Constrained Pattern Matching*

Yongwook Choi and Wojciech Szpankowski
Department of Computer Science
Purdue University
W. Lafayette, IN 47907 U.S.A.
Email: ywchoi@purdue.edu, spa@cs.purdue.edu

April 29, 2008

Abstract

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 paper, 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 over-represented patterns in neuronal data (spike trains), which satisfy structural constraints that match the framework of (d, k) binary sequences. Throughout this paper we use techniques of analytic algorithmics 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 Terms: Pattern matching, constrained sequences, languages, autocorrelation polynomials, neuronal spike trains, complex asymptotics

*The work was supported in part by the NSF Grants CCF-0513636, and DMS-0503742, the NIH Grant R01 GM068959-01, AFOSR Grant 073071, and NSA Grant 07G-044.

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 paper, we restrict our goal to study and understand some aspects of pattern matching in constrained sequences. Although our methods work for a large class of constrained systems, we further restrict our analysis to the so-called (d, k) sequences in which runs of zeros cannot be smaller than d nor bigger than k, where $0 \leq d < k$. Such sequences have proved to be very 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 precisely 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 each two consecutive ones (corresponding to firing times).

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 [14, 20, 22]. We study this problem in a probabilistic framework, that is, we assume that a sequence is generated by a (biased) 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 [2, 8, 14, 17, 22, 23]. For instance, Guibas and Odlyzko [8] (cf. also [10, 20]) revealed the fundamental role played by autocorrelation languages and their associated polynomials in the analysis of pattern matching. Régnier and Szpankowski [19, 20] established that the number of occurrences of a given pattern is asymptotically normal under a diversity of probabilistic models that include Markov chains. Nicodème, Salvy, and Flajolet [17] 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 [2] studied occurrences of a generalized pattern using (in a nutshell) the de Bruijn graph representation that allowed the authors to establish the central limit theorem, but without explicit mean and variance. Recent surveys on pattern matching can be found in Lothaire [14] (Chaps. 6 and 7). To the best of our knowledge, none of these works deal 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 [4, 13, 15, 24]. Pattern matching in constrained sequences can in principle be analyzed by various versions of the de Bruijn graph [2, 7] or automaton approach [2, 17]. This is an elegant and general approach but it sometimes leads to complicated analyses and is computationally extensive. In our constrained pattern matching, for example, 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 a matrix representation of the underlying de Bruijn graph and hence its largest eigenvalue (cf. [2, 7]). 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 paper, we take the view of combinatorics on words. 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 asymptotic formulas for the probability that there are r occurrences of w in a (d,k) sequence generated by a binary memoryless source. 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 paper 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 till 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. 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

$$\mathcal{A}_{d,k} = \{\underbrace{0\ldots0}_{d},\cdots,\underbrace{0\ldots0}_{k}\}$$

as a set of runs of zeros of length between d and k. The *extended alphabet* is then [15]

$$\mathcal{B}_{d,k} = \mathcal{A}_{d,k}\cdot\{1\} = \{\underbrace{0\ldots0}_{d}1,\cdots,\underbrace{0\ldots0}_{k}1\}.$$

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 only two symbols over $\mathcal{B}_{d,k}$, namely 001 and 0001. We shall also relax this assumption later.

As in the classical pattern matching, a special set (language) plays an important role. We define it next. Let $w = w_1 \dots w_m \in \{0,1\}^m$ with $w_1 = 0$ and $w_m = 1$. 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,

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

where $\beta_i^j = \beta_i \cdots \beta_j$ and $\beta_i^j = \epsilon$ if $i > j$.

As in [3, 11, 20], 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 $r \ge 1$ [20, 22]

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

$$\mathcal{T}_0 \cdot \{w\} = \mathcal{R} \cdot \mathcal{S}. \tag{2}$$

In order to find relationships between the languages $\mathcal{R}$, $\mathcal{M}$, and $\mathcal{U}$, we extend the approach from [20] to yield

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

$$\mathcal{U} \cdot \mathcal{B} = \mathcal{M} + \mathcal{U} - \{\epsilon\}, \tag{4}$$

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

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 $q = 1 - p$. Among others, we compute the probability that a randomly generated sequence is a (d, k) sequence. We actually derive the conditional probability distribution of the number of occurrences of w in 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:

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

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

$$R(z) = z^m P(w) \cdot U(z), \tag{8}$$

where $P(w)$ is the probability of w, and

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

In particular, from (1),(2) and above, one finds

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

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

where

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

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

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

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,

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

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

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

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

$$\mathbf{E}[O_n(\mathcal{D}_n)(O_n(\mathcal{D}_n)-1)] = \frac{[z^n]T_{uu}(z,1)}{[z^n]T(z,1)} \tag{15}$$

is the second factorial moment, where $T_{uu}(z,1)$ is the second derivative with respect to u at $u=1$.

2.1 Analytical Results

Our main analytical results are summarized in the following two theorems. In Theorem 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 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 asymptotically*

$$P(\mathcal{D}_n) = \frac{1}{B'(\rho)}\lambda^{n+1} + O(\omega^n) \tag{16}$$

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 becomes

$$\mathbf{Var}[O_n(\mathcal{D}_n)] = (n-m+1)P(w)\left[\frac{(1-2m)P(w)}{B'(\rho)^2}\lambda^{-2m+2}\right.$$
$$\left. + \frac{P(w)B''(\rho)}{B'(\rho)^3}\lambda^{-2m+1} + \frac{2S(\rho)-1}{B'(\rho)}\lambda^{-m+1}\right] + O(1) \tag{17}$$

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 *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 the unique positive real root of $B(z) = 1$. Then, for large n, the followings hold:*

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

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

(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.

(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,

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

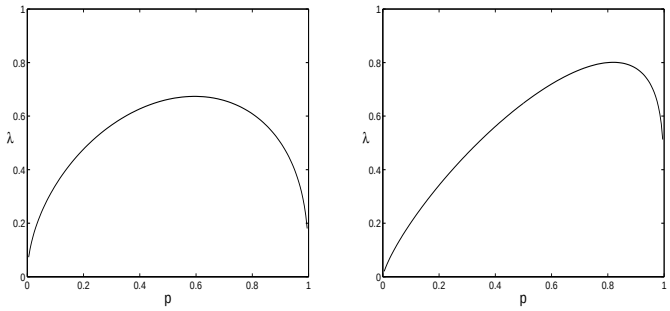
and

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

where $I(a) = -\log\rho - h_a(z_a)$, which is positive, and

$$c_1 = \frac{\rho B'(\rho)g(z_a)}{\tau_a}$$

with $g(z) = \frac{P(w)z^{m-1}}{D(z)^2M(z)}$ and $\tau_a^2 = h_a''(z_a)$. The constant c_2 is explicitly computed in (32).



(a) (1,2) sequences (b) (3,6) sequences

Figure 1: λ versus p.

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 [1].

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 $[z^n]T_r(2z)$. 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. In particular, Shannon entropy is

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

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

Remark 3. 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

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

and one can easily derive generating functions and asymptotic expressions from the above.

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 a $(1, 4)$ sequence, 0010001). Again, let $w = w_1 \dots w_m \in \{0, 1\}^m$ with $w_1 = 0$ and $w_m = 1$, 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

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

Using the languages $\mathcal{T}_r$, $\mathcal{R}$, $\mathcal{M}$, and $\mathcal{U}$ defined above, 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 $\mathcal{Z} = \{\epsilon, 0, 00, \cdots, 0^{k+1-|\beta_1|}\}$, and 0^k denotes a run of zeros of length k. Applying the same techniques as above we can derive the generating functions and asymptotic results.

Remark 5. We can also extend our results to a set of patterns $\mathcal{W} = \{w_1, w_2, \ldots, w_K\}$ such that w_i $(1 \le i \le K)$ is not a substring of another pattern w_j $(1 \le j \le K, i \ne j)$ over alphabet $\mathcal{B}_{d,k}$. Let us introduce some new languages. In particular, for any given two strings u and v, let

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

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

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

For $1 \le i, j \le K$, 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 $r \ge 1$,

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

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 $\mathcal{T}_r (r \ge 1)$ and $\mathcal{T}_0$ are represented as follows:

$$\mathcal{T}_r \ = \ \sum_{1 \le i,j \le K} \mathcal{R}_i \cdot \mathcal{M}_{ij}^{[r-1]} \cdot \mathcal{U}_j, \tag{18}$$

$$\mathcal{T}_0 \cdot \{w_j\} \ = \ \mathcal{R}_j + \sum_{1 \le i \le K} \mathcal{R}_i \cdot (\mathcal{S}_{ij} - \{\epsilon\}) \tag{19}$$

for $1 \le j \le K$. The languages $\mathcal{M}_{ij}$, $\mathcal{U}_i$, and $\mathcal{R}_j$ satisfy following relationships [19] for $1 \le i, j \le K$.

$$\sum_{k \ge 1} \mathcal{M}_{ij}^{[k]} \ = \ \mathcal{B}^* \cdot \{w_j\} + \mathcal{S}_{ij} - \{\epsilon\}, \tag{20}$$

$$\mathcal{U}_i \cdot \mathcal{B} \ = \ \sum_{1 \le j \le K} \mathcal{M}_{ij} + \mathcal{U}_i - \{\epsilon\}, \tag{21}$$

$$\mathcal{B} \cdot \mathcal{R}_j - (\mathcal{R}_j - \{w_j\}) \ = \ \sum_{1 \le i \le K} \{w_i\} \cdot \mathcal{M}_{ij}. \tag{22}$$

As before, the language relationships (20)–(22) are translated into generating functions [19]:

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

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 $\overrightarrow{W}(z)$, $\overrightarrow{R}(z)$, $\overrightarrow{U}(z)$, and $\overrightarrow{1}$ are column vectors of length K such that

$$\overrightarrow{W}(z)=(z^{|w_1|}P(w_1),\cdots,z^{|w_K|}P(w_K))^t,\quad \overrightarrow{R}(z)=(R_1(z),\cdots,R_K(z))^t,$$

$$\overrightarrow{U}(z)=(U_1(z),\cdots,U_K(z))^t,\text{ and }\overrightarrow{1}=(1,\cdots,1)^t.$$

¿From (18)–(19) and above, one finds

$$T_0(z)=\frac{\overrightarrow{R}^t(z)\cdot\mathbf{S}(z)\cdot\overrightarrow{1}}{\overrightarrow{W}^t(z)\cdot\overrightarrow{1}},$$

$$T_r(z)=\overrightarrow{W}^t(z)\cdot(\mathbf{D}(z)+(B(z)-1)\mathbf{I})^{r-1}\cdot\mathbf{D}(z)^{-(r+1)}\cdot\overrightarrow{1},$$

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

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

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

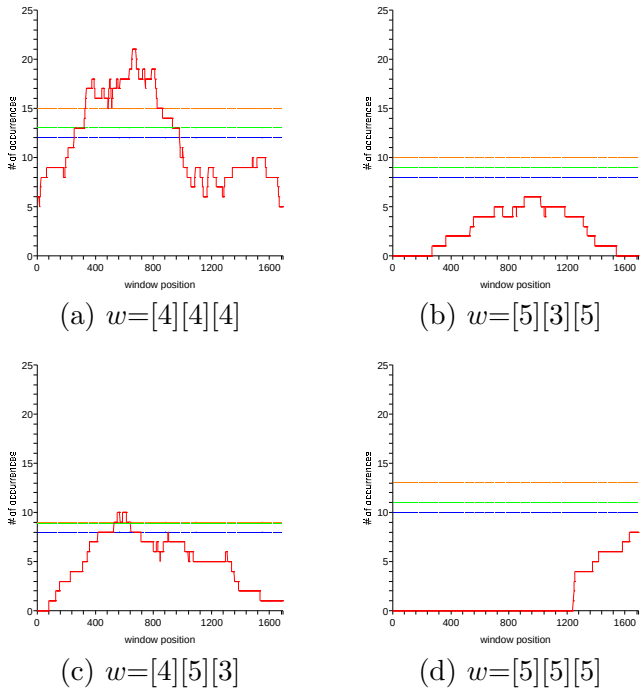
and

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

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

We now apply our theoretical results to statistical inference of some biological data. As a potential application of our main results, we use part (iii) of Theorem 2 to detect under- and over-represented structures in neuronal data (spike trains), and to obtain justifiably accurate statistical inferences about their biological properties and functions. 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



(a) w=[4][4][4] (b) w=[5][3][5]

(c) w=[4][5][3] (d) w=[5][5][5]

Figure 2: Number of occurrences of w within a window of size 500; here $[i]$ stands for the pattern $0\cdots01$ with $i-1$ zeros.

and describe the amount and representation of the information that is contained in this data [18]. 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 above.

For experiments, we use single-electrode data from cortical neurons under random current injection. The details can be found in [12, 25]. 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 paper.

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 we easily conclude

that for α_{th} in the range of the large deviations domain, the threshold is $O_{th} = na_{th}$, where

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

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

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 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 $B(z)$ has a variable p in each of its coefficients. 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 q. 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 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 window of size 500; here we use a short-hand notation $[i]$ for a pattern $\underbrace{0\cdots0}_{i-1}1$. 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 1 and 2 of previous section. In Section 3.1, we asymptotically derive $P(\mathcal{D}_n)$ and the first two moments of $O_n(\mathcal{D}_n)$. In Sections 3.2–3.4, using our expression (13) for the bivariate generating function we estimate asymptotically $P(O_n(\mathcal{D}_n) = r)$ for various ranges of r.

3.1 Moments

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

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

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 [22] 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 $\omega < \lambda$. In Lemma 2 of Appendix A we 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.

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

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 (17).

3.2 Distribution for $r = O(1)$

We prove here part (i) of Theorem 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 $a_{r+1} = \dfrac{P(w)(B(\tau)-1)^{r-1}}{D'(\tau)^{r+1}}$.

By Lemma 3 of Appendix B, we know 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(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. 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

$$T_n(u) = \mathbf{E}[u^{O_n(D_n)}] = \frac{[z^n]T(z,u)}{[z^n]T(z,1)}, \tag{23}$$

and

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

By Goncharov's theorem [22], it suffices to prove the following

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

for all $\nu = it'$ where $-\infty < t' < \infty$. But we prove more generally for all complex ν.

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

$$[z^n]T(z,u) = c(u)\lambda^{n+1}(u) + O(\omega^n(u)), \tag{24}$$

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

and $\lambda(u) = 1/\rho(u)$ where $|\omega(u)| < \lambda(u)$. Thus, by (23),

$$P(\mathcal{D}_n)T_n(u) = c(u)\lambda^{n+1}(u) + O(\omega^n(u)) \tag{25}$$

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

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

$$\lambda(e^t) = \lambda(1) + \lambda'(1)t + \frac{\lambda''(1) + \lambda'(1)}{2}t^2 + O(t^3). \tag{26}$$

Now let us find $\lambda'(1)$ and $\lambda''(1)$. From (14) and (24), we observe that

$$\mu_n[z^n]T(z,1) = [z^n]T_u(z,1), \tag{27}$$

and

$$\begin{aligned} [z^n]T_u(z,1) &= [z^n]\left.\frac{\partial T(z,u)}{\partial u}\right|_{u=1} \\ &= (n+1)c(1)\lambda^n(1)\lambda'(1) + c'(1)\lambda^{n+1}(1) + O(n\omega^n(1)). \end{aligned} \tag{28}$$

By (24),(27), and (28), we obtain

$$\mu_n\left(c(1)\lambda^{n+1}(1) + O(\omega^n(1))\right) = (n+1)c(1)\lambda^n(1)\lambda'(1) + c'(1)\lambda^{n+1}(1) + O(n\omega^n(1)).$$

Thus, we get

$$\lambda'(1) = \frac{\mu_n}{n+1}\lambda(1) - \frac{c'(1)}{(n+1)c(1)}\lambda(1) + O\left(n\xi^n(1)\right) \tag{29}$$

where $\xi(u) = \omega(u)/\lambda(u)$. We note that $|\xi(u)| < 1$. Similarly,

$$[z^n]T_{uu}(z,1) = n(n+1)c(1)\lambda^{n-1}(1)\lambda'^2(1) + 2(n+1)c'(1)\lambda^n(1)\lambda'(1)$$
$$+(n+1)c(1)\lambda^n(1)\lambda''(1) + c''(1)\lambda^{n+1}(1) + O(n^2\omega^n(1)).$$

Again, from (15) and (24), we observe that

$$(\sigma_n^2 + \mu_n^2 - \mu_n)[z^n]T(z,1) = [z^n]T_{uu}(z,1),$$

and from (29), after some algebra, we finally arrive at

$$\lambda''(1) = \left(\frac{\sigma_n^2 - \mu_n}{n+1} + \frac{\mu_n^2}{(n+1)^2}\right)\lambda(1) + O\left(\frac{1}{n}\right). \tag{30}$$

Using (25),(26),(29), and (30) we get

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

In the last equality we use the fact that $t = O\left(\frac{1}{\sqrt{n}}\right)$. Therefore,

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

and

$$\begin{aligned}
\lim_{n\to\infty} e^{-\nu\mu_n/\sigma_n}T_n(e^{\nu/\sigma_n}) &= \lim_{n\to\infty}\frac{e^{-\nu\mu_n/\sigma_n}P(\mathcal{D}_n)T_n(e^{\nu/\sigma_n})}{P(\mathcal{D}_n)} \\
&= \lim_{n\to\infty}\left(1 + \frac{\sigma_n^2}{2(n+1)}t^2 + O(t^3)\right)^{n+1} \\
&= \lim_{n\to\infty}\left(1 + \frac{\nu^2}{2(n+1)} + O(t^3)\right)^{n+1}.
\end{aligned}$$

We find upper and lower bounds as follows:

$$\begin{aligned}
\lim_{n\to\infty}\left(1+\frac{\nu^2}{2(n+1)}+O(t^3)\right)^{n+1} &= \lim_{n\to\infty}\exp\left((n+1)\ln\left(1+\frac{\nu^2}{2(n+1)}+O(t^3)\right)\right)\\
&\le \lim_{n\to\infty}\exp\left((n+1)\left(\frac{\nu^2}{2(n+1)}+O(t^3)\right)\right)\\
&= \lim_{n\to\infty}\exp\left(\frac{\nu^2}{2}+O\left(\frac{1}{\sqrt{n}}\right)\right)\\
&= \exp\left(\frac{\nu^2}{2}\right),\\
\lim_{n\to\infty}\left(1+\frac{\nu^2}{2(n+1)}+O(t^3)\right)^{n+1} &\ge \lim_{n\to\infty}\left(1+\frac{\nu^2}{2(n+1)}\right)^{n+1}\\
&= \exp\left(\frac{\nu^2}{2}\right).
\end{aligned}$$

Therefore,

$$\lim_{n\to\infty} e^{-\nu\mu_n/\sigma_n}T_n(e^{\nu/\sigma_n}) = \exp\left(\frac{\nu^2}{2}\right),$$

as desired to establish Theorem 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, 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 $P(O_n(D_n) = na)$ asymptotically when na is an integer. Clearly,

$$P(O_n(D_n) = na) = [u^{na}]T_n(u) = \frac{[z^n][u^{na}]T(z,u)}{[z^n]T(z,1)}. \tag{31}$$

By (13),

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

Hence, Cauchy's coefficient formula leads to [22]

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

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* [22]. 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

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

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

Lemma 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 C.*

(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 $\mathcal{C}_0 = \{z \in \mathcal{C} : |arg(z)| \le \theta_0\}$ and $\mathcal{C}_1 = \{z \in \mathcal{C} : |arg(z)| \ge \theta_0\}$ for some θ_0. That is,

$$[z^n][u^{na}]T(z,u) = \frac{1}{2\pi i}\int_{\mathcal{C}_0} e^{nh_a(z)}g(z)dz + \frac{1}{2\pi i}\int_{\mathcal{C}_1} e^{nh_a(z)}g(z)dz.$$

Let

$$I_0 = \frac{1}{2\pi i}\int_{\mathcal{C}_0} e^{nh_a(z)}g(z)dz$$

and

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

We will 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 $z = z_a e^{i\theta}$,

$$\begin{aligned} I_0 &= \frac{1}{2\pi}\int_{-\theta_0}^{+\theta_0} e^{nh_a(z_a e^{i\theta})}g(z_a e^{i\theta})z_a e^{i\theta}d\theta \\ &= \frac{z_a}{2\pi}\int_{-\theta_0}^{+\theta_0} \exp(nh_a(z_a e^{i\theta}) + i\theta)g(z_a e^{i\theta})d\theta. \end{aligned}$$

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

$$\tau_a^2 = h_a''(z_a) \quad \text{(cf. part (ii) of Lemma 1)},$$

$$\beta_a = \frac{h_a^{(3)}(z_a)}{3!\tau_a^3}, \text{ and } \gamma_a = \frac{h_a^{(4)}(z_a)}{4!\tau_a^4}.$$

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

$$\begin{aligned} h_a(z_a e^{i\theta}) &= h_a(z_a) - \frac{\tau_a^2 z_a^2}{2}\theta^2 - \left(\beta_a\tau_a^3 z_a^3 + \frac{\tau_a^2 z_a^2}{2}\right)i\theta^3 \\ &\quad + \left(\gamma_a\tau_a^4 z_a^4 + \frac{3}{2}\beta_a\tau_a^3 z_a^3 + \frac{7}{24}\tau_a^2 z_a^2\right)\theta^4 + O(\theta^5) \qquad (\because h_a'(z_a) = 0). \end{aligned}$$

Similarly,

$$g(z_a e^{i\theta}) = g(z_a) + g'(z_a) z_a i\theta - \frac{g''(z_a) z_a^2 + g'(z_a) z_a}{2}\theta^2 + O(\theta^3).$$

When $|\theta| \le \theta_0$, $n\theta^k \to 0$ $(k \ge 3)$ as $n \to \infty$. Thus,

$$\begin{aligned} e^{nh_a(z_a e^{i\theta}) + i\theta} &= \exp\left(nh_a(z_a) - \frac{\tau_a^2 z_a^2}{2} n\theta^2 + \alpha(\theta) \right) \\ &= \exp\left(nh_a(z_a) - \frac{\tau_a^2 z_a^2}{2} n\theta^2 \right) \left(1 + \alpha(\theta) + \frac{\alpha(\theta)^2}{2!} + \frac{\alpha(\theta)^3}{3!} + \cdots \right) \end{aligned}$$

where

$$\alpha(\theta) = i\theta - \left(\beta_a \tau_a^3 z_a^3 + \frac{\tau_a^2 z_a^2}{2} \right) in\theta^3 + \left(\gamma_a \tau_a^4 z_a^4 + \frac{3}{2}\beta_a \tau_a^3 z_a^3 + \frac{7}{24}\tau_a^2 z_a^2 \right) n\theta^4 + O(n\theta^5).$$

Therefore we have

$$\begin{aligned} I_0 &= \frac{z_a}{2\pi} \int_{-\theta_0}^{+\theta_0} \exp(nh_a(z_a e^{i\theta}) + i\theta) g(z_a e^{i\theta}) d\theta \\ &= \frac{z_a e^{nh_a(z_a)}}{2\pi} \int_{-\theta_0}^{+\theta_0} \exp\left(-n\frac{\tau_a^2 z_a^2}{2}\theta^2 \right) \left(1 + \alpha(\theta) + \frac{\alpha(\theta)^2}{2!} + \frac{\alpha(\theta)^3}{3!} + \cdots \right) g(z_a e^{i\theta}) d\theta. \end{aligned}$$

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

$$\begin{aligned} \alpha(\theta) = \eta(\omega) &= \frac{i\omega}{\tau_a z_a \sqrt{n}} - \left(\beta_a + \frac{1}{2\tau_a z_a} \right) \frac{i\omega^3}{\sqrt{n}} + \left(\gamma_a + \frac{3}{2}\frac{\beta_a}{\tau_a z_a} + \frac{7}{24}\frac{1}{\tau_a^2 z_a^2} \right) \frac{\omega^4}{n} + O\left(\frac{\omega^5}{n\sqrt{n}} \right), \\ g\left(z_a e^{\frac{i\omega}{\tau_a z_a \sqrt{n}}} \right) &= g(z_a) + \frac{g'(z_a)}{\tau_a} \frac{i\omega}{\sqrt{n}} - \left(\frac{g''(z_a)}{2\tau_a^2} + \frac{g'(z_a)}{2\tau_a^2 z_a} \right) \frac{\omega^2}{n} + O\left(\frac{\omega^3}{n\sqrt{n}} \right), \end{aligned}$$

and

$$I_0 = \frac{e^{nh_a(z_a)}}{2\pi\tau_a\sqrt{n}} \int_{-\omega_o}^{+\omega_o} \exp\left(-\frac{\omega^2}{2} \right) \left(1 + \eta(\omega) + \frac{\eta(\omega)^2}{2!} + \frac{\eta(\omega)^3}{3!} + \cdots \right) g\left(z_a e^{\frac{i\omega}{\tau_a z_a \sqrt{n}}} \right) d\omega$$

where $\omega_o = \tau_a z_a n^{\frac{1}{10}}$.

Each term of odd degree of ω in

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

contributes nothing to the integral. Thus

$$\begin{aligned} I_0 &= \frac{e^{nh_a(z_a)}}{2\pi\tau_a\sqrt{n}} \int_{-\omega_o}^{+\omega_o} \exp\left(-\frac{\omega^2}{2} \right) \left(A + B\omega^2 + C\omega^4 + D\omega^6 + O\left(\frac{1}{n^2} \right) \right) d\omega \\ &= \frac{e^{nh_a(z_a)}}{2\pi\tau_a\sqrt{n}} \left[\int_{-\infty}^{+\infty} \exp\left(-\frac{\omega^2}{2} \right) \left(A + B\omega^2 + C\omega^4 + D\omega^6 + O\left(\frac{1}{n^2} \right) \right) d\omega + O\left(e^{-\frac{1}{2}\omega_o^2} \right) \right] \end{aligned}$$

where

$$A = g(z_a), \quad B = -\frac{1}{n}\left(\frac{g''(z_a)}{2\tau_a^2} + \frac{3g'(z_a)}{2\tau_a^2 z_a} + \frac{g(z_a)}{2\tau_a^2 z_a^2}\right),$$

$$C = \frac{1}{n}\left(g'(z_a)\left(\frac{\beta_a}{\tau_a} + \frac{1}{2\tau_a^2 z_a}\right) + g(z_a)\left(\gamma_a + \frac{5\beta_a}{2\tau_a z_a} + \frac{19}{24\tau_a^2 z_a^2}\right)\right),$$

and

$$D = -\frac{g(z_a)}{2n}\left(\beta_a + \frac{1}{2\tau_a z_a}\right)^2.$$

Using the fact that

$$\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} x^{2k} = \frac{\Gamma(2k)}{2^{k-1}\Gamma(k)}\sqrt{2\pi},$$

we finally obtain

$$\begin{aligned} I_0 &= \frac{e^{nh_a(z_a)}}{\tau_a\sqrt{2\pi n}}\left(A + B + 3C + 15D + O\left(\frac{1}{n^2}\right) + O\left(e^{-\frac{\tau_a^2 z_a^2}{2}n^{1/5}}\right)\right) \\ &= \frac{g(z_a)e^{nh_a(z_a)}}{\tau_a\sqrt{2\pi n}}\left[1 + \frac{1}{n}\left(\frac{3\beta_a g'(z_a)}{\tau_a g(z_a)} - \frac{g''(z_a)}{2\tau_a^2 g(z_a)} + 3\gamma_a - \frac{15\beta_a^2}{2}\right) + O\left(\frac{1}{n^2}\right)\right]. \end{aligned}$$

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. For this, we need to first consider $M(z)$, the probability generating function of non-empty language $\mathcal{M}$. Clearly, all coefficients of $M(z)$ are non-negative, and $M(z)$ is aperiodic by Lemma 4 in Appendix D. By the non-negativity 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

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

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

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

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

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

and this leads to

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

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,

$$\frac{1}{2\pi}\left|\int_{-\pi}^{-\theta_0} e^{nh_a(z_a e^{i\theta})} g(z_a e^{i\theta}) z_a e^{i\theta} d\theta\right| = O\left(I_0 \cdot e^{-cn^{1/5}}\right).$$

Thus,

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

that is, $|I_1|$ is exponentially smaller than I_0.

Putting everything together, we obtain

$$\begin{aligned} [z^n][u^{na}]T(z,u) &= I_0 + I_1 = I_0\left(1 + O\left(e^{-cn^{1/5}}\right)\right) \\ &= \frac{g(z_a)e^{nh_a(z_a)}}{\tau_a\sqrt{2\pi n}}\left[1 + \frac{1}{n}\left(\frac{3\beta_a g'(z_a)}{\tau_a g(z_a)} - \frac{g''(z_a)}{2\tau_a^2 g(z_a)} + 3\gamma_a - \frac{15\beta_a^2}{2}\right) + O\left(\frac{1}{n^2}\right)\right]. \end{aligned}$$

Finally, we are ready to compute $P(O_n(D_n) = na)$. By (16),(31), and the above,

$$\begin{aligned} &P(O_n(D_n) = na) = \frac{[z^n][u^{na}]T(z,u)}{[z^n]T(z,1)} \\ &= \frac{\rho B'(\rho) g(z_a) e^{-nI(a)}}{\tau_a\sqrt{2\pi n}}\left[1 + \frac{1}{n}\left(\frac{3\beta_a g'(z_a)}{\tau_a g(z_a)} - \frac{g''(z_a)}{2\tau_a^2 g(z_a)} + 3\gamma_a - \frac{15\beta_a^2}{2}\right) + O\left(\frac{1}{n^2}\right)\right] \quad (32) \end{aligned}$$

where $I(a) = -\log\rho - h_a(z_a)$, which is positive. This establishes part (iii) of Theorem 2, where the constant c_2 can be extracted from the above.

Appendix

A The Root of $B(z) = 1$

The lemma below shows that a unique positive real root of the equation $B(z) = 1$ has the smallest modulus among all complex roots, which is needed to derive the asymptotic formula for $P(\mathcal{D}_n)$ and the moments of $O_n(\mathcal{D}_n)$ in Section 3.1.

Lemma 2 *The equation $B(z) = 1$ has one positive real root ρ that is greater than 1. All other roots ρ' satisfy $|\rho'| > \rho$.*

Proof: By definition, $B(z) := p^d q z^{d+1} + p^{d+1} q z^{d+2} + \cdots + p^k q z^{k+1}$. Let $f(z) = 1 - B(z)$. Then, we observe that $f(1) = 1 - B(1) > 0$ and $\lim_{z\to\infty} f(z) = -\infty$. We also see that $f'(z) = -B'(z) < 0$ for $z > 0$, that is, $f(z)$ is a decreasing function. Therefore, $f(z) = 0$ has one real root on $(1, \infty)$.

Let ρ be the real root, and let $h(z) = 1$ and $g(z) = -B(z)$. Now let's consider a closed contour $C = \{z : |z| = \rho - \epsilon\}$ where ϵ is an arbitrarily small positive constant. At points

on C we have

$$\begin{aligned}
|g(z)| &\leq p^d q|z|^{d+1} + p^{d+1}q|z|^{d+2} + \cdots + p^k q|z|^{k+1} \\
&= p^d q(\rho-\epsilon)^{d+1} + p^{d+1}q(\rho-\epsilon)^{d+2} + \cdots + p^k q(\rho-\epsilon)^{k+1} \\
&< p^d q\rho^{d+1} + p^{d+1}q\rho^{d+2} + \cdots + p^k q\rho^{k+1} \\
&= 1 \\
&= |h(z)|.
\end{aligned}$$

Thus, by Rouché's theorem [9] $f(z)$ and $h(z)$ have the same number of zeros inside C, that is, $f(z)$ has no root inside C. Therefore, All other complex roots ρ' satisfy $|\rho'| \geq \rho$.

Suppose that another complex root ρ' satisfies $|\rho'| = \rho$, that is, $\rho' = \rho e^{i\theta}$ for some θ. Then

$$\begin{aligned}
|1| &= |B(\rho')| \\
&= |p^d q\rho^{d+1}e^{i(d+1)\theta} + p^{d+1}q\rho^{d+2}e^{i(d+2)\theta} + \cdots + p^k q\rho^{k+1}e^{i(k+1)\theta}| \\
&\leq p^d q\rho^{d+1}|e^{i(d+1)\theta}| + p^{d+1}q\rho^{d+2}|e^{i(d+2)\theta}| + \cdots + p^k q\rho^{k+1}|e^{i(k+1)\theta}| \\
&= p^d q\rho^{d+1} + p^{d+1}q\rho^{d+2} + \cdots + p^k q\rho^{k+1} \\
&= 1.
\end{aligned}$$

But, in the third line, the equality holds only when $\theta = 2\pi j$ for some integer j. Thus ρ' must be a real root, which is ρ. Therefore, All other roots ρ' satisfy $|\rho'| > \rho$. ∎

B The Root of $D(z) = 0$

The lemma below shows the existence of the positive real root of the equation $D(z) = 0$ which is needed in the proof of Theorem 2(i).

Lemma 3 *The equation $D(z) = 0$ has at least one positive real root τ, which is greater than ρ.*

Proof: For $0 \leq z \leq \rho$, we observe that $D(z) := S(z)(1 - B(z)) + z^m P(w) > 0$. It follows from the fact that $S(z) > 0$ and $1 - B(z) > 0$ (for $0 < z < \rho$), $D(0) = 1$, and $D(\rho) = \rho^m P(w)$.

Now let us consider when $z > \rho$. Let m' be the length of the pattern over the extended alphabet $\mathcal{B}_{d,k}$. Notice that $d + 1 \leq \frac{m}{m'} \leq k + 1$. Firstly, let us assume $m' = 1$, that is, $d + 1 \leq m \leq k + 1$. Then, since $S(z) = 1$ and $P(w) = p^{m-1}q$,

$$D(z) = 1 - B(z) + p^{m-1}qz^m.$$

Thus, $D(z) \to -\infty$ as $z \to \infty$ because $B(z)$ has at least two terms, and one term in $B(z)$ cancels out $p^{m-1}qz^m$. Therefore, there exists at least one real root on (ρ, ∞).

Now we can assume that $m' \geq 2$, and first consider when $\frac{m}{m'}$ is either $d+1$ or $k+1$, that is, the pattern is periodic. If $\frac{m}{m'} = d+1$, then

$$\begin{aligned} D(z) &= S(z)(1-B(z)) + P(w)z^m \\ &= S(z)(1-p^d q z^{d+1}) + P(w)z^m - S(z)(B(z) - p^d q z^{d+1}) \\ &= \left(1 + p^d q z^{d+1} + \left(p^d q z^{d+1}\right)^2 + \cdots + \left(p^d q z^{d+1}\right)^{m'-1}\right)(1 - p^d q z^{d+1}) + \left(p^d q z^{d+1}\right)^{m'} \\ &\quad -S(z)(B(z) - p^d q z^{d+1}) \\ &= 1 - S(z)(B(z) - p^d q z^{d+1}). \end{aligned}$$

Thus, again $D(z) \to -\infty$ as $z \to \infty$. The same is true when $\frac{m}{m'} = k+1$. Therefore, there exists at least one real root.

Next, we consider when $d+1 < \frac{m}{m'} < k+1$, and we show that $D(z_o) \leq 0$ for some positive z_o. Let us define two integers, ℓ and u. Let ℓ be the largest integer less than $\frac{m}{m'}$. Similarly, let u be the smallest integer larger than $\frac{m}{m'}$.

$$\begin{aligned} D(z) &\leq 1 - B(z) + z^m P(w) \\ &= 1 - (p^{\ell-1} q z^\ell + p^{u-1} q z^u) + p^{m-m'} q^{m'} z^m - (B(z) - p^{\ell-1} q z^\ell - p^{u-1} q z^u) \\ &= (1 - p^{\ell-1} q z^\ell)(1 - p^{u-1} q z^u) - p^{\ell+u-2} q^2 z^{\ell+u} + p^{m-m'} q^{m'} z^m - (B(z) - p^{\ell-1} q z^\ell - p^{u-1} q z^u). \end{aligned}$$

If $m = \ell + u$, then m' must be 2. Thus,

$$D(z) \leq (1 - p^{\ell-1} q z^\ell)(1 - p^{u-1} q z^u) - (B(z) - p^{\ell-1} q z^\ell - p^{u-1} q z^u),$$

and either $z_o = \left(p^{\ell-1} q\right)^{-\frac{1}{\ell}}$ or $z_o = \left(p^{u-1} q\right)^{-\frac{1}{u}}$ makes $D(z_o) \leq 0$.

If $m \neq \ell + u$, then we choose z_o as the root of the equation $p^{m-m'} q^{m'} z^m = p^{\ell+u-2} q^2 z^{\ell+u}$. That is,

$$z_o = \left(p^{\ell+u-2-m+m'} q^{2-m'}\right)^{\frac{1}{m-\ell-u}}.$$

Then,

$$p^{\ell-1} q z_o^\ell = p^{\ell-1} q \left(p^{\ell+u-2-m+m'} q^{2-m'}\right)^{\frac{\ell}{m-\ell-u}} = \left(\frac{q}{p}\right)^{\frac{m-\ell m'-(u-\ell)}{m-\ell-u}}.$$

Similarly,

$$p^{u-1} q z_o^u = p^{u-1} q \left(p^{\ell+u-2-m+m'} q^{2-m'}\right)^{\frac{u}{m-\ell-u}} = \left(\frac{p}{q}\right)^{\frac{um'-m-(u-\ell)}{m-\ell-u}}.$$

Thus,

$$\begin{aligned} D(z_o) &\leq (1 - p^{\ell-1} q z_o^\ell)(1 - p^{u-1} q z_o^u) - (B(z_o) - p^{\ell-1} q z_o^\ell - p^{u-1} q z_o^u) \\ &\leq (1 - p^{\ell-1} q z_o^\ell)(1 - p^{u-1} q z_o^u) \\ &= \left(1 - \left(\frac{q}{p}\right)^x\right)\left(1 - \left(\frac{p}{q}\right)^y\right) \end{aligned}$$

where $x = \frac{m-\ell m'-(u-\ell)}{m-\ell-u}$ and $y = \frac{um'-m-(u-\ell)}{m-\ell-u}$. We can see that both numerators in x and y are positive. Indeed, we consider two cases. First, if $\frac{m}{m'}$ is an integer, then $\ell = \frac{m}{m'} - 1$,

$u = \frac{m}{m'} + 1$, and $u - \ell = 2$. Thus, $m - \ell m' - (u - \ell) = m' - 2 \geq 0$ and $um' - m - (u - \ell) = m' - 2 \geq 0$. Otherwise, if $\frac{m}{m'}$ is not an integer, then $m - \ell m' \geq 1$, $um' - m \geq 1$, and $u - \ell = 1$. Thus $m - \ell m' - (u - \ell) = m - \ell m' - 1 \geq 0$ and $um' - m - (u - \ell) = um' - m - 1 \geq 0$. Both x and y have the same denominators, and consequently both have the same sign.

Let us assume that both x and y are positive. Then, $D(z_o) \leq 0$ because

$$1 - \left(\frac{q}{p}\right)^x \leq 0 \text{ and } 1 - \left(\frac{p}{q}\right)^y \geq 0 \quad \text{if } \frac{q}{p} \geq 1,$$

and

$$1 - \left(\frac{q}{p}\right)^x \geq 0 \text{ and } 1 - \left(\frac{p}{q}\right)^y \leq 0 \quad \text{if } \frac{q}{p} < 1.$$

Similarly, it is also true that $D(z_o) \leq 0$ when both x and y are negative.

Hence, there always exists a positive z_o such that $D(z_o) \leq 0$, and consequently there always exists at least one positive real root on $(\rho, z_o]$.

Therefore, there exists at least one positive real root, which is greater than ρ. ■

C Proof of Lemma 1

Here we present the proof of Lemma 1 used for the large deviations results of 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 $\mathcal{S}$.

Now we show the existence of the real root z_a. By the definition of $h_a(z)$,

$$\begin{aligned} h_a'(z) &= \frac{aM'(z)}{M(z)} - \frac{1}{z} \\ &= \frac{-D(z)^2 + (azB'(z) + 1 - B(z))D(z) + az(1 - B(z))D'(z)}{zD(z)(D(z) - 1 + B(z))}. \end{aligned}$$

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,

$$\begin{aligned} f_a(z) &= -D(z)^2 + (azB'(z) + 1 - B(z))D(z) + az(1 - B(z))D'(z) \\ &= \{(1 - S(z))S(z) + azS'(z)\}(1 - B(z))^2 \\ &\quad + z^m P(w)\{azB'(z) + (1 - 2S(z) + am)(1 - B(z)) - z^m P(w)\}. \end{aligned}$$

We find that

$$f_a(\rho) = \rho^m P(w)\left(a\rho B'(\rho) - \rho^m P(w)\right) > 0 \tag{33}$$

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 (33) and above, $f_a(z)$ has at least one real root z_a between 0 and ρ. Therefore, there exists a real root z_a of $h'_a(z) = 0$.

Secondly, we consider when the pattern is self-overlapping, that is, $S(z) \not\equiv 1$. Then, there exist non-constant terms in $S(z)$. Let r $(0 < r < m)$ be the smallest degree among them. That is,

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

where 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, we get the same result. Therefore, there exists at least one real root between 0 and ρ. 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 1. Let z_a be the real root of $h'_a(z) = 0$. Then, by definition,

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

and this leads to

$$\frac{M'(z_a)}{M(z_a)} = \frac{1}{az_a}.$$

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,

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

where X is a random variable which has the following distribution function:

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

Therefore, in other words, z_a is the real value that makes $\mathbf{E}[X] = \frac{1}{a}$.

Now let us compute $h_a''(z_a)$. We have

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

Therefore, $h_a''(z_a) > 0$ because definitely the distribution is not concentrated at one value. This proves part (ii) of Lemma 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 1.

D Aperiodicity of $M(z)$

The lemma below shows that the probability generating function of a language $\mathcal{M}$ is aperiodic, which is useful to derive the large deviations results in Section 3.4.

Lemma 4 *$M(z)$ is aperiodic if the length of the pattern w over the extended alphabet $\mathcal{B}_{d,k}$ is greater than 1.*

Proof: Let $\mathcal{B}_{d,k} = \{\beta_d, \beta_{d+1}, \cdots, \beta_k\}$ and ℓ be the length of w over $\mathcal{B}_{d,k}$ ($\ell \geq 2$). We consider two cases - when some super-symbols of $\mathcal{B}_{d,k}$ do not appear in w and when all symbols appear in w.

Let us prove the first case. Let β_i be the symbol that does not appear in w. Then, definitely 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, that is, when all symbols of $\mathcal{B}_{d,k}$ do appear in w. For this, we consider three sub cases and find two words in $\mathcal{M}$, which differ by 1 in length for each case:

Case (i) $|\mathcal{B}_{d,k}| \geq 3$:
Let $u_1 = \underbrace{\beta_d \cdots \beta_d}_{\ell}\beta_d\underbrace{\beta_d \cdots \beta_d}_{\ell} w$ 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 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 β_i and β_j are the only symbols in $\mathcal{B}_{d,k}$.

Case (iii) $|\mathcal{B}_{d,k}| = 2$ and $\ell = 2$:
There are only two cases. That is, $w = \beta_d\beta_k$ or $w = \beta_k\beta_d$. Definitely, for both cases, $\beta_d w$ and $\beta_k w$ are in $\mathcal{M}$.

In summary, in $\mathcal{M}$, there always exist two words whose lengths differ by 1. Therefore $M(z)$ is aperiodic. ■

References

[1] S. Aviran, P. Siegel, and J. Wolf, Optimal Parsing Trees for Run-Length Coding of Biased Data, *IEEE Intl. Symposium on Information Theory*, 1495-1499, Seattle, 2006.

[2] E. Bender and F. Kochman, The Distribution of Subword Counts is Usually Normal, *European Journal of Combinatorics*, 14, 265–275, 1993.

[3] Y. Choi and W. Szpankowski, Pattern Matching in Constrained Sequences, *IEEE Intl. Symposium on Information Theory*, 2606–2610, Nice, 2007.

[4] A. Dembo and I. Kontoyiannis, Source coding, large deviations, and approximate pattern matching, *IEEE Transactions on Information Theory*, 48, 1590–1615, 2002.

[5] E. Drinea and M. Mitzenmacher, On Lower Bounds for the Capacity of Deletion Channels, *IEEE Transactions on Information Theory*, 52, 4648–4657, 2006.

[6] J. Fan, T. L. Poo, and B. Marcus, Constraint Gain, *IEEE Transactions on Information Theory*, 50, 1989–2001, 2004.

[7] P. Flajolet, W. Szpankowski, and B. Vallée, Hidden Word Statistics, *Journal of the ACM*, 53, 1–37, 2006.

[8] L. Guibas and A. M. Odlyzko, Periods in Strings, *J. Combinatorial Theory*, 30, 19–42, 1981.

[9] P. Henrici, *Applied and Computational Complex Analysis*, Vols. 1–3, John Wiley & Sons, New York, 1977.

[10] P. Jacquet and W. Szpankowski, Autocorrelation on Words and Its Applications - Analysis of Suffix Trees by String-Ruler Approach, *J. Combinatorial Theory*, 66, 237–269, 1994.

[11] P. Jacquet and W. Szpankowski, On (d, k) Sequences Not Containing a Given Word, *IEEE Intl. Symposium on Information Theory*, 1486–1489, Seattle, 2006.

[12] R. Jolivet, A. Rauch, H.-R. Luscher, and W. Gerstner, Predicting spike timing of neocortical pyramidal neurons by simple threshold models, *Journal of Computational Neuroscience* 21(1):35–49, 2006.

[13] V. Kolesnik and V. Krachkovsky, Generating Functions and Lower Bounds on Rates for Limited Error-Correcting Codes, *IEEE Trans. Information Theory*, 37, 778–788, 1991.

[14] M. Lothaire, (Editor) *Applied Combinatorics on Words*, Cambridge University Press, 2005.

[15] B. Marcus, R. Roth, and P. Siegel, Constrained Systems and Coding for Recording Channels, Chap. 20 in *Handbook of Coding Theory* (eds. V. S. Pless and W. C. Huffman), Elsevier Science, 1998.

[16] B. Moision, A. Orlitsky, and P. Siegel, On codes that avoid specific differences, *IEEE Trans. Information Theory*, 47, 433–442, 2001.

[17] P. Nicodème, B. Salvy, and P. Flajolet, Motif Statistics, *European Symposium on Algorithms*, *Lecture Notes in Computer Science*, No. 1643, 194–211, 1999.

[18] L. Paninski, Estimation of Entropy and Mutual Information, *Neural Computation*, 15, 1191–1253, 2003.

[19] M. Régnier and W. Szpankowski, On the Approximate Pattern Occurrences in a Text, *Proc. Compression and Complexity of SEQUENCE'97*, IEEE Computer Society, 253–264, Positano, 1997.

[20] M. Régnier and W. Szpankowski, On pattern frequency occurrences in a Markovian sequence, *Algorithmica*, 22:631–649, 1998.

[21] G. Reinert, S. Schbath, and M. Waterman, Probabilistic and statistical properties of words: an overview, *J. Comput. Biol.*, 7, 1–46, 2000.

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

[23] M. Waterman, *Introduction to Computational Biology*, Chapman & Hall, London, 1995.

[24] E. Zehavi and J. Wolf, On runlength codes, *IEEE Transactions on Information Theory*, 34, 45–54, 1988.

[25] http://lcn.epfl.ch/QuantNeuronMod2007/challenge.html, Quantitative Single-Neuron Modeling: Competition, 2007.