Large Deviations for 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

Abstract—In the constrained pattern matching one searches for a given pattern in a constrained sequence, which finds applications in communication, magnetic recording, and biology. We concentrate on 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 our previous paper [2] we established the central limit theorem (CLT) for the number of occurrences of a given pattern in such sequences. Here, we present precise large deviations results; they are often of greater values in diverse applications. In particular, we 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. Among others, we obtain justifiably accurate statistical inferences about their biological properties and functions. Throughout, we use techniques of analytic information theory such as combinatorial calculus, generating functions, and complex asymptotics.

I. 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. For example, spike trains of neuronal data satisfy certain structural constraints due to the very nature of the biological mechanisms that produce them.

In our previous paper [2] we set up the goal to understand some aspects of pattern matching in constrained sequences. Although our methods work for a large class of constrained systems, we restrict here our analysis to the so-called $(d,k)$ sequences in which runs of zeros can neither 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. For example, the spike trains of neuronal data (recorded from different neurons in the brain of an animal) discussed above seem to satisfy structural constraints that 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 will need to contain at least a certain number of zeros (corresponding to no activity) between each two consecutive ones (corresponding to firing times).

In those applications, one often looks for statistically significant patterns, that is, under- and over-represented patterns. Therefore, as in our previous work [2] we study here the following problem: given a word $w$ how many times it occurs in a $(d,k)$ sequence. We explore this problem in a probabilistic framework, that is, we assume that a sequence is generated by a (biased) binary 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 our previous work [2] we derived the central Limit Theorem (CLT), while here we present precise large deviations results that often are more useful in practice.

In the information theory community, $(d,k)$ sequences were analyzed since Shannon. Recently, there has been a resurgence of interest in constrained pattern matching [3], [8], [10] due to various applications in recording and biology. Pattern matching problems were also studied in computer science with the main contributions coming from Guibas, Odlyzko, and others [1], [4], [5], [9], [13], [15]. In this paper, we take the view of combinatorics on words and construct languages representing $(d,k)$ sequences containing a given number of $w$ occurrences. Using generating functions and complex asymptotics, we present precise asymptotics for large deviations for the number of occurrences. We apply it to find under- and over-represented patterns in spike trains of neuronal data. To the best of our knowledge these are novel results obtained through tools that belong to analytic information theory.

II. MAIN RESULTS

We start with reviewing some facts from [2] to make the paper self-contained and to introduce gently the reader into the main methodology developed here. We focus on restricted $(d,k)$ sequences that start with 0 and end with 1. We use language representation to derive the probability distribution of the number of occurrences of a given pattern $w$ (a $(d,k)$ sequence itself) in a $(d,k)$ sequence generated by a binary memoryless source.

As in [2], [10], we first define the so called extended alphabet as

$$\mathcal{B}_{d,k} = \{0 \ldots 1, \ldots, 0 \ldots 0 \ldots 1\}.$$ 

Let now $w = w_1 \ldots w_m \in \{0,1\}^m$ with $w_1 = 0$ and $w_m = 1$. The same pattern over $\mathcal{B}_{d,k}$ becomes $w = \beta_1 \ldots \beta_m'$, where $\beta_i \in \mathcal{B}_{d,k}$ and $\sum_{i=1}^{m'} |\beta_i| = m$. Recall that $S$ denotes the autocorrelation set of $w$ over $\mathcal{B}_{d,k}$ defined as

$$S = \{\beta_{m' + l + 1} : \beta_{1} = \beta_{m' - l + 1}, \quad 1 \leq l \leq m'\}.$$
where $\beta_i = \beta_i \cdots \beta_j$ and $\beta_i = \epsilon$ if $i > j$.

As in [2, 6, 13], we shall use language approach and we define four languages, $T_r^{(d,k)}$, $R^{(d,k)}$, $M^{(d,k)}$, and $U^{(d,k)}$ as follows:

(i) $T_r^{(d,k)}$ as the set of all $(d, k)$ sequences (over the extended alphabet $B_{d, k}$) containing exactly $r$ occurrences of $w$;
(ii) $R^{(d,k)}$ as the set of all $(d, k)$ sequences (over the extended alphabet $B_{d, k}$) containing only one occurrence of $w$, located at the right end;
(iii) $U^{(d,k)}$ defined as

\[ U^{(d,k)} = \{ u : w \cdot u \in T_r^{(d,k)} \} . \]

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

\[ M^{(d,k)} = \{ v : w \cdot v \in T_r^{(d,k)} \text{ and } w \text{ occurs at the right end of } w \cdot v \} . \]

To simplify our notation, we drop the upper index $(d, k)$ unless it is necessary. Then as in [2, 13, 15]

\[ T_0 \cdot \{ w \} = R \cdot S, \quad \text{(2)} \]

and

\[ T_r = R \cdot M^{r-1} \cdot U, \quad \text{(1)} \]

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

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

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

where $B^*$ is the set of all restricted $(d, k)$ sequences, that is, $B^* = \{ \epsilon \} + B + B^2 + B^3 + \cdots$. Similarly, $\mathcal{M}^* = \sum_{i=0}^{\infty} M^i$, where $\mathcal{M}^0 = \{ \epsilon \}$.

Throughout, we assume that a binary sequence is generated by a memoryless source with $p$ being the probability of emitting a ‘0’ and $q = 1 - p$. The main tool of analytic information theory is generating function defined for a language $L$ as

\[ L(z) := \sum_{u \in L} P(u)z^{|u|}, \]

where $P(u)$ is the probability of $u$ and $|u|$ is the length of $u$. Recall that the autocorrelation polynomial $S(z)$ is the probability generating function for the autocorrelation sequence $S$. In general, we write $[z^n]L(z)$ for the coefficient of $L(z)$ at $z^n$.

The language relationships (3)-(5) are easily translated into probability generating functions. For example,

\[ M(z) - 1 = \frac{B(z) - 1}{z^m P(w) + S(z)(1 - B(z))}, \]

and

\[ U(z) = (1 - M(z))/(1 - B(z)) \]

as well as

\[ R(z) = \frac{z^m P(w)U(z)}{B'(\rho)}, \]

where $B(z) = zq(zp)^{\frac{d}{1 - zp}} - \frac{(zp)^{k+1}}{1 - zp}$. (6)

This finally leads to

\[ T_0(z) = \frac{S(z)}{D(z)}, \quad \text{(7)} \]

\[ T_r(z) = \frac{z^m P(w)D(z)}{D(z)^{r+1}}, \quad \text{(8)} \]

where $D(z) = S(z)(1 - B(z)) + z^m P(w)$. (9)

Our goal is to compute a large deviations estimate of the number of $w$ occurrences. This is achieved by defining $O_n$ as a random variable representing the number of occurrences of $w$ in a sequence of length $n$. We define two generating functions

\[ T_r(z) = \sum_{n \geq 0} P(O_n = r, D_n)z^n, \]

\[ T(z, u) = \sum_{r \geq 0} \sum_{n \geq 0} P(O_n = r, D_n)z^n u^r, \]

where $D_n$ is the event that a randomly generated binary sequence of length $n$ is a $(d, k)$ sequence. Observe that

\[ P(D_n) = [z^n]T(z, 1) \]

is the probability that a randomly generated sequence of length $n$ is a $(d, k)$ sequence. Clearly, from (7)-(9) we find

\[ T(z, u) = R(z) \frac{u}{1 - uM(z)} U(z) + T_0(z). \]

We must realize that $T(z, u)$ is not a bivariate probability generating function since $[z^n]T(z, 1) \neq 1$. Therefore, we must introduce a conditional probability generating function. Let $O_n(D_n)$ be a short-hand notation for the conditional number of occurrences of $w$ in a $(d, k)$ sequence. More formally,

\[ P(O_n(D_n) = r) = P(O_n = r | D_n). \]

Then, the probability generating function of $O_n(D_n)$ is

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

and the expected value of $O_n(D_n)$ becomes a conditional expected value defined as

\[ \mathbf{E}[O_n(D_n)] = \frac{[z^n]T_0(z, 1)}{[z^n]T(z, 1)}, \]

where $T_0(z, 1)$ is the derivative of $T(z, u)$ at $u = 1$.

To formulate our main result we need one result from [2], namely,

\[ P(D_n) = \frac{1}{B'(\rho)} \lambda^{n+1} + O(\omega^n) \]

for some $\omega < \lambda$, and

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

where $\lambda = 1/\rho$ and $\rho$ is the smallest positive real root of $B(z) = 1$.

In this paper, our main result deals with large deviations and is presented next.

**Theorem 1:** Let $\tau := \tau(p, w)$ be the smallest positive real root of $D(z) = 0$ (cf. (9)) and $\rho := \rho(p)$ be the unique positive real root of $B(z) = 1$. Define $r = (1 + \delta)\mathbf{E}[O_n(D_n)]$ for
some $\delta > 0$, and then for $a$ such that $na = (1 + \delta)E[O_n(D_n)]$ set

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

Let also $z_a$ be a unique real root $z_a$ of the equation $h_a'(z) = 0$ such that $z_a \in (0, \rho)$. Then

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

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

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

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

with $g(z) = P(w)z^{m-1}/D(z)^{m+2}M(z)$ and $\tau_a = h''_a(z_a)$. The constant $c_2$ is explicitly computed in (13)).

As a potential application of our main results, we use Theorem 1 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 a constrained sequence. 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 [12]. 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.

In our experiments, we used single-electrode data from cortical neurons under random current injection. The details can be found in [7], [16]. This spike timing data can be transformed to a $(d, k)$ sequence by setting the time resolution and dividing time into $bins$ of equal 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 1; otherwise it is represented by a 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

$$P(O_n(D_n) \geq O_{th}) \leq \alpha_{th},$$

where $\alpha_{th}$ is a given probability threshold. From Theorem 1 we easily conclude that for $\alpha_{th}$ 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.

In our experiment, we set the size of bin to 3 ms and obtained a $(d, k) = (1, 6)$ sequence of length 2193. Figure 1 shows the number of occurrences for various patterns $w$ within a window of size $W = 500$: here we use a short-hand notation $[i]$ for a pattern $0 \cdots 01$. 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.

III. ANALYSIS

In this section we prove our main large deviations result, that is, we compute $P(O_n(D_n) = r)$ for $r = (1 + \delta)E[O_n(D_n)]$ with $\delta > 0$. Let $a$ be a real constant such that $na =
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 $z = zn e^{i \theta}$,

$$I_0 = \frac{1}{2\pi} \oint_{C} e^{nh_n(z_n e^{i \theta})} g(z_n e^{i \theta}) dz_n e^{i \theta} d\theta$$

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

$$\tau_a^2 = h''(z_a)$$ (cf. part (ii) of Lemma 1),

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

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

$$h_n(z_a e^{i \theta}) = h_n(z_a) - \frac{\tau_a^2}{2} z_n^2 \theta^2 - \left(\beta_a \tau_a z_n^3 + \frac{\tau_n^2}{2} z_n^2 \right) i \theta^3$$

since $h_n'(z_a) = 0$. Similarly,

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

When $|\theta| \leq \theta_0$, $n \theta^k \rightarrow 0$ ($k \geq 3$) as $n \rightarrow \infty$, and then

$$e^{nh_n(z_n e^{i \theta})} = \exp \left( nh_n(z_n) - \frac{\tau_a^2}{2} z_n^2 \theta^2 \right)$$

$$\times \left(1 + \alpha(\theta) + \frac{\alpha(\theta)^2}{2!} + \frac{\alpha(\theta)^3}{3!} + \cdots\right),$$

where

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

After some tedious algebra we find

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

$$= \frac{z_n e^{nh_n(z_n)}}{2\pi} \int_{-\theta_0}^{\theta_0} \exp \left( -nh_n(z_n e^{i \theta}) - \frac{\tau_n^2}{2} z_n^2 \theta^2 \right)$$

$$\times \left(1 + \alpha(\theta) + \frac{\alpha(\theta)^2}{2!} + \frac{\alpha(\theta)^3}{3!} + \cdots\right) g(z_n e^{i \theta}) d\theta$$

With the change of variable $\theta = -\frac{\theta}{\tau_a \sqrt{2n}}$, and using

$$\int_{-\infty}^{+\infty} e^{-x^2} x^k = \frac{\Gamma(2k)}{2^{k-1} \Gamma(k) \sqrt{2\pi}}, \text{ after some algebra we obtain}$$

$$I_0 = \frac{g(z_n)}{\tau_n \sqrt{2n}} \left[1 + \frac{1}{n} \left(\frac{1}{2a} g'(z_n) + \frac{3\beta_a g'(z_n)}{\tau_a g(z_n)} + 3\gamma_a - \frac{15}{2} \beta_a^2\right) + O\left(\frac{1}{n^2}\right)\right].$$
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 \( M \). Clearly, all coefficients of \( M(z) \) are non-negative, and \( M(z) \) is aperiodic, as can be proved. By the non-negativity of coefficients and aperiodicity, the function \( |M(z_{a}e^{i\theta})| \) is uniquely maximum at \( \theta = 0 \). It also infinitely differentiable at \( \theta = 0 \). Consequently, there exists an angle \( \theta_1 \in (0, \pi) \) such that

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

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

\[
|M(z_{a}e^{i\theta})| \leq \frac{M(z_{a}e^{i\theta_0})}{z_{a}^{n}} \quad \text{for} \quad \theta \in [\theta_0, \pi],
\]

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

\[
\left| e^{nh_{a}(z_{a}e^{i\theta})} \right| = \left| \frac{M(z_{a}e^{i\theta})}{z_{a}^{n}} \right| \leq \frac{M(z_{a}e^{i\theta_0})}{z_{a}^{n}} = e^{nh_{a}(z_{a}e^{i\theta_0})},
\]

and this leads to

\[
\frac{1}{2\pi} \left| \int_{\theta_0}^{\pi} e^{nh_{a}(z_{a}e^{i\theta})} g(z_{a}e^{i\theta})z_{a}e^{i\theta} d\theta \right| \\
\leq \frac{z_{a} \cdot \max(g)}{2\pi} \int_{\theta_0}^{\pi} e^{nh_{a}(z_{a}e^{i\theta})} d\theta \\
\leq \frac{z_{a} \cdot \max(g)}{2\pi} \int_{\theta_0}^{\pi} e^{nh_{a}(z_{a}e^{i\theta_0})} d\theta \\
= \frac{z_{a}(\pi - \theta_0) \cdot \max(g)}{2\pi} \cdot \exp \left( nh_{a}(z_{a}) - O(n^{1/5}) \right) \\
= O(I_0 \cdot e^{-cn^{1/5}}),
\]

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,

\[
\frac{1}{2\pi} \left| \int_{-\pi}^{-\theta_0} e^{nh_{a}(z_{a}e^{i\theta})} g(z_{a}e^{i\theta})z_{a}e^{i\theta} d\theta \right| = O(I_0 \cdot e^{-cn^{1/5}}).
\]

Thus,

\[
|I_1| \leq \frac{1}{2\pi} \left| \int_{\theta_0}^{\pi} e^{nh_{a}(z_{a}e^{i\theta})} g(z_{a}e^{i\theta})z_{a}e^{i\theta} d\theta \right| \\
+ \frac{1}{2\pi} \left| \int_{-\pi}^{-\theta_0} e^{nh_{a}(z_{a}e^{i\theta})} g(z_{a}e^{i\theta})z_{a}e^{i\theta} d\theta \right| \\
= O(I_0 \cdot e^{-cn^{1/5}}),
\]

that is, \( |I_1| \) is exponentially smaller than \( I_0 \).

Putting everything together, we obtain

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

\[
= \frac{g(z_{a})e^{nh_{a}(z_{a})}}{\tau_{a} \sqrt{2\pi n}} \left[ 1 + \frac{1}{n} \left( -\frac{1}{2\tau_{a}^{2}} \frac{g''(z_{a})}{g(z_{a})} \right) + \frac{3\beta_{a} g'(z_{a})}{\tau_{a} g(z_{a})} + 3\gamma_{a} - \frac{15}{2} \frac{\beta_{a}^{2}}{\tau_{a}^{2}} + O \left( \frac{1}{n^{2}} \right) \right].
\]

Finally, we are ready to compute \( P(O_{n}(D_{n}) = na) \). By (11),(12) and the above,

\[
P(O_{n}(D_{n}) = na) = \frac{[z^{n}]u^{na}T(z, u)}{[z^{n}]T(z, 1)} (13)
\]

\[
= \rho B'(\rho)g(z_{a})e^{-nI(a)} \cdot \frac{1}{\tau_{a} \sqrt{2\pi n}} \left[ 1 + \frac{1}{n} \left( -\frac{1}{2\tau_{a}^{2}} \frac{g''(z_{a})}{g(z_{a})} \right) + \frac{3\beta_{a} g'(z_{a})}{\tau_{a} g(z_{a})} + 3\gamma_{a} - \frac{15}{2} \frac{\beta_{a}^{2}}{\tau_{a}^{2}} + O \left( \frac{1}{n^{2}} \right) \right],
\]

where \( I(a) = -\log \rho - h_{a}(z_{a}) \), which can be proved to be positive. This establishes Theorem 1, where the constant \( c_2 \) can be extracted from the above.

ACKNOWLEDGMENT

This work was supported in part by the NSF Grants CCF-0513636, and DMS-0503742, the NIH Grant R01 GM068959-01, NSA Grant 07G-044, and AFOSR Grant FA8655-07-1-3071.

REFERENCES


