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, often used 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 needs 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 searches 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 are often 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 pattern 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 only 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

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

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 $\mathcal{S}$ denotes the *autocorrelation set* of w over $\mathcal{B}_{d,k}$ defined as

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

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

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

(i) $\mathcal{T}_r^{(d,k)}$ as the set of all (d,k) sequences (over the extended alphabet $\mathcal{B}_{d,k}$) containing exactly r occurrences of w;

(ii) $\mathcal{R}^{(d,k)}$ as the set of all (d,k) sequences (over the extended alphabet $\mathcal{B}_{d,k}$) containing only one occurrence of w, located at the right end;

(iii) $\mathcal{U}^{(d,k)}$ defined as

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

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

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

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

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

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

and

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

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

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

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 $\mathcal{M}^0 = \{\epsilon\}$.

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 $\mathcal{L}$ as

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

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 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 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 $R(z) = z^m P(w) U(z)$, where

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

This finally leads to

$$T_0(z) = \frac{S(z)}{D(z)}, \quad (6)$$

$$T_r(z) = \frac{z^m P(w)(D(z) + B(z) - 1)^{r-1}}{D(z)^{r+1}}, \quad (7)$$

where

$$D(z) = S(z)(1 - B(z)) + z^m P(w). \quad (8)$$

Our goal is to compute a large deviations estimate of the number of w occurrences. To this end, we define O_n as a random variable representing the number of occurrences of w in a (regular) binary sequence of length n. We define two generating functions

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

$$T(z,u) = \sum_{r \geq 0} T_r(z) u^r = \sum_{r \geq 0} \sum_{n \geq 0} P(O_n = r, \mathcal{D}_n) z^n u^r, \quad (9)$$

where $\mathcal{D}_n$ is the event that a randomly generated binary sequence of length n is a (d,k) sequence. Observe that

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

is the probability that a randomly generated sequence of length n is a (d,k) sequence. Clearly, from (1) and (9) we find

$$T(z,u) = R(z)\frac{u}{1 - uM(z)}U(z) + T_0(z). \quad (10)$$

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

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

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

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

and the expected value of $O_n(\mathcal{D}_n)$ becomes a conditional expected value defined as

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

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

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

$$P(\mathcal{D}_n) = \frac{1}{B'(\rho)}\lambda^{n+1} + O(\omega^n) \quad (11)$$

for some $\omega < \lambda$, and

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

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

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. (8)) and $\rho := \rho(p)$ be the the unique positive real root of $B(z) = 1$. Define $r = (1+\delta)\mathbf{E}[O_n(\mathcal{D}_n)]$ for some $\delta > 0$, and then for a such that $r = na$ set

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

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

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

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

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

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

with $g(z) = \frac{P(w) z^{m-1}}{D(z)^2 M(z)}$ and $\tau_a^2 = h''_a(z_a)$. 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.

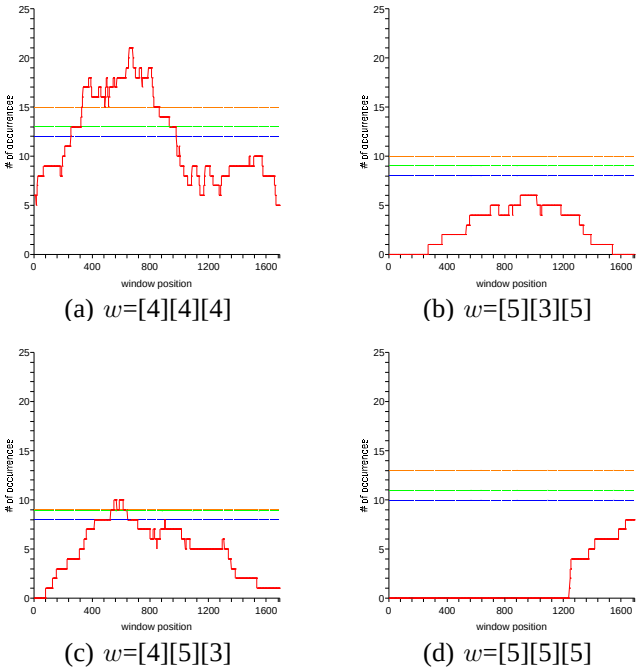
For experiments, we used single-electrode data from cortical neurons under random current injection. The details can be found in [7], [16]. This spike timing data can be transformed to a (d, k) sequence by setting the time resolution and dividing time into *bins* of the same size. Each time bin is represented by a bit 0 or 1. If there is a spike in a certain time bin, it is represented by a 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

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

where α_{th} is a given probability threshold. From Theorem 1 we easily conclude that for α_{th} in the range of the large deviations domain, the threshold is $O_{th} = na_{th}$, where

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



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

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

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

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). 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 1 shows the number of occurrences for various patterns within a window of size 500; here we use a short-hand notation $[i]$ for a pattern $\underbrace{0 \cdots 0}_{i-1}1$. 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 the constraints were ignored.

III. Analysis

In this section we prove Theorem 1, that is, we compute $P(O_n(\mathcal{D}_n) = r)$ 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 we compute $P(O_n(D_n) = na)$ asymptotically when na is an integer. Clearly,

$$P(O_n(D_n) = na) = [u^{na}]\mathbf{E}[u^{O_n(D_n)}] = \frac{[z^n][u^{na}]T(z,u)}{[z^n]T(z,1)}. \tag{12}$$

By (10),

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

Hence, Cauchy's coefficient formula leads to [15]

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

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* [15]. 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

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

In the lemma below, we characterize some properties of $h_a(z)$ that are needed to estimate the integral. The proof of this lemma will be presented in the final version of the paper.

Lemma 1: Under the conditions of this paper, the following holds:
(i) There exists a unique real root z_a of the equation $h_a'(z) = 0$ that satisfies $0 < z_a < \rho$.
(ii) We have $h_a''(z_a) > 0$.
(iii) The following is true $h_a(z_a) < -\log\rho$.

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 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,

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

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

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

since $h_a'(z_a) = 0$. 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 $n \to \infty$, and then

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

where

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

Therefore we have

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

With the change of variable $\theta = \frac{\omega}{\tau_a z_a \sqrt{n}}$, and using $\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} x^{2k} = \frac{\Gamma(2k)}{2^{k-1}\Gamma(k)}\sqrt{2\pi}$, after some algebra we obtain

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

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. 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, 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

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

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

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

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,

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

Thus,

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

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

Putting everything together, we obtain

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

Finally, we are ready to compute $P(O_n(D_n) = na)$. By (11),(12) and the above,

$$\begin{aligned}
P(O_n(D_n) = na) &= \frac{[z^n][u^{na}]T(z,u)}{[z^n]T(z,1)} \\
&= \frac{\rho B'(\rho) g(z_a) e^{-nI(a)}}{\tau_a\sqrt{2\pi n}}\left[1+\frac{1}{n}\left(\frac{3\beta_a}{\tau_a}\frac{g'(z_a)}{g(z_a)} - \frac{1}{2{\tau_a}^2}\frac{g''(z_a)}{g(z_a)}\right.\right. \\
&\quad \left.\left. + 3\gamma_a - \frac{15}{2}{\beta_a}^2\right) + O\left(\frac{1}{n^2}\right)\right], \qquad (13)
\end{aligned}$$

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

[1] E. Bender, and F. Kochman, The distribution of subword counts is usually normal. *European Journal of Combinatorics*, **14**, 265–275, 1993.

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

[3] A. Dembo and I. Kontoyiannis. Source coding, large deviations, and approximate pattern matching. *IEEE Trans. Inform. Theory*, 48, 1590–1615, 2002.

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

[5] L. Guibas and A. M. Odlyzko. Periods in strings. *J. Combinatorial Theory*, 30:19–43, 1981.

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

[7] 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.

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

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

[10] 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.

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

[12] L. Paninski", Estimation of Entropy and Mutual Information, *Neural Computation*, 1191-1253, 2003.

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

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

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

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