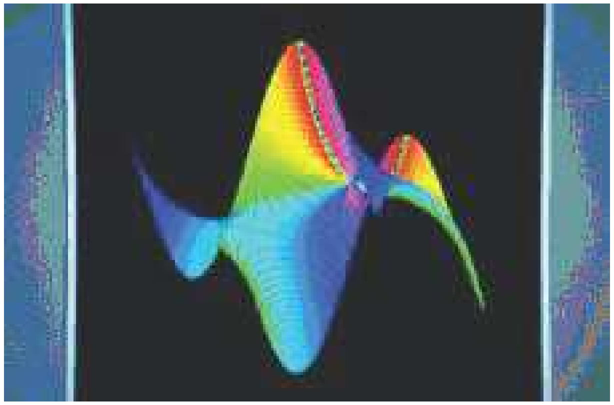
Constrained Pattern Matching*

Yongwook Choi and **Wojciech Szpankowski**
Department of Computer Science
Purdue University

April 13, 2008



Analysis of Algorithm 2008

*This research is supported by NSF and NIH.

Outline

1. Pattern Matching and Constrained Pattern Matching Problems
2. Combinatorial Approach and Language Representation
3. Number of Pattern Occurrences and Analytical Results
4. Experimental Results
5. Proof Sketch of Large Deviation Result

Pattern Matching

Let $\mathcal{W}$ and T be (set of) strings generated over a finite alphabet $\mathcal{A}$.

We call $\mathcal{W}$ the pattern and T the text. The text T is of length n and is generated by a probabilistic source.

The pattern $\mathcal{W}$ can be a single string

$$\mathcal{W} = w_1 \dots w_m, \quad w_i \in \mathcal{A}$$

or a set of strings

$$\mathcal{W} = \{\mathcal{W}_1, \dots, \mathcal{W}_d\}$$

with $\mathcal{W}_i \in \mathcal{A}^{m_i}$ being a set of strings of length m_i.

Questions

- How many times does $\mathcal{W}$ occur in T ?
- What is the probability that $\mathcal{W}$ occurs exactly r times in T ?

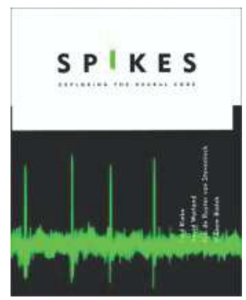
Constrained Pattern Matching

There are constraints on the text T. (e.g., (d, k) sequences, regular expression)

A (d, k) sequence is a binary sequence in which any run of zeros must be of length at least d and at most k.

Example: $(2, 4)$ sequence - $0010001001000100001000100100001000$

(d, k) sequences are useful for digital recording and biology.



Questions

- How many times does $\mathcal{W}$ occur in a (d, k) sequence, T ?
- What is the conditional probability that $\mathcal{W}$ occurs exactly r times in a (d, k) sequence, T ?

Outline

1. Pattern Matching and Constrained Pattern Matching Problems
2. **Combinatorial Approach and Language Representation**
3. Number of Pattern Occurrences and Analytical Results
4. Experimental Results
5. Proof Sketch of Large Deviation Result

Combinatorial Approach

We use a combinatorial approach, based on (M. Régnier & W. Szpankowski, Algorithmica, 1998), (P. Jacquet & W. Szpankowski, ISIT, 2006).

- Construct languages and their relationships
- Translate the language relationships into generating functions

A language, say $\mathcal{L}$, is a collection of words, and its probability generating function is defined as

$$L(z) = \sum_{u \in \mathcal{L}} P(u) z^{|u|} = \sum_{n \geq 0} z^n L_n, \qquad [z^n]L(z) = L_n$$

where $P(u)$ is the probability of u.

Define

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

that is, a set of runs of zeros of length between d and k.

Combinatorial Approach

Define

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

as an *extended alphabet*.

The probability generating function of $\mathcal{B}_{d,k}$ is

$$B(z) \quad = \quad p^d q z^{d+1} + p^{d+1} q z^{d+2} + \cdots + p^k q z^{k+1} = zq \frac{(zp)^d - (zp)^{k+1}}{1 - zp},$$

where p is the probability of emitting a '0' and $q = 1 - p$.

We consider only restricted (d, k) sequences, which are (d, k) sequences that start with '0' and end with '1'.

Observe that the set of all restricted (d, k) sequences is

$$B^*_{d,k} = \{\epsilon\} + \mathcal{B}_{d,k} + \mathcal{B}^2_{d,k} + \mathcal{B}^3_{d,k} + \cdots, \quad \text{and} \quad B^*(z) = \frac{1}{1 - B(z)}.$$

Note: We only consider occurrences of the pattern w **over** $\mathcal{B}_{d,k}$, not over the binary alphabet.

Example: $w = 01$ occurs only once in a sequence 001010001.

Autocorrelation Set

Let $w = \beta_1 \ldots \beta_m$, where $\beta_i \in \mathcal{B}_{d,k}$.

We define the *autocorrelation set* of w over $\mathcal{B}_{d,k}$ as

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

where $\beta_i^j = \beta_i \cdots \beta_j$. Its probability generating function $S(z)$ is called the autocorrelation polynomial. (as in L. Guibas & A.M. Odlyzko, 1981)

Example: Let $w = 0100101$ over $\mathcal{B} = \{01, 001, 0001\}$.
Then

$$\mathcal{S} = \{\varepsilon, 00101\}$$

since

01	001	01		
		01	001	01.

Note that $S(z) = 1 + P(00101)z^5$.

Language $\mathcal{T}_r$

$\mathcal{T}_r$ – the set of all **restricted** (d,k) **sequences** containing exactly r occurrences of w. (M. Régnier and W. Szpankowski, 1998)

We define some languages: $\mathcal{R}$, $\mathcal{U}$, and $\mathcal{M}$

(i) We define $\mathcal{R}$ as the set of all restricted (d,k) sequences containing only one occurrence of w, located at the right end.

(ii) We also define $\mathcal{U}$ as

$$\mathcal{U} = \{u : \ w \cdot u \in \mathcal{T}_1\},$$

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

(iii) $\mathcal{M}$ is defined as

$$\mathcal{M} = \{u : \ w \cdot u \in \mathcal{T}_2 \text{ and } w \text{ occurs at the right of } w \cdot u\},$$

that is, $\mathcal{M}$ is a language such that any word in $\{w\} \cdot \mathcal{M}$ has exactly two occurrences of w at the left and right ends.

Example: Let $w = 0100101$. Notice $010100101 \in \mathcal{R}$, and $01 \in \mathcal{U}$.
Observe $00101 \notin \mathcal{U}$, but $00101 \in \mathcal{M}$ because $010010100101 \in \mathcal{T}_2$.

Language Relationships and Generating Functions

The following holds:

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

Then, the above language relationships translate into

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

$$U(z) = \frac{M(z) - 1}{B(z) - 1}, \qquad R(z) = z^m P(w) \cdot U(z)$$

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

In particular, we find

$$T_0(z) = \frac{S(z)}{D(z)}, \qquad T_r(z) = \frac{z^m P(w)(D(z) + B(z) - 1)^{r-1}}{D(z)^{r+1}},$$

where $S(z)$ is the autocorrelation polynomial for w and

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

Outline

1. Pattern Matching and Constrained Pattern Matching Problems
2. Combinatorial Approach and Language Representation
3. **Number of Pattern Occurrences and Analytical Results**
4. Experimental Results
5. Proof Sketch of Large Deviation Result

Number of Occurrences

Let O_n be a random variable representing the number of occurrences of w in a (regular) binary sequence of length n.

The probability generating function of $\mathcal{T}_r$,

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

where

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

Define the bivariate generating function as

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

The probability that a randomly generated sequence of length n is a (d, k) sequence is

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

Number of Occurrences

Introduce a short-hand notation $O_n(\mathcal{D}_n)$ for the conditional number of occurrences of w in a (d, k) sequence,

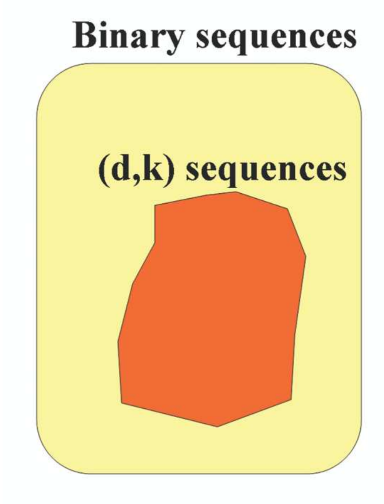
$$P(O_n(\mathcal{D}_n) = r) = P(O_n = r \,|\, \mathcal{D}_n).$$

The probability generating function of $O_n(\mathcal{D}_n)$,

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

The mean and second factorial moment of $O_n(\mathcal{D}_n)$ can be computed by

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



Main Results

Theorem 1. *Let* $\rho := \rho(p) = 1/\lambda$ *be the unique positive real root of*

$$1 - B(z) = 0.$$

Then

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

is the probability of generating a (d, k) *sequence for some* $\omega < \lambda$. *Furthermore, the mean is*

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

and the variance becomes

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

$$\left. + \frac{P(w)B''(\rho)}{B'(\rho)^3}\lambda^{-2m+1} + \frac{2S(\rho)-1}{B'(\rho)}\lambda^{-m+1}\right] + O(1).$$

Main Results

Theorem 2. *Let* $\tau := \tau(p, w)$ *be the smallest real root of*

$$D(z) = 0, \quad \text{(cf. } D(z) = S(z)(1 - B(z)) + z^m P(w)\text{)}$$

and $\rho := \rho(p)$ *be the unique positive real root of* $B(z) = 1$.

(i) *For* $r = O(1)$,

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

for large n *and* $r \geq 1$.

(ii) *(Central limit) 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.*

Main Results

(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

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

Let also z_a be 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

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

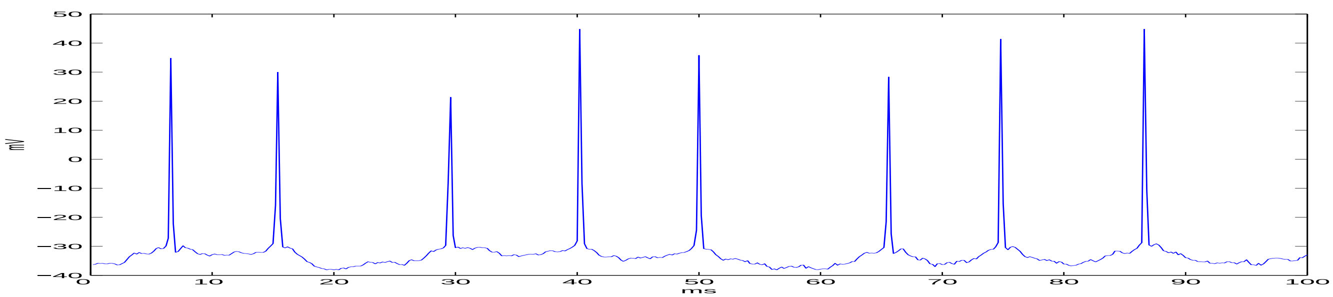
and the constants c_1 and c_2 are explicitly computable.

Outline

1. Pattern Matching and Constrained Pattern Matching Problems
2. Combinatorial Approach and Language Representation
3. Number of Pattern Occurrences and Analytical Results
4. **Experimental Results**
5. Proof Sketch of Large Deviation Result

Experimental Results

Spike trains of neuronal data satisfy structural constraints that exactly match the framework of (d,k) binary sequences.

spike train : 

(d,k) sequence : $0100010000000100001000100000001000100000010000\cdots$

Question: How can we classify a pattern as significant?

We use the large deviations results to detect under- and over-represented patterns.

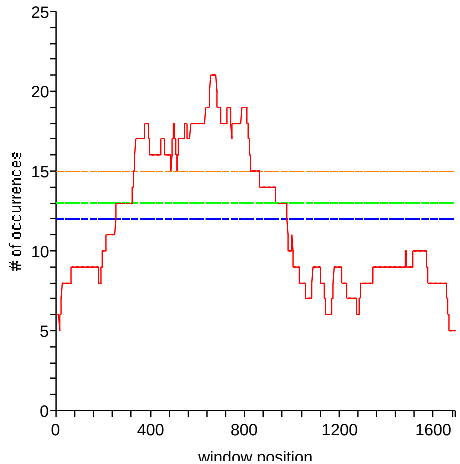
The threshold, O_{th}, above which pattern occurrences will be classified as statistically significant, is defined as the minimum O_{th} such that

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

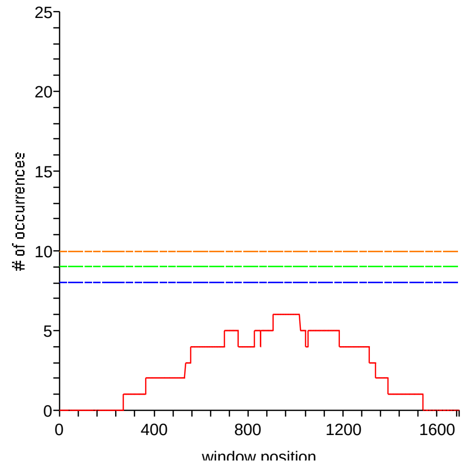
where α_{th} is a given probability threshold (e.g. $\alpha_{th} = 10^{-6}, 10^{-8}$).

Experimental Results

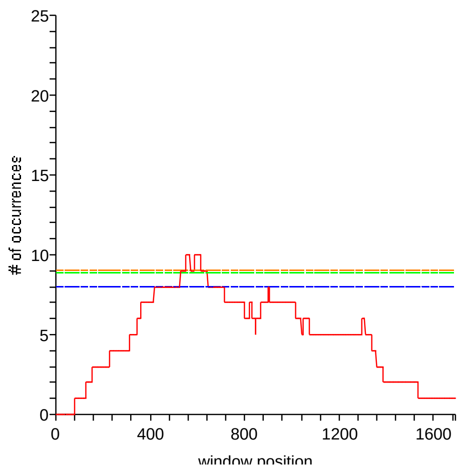
Number of occurrences of w within a window of size 500; here $[i] = \underbrace{0 \cdots 0}_{i-1} 1$.



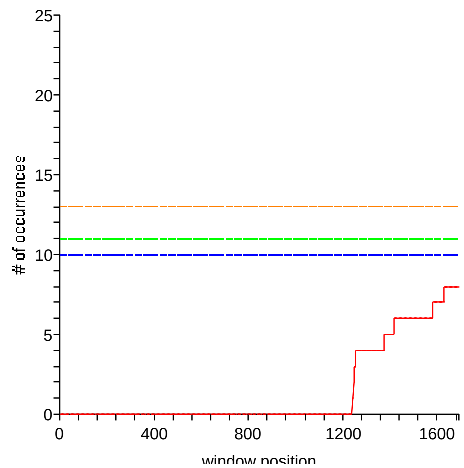
(a) w=(4)(4)(4)



(b) w=(5)(3)(5)



(c) w=(4)(5)(3)



(d) w=(5)(5)(5)

Outline

1. Pattern Matching and Constrained Pattern Matching Problems
2. Combinatorial Approach and Language Representation
3. Number of Pattern Occurrences and Analytical Results
4. Experimental Results
5. **Proof Sketch of Large Deviation Result**

Analysis : Large Deviation Result

Theorem 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

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

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

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

where

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

and the constants c_1 and c_2 are explicitly computable.

Analysis : Sketch of the Proof

1. Generating functions and Cauchy coefficient formula

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

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

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

where

$$h_a(z) = a\log M(z) - \log z \quad \text{and} \quad g(z) = \frac{P(w)z^{m-1}}{D(z)^2M(z)}.$$

Analysis : Sketch of the Proof

2. Saddle point contour

Let z_a a unique real root of the equation $h'_a(z) = 0$. We evaluate the integral on $\mathcal{C} = \{z : |z| = z_a\}$

3. Contour split

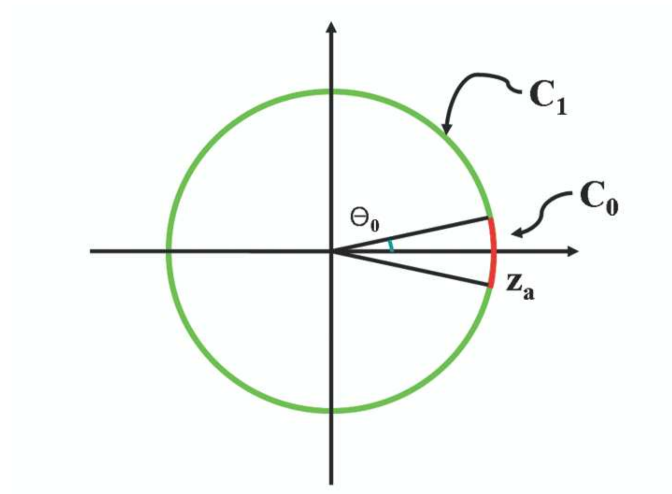
We split $\mathcal{C}$ into $\mathcal{C}_0$ and $\mathcal{C}_1$ where

$$\mathcal{C}_0 = \{z \in \mathcal{C} : |arg(z)| \leq \theta_0\}$$

and

$$\mathcal{C}_1 = \{z \in \mathcal{C} : |arg(z)| \geq \theta_0\}$$

for $\theta_0 = n^{-2/5}$.



$$\begin{aligned} & [z^n][u^{na}]T(z,u) \\ = & \; I_0 + I_1 \\ = & \; \frac{1}{2\pi i}\int_{\mathcal{C}_0} e^{nha(z)}g(z)dz + \frac{1}{2\pi i}\int_{\mathcal{C}_1} e^{nha(z)}g(z)dz. \end{aligned}$$

Analysis : Sketch of the Proof

4. Approximation of I_0

Using change of variables and Taylor series expansion, we get

$$\begin{aligned} I_0 &= \frac{1}{2\pi i}\int_{\mathcal{C}_0} e^{nha(z)}g(z)dz = \frac{1}{2\pi}\int_{-\theta_0}^{+\theta_0} e^{nha(z_a e^{i\theta})}g(z_a e^{i\theta})z_a e^{i\theta}d\theta \\ &\sim \frac{e^{nha(z_a)}}{2\pi\tau_a\sqrt{n}}\int_{-\infty}^{+\infty}\exp\left(-\frac{\omega^2}{2}\right)F(w)d\omega = \frac{g(z_a)e^{nha(z_a)}}{\tau_a\sqrt{2\pi n}}\left(1+\frac{c_2}{n}+O\left(\frac{1}{n^2}\right)\right) \end{aligned}$$

5. Elimination of I_1

We show that I_1 is exponentially smaller than I_0.

$M(z)$ is the probability generating function of language $\mathcal{M}$. By its non-negativity of coefficients and aperiodicity, $|M(z_a e^{i\theta})|$ is uniquely maximum at $\theta = 0$. For $\theta \in [\theta_0, \pi]$,

$$\left|e^{nha(z_a e^{i\theta})}\right| = \frac{\left|M(z_a e^{i\theta})\right|^{na}}{z_a^n} \leq \frac{\left|M(z_a e^{i\theta_0})\right|^{na}}{z_a^n} = \left|e^{nha(z_a e^{i\theta_0})}\right|.$$

Analysis : Sketch of the Proof

6. Putting together

$$P(O_n(D_n) = na) = \frac{[z^n][u^{na}]T(z,u)}{[z^n]T(z,1)} = \frac{I_0 + I_1}{P(\mathcal{D}_n)} = \frac{I_0\left(1 + O\left(e^{-cn^{1/5}}\right)\right)}{P(\mathcal{D}_n)}$$

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

where

$$I(a) = -\log\rho - h_a(z_a).$$