$

$$\widehat{K}_n = \{\omega\in\Omega : (\omega_{n-k-2L+1},\ldots,\omega_{n-k}) = (w_{1,m+1},\ldots,w_{2L,m+1})\}, \tag{2.26}$$

and, for $l = 1,\ldots,m+1$ and for $i = 1,\ldots,2L$,,

$$\widehat{G}_n^{(i,l)} = \{\omega\in\Omega : (\omega_{n-2k-i+1},\ldots,\omega_{n-k-2L}) \neq (w_{1,l},\ldots,w_{k-2L+i,l})\}, \tag{2.27}$$

$$\widehat{G}_n = \bigcap_{l=1}^{m+1}\bigcap_{i=1}^{2L}\widehat{G}_n^{(i,l)}. \tag{2.28}$$

We will also use H_n^c where H_n is defined in (2.18). Concerning the event H_n^c, we can write, for $n \geq 3k$, that the event $H_n^c \cap \widehat{F}_n$ is equivalent to $\{V_{1,n} - V_{1,n-1} = 0,\ N_{1,n} - N_{1,n-1} = 1\}$. Moreover the probabilities $P(\widetilde{G}_n \cap \widetilde{K}_n \cap \widehat{F}_n)$ and $P(\widehat{F}_n)$ do not depend on n for $n \geq 3k$; the events $\widehat{G}_n$, $\widehat{F}_n$ and $\widehat{K}_n$ are independent. Concerning the event $\widehat{G}_n$, we can say that Lemma 2.4 allow us to limit the range of the index i in formula (2.28), *i.e.* $i = 1,\ldots,2L$.

We can now use the arguments in Remark 2.1 as follows: in view of Lemma 2.4 and the above definition of $\widehat{G}_n$, we will show that

$$\widehat{G}_n \cap \widehat{K}_n \cap \widehat{F}_n \subset H_n^c \cap \widehat{F}_n, \tag{2.29}$$

whenever $n \geq 3k$. In fact, $\widehat{F}_n \cap \widehat{K}_n$ implies that the special word $\mathbf{w}_{m+1}$ appears at time $n - 2L$, which in particular means $n - 2L \in \mathcal{T}$; moreover, the concomitant occurrence of the event $\widehat{G}_n$ guarantees that $n - 2L \in \mathcal{R}$ by using an argument similar to the proof of Lemma 2.5. The event $\{n - 2L \in \mathcal{R}\}$ implies $\{n \notin \mathcal{R}\}$, whence (2.29) easily follows.

Now notice that, for $i \geq 0$, $P(\widehat{F}_{3k+i}) = P(\widehat{F}_{3k})$. Similarly it happens for $\widehat{K}_{3k+i}$, $\widehat{G}_{3k+i}$, and $\widehat{F}_{3k+i} \cap \widehat{K}_{3k+i} \cap \widehat{G}_{3k+i}$ and we set $p_{\widehat{F}} = P(\widehat{F}_{3k})$,$p_{\widehat{K}} = P(\widehat{K}_{3k})$, $p_{\widehat{G}} = P(\widehat{G}_{3k})$, and $p_{\widehat{F}\cap\widehat{K}\cap\widehat{G}} = P(\widehat{F}_{3k} \cap \widehat{K}_{3k} \cap \widehat{G}_{3k})$. Independence among $\widehat{F}_{3k}$, $\widehat{K}_{3k}$ and $\widehat{G}_{3k}$ immediately yields $p_{\widehat{F}\cap\widehat{K}\cap\widehat{G}} = p_{\widehat{F}} p_{\widehat{K}} p_{\widehat{G}}$.

By ergodic theorem we claim

$$\lim_{n\to\infty} \frac{V_{1,n}}{N_{1,n}} = \frac{1}{p_{\widehat{F}}} \lim_{n\to\infty} P(\widehat{F}_n \cap H_n) \ \ a.s. \tag{2.30}$$

As to the r.h.s. of previous formula, we can write

$$\frac{\lim_{n\to\infty}(P(\widehat{F}_n) - P(\widehat{F}_n \cap H_n^c))}{p_{\widehat{F}}} \leq 1 - \frac{p_{\widehat{F}\cap\widehat{K}\cap\widehat{G}}}{p_{\widehat{F}}} = 1 - p_{\widehat{K}} p_{\widehat{G}} =$$

$$= 1 - \frac{1}{N^{2L}} p_{\widehat{G}} \leq 1 - \frac{1}{N^{2L}} + \frac{1}{N^{2L}} \sum_{l=1}^{m+1} \sum_{i=1}^{2L} \frac{1}{N^{k-2L+i+1}} \leq$$

$$\leq 1 - \frac{1}{N^{2L}} + \sum_{l=1}^{m+1} \sum_{i=1}^{2L} \frac{1}{N^{k+2}} \leq 1 - \frac{1}{N^{2L}} + \frac{2L(m+1)}{N^k}.$$

□

We are now in a position to give the proof of our main result.

Proof of Theorem 2.3. Since the probability of the occurrence in a given position of a single word $(a_1, \ldots, a_k)$ is a constant, equal to $1/N^k$, we have that, for $l = 1, \ldots, m+1$,

$$\lim_{n\to\infty} \frac{N_{l,n}}{n} = \frac{1}{N^k} \ \ a.s., \tag{2.31}$$

hence

$$\lim_{n\to\infty} \frac{N_{l,n}}{\sum_{k=1}^{m+1} N_{k,n}} = \frac{1}{m+1} \ \ a.s. \ . \tag{2.32}$$

In order to achieve the proof, we will use the inequality (2.16) of Lemma 2.5 and (2.24) of Lemma 2.6. For what concerns the remaining indexes, $l = 2, \ldots, m$, it is enough taking into account the obvious inequality

$$N_{l,n} \geq V_{l,n}. \tag{2.33}$$

By employing the inequalities (2.16), (2.24) and (2.33) in equation (2.9) we can write

$$\lim_{n\to\infty} \frac{N_{m+1,n}(1 - \frac{2L(m+1)}{N^{k-2L}})}{N_{1,n}(1 - \frac{1}{N^{2L}} + \frac{2L(m+1)}{N^k}) + \sum_{h=2}^{m+1} N_{h,n}} \leq P(E_{m+1}) \ \ a.s., \tag{2.34}$$

thus (2.34) gives a lower bound for the probability of winning for $\mathbf{w}_{m+1}$.

Finally, by (2.32), we obtain that

$$\lim_{n\to\infty} \frac{N_{m+1,n}(1 - \frac{2L(m+1)}{N^{k-2L}})}{N_{1,n}(1 - \frac{1}{N^{2L}} + \frac{2L(m+1)}{N^k}) + \sum_{h=2}^{m+1} N_{h,n}} =$$

$$\frac{1 - \frac{2L(m+1)}{N^{k-2L}}}{m + 1 - \frac{1}{N^{2L}} + \frac{2L(m+1)}{N^k}} = \frac{1}{m+1} \left(\frac{1 - \frac{2L(m+1)}{N^{k-2L}}}{1 - \frac{1}{(m+1)N^{2L}} + \frac{2L}{N^k}} \right) > \frac{1}{m+1},$$

the last inequality following from the hypothesis (2.13). This ends the proof. □

Remark 2.7. *In the proof of Theorem 2.3 we obtain the explicit bound*

$$P(E_{m+1}) \geq \frac{1}{m+1}\left(\frac{1-\frac{2L(m+1)}{N^{k-2L}}}{1-\frac{1}{(m+1)N^{2L}}+\frac{2L}{N^k}}\right).$$

We notice furthermore that, in the statement of Theorem 2.3, the condition (2.13) can be replaced by the simpler inequality $k \geq 22(1+\log_N m)$. In fact, at the cost of elementary but rather tedious manipulations, one can show that the latter implies (2.13). We notice in this respect that the latter estimate is of the right order. In fact, for given N and k, the number of distinct words of length k on the alphabet $\mathcal{A}_N$ is obviously N^k; then, since we need to find at least $(m+1)$ distinct words, we must necessarily have $k \geq \log_N(m+1)$. On the other hand

$$\lim_{m\to\infty} \frac{22(1+\log_N m)}{\log_N(m+1)} = 22.$$

This limit does not depend on N and shows that the condition in Theorem 2.3 is quite efficient.

3 Discussion and final remarks

We conclude the paper with some comments about our result and about the method that we use. First of all, our construction is based on rather general probabilistic arguments. This has allowed us to formulate a general result. In fact, in Theorem 2.3, we have no limitation on the choice of N and m, provided that k is large enough. Notice, for instance, that the methods used in [1, 2] are specific for the cases $N=2$, $m=1,2$.

Our procedure could be easily extended to deal with cases where stochastic independence among the letters fails.

Let us denote by $\pi_{\mathbf{w}}$ the probability that a word $\mathbf{w}$ of length k occurs as soon as possible, namely in the first k drawings. In the independence setting this probability is $1/N^k$. Now we point out that, essentially, we have used three hypotheses to prove our result. The latter can be conveniently summarized as follows

a) The process of generation of random sequences of letters is ergodic.

b) For any word $\mathbf{w}$ of length k the probability $\pi_{\mathbf{w}}$ belongs to $\{0, C_k\}$, where C_k is a positive constant, i.e. some words are *forbidden* and all the remaining words share the same probability C_k.

Besides the probabilistic hypotheses a) and b), our method requires the construction of a word satisfying suitable conditions (see Step 1, Step 2, and Lemma 1, Lemma 2, Lemma 3 in the previous section). More precisely we assume

c) Given $\mathbf{w}_1, \dots, \mathbf{w}_m \in \mathcal{A}_N^k$, one can construct a word $\mathbf{w}_{m+1} \in \mathcal{A}_N^k$ such that

   c1) $\pi_{\mathbf{w}_{m+1}} = C_k$;

   c2) For $l = 1, \dots, m+1$. If $i < k$ and

$$(w_{k-i+1,l}, w_{k-i+2,l}, \dots, w_{k,l}) = (w_{1,m+1}, w_{2,m+1}, \dots, w_{i,m+1}),$$

   then $i \leq 2L$.

   c3) $w_{i+2L,m+1} = w_{i,1}$ for $i = 1, \dots, k-2L$.

Concerning item c), we point out that $\mathbf{w}_{m+1}$ should be composed of two different parts. The second part is fixed (it coincides with the initial part of word $\mathbf{w}_1$). On the contrary, several possible choices for determining the first part of $\mathbf{w}_{m+1}$ are possible. We only need, in fact, that the first letters of $\mathbf{w}_{m+1}$ give rise to a sub-word, of a convenient length, which does not coincide with any sub-word extracted from $\mathbf{w}_1, \dots, \mathbf{w}_m$.

Under the case of independence, that we considered along the paper, condition c1) is trivially satisfied and there are many different words that satisfy item c). When independence is dropped, c1) is not trivial anymore. We can still expect however that one can find different words that satisfy c1) besides satisfying c2), c3). It is just this possibility of different *employable* words which could be useful in a possible extension of our work beyond the case of independence.

An instance where all the conditions a), b) and c) hold is the model of random drawings with *delayed replacement* of letters from an urn: two consecutive letters can not be equal. For such a model, validity of b) is obvious with $C_k = \frac{1}{N(N-1)^k}$ and a) holds when $N \geq 3$. In fact, in such a case, the sequence of letters drawn is an irreducible, aperiodic Markov chain. Finally, some words $\mathbf{w}_{m+1}$ satisfying condition c) can be constructed, provided $N \geq 4$.

References

[1] Robert Chen and Alan Zame. On fair coin-tossing games. *J. Multivariate Anal.*, 9(1):150–156, 1979. MR-0530646

[2] Robert W. Chen, Alan Zame, and Burton Rosenberg. On the first occurrence of strings. *Electron. J. Combin.*, 16(1):Research Paper 29, 16, 2009. MR-2482097

[3] William Feller. *An introduction to probability theory and its applications. Vol. I.* Third edition. John Wiley & Sons Inc., New York, 1968. MR-0228020

[4] Shuo-Yen Robert Li. A martingale approach to the study of occurrence of sequence patterns in repeated experiments. *Ann. Probab.*, 8(6):1171–1176, 1980. MR-0602390

[5] S. Robin and J. J. Daudin. Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Probab.*, 36(1):179–193, 1999. MR-1699643

[6] V. T. Stefanov and A. G. Pakes. Explicit distributional results in pattern formation. *Ann. Appl. Probab.*, 7(3):666–678, 1997. MR-1459265

[7] V. T. Stefanov, S. Robin, and S. Schbath. Waiting times for clumps of patterns and for structured motifs in random sequences. *Discrete Appl. Math.*, 155(6-7):868–880, 2007. MR-2309853

Acknowledgments. We like to thank two anonymous referees for helpful comments on an earlier version of this paper.