

Constrained Pattern Matching

YONGWOOK CHOI and WOJCIECH SZPANKOWSKI, Purdue University

Constrained sequences are strings satisfying certain additional structural restrictions (e.g., some patterns are forbidden). They find applications in communication, digital recording, and biology. In this article, we restrict our attention to the so-called (d, k) constrained binary sequences in which any run of zeros must be of length at least d and at most k , where $0 \leq d < k$. In many applications, one needs to know the number of occurrences of a given pattern w in such sequences, for which we coin the term *constrained pattern matching*. For a given word w , we first estimate the mean and the variance of the number of occurrences of w in a (d, k) sequence generated by a memoryless source. Then we present the central limit theorem and large deviations results. As a by-product, we enumerate asymptotically the number of (d, k) sequences with exactly r occurrences of w , and compute Shannon entropy of (d, k) sequences with a given number of occurrences of w . We also apply our results to detect under- and overrepresented patterns in neuronal data (spike trains), which satisfy structural constraints that match the framework of (d, k) binary sequences. Throughout this article we use techniques of analytic combinatorics such as combinatorial calculus, generating functions, and complex asymptotics.

Categories and Subject Descriptors: F.2.2 [Analysis of Algorithms and Problem Complexity]: Nonnumerical Algorithms and Problems—Computations on discrete structures; G.2.1 [Discrete Mathematics]: Combinatorics—Generating functions; counting problems

General Terms: Algorithms

Additional Key Words and Phrases: Autocorrelation polynomials, complex asymptotics, constrained sequences, languages, neuronal spike trains, pattern matching

ACM Reference Format:

Choi, Y. and Szpankowski, W. 2011. Constrained pattern matching. ACM Trans. Algor. 7, 2, Article 25 (March 2011), 19 pages.

DOI = 10.1145/1921659.1921671 <http://doi.acm.org/10.1145/1921659.1921671>

1. INTRODUCTION

The main idea of *constrained pattern matching* is to search for special structures (patterns) in a constrained sequence. In digital communication systems such as magnetic and optical recording, the main purpose of constrained pattern matching is to improve the performance by matching system characteristics to those of the channel. In biology, constrained sequences are in abundance (e.g., spike trains of neuronal data). In this article, our goal is to study and understand some aspects of pattern matching in constrained sequences. Although our methods seem to work for a larger class of constrained systems, we further restrict our analysis to the so-called (d, k) sequences in

This work was supported in part by the NSF grants CCF-0513636, DMS-0503742, DMS-0800568, and CCF-0830140, NIH grant R01 GM068959-01, NSA grant H98230-08-1-0092, and the AFOSR grant FA8655-08-1-3018.

Authors' addresses: Y. Choi, Department of Computer Science, Purdue University, West Lafayette, IN 47907; email: ywchoi@purdue.edu; W. Szpankowski, Department of Computer Science, Purdue University, West Lafayette, IN 47907; email: spa@cs.purdue.edu.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

© 2011 ACM 1549-6325/2011/03-ART25 \$10.00

DOI 10.1145/1921659.1921671 <http://doi.acm.org/10.1145/1921659.1921671>

which runs of zeros cannot be smaller than d nor bigger than k , where $0 \leq d < k$. Such sequences have proved useful for digital recording and biology. In digital recording, they have been widely used in hard disk drives and digital optical discs such as CD, DVD, and Blu-ray. In biology, for example, the spike trains of neuronal data, recorded from different neurons in the brain of an animal, seem to satisfy structural constraints that exactly match the framework of (d, k) binary sequences. Indeed, refractoriness requires that a neuron cannot fire two spikes in too short a time; this translates into the constraint that the induced binary spike train needs to contain at least a certain number of zeros (corresponding to no activity) between any two consecutive ones (corresponding to spikes).

In these applications, one often requires that some given words do not occur or occur only a few times in a (d, k) sequence. Therefore, we study here the following problem: given a word w or a set of words \mathcal{W} , how many times it occurs in a (d, k) sequence. For such a problem we coin the term *constrained pattern matching* as an extension of standard pattern matching [Lothaire 2005; Régnier and Szpankowski 1998; Szpankowski 2001]. We study this problem in a probabilistic framework, that is, we assume that a sequence is generated by a memoryless source and derive the (conditional) distribution of the number of occurrences of w in a (d, k) sequence. We need the conditional distribution since naturally only a small fraction of binary sequences satisfies the (d, k) constraints.

In the (standard) pattern matching problem, one asks for pattern occurrences in a binary string also known as *text* without any additional restrictions on the text. In a probabilistic framework, one determines the distribution of the number of pattern occurrences. The first analysis of such pattern matching goes back at least to Feller, and enormous progress in this area has been reported since then [Bender and Kochman 1993; Guibas and Odlyzko 1981; Lothaire 2005; Nicodème et al. 1999; Régnier and Denise 2004; Szpankowski 2001; Waterman 1995]. For instance, Guibas and Odlyzko [1981] (also refer to Jacquet and Szpankowski [1994] and Régnier and Szpankowski [1998]) revealed the fundamental role played by autocorrelation languages and their associated polynomials in the analysis of pattern matching. Régnier and Szpankowski [1997, 1998] established that the number of occurrences of a given pattern is asymptotically normal under a diversity of probabilistic models that include Markov chains. Nicodème et al. [1999] showed generally that the number of places in a random text at which a “motif” (i.e., a general regular expression pattern) terminates is asymptotically normally distributed. Bender and Kochman [1993] studied occurrences of a generalized pattern using the de Bruijn graph representation that allowed the authors to establish the central limit theorem, but without computationally explicit mean and variance. Finally, in Flajolet et al. [2006] the number of *subsequence* occurrences were studied. Recent surveys on pattern matching can be found in Lothaire [2005, Chapters 6 and 7]. To the best of our knowledge, none of these works deals with pattern matching in constrained sequences such as (d, k) sequences.

In the information theory community, (d, k) sequences were analyzed since Shannon with some recent contributions [Dembo and Kontoyiannis 2002; Kolesnik and Krachkovsky 1991; Marcus et al. 1988; Zehavi and Wolf 1988]. Pattern matching in constrained sequences can in principle be analyzed by various versions of the de Bruijn graph [Bender and Kochman 1993; Flajolet et al. 2006] or automaton approach [Bender and Kochman 1993; Nicodème et al. 1999]. This is an elegant and general approach but it sometimes leads to complicated and rather un insightful analyses that are computationally extensive. For example, in our constrained pattern matching problem, one must build a de Bruijn graph over *all* strings of length equal to the longest string in the set \mathcal{W} . The (d, k) constraints are built into the graph as *forbidden strings* (i.e., runs of zeros of length smaller than d or larger than k), which result in forbidden

edges of the graph. Based on this method, one represents the number of pattern occurrences as a product of matrices representing the underlying de Bruijn graph and hence also its largest eigenvalue (refer to Bender and Kochman [1993] and Flajolet et al. [2006]). In general, this matrix is of a large dimension and such a solution is not easily interpretable in terms of the original patterns.

In this article, we take the view of combinatorics on words as in Lothaire [2005] and Régnier and Szpankowski [1998]. We first construct languages representing (d, k) sequences containing exactly r occurrences of a given pattern w or a set of patterns \mathcal{W} . Using generating functions and complex asymptotics, we present simple and precise asymptotics for the mean and variance of the number of pattern occurrences. In particular, we estimate the probability that a randomly generated sequence is a (d, k) sequence. We also compute the probability that there are r occurrences of w in a (d, k) sequence generated by a binary memoryless source. In fact, we present the asymptotics for different ranges of r including central limit and large deviation regimes. Furthermore, we enumerate (d, k) sequences that contain exactly r occurrences of w and compute Shannon entropy when the binary source is unbiased. To put the theory into practice, we show how these theoretical results can be applied to analyze neuronal spike data.

The article is organized as follows. In Section 2, after introducing some preliminary notions and definitions we present our main analytical and experimental results. Most proofs are delayed until Section 3 where we use analytic tools such as generating functions, singularity analysis, and the saddle point method to establish our main results.

2. MAIN RESULTS

To simplify our presentation, we first derive all results for *restricted* (d, k) sequences unless stated otherwise. In general, a restricted (d, k) sequence is a (d, k) sequence that starts with 0 and ends with 1. We will relax this assumption later. We aim at finding the probability distribution of the number of occurrences of a given pattern w in a (d, k) sequence generated by a binary memoryless source. Here w is also a (d, k) sequence, and pattern overlapping is allowed.

Let us start with language representations. Define

$$\mathcal{A}_{d,k} = \{\underbrace{0\dots 0}_d, \dots, \underbrace{0\dots 0}_k\},$$

as a set of runs of zeros of length between d and k . For $d = 0$, the first word of $\mathcal{A}_{d,k}$ is an empty word ϵ . The *extended alphabet* is then [Marcus et al. 1988]

$$\mathcal{B}_{d,k} = \mathcal{A}_{d,k} \cdot \{1\} = \{\underbrace{0\dots 0}_d 1, \dots, \underbrace{0\dots 0}_k 1\}.$$

Observe that restricted (d, k) sequences are built over $\mathcal{B}_{d,k}$, and we count pattern occurrences also over $\mathcal{B}_{d,k}$. For example, $w = 01$ does *not* occur in a $(1, 4)$ sequence 0010001 since it contains two symbols 001 and 0001 over $\mathcal{B}_{d,k}$. We shall also relax this assumption later.

As in the classical pattern matching, a special set (language) plays an important role. Let $w = w_1 \dots w_m \in \{0, 1\}^m$ with $w_m = 1$ so that w is a restricted (d, k) sequence itself. Over $\mathcal{B}_{d,k}$ we have $w = \beta_1 \dots \beta_{m'}$, where $\beta_i \in \mathcal{B}_{d,k}$ and $\sum_{i=1}^{m'} |\beta_i| = m$. Let \mathcal{S} denote the *autocorrelation set* of w over $\mathcal{B}_{d,k}$, that is,

$$\mathcal{S} = \{\beta_{\ell+1}^{m'} : \beta_1^\ell = \beta_{m'-\ell+1}^{m'}\}, \quad 1 \leq \ell \leq m',$$

where $\beta_i^j = \beta_i \cdots \beta_j$ and $\beta_i^j = \epsilon$ if $i > j$. Notice that, by definition, the autocorrelation set always contains the empty word ϵ .

As in Choi and Szpankowski [2007], Jacquet and Szpankowski [2006], and Régnier and Szpankowski [1998], we introduce four languages, $\mathcal{T}_r^{(d,k)}$, $\mathcal{R}^{(d,k)}$, $\mathcal{U}^{(d,k)}$, and $\mathcal{M}^{(d,k)}$ as follows:

- (i) $\mathcal{T}_r^{(d,k)}$ as the set of all (d, k) sequences containing exactly r occurrences of w ;
- (ii) $\mathcal{R}^{(d,k)}$ as the set of all (d, k) sequences containing only one occurrence of w , located at the right end;
- (iii) $\mathcal{U}^{(d,k)}$ defined as

$$\mathcal{U}^{(d,k)} = \{u : w \cdot u \in \mathcal{T}_1^{(d,k)}\},$$

that is, a word $u \in \mathcal{U}^{(d,k)}$ if u is a (d, k) sequence and $w \cdot u$ has exactly one occurrence of w at the left end of $w \cdot u$;

- (iv) $\mathcal{M}^{(d,k)}$ defined as

$$\mathcal{M}^{(d,k)} = \{v : w \cdot v \in \mathcal{T}_2^{(d,k)} \text{ and } w \text{ occurs at the right end of } w \cdot v\},$$

that is, any word in $\{w\} \cdot \mathcal{M}^{(d,k)}$ has exactly two occurrences of w , one at the left end and the other at the right end.

To simplify our notation, we drop the upper index (d, k) unless it is necessary. It is easy to see that for $r \geq 1$ [Régnier and Szpankowski 1998; Szpankowski 2001]

$$\mathcal{T}_r = \mathcal{R} \cdot \mathcal{M}^{r-1} \cdot \mathcal{U}, \quad (1)$$

$$\mathcal{T}_0 \cdot \{w\} = \mathcal{R} \cdot \mathcal{S}. \quad (2)$$

In order to find relationships between the languages \mathcal{R} , \mathcal{M} , and \mathcal{U} , we extend the approach from Régnier and Szpankowski [1998] to yield

$$\mathcal{M}^* = \mathcal{B}^* \cdot \{w\} + \mathcal{S}, \quad (3)$$

$$\mathcal{U} \cdot \mathcal{B} = \mathcal{M} + \mathcal{U} - \{\epsilon\}, \quad (4)$$

$$\{w\} \cdot \mathcal{M} = \mathcal{B} \cdot \mathcal{R} - (\mathcal{R} - \{w\}), \quad (5)$$

where \mathcal{B}^* is the set of all restricted (d, k) sequences, that is,

$$\mathcal{B}^* = \{\epsilon\} + \mathcal{B} + \mathcal{B}^2 + \mathcal{B}^3 + \cdots.$$

Similarly, $\mathcal{M}^* = \sum_{i=0}^{\infty} \mathcal{M}^i$, where $\mathcal{M}^0 = \{\epsilon\}$. For example, (3) indicates that any word in language \mathcal{M}^* is either in \mathcal{S} (if the length of the word from \mathcal{M}^* is smaller than that of w) or it must end with w .

At this point we need to set up the probabilistic framework. Throughout, we assume that a binary sequence is generated by a memoryless source with p being the probability of emitting a “0” and $q = 1 - p$. We find the conditional probability distribution of the number of occurrences of w in a (d, k) sequence. We also compute the probability that a randomly generated sequence is a (d, k) sequence.

We start by defining for a language \mathcal{L} its *probability generating function* $L(z)$ as

$$L(z) = \sum_{u \in \mathcal{L}} P(u)z^{|u|},$$

where $P(u)$ is the probability of a word u and $|u|$ is the length of u over the binary alphabet. In particular, the *autocorrelation polynomial* $S(z)$ is the probability generating function for the autocorrelation language \mathcal{S} . In general, we write $[z^n]L(z)$ for the coefficient of $L(z)$ at z^n .

The language relationships (3)–(5) are translated into probability generating functions

$$\frac{1}{1 - M(z)} = \frac{1}{1 - B(z)} \cdot z^m P(w) + S(z), \quad (6)$$

$$U(z) = \frac{M(z) - 1}{B(z) - 1}, \quad (7)$$

$$R(z) = z^m P(w) \cdot U(z), \quad (8)$$

where

$$\begin{aligned} B(z) &= p^d q z^{d+1} + p^{d+1} q z^{d+2} + \dots + p^k q z^{k+1} \\ &= zq \frac{(zp)^d - (zp)^{k+1}}{1 - zp}. \end{aligned} \quad (9)$$

In particular, from (1),(2) and the preceding, one finds

$$T_0(z) = \frac{S(z)}{D(z)}, \quad (10)$$

$$T_r(z) = \frac{z^m P(w)(D(z) + B(z) - 1)^{r-1}}{D(z)^{r+1}}, \quad (11)$$

where

$$D(z) = S(z)(1 - B(z)) + z^m P(w). \quad (12)$$

Let O_n be a random variable representing the number of occurrences of w in a (regular) binary sequence of length n . Then, the generating function $T_r(z)$ for (d, k) sequences is defined as

$$T_r(z) = \sum_{n \geq 0} P(O_n = r, \mathcal{D}_n) z^n,$$

where \mathcal{D}_n is the event that a randomly generated binary sequence of length n is a (d, k) sequence. Let us also define the bivariate generating function $T(z, u)$ as

$$T(z, u) = \sum_{r \geq 0} T_r(z) u^r = \sum_{r \geq 0} \sum_{n \geq 0} P(O_n = r, \mathcal{D}_n) z^n u^r.$$

From (1) and (2), we find

$$T(z, u) = R(z) \frac{u}{1 - uM(z)} U(z) + T_0(z). \quad (13)$$

Observe that $T(z, u)$ is *not* a bivariate *probability* generating function since $[z^n]T(z, 1) \neq 1$. But we can easily make it a *conditional* probability generating function. First, define

$$P(\mathcal{D}_n) = [z^n]T(z, 1),$$

as the probability that a randomly generated sequence of length n is a (d, k) sequence. We also introduce a short-hand notation $O_n(\mathcal{D}_n)$ for the conditional number of occurrences of w in a (d, k) sequence. More formally,

$$P(O_n(\mathcal{D}_n) = r) = P(O_n = r \mid \mathcal{D}_n).$$

Therefore, the probability generating function of $O_n(\mathcal{D}_n)$ is

$$\mathbf{E}[u^{O_n(\mathcal{D}_n)}] = \frac{[z^n]T(z, u)}{[z^n]T(z, 1)}.$$

Thus, the expected value of $O_n(\mathcal{D}_n)$ is

$$\mathbf{E}[O_n(\mathcal{D}_n)] = \frac{[z^n]T_u(z, 1)}{[z^n]T(z, 1)}, \quad (14)$$

where $T_u(z, 1)$ is the derivative of $T(z, u)$ at $u = 1$, and

$$\mathbf{E}[O_n(\mathcal{D}_n)(O_n(\mathcal{D}_n) - 1)] = \frac{[z^n]T_{uu}(z, 1)}{[z^n]T(z, 1)},$$

is the second factorial moment, where $T_{uu}(z, 1)$ is the second derivative with respect to u at $u = 1$ [Flajolet and Sedgewick 2008; Szpankowski 2001].

2.1. Analytical Results

Our main analytical results are summarized in the following two theorems. In Theorem 2.1 we present asymptotics for $P(\mathcal{D}_n)$ and the first two moments of $O_n(\mathcal{D}_n)$. The proof is presented in Section 3.1.

THEOREM 2.1. *Let $\rho := \rho(p)$ be the unique positive real root of $B(z) = 1$ where $B(z)$ is defined in (9), and let $\lambda = 1/\rho$. Then the probability of generating a (d, k) sequence is equal to*

$$P(\mathcal{D}_n) = \frac{1}{B'(\rho)}\lambda^{n+1} + O(\omega^n), \quad (15)$$

for some $\omega < \lambda$. Furthermore,

$$\mathbf{E}[O_n(\mathcal{D}_n)] = \frac{(n-m+1)P(w)}{B'(\rho)}\lambda^{-m+1} + O(1),$$

and the variance is

$$\begin{aligned} \mathbf{Var}[O_n(\mathcal{D}_n)] &= (n-m+1)P(w) \left[\frac{(1-2m)P(w)}{B'(\rho)^2}\lambda^{-2m+2} \right. \\ &\quad \left. + \frac{P(w)B''(\rho)}{B'(\rho)^3}\lambda^{-2m+1} + \frac{2S(\rho)-1}{B'(\rho)}\lambda^{-m+1} \right] + O(1), \end{aligned} \quad (16)$$

for large n .

The next theorem presents the asymptotic formulas of $P(O_n(\mathcal{D}_n) = r)$ for three different ranges of r , including central limit and large deviations regimes. The results are derived in Sections 3.2–3.4 by using analytical tools.

THEOREM 2.2. *Let $\tau := \tau(p, w)$ be the smallest positive real root of $D(z) = 0$ where $D(z)$ is defined in (12), and $\rho := \rho(p) < \tau$ be the unique positive real root of $B(z) = 1$. Then, for large n , the following holds.*

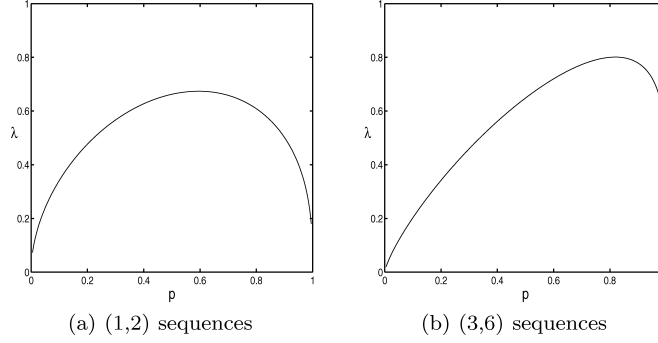
(i) For $r = O(1)$ with $r \geq 1$,

$$P(O_n(\mathcal{D}_n) = r) \sim \frac{P(w)B'(\rho)(1-B(\tau))^{r-1}}{D'(\tau)^{r+1}\tau^{r-m}} \binom{n-m+r}{r} \left(\frac{\rho}{\tau}\right)^{n+1}.$$

(ii) [central limit theorem] For $r = \mathbf{E}[O_n(\mathcal{D}_n)] + x\sqrt{\mathbf{Var}[O_n(\mathcal{D}_n)]}$ with $x = O(1)$,

$$\frac{O_n(\mathcal{D}_n) - \mathbf{E}[O_n(\mathcal{D}_n)]}{\sqrt{\mathbf{Var}[O_n(\mathcal{D}_n)]}} \xrightarrow{d} N(0, 1),$$

where $N(0, 1)$ is the standard normal distribution.

Fig. 1. λ versus p .

(iii) [large deviations] For $r = (1 + \delta)\mathbf{E}[O_n(\mathcal{D}_n)]$ with $\delta > 0$, let a be a real constant such that $na = (1 + \delta)\mathbf{E}[O_n(\mathcal{D}_n)]$, and let

$$h_a(z) = a \log M(z) - \log z.$$

We denote by z_a the unique real root of the equation $h'_a(z) = 0$ such that $z_a \in (0, \rho)$. Then,

$$P(O_n(\mathcal{D}_n) = na) = \frac{c_1 \cdot e^{-nI(a)}}{\sqrt{2\pi n}} \left[1 + \frac{c_2}{n} + O\left(\frac{1}{n^2}\right) \right],$$

and

$$P(O_n(\mathcal{D}_n) \geq na) = \frac{c_1 \cdot e^{-nI(a)}}{\sqrt{2\pi n(1 - M(z_a))}} \left[1 + O\left(\frac{1}{n}\right) \right],$$

where $I(a) = -\log \rho - h_a(z_a) > 0$ with

$$c_1 = \frac{\rho B'(\rho)g(z_a)}{\tau_a}, \quad g(z) = \frac{P(w)z^{m-1}}{D(z)^2 M(z)}, \quad \text{and} \quad \tau_a^2 = h''_a(z_a).$$

The constant c_2 is explicitly computed in (31).

Remark 1. In Figure 1, we plot $\lambda = 1/\rho$ versus p for various (d, k) sequences. Observe that the probability $P(\mathcal{D}_n) \asymp \lambda^n$ is asymptotically maximized for some $p \neq 0.5$ (biased source) which may be used to design a better run-length coding as in Aviran et al. [2006].

Remark 2. When the binary source is unbiased ($p = q = \frac{1}{2}$), we can count the number $N_n(r)$ of (d, k) sequences of length n that contain w exactly r times, by computing $[z^n]T_r(2z)$ (r is fixed with respect to n .) In fact, $N_n(r) = 2^n P(O_n = r, \mathcal{D}_n)$ and one finds asymptotics of $N_n(r)$ from part (i) of Theorem 2.2. In particular, Shannon entropy is

$$C(r) = \lim_{n \rightarrow \infty} \frac{\log_2 N_n(r)}{n} = \log_2 \left(\frac{2}{\tau} \right),$$

where $\tau = \tau(1/2, w)$ is defined in Theorem 2.2 for $p = 1/2$.

Remark 3. Up to now we considered only *restricted* (d, k) sequences. A small modification can extend this analysis to *all* (d, k) sequences. Let \mathcal{T}_r^{all} be the set of all (d, k) sequences containing exactly r occurrences of w . Then for $d \geq 1$

$$\mathcal{T}_r^{all} = \{\epsilon, 1\} \cdot \mathcal{T}_r \cdot (\{\epsilon\} + \mathcal{A}_{d,k}),$$

and generating functions and asymptotic expressions can be derived from the precedings.

Remark 4. We counted the occurrences of the pattern w over the alphabet $\mathcal{B}_{d,k}$. We can extend this analysis to count the occurrences over a binary alphabet (e.g., $w = 01$ occurs twice in the (1, 4) sequence, 0010001). Again, let $w = w_1 \dots w_m \in \{0, 1\}^m$ with $w_m = 1$, and w be represented over $\mathcal{B}_{d,k}$, that is, $w = \beta_1 \dots \beta_m$ where $\beta_i \in \mathcal{B}_{d,k}$. Then the autocorrelation set \mathcal{S}_2 over the binary alphabet is defined as

$$\mathcal{S}_2 = \{w_{\ell+1}^m : w_1^\ell = w_{m-\ell+1}^m\}, \quad 1 \leq \ell \leq m.$$

Using the languages \mathcal{T}_r , \mathcal{R} , \mathcal{M} , and \mathcal{U} defined earlier, we find

$$\begin{aligned} \mathcal{T}_r &= \mathcal{R} \cdot \mathcal{M}^{r-1} \cdot \mathcal{U}, \\ \mathcal{T}_0 \cdot \mathcal{Z} \cdot \{w\} &= \mathcal{R} \cdot \mathcal{S}_2, \\ \mathcal{M}^* &= \mathcal{B}^* \cdot \mathcal{Z} \cdot \{w\} + \mathcal{S}_2, \\ \mathcal{U} \cdot \mathcal{B} &= \mathcal{M} + \mathcal{U} - \{\epsilon\}, \\ \mathcal{Z} \cdot \{w\} \cdot \mathcal{M} &= \mathcal{B} \cdot \mathcal{R} - (\mathcal{R} - \mathcal{Z} \cdot \{w\}), \end{aligned}$$

where $\mathcal{Z} = \{\epsilon, 0, 00, \dots, 0^{k+1-|\beta_1|}\}$, and 0^k denotes a run of zeros of length k . We replace \mathcal{S} by \mathcal{S}_2 and $\{w\}$ by $\mathcal{Z} \cdot \{w\}$ in (1)–(5). The key idea is that any word in $\mathcal{Z} \cdot \{w\}$ is now considered to have an occurrence of w . Applying the same techniques as before we can derive the generating functions and asymptotic results.

Remark 5. We can also extend our results to a set of patterns $\mathcal{W} = \{w_1, w_2, \dots, w_K\}$ such that w_i ($1 \leq i \leq K$) is not a substring of another pattern w_j ($1 \leq j \leq K, i \neq j$) over alphabet $\mathcal{B}_{d,k}$. (An extension to a general set of words is possible through to novel ideas discussed in Bassino et al. [2007].) Let us introduce some new languages. In particular, for any given two strings u and v , let

$$\mathcal{S}_{u,v} = \{v_{k+1}^{|v|} : u_{|u|-k+1}^{|u|} = v_1^k\}, \quad 1 \leq k \leq \min\{|u|, |v|\},$$

be the *correlation set*. Now we define a correlation set over $\mathcal{B}_{d,k}$ for patterns in \mathcal{W} . Let $w_i = \beta_{i_1} \dots \beta_{i_s}$ and $w_j = \beta_{j_1} \dots \beta_{j_m}$. Then \mathcal{S}_{ij} , the correlation set for w_i and w_j over $\mathcal{B}_{d,k}$, is defined as

$$\mathcal{S}_{ij} = \{\beta_{j_{\ell+1}}^{j_m} : \beta_{i_s-\ell+1}^{i_s} = \beta_{j_1}^{j_\ell}\}, \quad 1 \leq \ell \leq \min\{s, m\}.$$

For $1 \leq i, j \leq K$, we introduce new languages as follows (again, we drop the upper index (d, k)):

- (i) \mathcal{R}_i as the set of all (d, k) sequences (over $\mathcal{B}_{d,k}$) containing only one occurrence of w_i , located at the right end;
- (ii) \mathcal{U}_i is defined as $\{u : w_i \cdot u \in \mathcal{T}_1\}$, that is, a word $u \in \mathcal{U}_i$ if u is a (d, k) sequence and $w_i \cdot u$ has exactly one occurrence of w_i at the left end of $w_i \cdot u$;
- (iii) $\mathcal{M}_{ij}^{[r]}$ defined as, for $r \geq 1$,

$$\mathcal{M}_{ij}^{[r]} = \{v : w_i \cdot v \in \mathcal{T}_{r+1} \text{ and } w_j \text{ occurs at the right end of } w_i \cdot v\},$$

that is, any word in $\{w_i\} \cdot \mathcal{M}_{ij}^{[r]}$ is a (d, k) sequence and has one occurrence of w_i at the left end, one occurrence of w_j at the right end, and $r - 1$ occurrences from \mathcal{W} elsewhere. We write $\mathcal{M}_{ij} = \mathcal{M}_{ij}^{[1]}$.

We can see that \mathcal{T}_r ($r \geq 1$) and \mathcal{T}_0 are represented as follows:

$$\mathcal{T}_r = \sum_{1 \leq i, j \leq K} \mathcal{R}_i \cdot \mathcal{M}_{ij}^{[r-1]} \cdot \mathcal{U}_j, \quad (17)$$

$$\mathcal{T}_0 \cdot \{w_j\} = \mathcal{R}_j + \sum_{1 \leq i \leq K} \mathcal{R}_i \cdot (\mathcal{S}_{ij} - \{\epsilon\}), \quad (18)$$

for $1 \leq j \leq K$. The languages \mathcal{M}_{ij} , \mathcal{U}_i , and \mathcal{R}_j satisfy following relationships [Régnier and Szpankowski 1997] for $1 \leq i, j \leq K$.

$$\sum_{k \geq 1} \mathcal{M}_{ij}^{[k]} = \mathcal{B}^* \cdot \{w_j\} + \mathcal{S}_{ij} - \{\epsilon\} \quad (19)$$

$$\mathcal{U}_i \cdot \mathcal{B} = \sum_{1 \leq j \leq K} \mathcal{M}_{ij} + \mathcal{U}_i - \{\epsilon\} \quad (20)$$

$$\mathcal{B} \cdot \mathcal{R}_j - (\mathcal{R}_j - \{w_j\}) = \sum_{1 \leq i \leq K} \{w_i\} \cdot \mathcal{M}_{ij} \quad (21)$$

As before, the language relationships (19)–(21) are translated into generating functions [Régnier and Szpankowski 1997]. We have

$$\begin{aligned} (\mathbf{I} - \mathbf{M}(z))^{-1} &= \mathbf{S}(z) + \frac{1}{1 - B(z)} \vec{\mathbf{1}} \cdot \vec{W}^t(z), \\ \vec{U}(z) &= \frac{1}{1 - B(z)} (\mathbf{I} - \mathbf{M}(z)) \cdot \vec{\mathbf{1}}, \\ \vec{R}^t(z) &= \frac{1}{1 - B(z)} \vec{W}^t(z) \cdot (\mathbf{I} - \mathbf{M}(z)), \end{aligned}$$

where $\mathbf{M}(z)$ and $\mathbf{S}(z)$ are $K \times K$ matrices such that $M_{ij}(z)$ and $S_{ij}(z)$ are the (i, j) -elements in $\mathbf{M}(z)$ and $\mathbf{S}(z)$, respectively. Furthermore, \mathbf{I} is the $K \times K$ identity matrix, and $\vec{W}(z)$, $\vec{R}(z)$, $\vec{U}(z)$, and $\vec{\mathbf{1}}$ are column vectors of length K such that

$$\begin{aligned} \vec{W}(z) &= (z^{|w_1|} P(w_1), \dots, z^{|w_K|} P(w_K))^t, \quad \vec{R}(z) = (R_1(z), \dots, R_K(z))^t, \\ \vec{U}(z) &= (U_1(z), \dots, U_K(z))^t, \quad \text{and } \vec{\mathbf{1}} = (1, \dots, 1)^t. \end{aligned}$$

From (17)–(18) and the preceding, one finds

$$\begin{aligned} T_0(z) &= \frac{\vec{R}^t(z) \cdot \mathbf{S}(z) \cdot \vec{\mathbf{1}}}{\vec{W}^t(z) \cdot \vec{\mathbf{1}}}, \\ T_r(z) &= \vec{W}^t(z) \cdot (\mathbf{D}(z) + (B(z) - 1)\mathbf{I})^{r-1} \cdot \mathbf{D}(z)^{-(r+1)} \cdot \vec{\mathbf{1}}, \end{aligned}$$

where $\mathbf{D}(z) = \vec{\mathbf{1}} \cdot \vec{W}^t(z) + (1 - B(z))\mathbf{S}(z)$.

Using this and following the footsteps of our previous analysis, as presented in the next section, one easily shows that for large n

$$\mathbf{E}[O_n(\mathcal{D}_n)] = \sum_{i=1}^K \frac{(n - |w_i| + 1)P(w_i)}{B(\rho)} \lambda^{-|w_i|+1} + O(1),$$

and

$$\mathbf{Var}[O_n(\mathcal{D}_n)] = n\alpha + O(1),$$

where $\rho := \rho(p)$ is the unique positive real root of $B(z) = 1$, $\lambda = 1/\rho$, and α is an explicitly computable constant. Furthermore, technically more challenging analysis allows us to conclude that $O_n(\mathcal{D}_n)$ satisfies the central limit theorem.

2.2. Experimental Results

In this section, we illustrate how our theoretical results can be used to obtain justifiably accurate statistical inferences from biological data. We apply Theorem 2.2 to

detect under- and overrepresented structures in neuronal data (spike trains). We shall first argue that neuronal data are best represented by constrained sequences. Indeed, current technology allows for the simultaneous recording of the spike trains from one hundred (or more) different neurons in the brain of a live animal. Such experiments have produced enormous amounts of extremely valuable data, and one of the core research areas of activity in neuroscience is devoted to developing accurate and precise statistical tools to quantify and describe the amount and representation of the information that is contained in this data [Paninski 2003]. Because of the very nature of the biological mechanisms that produce them, spike train data satisfy structural constraints that match the framework of (d, k) binary sequences, as discussed in the Introduction.

For experiments, we use single-electrode data from cortical neurons under random current injection. The details can be found in Jolivet et al. [2006, 2008]. This spike timing data can be transformed into a (d, k) sequence by setting the time resolution and dividing time into *bins* of the same size. Each time bin is represented by a bit 0 or 1. If there is a spike in a certain time bin, it is represented by a bit 1; otherwise it is represented by a bit 0. A fundamental question is how one classifies an occurrence of a pattern as significant. Here, the connotation of “significant” is used for observed data that is interesting, surprising, suspicious, or—perhaps most importantly—meaningful. We classify a pattern as significant if it is unlikely to occur fortuitously, that is, in a randomly generated instance of the problem. Thus, we compare experimental data to the reference model, which in our case is the probabilistic model developed in this article.

Having this in mind, and using our large deviations results, we derive a threshold, O_{th} , above which pattern occurrences will be classified as statistically significant. The threshold is defined as the minimum O_{th} such that

$$P(O_n(\mathcal{D}_n) \geq O_{th}) \leq \alpha_{th},$$

where α_{th} is a given probability threshold. From part (iii) of Theorem 2.2 we easily conclude that for α_{th} in the range of the large deviations domain, the threshold is $O_{th} = n\alpha_{th}$, where

$$\alpha_{th} \approx I^{-1}(\log(1/\alpha_{th})/n),$$

and $I^{-1}(\cdot)$ is the inverse function of $I(a)$ defined in Theorem 2.2.

To set up our reference model, we need to fix the parameters d , k , and p . First we can find d and k by observing the binary sequence (e.g., by finding the minimum and maximum length of runs of zeros between spikes in the sequence). Then we can find p by solving the following simultaneous equations with variables ρ and p .

$$B(\rho) = 1 \quad \text{and} \quad 1 - p = \frac{1}{\rho B(\rho)}$$

Note that the coefficients of $B(z)$ depend on p . The second equation follows from the fact that $\rho B(\rho)$ captures the average length of symbols of $\mathcal{B}_{d,k}$ in a (d, k) sequence, and thus its reciprocal represents $1 - p$. In other words, we estimate p indirectly through the estimation of d and k . One might be tempted to estimate p by just counting the total number of 0's and dividing it by the length of the sequence. But this could lead to a poor estimate if a large portion of (d, k) sequence set is not typical.

In our experiment, we set the size of the bin to 3 ms and obtained a $(d, k) = (1, 6)$ sequence of length 2193 with $p = 0.752686$. Figure 2 shows the number of occurrences for various patterns w within a (sliding) window of size 500; here we use a short-hand

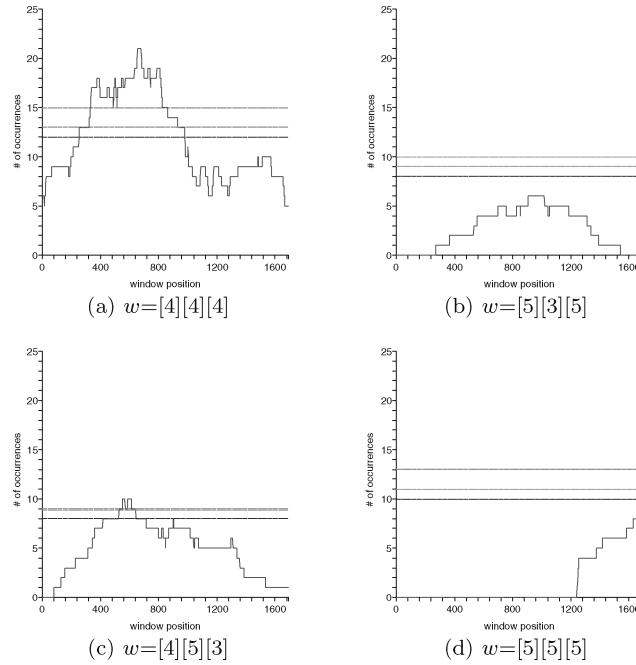


Fig. 2. Number of occurrences of w within a window of size 500; here $[i]$ stands for the pattern $0 \dots 01$ with $i - 1$ zeros.

notation $[i]$ for a pattern

$$\underbrace{0 \dots 0}_{i-1} 1.$$

The three horizontal lines represent thresholds for $\alpha_{th} = 10^{-6}$, 10^{-7} , and 10^{-8} , respectively. As expected, the thresholds vary with the structure of w . If the number of occurrences exceeds the threshold at some position, we claim the pattern occurrence is statistically significant in that window. This observation can be used as a starting point for interpretation of neural signals, although there is still a huge gap between patterns of spike trains and their meaning in a real nervous system. In passing we observe that one would have obtained quite different threshold values if constraints were ignored.

3. ANALYSIS

In this section, we prove Theorems 2.1 and 2.2 of the previous section. In Section 3.1, we derive $P(\mathcal{D}_n)$ and the first two moments of $O_n(\mathcal{D}_n)$. In Sections 3.2–3.4 we estimate asymptotically $P(O_n(\mathcal{D}_n) = r)$ for various ranges of r .

3.1. First and Second Moments

We first obtain asymptotic formulas for the mean and the variance of $O_n(\mathcal{D}_n)$. From (6)–(13), we find

$$T(z, 1) = \frac{1}{1 - B(z)}, \quad T_u(z, 1) = \frac{z^m P(w)}{(1 - B(z))^2},$$

25:12

Y. Choi and W. Szpankowski

and

$$T_{uu}(z, 1) = \frac{2z^m P(w)M(z)}{U(z)(1-B(z))^3} = \frac{2z^m P(w)D(z)}{(1-B(z))^3} - \frac{2z^m P(w)}{(1-B(z))^2}.$$

By Cauchy's coefficient formula and Cauchy's residue theorem [Flajolet and Sedgewick 2008; Szpankowski 2001] we immediately obtain

$$P(\mathcal{D}_n) = [z^n]T(z, 1) = [z^n]\frac{1}{1-B(z)} = \frac{1}{B'(\rho)}\lambda^{n+1} + O(\omega^n),$$

where ρ is the unique positive real root of $B(z) = 1$, $\lambda = 1/\rho$, and $\omega < \lambda$. We can easily prove that there always exists a unique positive real root of $B(z) = 1$, which is greater than 1, and its modulus is the smallest among all complex roots.

To find moments, we proceed as follows.

$$[z^n]T_u(z, 1) = [z^n]\frac{z^m P(w)}{(1-B(z))^2} = \frac{P(w)}{B'(\rho)^2} \left((n-m+1)\lambda + \frac{B''(\rho)}{B'(\rho)} \right) \lambda^{n-m+1} + O(\omega^n)$$

Thus

$$\mathbf{E}[O_n(\mathcal{D}_n)] = \frac{[z^n]T_u(z, 1)}{[z^n]T(z, 1)} = \frac{(n-m+1)P(w)}{B'(\rho)}\lambda^{-m+1} + O(1).$$

Similarly,

$$\mathbf{Var}[O_n(\mathcal{D}_n)] = \frac{[z^n]T_{uu}(z, 1)}{[z^n]T(z, 1)} + \mathbf{E}[O_n(\mathcal{D}_n)] - \mathbf{E}[O_n(\mathcal{D}_n)]^2.$$

After some algebra, we establish the formula on the variance in (16).

3.2. Distribution for $r = O(1)$

We prove here part (i) of Theorem 2.2, that is, we estimate $P(O_n(\mathcal{D}_n) = r)$ for $r = O(1)$. By Cauchy's coefficient formula and Cauchy's residue theorem,

$$\begin{aligned} P(O_n = r, \mathcal{D}_n) &= [z^n]T_r(z) = [z^{n-m}]\frac{P(w)(D(z) + B(z) - 1)^{r-1}}{D(z)^{r+1}} \\ &= \sum_{j=1}^{r+1} (-1)^j a_j \binom{n-m+j-1}{j-1} \left(\frac{1}{\tau}\right)^{n-m+j} + O(t^n), \end{aligned}$$

where $\tau < t^{-1}$ is the smallest positive real root of $D(z) = 0$, and

$$a_{r+1} = \frac{P(w)(B(\tau) - 1)^{r-1}}{D'(\tau)^{r+1}}.$$

We can easily prove that there exists at least one positive real root of $D(z) = 0$, which is greater than ρ .

Finally, we find

$$\begin{aligned} P(O_n(\mathcal{D}_n) = r) &= \frac{P(O_n = r, \mathcal{D}_n)}{P(\mathcal{D}_n)} \\ &\sim \frac{P(w)B'(\rho)(1-B(\tau))^{r-1}}{D'(\tau)^{r+1}} \binom{n-m+r}{r} \frac{\rho^{n+1}}{\tau^{n-m+r+1}}, \end{aligned}$$

as desired for Theorem 2.2(i).

3.3. Distribution for $r = \mathbf{E}[O_n(\mathcal{D}_n)] + x\sqrt{\mathbf{Var}[O_n(\mathcal{D}_n)]}$

We now establish the central limit theorem, that is, part (ii) of Theorem 2.2. We estimate $P(O_n(\mathcal{D}_n) = r)$ for $r = \mathbf{E}[O_n(\mathcal{D}_n)] + x\sqrt{\mathbf{Var}[O_n(\mathcal{D}_n)]}$ with $x = O(1)$. Define

$$T_n(u) = \mathbf{E}[u^{O_n(\mathcal{D}_n)}] = \frac{[z^n]T(z, u)}{[z^n]T(z, 1)}, \quad (22)$$

and

$$\mu_n = \mathbf{E}[O_n(\mathcal{D}_n)], \quad \sigma_n = \sqrt{\mathbf{Var}[O_n(\mathcal{D}_n)]}.$$

By Goncharov's theorem [Szpankowski 2001], it suffices to prove the following

$$\lim_{n \rightarrow \infty} e^{-v\mu_n/\sigma_n} T_n(e^{v/\sigma_n}) = e^{v^2/2}$$

for all $v = it'$ where $-\infty < t' < \infty$. However, we shall observe that the preceding holds for complex v .

Let $\rho(u)$ be the smallest real root of $1 - uM(z) = 0$. We can easily find out that the pole of $T_0(z)$ is always greater than $\rho(u)$. Then, by Cauchy's coefficient formula and Cauchy's residue theorem, from (13) we get

$$[z^n]T(z, u) = c(u)\lambda^{n+1}(u) + O(\omega^n(u)), \quad (23)$$

$$c(u) = \frac{R(\rho(u))U(\rho(u))}{M'(\rho(u))},$$

and $\lambda(u) = 1/\rho(u)$ where $|\omega(u)| < \lambda(u)$. Thus, by (22),

$$P(\mathcal{D}_n)T_n(u) = c(u)\lambda^{n+1}(u) + O(\omega^n(u)), \quad (24)$$

since $P(\mathcal{D}_n) = [z^n]T(z, 1)$.

Let $u = e^t$ and $t = v/\sigma_n$. Since $t \rightarrow 0$ and $u \rightarrow 1$ as $n \rightarrow \infty$, using Taylor series around $t = 0$ we get

$$\lambda(e^t) = \lambda(1) + \lambda'(1)t + \frac{\lambda''(1) + \lambda'(1)}{2}t^2 + O(t^3). \quad (25)$$

Now let us find $\lambda'(1)$ and $\lambda''(1)$. From (23), we observe that

$$[z^n]T_u(z, 1) = (n+1)c(1)\lambda^n(1)\lambda'(1) + c'(1)\lambda^{n+1}(1) + O(n\omega^n(1)). \quad (26)$$

By (14), (23), and (26), we obtain

$$\lambda'(1) = \frac{\mu_n}{n+1}\lambda(1) - \frac{c'(1)}{(n+1)c(1)}\lambda(1) + O(n\xi^n(1)), \quad (27)$$

where $\xi(u) = \omega(u)/\lambda(u)$. We note that $|\xi(u)| < 1$. Similarly, after some algebra, we arrive at

$$\lambda''(1) = \left(\frac{\sigma_n^2 - \mu_n}{n+1} + \frac{\mu_n^2}{(n+1)^2} \right) \lambda(1) + O\left(\frac{1}{n}\right). \quad (28)$$

Using (24), (25), (27), (28), and the fact that $t = O(\frac{1}{\sqrt{n}})$, we get

$$\begin{aligned} P(\mathcal{D}_n)T_n(e^t) &= c(u)\lambda^{n+1}(u) \cdot (1 + O(\xi^n(u))) \\ &= c(u)\lambda^{n+1}(1) \left[1 + \frac{\mu_n}{n+1}t + \frac{1}{2} \left(\frac{\sigma_n^2}{n+1} + \frac{\mu_n^2}{(n+1)^2} \right) t^2 + O(t^3) \right]^{n+1} \cdot (1 + O(\xi^n(u))). \end{aligned}$$

25:14

Y. Choi and W. Szpankowski

Therefore,

$$e^{-\nu\mu_n/\sigma_n}P(\mathcal{D}_n)T_n(e^{\nu/\sigma_n}) = c(u)\lambda^{n+1}(1)\left(1 + \frac{\sigma_n^2}{2(n+1)}t^2 + O(t^3)\right)^{n+1} \cdot (1 + O(\xi^n(u))),$$

and

$$\lim_{n \rightarrow \infty} e^{-\nu\mu_n/\sigma_n}T_n(e^{\nu/\sigma_n}) = \lim_{n \rightarrow \infty} \left(1 + \frac{\nu^2}{2(n+1)} + O(t^3)\right)^{n+1} = \exp\left(\frac{\nu^2}{2}\right).$$

This completes the proof of Theorem 2.2(ii).

3.4. Distribution for $r = (1 + \delta)\mathbf{E}[O_n(\mathcal{D}_n)]$

Finally we establish the large deviations results in part (iii) of Theorem 2.2, that is, we compute $P(O_n(\mathcal{D}_n) = r)$ for $r = (1 + \delta)\mathbf{E}[O_n(\mathcal{D}_n)]$ for some $\delta > 0$. Let a be a real constant such that $na = (1 + \delta)\mathbf{E}[O_n(\mathcal{D}_n)]$, and we compute $P(O_n(\mathcal{D}_n) = na)$ asymptotically when na is an integer. Clearly,

$$P(O_n(\mathcal{D}_n) = na) = [u^{na}]T_n(u) = \frac{[z^n][u^{na}]T(z, u)}{[z^n]T(z, 1)}. \quad (29)$$

By (13),

$$\begin{aligned} [u^{na}]T(z, u) &= [u^{na}] \left(T_0(z) + uR(z)U(z) \sum_{i=0}^{\infty} (uM(z))^i \right) \\ &= R(z)U(z)M(z)^{na-1} = \frac{P(w)z^n}{D(z)^2} M(z)^{na-1}. \end{aligned}$$

Hence, Cauchy's coefficient formula leads to [Flajolet and Sedgewick 2008; Szpankowski 2001]

$$[z^n][u^{na}]T(z, u) = \frac{1}{2\pi i} \oint \frac{P(w)z^n}{D(z)^2} M(z)^{na-1} \frac{1}{z^{n+1}} dz,$$

where the integration is done along any contour around zero in the convergence circle.

In order to derive large deviation results, we need to apply the *saddle point method* [Flajolet and Sedgewick 2008; Szpankowski 2001]. Therefore, we define the function $h_a(z)$ of complex variable z as

$$h_a(z) = a \log M(z) - \log z,$$

such that

$$[z^n][u^{na}]T(z, u) = \frac{1}{2\pi i} \oint e^{nh_a(z)} g(z) dz,$$

where

$$g(z) = \frac{P(w)z^{n-1}}{D(z)^2 M(z)}.$$

In the lemma that follows, we characterize some properties of $h_a(z)$ that are needed to estimate the integral. The proof can be found in Appendix A.

LEMMA 3.1. (i) *There exists a unique real root z_a of the equation $h'_a(z) = 0$ that satisfies $0 < z_a < \rho$ for some constant a described in Appendix A.*

(ii) $h''_a(z_a) > 0$.

(iii) $h_a(z_a) < -\log \rho$.

Let z_a be the unique positive real root of the equation $h'_a(z) = 0$. We evaluate the integral on $\mathcal{C} = \{z : |z| = z_a\}$, and we first split \mathcal{C} into \mathcal{C}_0 and \mathcal{C}_1 where $\mathcal{C}_0 = \{z \in \mathcal{C} : |\arg(z)| \leq \theta_0\}$ and $\mathcal{C}_1 = \{z \in \mathcal{C} : |\arg(z)| \geq \theta_0\}$ for some θ_0 . That is,

$$[z^n][u^{na}]T(z, u) = I_0 + I_1,$$

where

$$I_0 = \frac{1}{2\pi i} \int_{\mathcal{C}_0} e^{nh_a(z)} g(z) dz, \quad \text{and} \quad I_1 = \frac{1}{2\pi i} \int_{\mathcal{C}_1} e^{nh_a(z)} g(z) dz.$$

We compute I_0 first and we later show that $|I_1|$ is exponentially smaller than I_0 .

Now we set $\theta_0 = n^{-2/5}$ and compute I_0 with the change of variable $z = z_a e^{i\theta}$,

$$I_0 = \frac{z_a}{2\pi} \int_{-\theta_0}^{+\theta_0} \exp(nh_a(z_a e^{i\theta}) + i\theta) g(z_a e^{i\theta}) d\theta.$$

To simplify the notation, let us define some variables as follows.

$$\tau_a^2 = h''_a(z_a) \quad (\text{refer to part (ii) of Lemma 3.1}), \quad \beta_a = \frac{h_a^{(3)}(z_a)}{3!\tau_a^3}, \quad \text{and} \quad \gamma_a = \frac{h_a^{(4)}(z_a)}{4!\tau_a^4}$$

Using Taylor series around $\theta = 0$, we arrive at

$$h_a(z_a e^{i\theta}) = h_a(z_a) - \frac{\tau_a^2 z_a^2}{2} \theta^2 - \left(\beta_a \tau_a^3 z_a^3 + \frac{\tau_a^2 z_a^2}{2} \right) i\theta^3 + O(\theta^4).$$

When $|\theta| \leq \theta_0$, $n\theta^k \rightarrow 0$ ($k \geq 3$) as $n \rightarrow \infty$. Thus, using Taylor series, we get

$$e^{nh_a(z_a e^{i\theta}) + i\theta} = \exp\left(nh_a(z_a) - \frac{\tau_a^2 z_a^2}{2} n\theta^2\right) \left(1 + \alpha(\theta) + \frac{\alpha(\theta)^2}{2!} + \frac{\alpha(\theta)^3}{3!} + \dots\right),$$

where

$$\alpha(\theta) = i\theta - \left(\beta_a \tau_a^3 z_a^3 + \frac{\tau_a^2 z_a^2}{2} \right) i n\theta^3 + O(n\theta^4).$$

With the change of variable $\theta = \frac{\omega}{\tau_a z_a \sqrt{n}}$, we rewrite

$$\alpha(\theta) = \eta(\omega) = \frac{i\omega}{\tau_a z_a \sqrt{n}} - \left(\beta_a + \frac{1}{2\tau_a z_a} \right) \frac{i\omega^3}{\sqrt{n}} + O\left(\frac{\omega^4}{n}\right)$$

and

$$I_0 = \frac{e^{nh_a(z_a)}}{2\pi \tau_a \sqrt{n}} \int_{-\omega_0}^{+\omega_0} \exp\left(-\frac{\omega^2}{2}\right) \left(1 + \eta(\omega) + \frac{\eta(\omega)^2}{2!} + \dots\right) g\left(z_a e^{\frac{i\omega}{\tau_a z_a \sqrt{n}}}\right) d\omega,$$

where $\omega_0 = \tau_a z_a n^{1/10}$. We observe that each term of odd degree of ω in

$$\left[\left(1 + \eta(\omega) + \frac{\eta(\omega)^2}{2!} + \frac{\eta(\omega)^3}{3!} + \dots\right) g\left(z_a e^{\frac{i\omega}{\tau_a z_a \sqrt{n}}}\right) \right]$$

contributes nothing to the integral. Thus, using Taylor series of $g\left(z_a e^{\frac{i\omega}{\tau_a z_a \sqrt{n}}}\right)$ around $\omega = 0$, we arrive at

$$I_0 = \frac{e^{nh_a(z_a)}}{2\pi \tau_a \sqrt{n}} \int_{-\omega_0}^{+\omega_0} \exp\left(-\frac{\omega^2}{2}\right) \left(A + B\omega^2 + C\omega^4 + D\omega^6 + O\left(\frac{1}{n^2}\right)\right) d\omega$$

where A , B , C , and D are explicitly computable constants. Using the fact that, as $\omega_0 \rightarrow \infty$,

$$\int_{\omega_0}^{+\infty} x^k e^{-\frac{x^2}{2}} dx = O(e^{-\frac{1}{2}\omega_0^2}), \quad \text{and} \quad \int_{-\infty}^{+\infty} x^{2k} e^{-\frac{x^2}{2}} dx = \frac{\Gamma(2k)}{2^{k-1}\Gamma(k)} \sqrt{2\pi},$$

we finally obtain

$$I_0 = \frac{g(z_a) e^{nh_a(z_a)}}{\tau_a \sqrt{2\pi n}} \left[1 + \frac{1}{n} \left(\frac{3\beta_a g'(z_a)}{\tau_a g(z_a)} - \frac{g''(z_a)}{2\tau_a^2 g(z_a)} + 3\gamma_a - \frac{15\beta_a^2}{2} \right) + O\left(\frac{1}{n^2}\right) \right]. \quad (30)$$

It is easy to see that the main contribution to the large deviations comes from I_0 . Thus, we only need to show that I_1 is small.

We compute a bound on $|I_1|$, and we show that it is exponentially smaller than I_0 . We can rewrite $I_1 = I_1^+ + I_1^-$ where

$$I_1^+ = \frac{z_a}{2\pi} \int_{\theta_0}^{\pi} e^{nh_a(z_a e^{i\theta})} g(z_a e^{i\theta}) e^{i\theta} d\theta, \quad \text{and} \quad I_1^- = \frac{z_a}{2\pi} \int_{\theta_0}^{\pi} e^{nh_a(z_a e^{i\theta})} g(z_a e^{i\theta}) e^{i\theta} d\theta.$$

First we need to consider $M(z)$, the probability generating function of nonempty language \mathcal{M} . Clearly, all coefficients of $M(z)$ are nonnegative, and $M(z)$ is aperiodic by Lemma B.1 in Appendix B. By the nonnegativity of coefficients and aperiodicity, the function $|M(z_a e^{i\theta})|$ is uniquely maximum at $\theta = 0$. It is also infinitely differentiable at $\theta = 0$. Consequently, there exists an angle $\theta_1 \in (0, \pi)$ such that

$$|M(z_a e^{i\theta})| \leq |M(z_a e^{i\theta_1})| \quad \text{for } \theta \in [\theta_1, \pi],$$

and $|M(z_a e^{i\theta})|$ is decreasing for $\theta \in [0, \theta_1]$. Thus, for large n ,

$$|M(z_a e^{i\theta})| \leq |M(z_a e^{i\theta_0})| \quad \text{for } \theta \in [\theta_0, \pi],$$

since $\theta_0 = n^{-2/5} < \theta_1$. Therefore, for $\theta \in [\theta_0, \pi]$,

$$|e^{nh_a(z_a e^{i\theta})}| = \frac{|M(z_a e^{i\theta})|^{na}}{z_a^n} \leq \frac{|M(z_a e^{i\theta_0})|^{na}}{z_a^n} = |e^{nh_a(z_a e^{i\theta_0})}|,$$

and this leads to

$$\begin{aligned} |I_1^+| &\leq \frac{z_a \cdot \max(g)}{2\pi} \int_{\theta_0}^{\pi} |e^{nh_a(z_a e^{i\theta})}| d\theta \leq \frac{z_a \cdot \max(g)}{2\pi} \int_{\theta_0}^{\pi} |e^{nh_a(z_a e^{i\theta_0})}| d\theta \\ &= \frac{z_a(\pi - \theta_0) \cdot \max(g)}{2\pi} \cdot \exp\left(nh_a(z_a) - \frac{\tau_a^2 z_a^2}{2} n^{\frac{1}{5}} + O(n^{-\frac{1}{5}})\right) = O\left(I_0 \cdot e^{-cn^{\frac{1}{5}}}\right), \end{aligned}$$

where $\max(g)$ is the maximum of $|g(z_a e^{i\theta})|$ for $\theta \in [\theta_0, \pi]$ and c is a positive constant. Similarly, $|I_1^-|$ is bounded. Thus,

$$[z^n][u^{na}]T(z, u) = I_0 + I_1 = I_0(1 + O(e^{-cn^{1/5}})).$$

Finally, by (15), (29), (30) and the preceding formula, we get,

$$\begin{aligned} P(O_n(D_n) = na) & \\ &= \frac{\rho B(\rho) g(z_a) e^{-nI(a)}}{\tau_a \sqrt{2\pi n}} \left[1 + \frac{1}{n} \left(\frac{3\beta_a g'(z_a)}{\tau_a g(z_a)} - \frac{g''(z_a)}{2\tau_a^2 g(z_a)} + 3\gamma_a - \frac{15\beta_a^2}{2} \right) + O\left(\frac{1}{n^2}\right) \right], \end{aligned} \quad (31)$$

where $I(a) = -\log \rho - h_a(z_a)$, which is positive. This establishes part (iii) of Theorem 2.2, where the constant c_2 can be extracted from the preceding equation.

APPENDIX

A. PROOF OF LEMMA 3.1

Here we present the proof of Lemma 3.1 discussed in Section 3.4.

We first describe the conditions on a . It is clear that $0 \leq a \leq 1$ since the number of occurrences cannot be greater than n , the length of a text. More precisely, a must satisfy one of the following conditions:

- (i) the pattern is not self-overlapping and $a < \frac{1}{m}$;
- (ii) the pattern is self-overlapping and $a < \frac{1}{r}$;

where m is the length of the pattern, and r is the length of the shortest nonempty word in the autocorrelation set S .

Now we show the existence of the real root z_a of the equation $h'_a(z) = 0$. By the definition of $h_a(z)$,

$$h'_a(z) = \frac{aM'(z)}{M(z)} - \frac{1}{z} = \frac{-D(z)^2 + (azB'(z) + 1 - B(z))D(z) + az(1 - B(z))D'(z)}{zD(z)(D(z) - 1 + B(z))}.$$

We notice that the denominator is always positive for $0 < z < \rho$.

Let us define $f_a(z)$, a function of a real variable z , as the numerator of $h'_a(z)$. That is, by the definition of $D(z)$ in (12),

$$f_a(z) = \{(1 - S(z))S(z) + azS'(z)\}(1 - B(z))^2 + z^m P(w)\{azB'(z) + (1 - 2S(z) + am)(1 - B(z)) - z^m P(w)\}.$$

We find that

$$f_a(\rho) = \rho^m P(w) (a\rho B'(\rho) - \rho^m P(w)) > 0, \quad (32)$$

since, for large n ,

$$a = (1 + \delta) \frac{\mathbf{E}[O_n(D_n)]}{n} = (1 + \delta) \frac{\rho^{m-1}}{B'(\rho)} P(w) \left(1 - O\left(\frac{1}{n}\right)\right) > \frac{\rho^{m-1}}{B'(\rho)} P(w).$$

Firstly, we consider when the pattern is not self-overlapping, that is, $S(z) \equiv 1$. Then,

$$f_a(z) = z^m P(w)\{azB'(z) + (am - 1)(1 - B(z)) - z^m P(w)\}.$$

The term of the smallest degree in $f_a(z)$ is $(am - 1)P(w)z^m$, and its coefficient is negative since $a < \frac{1}{m}$. Thus, there exists a sufficiently small $\epsilon > 0$ such that $f_a(\epsilon) < 0$. By this and (32), there exists a real root z_a of $h'_a(z) = 0$.

Secondly, we consider when the pattern is self-overlapping. That is,

$$S(z) = 1 + c_r z^r + (\text{higher order terms}), \quad (0 < r < m),$$

where r is the smallest degree among nonconstant terms and c_r is a positive constant. Then, the term of the smallest degree in $f_a(z)$ becomes $(ar - 1)c_r z^r$, and its coefficient is negative since $a < \frac{1}{r}$. Similarly to the first case, $f_a(\epsilon) < 0$ for sufficiently small $\epsilon > 0$, and we get the same result. The uniqueness comes from this result and part (ii) in the lemma because $h'_a(z)$ is continuous on $z \in [0, \rho]$.

Next, we prove part (ii) of Lemma 3.1. Let z_a be the real root of $h'_a(z) = 0$. Then, by definition,

$$h'_a(z_a) = \frac{aM'(z_a)}{M(z_a)} - \frac{1}{z_a} = 0. \quad (33)$$

On the other hand, we can write $M(z) = \sum_{i \geq 0} p_i z^i$ with $p_i \geq 0$ since $M(z)$ is the probability generating function of language \mathcal{M} . Then,

$$\frac{M'(z_a)}{M(z_a)} = \frac{\sum_{i \geq 0} i p_i z_a^{i-1}}{\sum_{i \geq 0} p_i z_a^i} = \frac{1}{z_a} \sum_{i \geq 0} i \frac{p_i z_a^i}{\sum_{j \geq 0} p_j z_a^j} = \frac{1}{z_a} \mathbf{E}[X], \quad (34)$$

where X is a random variable such that

$$Pr(X = i) = \frac{p_i z_a^i}{\sum_{j \geq 0} p_j z_a^j} \quad \text{for } i \geq 0.$$

From (33) and (34), we see that $\mathbf{E}[X] = \frac{1}{a}$. Now let us compute $h_a''(z_a)$. We have

$$\begin{aligned} h_a''(z_a) &= \frac{aM''(z)}{M(z)} - a \left(\frac{M'(z)}{M(z)} \right)^2 + \frac{1}{z_a^2} \\ &= \frac{a}{z_a^2} \mathbf{E}[X(X-1)] - a \left(\frac{\mathbf{E}[X]}{z_a} \right)^2 + \frac{1}{z_a^2} = \frac{a}{z_a^2} \mathbf{Var}[X]. \end{aligned}$$

Therefore, $h_a''(z_a) > 0$ because definitely the distribution is not concentrated at one value, hence $\mathbf{Var}[X] > 0$. This proves part (ii) of Lemma 3.1.

Finally, we know that $h_a(\rho) = -\log \rho$ and $h_a'(z) > 0$ for $z_a < z \leq \rho$. Therefore $h_a(z_a) < h_a(\rho) = -\log \rho$. This completes the proof of Lemma 3.1.

B. APERIODICITY OF $M(\mathbf{Z})$

The lemma that follows shows that $M(z)$, the probability generating function of language \mathcal{M} , is aperiodic. This is useful to derive the large deviations results in Section 3.4.

LEMMA B.1. *$M(z)$ is aperiodic if the length of the pattern w over the extended alphabet $\mathcal{B}_{d,k}$ is greater than 1.*

PROOF. To prove the aperiodicity of $M(z)$, we will show that there always exist two words in \mathcal{M} whose lengths differ by 1. Let $\mathcal{B}_{d,k} = \{\beta_d, \beta_{d+1}, \dots, \beta_k\}$ and ℓ be the length of w over $\mathcal{B}_{d,k}$ ($\ell \geq 2$). We consider the following two cases: (i) some supersymbol of $\mathcal{B}_{d,k}$ does not appear in w ; (ii) all symbols appear in w .

To prove the first case, let β_i be the symbol that does not appear in w . Then, clearly both $\beta_i \beta_d \beta_i w$ and $\beta_i \beta_{d+1} \beta_i w$ are in \mathcal{M} , and their difference in length is 1.

Now we prove the second case. We consider three subcases and find two words in \mathcal{M} , which differ by 1 in length for each subcase.

Case (i) $|\mathcal{B}_{d,k}| \geq 3$:

Let $u_1 = \underbrace{\beta_d \cdots \beta_d}_{\ell} \underbrace{\beta_d \cdots \beta_d}_{\ell} w$ and $u_2 = \underbrace{\beta_d \cdots \beta_d}_{\ell} \beta_{d+1} \underbrace{\beta_d \cdots \beta_d}_{\ell} w$. Then, w occurs in $w \cdot u_1$

only at the left and the right ends because the occurrence of w elsewhere implies that $w = \underbrace{\beta_d \cdots \beta_d}_{\ell}$, which contradicts the assumption that all symbols appear in w . Similarly,

w occurs ^{ℓ} in $w \cdot u_2$ only at the both ends. Otherwise, w must have only one or two kinds of symbols, which contradicts the assumption. Thus, both u_1 and u_2 are in \mathcal{M} .

Case (ii) $|\mathcal{B}_{d,k}| = 2$ and $\ell \geq 3$:

Let β_i be the symbol that appears more than once in w . Then, by the similar argument to the first case, $\underbrace{\beta_j \cdots \beta_j}_{\ell} \beta_i \underbrace{\beta_j \cdots \beta_j}_{\ell} w$ and $\underbrace{\beta_j \cdots \beta_j}_{\ell} \beta_j \underbrace{\beta_j \cdots \beta_j}_{\ell} w$ are in \mathcal{M} , and their lengths

differ by 1 because $|i - j| = 1$.

Case (iii) $|\mathcal{B}_{d,k}| = 2$ and $\ell = 2$:

In this case, w must be either $\beta_d\beta_k$ or $\beta_k\beta_d$. Thus, β_dw and β_kw are in \mathcal{M} . \square

REFERENCES

- AVIRAN, S., SIEGEL, P., AND WOLF, J. 2006. Optimal parsing trees for run-length coding of biased data. In *Proceedings of the IEEE International Symposium on Information Theory*. 1495–1499.
- BASSINO, F., CLEMENT, J., FAYOLLE, J., AND NICODEME, P. 2007. Counting occurrences for a finite set of words: An inclusion-exclusion approach. In *Proceedings of the Conference on Analysis of Algorithms*. 29–44.
- BENDER, E. A. AND KOCHMAN, F. 1993. The distribution of subword counts is usually normal. *Euro. J. Comb.* 14, 4, 265–275.
- CHOI, Y. AND SZPANKOWSKI, W. 2007. Pattern matching in constrained sequences. In *Proceedings of the IEEE International Symposium on Information Theory*. 2606–2610.
- DEMBO, A. AND KONTTOYIANNIS, I. 2002. Source coding, large deviations, and approximate pattern matching. *IEEE Trans. Inf. Theory* 48, 6, 1590–1615.
- FLAJOLET, P. AND SEDGEWICK, R. 2008. *Analytic Combinatorics*. Cambridge University Press.
- FLAJOLET, P., SZPANKOWSKI, W., AND VALLÉE, B. 2006. Hidden word statistics. *J. ACM* 53, 1, 147–183.
- GUIBAS, L. AND ODLYZKO, A. M. 1981. Periods in strings. *J. Combin. Theory* 30, 19–42.
- JACQUET, P. AND SZPANKOWSKI, W. 1994. Autocorrelation on words and its applications—Analysis of suffix trees by string-ruler approach. *J. Combin. Theory* 66, 237–269.
- JACQUET, P. AND SZPANKOWSKI, W. 2006. On (d, k) sequences not containing a given word. In *Proceedings of the IEEE International Symposium on Information Theory*. 1486–1489.
- JOLIVET, R., KOBAYASHI, R., RAUCH, A., NAUD, R., SHINOMOTO, S., AND GERSTNER, W. 2008. A benchmark test for a quantitative assessment of simple neuron models. *J. Neurosci. Methods* 169, 417–424.
- JOLIVET, R., RAUCH, A., LSCHER, H. R., AND GERSTNER, W. 2006. Predicting spike timing of neocortical pyramidal neurons by simple threshold models. *J. Comput. Neurosci.* 21, 35–49.
- KOLESNIK, V. D. AND KRACHKOVSKY, V. Y. 1991. Generating functions and lower bounds on rates for limited error-correcting codes. *IEEE Trans. Inf. Theory* 37, 3, 778–788.
- LOTHAIRE, M. 2005. *Applied Combinatorics on Words*. Cambridge University Press, New York.
- MARCUS, B., ROTH, R., AND SIEGEL, P. 1988. Constrained systems and coding for recording channels. In *Handbook of Coding Theory*, V. Pless and W.C. Huffman, Eds., Elsevier Science, Chapter 20.
- NICODÈME, P., SALVY, B., AND FLAJOLET, P. 1999. Motif statistics. In *Proceedings of the European Symposium on Algorithms*. Lecture Notes in Computer Science, vol. 1643. Springer, 194–211.
- PANINSKI, L. 2003. Estimation of entropy and mutual information. *Neural Comp.* 15, 6, 1191–1253.
- RÉGNIER, M. AND DENISE, A. 2004. Rare events and conditional events on random strings. *Discr. Math. Theor. Comput. Sci.* 6, 2, 191–214.
- RÉGNIER, M. AND SZPANKOWSKI, W. 1997. On the approximate pattern occurrences in a text. In *Compression and Complexity of SEQUENCES*. IEEE Computer Society. 253–264.
- RÉGNIER, M. AND SZPANKOWSKI, W. 1998. On pattern frequency occurrences in a Markovian sequence. *Algorithmica* 22, 631–649.
- SZPANKOWSKI, W. 2001. *Average Case Analysis of Algorithms on Sequences*. John Wiley & Sons, New York.
- WATERMAN, M. 1995. *Introduction to Computational Biology*. Chapman & Hall, London.
- ZEHAVI, E. AND WOLF, J. K. 1988. On runlength codes. *IEEE Trans. Inf. Theory* 34, 1, 45–54.

Received April 2008; revised February 2009; accepted March 2009