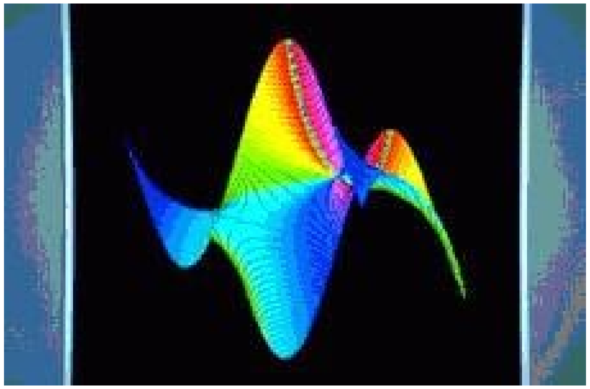
From Pattern Matching to Suffix Trees*

W. Szpankowski
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.

May 26, 2004



*Joint Work with P. Jacquet and M. Regnier

Outline of the Talk

1. Pattern Matching Problems
   - String Matching
   - Subsequence Matching
   - Self-Repetitive Pattern Matching
2. Suffix Trees and Its Parameters
3. Applications
4. Analysis of the Exact Pattern Matching
   - Languages and Generating Functions
   - Asymptotic Analysis
5. Analysis of Suffix Trees
   - Relation to the Exact Pattern Matching
   - Generating Function for the Depth
   - Analysis of Tries
   - Comparison of Tries and Suffix Trees

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.

We shall write

$$T_m^n = T_m \dots T_n.$$

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.

Basic Parameters

Two basic questions are:

- **how many times $\mathcal{W}$ occurs in T**,
- **how long one has to wait until $\mathcal{W}$ occurs in T**.

The following quantities are of interest:

$O_n(\mathcal{W})$ — the number of times $\mathcal{W}$ occurs in T:

$$O_n(\mathcal{W}) = \#\{i : \; T_{i-m+1}^{i} = \mathcal{W}, \; m \le i \le n\}.$$

$W_{\mathcal{W}}$ — the first time $\mathcal{W}$ occurs in T:

$$W_{\mathcal{W}} := \min\{n : \; T_{n-m+1}^{n} = \mathcal{W}\}.$$

Relationship:

$$W_{\mathcal{W}} > n \quad \Leftrightarrow \quad O_n(\mathcal{W}) = 0.$$

Various Pattern Matching

(Exact) String Matching

In the exact string matching the pattern $\mathcal{W} = w_1 \ldots w_m$ is a given string (i.e., consecutive sequence of symbols).

Generalized String Matching

In the generalized pattern matching a set of patterns (rather than a single pattern) is given, that is,

$$\mathcal{W} = (\mathcal{W}_0, \mathcal{W}_1, \ldots, \mathcal{W}_d), \quad \mathcal{W}_i \in \mathcal{A}^{m_i}$$

where $\mathcal{W}_i$ itself for $i \geq 1$ is a subset of $\mathcal{A}^{m_i}$ (i.e., a set of words of a given length m_i).
The set $\mathcal{W}_0$ is called the forbidden set.

Three cases to be considered:

$\mathcal{W}_0 = \emptyset$ — one is interested in the number of patterns from $\mathcal{W}$ occurring in the text.

$\mathcal{W}_0 \neq \emptyset$ — we study the number of $\mathcal{W}_i$, $i \geq 1$ pattern occurrences under the condition that no pattern from $\mathcal{W}_0$ occurs in the text.

$\mathcal{W}_i = \emptyset$, $i \geq 1$, $\mathcal{W}_0 \neq \emptyset$ — restricted pattern matching.

Pattern Matching Problems

Hidden Words or Subsequence Pattern Matching

In this case we search in text for a subsequence $\mathcal{W} = w_1 \ldots w_m$ rather than a string, that is, we look for indices $1 \leq i_1 < i_2 < \cdots < i_m \leq n$ such that

$$T_{i_1} = w_1, \; T_{i_2} = w_2, \cdots, \; T_{i_m} = w_m.$$

We also say that the word $\mathcal{W}$ is "hidden" in the text.

For example:

$$\begin{aligned} \mathcal{W} &= \mathbf{date} \\ T &= \mathbf{hidden\ pattern} \end{aligned}$$

occurs four times as a subsequence in the text as hidden pattern but not even once as a string.

Self-Repetitive Pattern Matching

In this case the pattern $\mathcal{W}$ is part of the text:

$$\mathcal{W} = T_1^m.$$

We may ask when the first m symbols of the text will occur again. This is important in Lempel-Ziv like compression algorithms and suffix trees.

Suffix Trees and Its Parameters

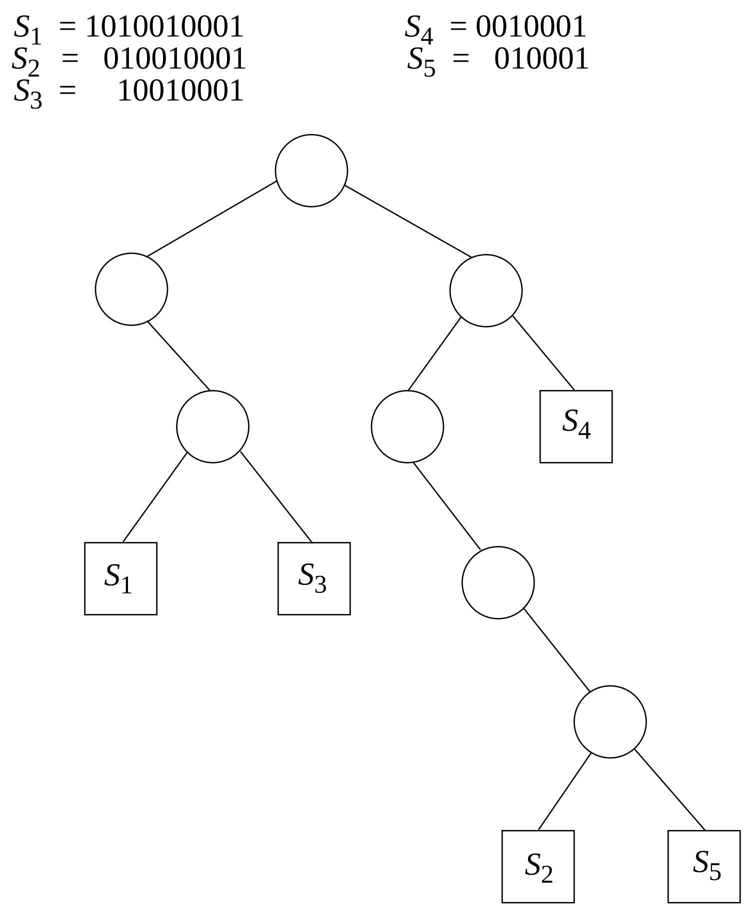


Figure 1: Suffix tree built from the first five suffixes $S_1, \ldots, S_5$ of $T = 1010010001\ldots$

Depth D_n – length of the path from the root to a randomly selected external node (suffix).

Height H_n – length of the longest path.

Path length L_n – sum of all paths to nodes.

Pattern Matching vs Suffix Tree Parameters

The depth D_n in a suffix tree can be defined in terms of the number of occurrence O_n of a pattern $\mathcal{W} = w$ as follows:

- Define $D_n(i)$, $1 \leq i \leq n$, to be the largest value of $k \leq n$ such that T_i^{i+k-1} occurs at least twice in the text T_1^n of length n; that is, $O_n(T_i^{i+k-1}) \geq 2$.
- The depth D_n is equal to $D_n(i)$ when i is randomly and uniformly selected between 1 and n.
- Thus for $w \in \mathcal{A}^k$ we have

$$\Pr(D_n(i) \geq k \;\&\; T_i^{i+k-1} = w) = \Pr(O_n(w) \geq 2 \;\&\; T_i^{i+k-1} = w),$$

and

$$\sum_{i=1}^{n} \Pr(O_n(w) = r \;\&\; T_i^{i+k-1} = w) = r\Pr(O_n(w) = r).$$

Probabilistic Sources

Throughout the talk I will assume that the text is generated by a random source.

Memoryless Source

The text is a realization of an independently, identically distributed sequence of random variables (i.i.d.), such that a symbol $s \in \mathcal{A}$ occurs with probability $P(s)$.

Markovian Source

The text is a realization of a stationary Markov sequence of order K, that is, probability of the next symbol occurrence depends on K previous symbols.

Exact Pattern Matching

Here is an incomplete list of results on string pattern matching (given a pattern $\mathcal{W}$ find statistics of its occurrences):

- Feller (1968),
- Guibas and Odlyzko (1978, 1981),
- Prum, Rodolphe, and Turckheim (1995) – Markovian model, limiting distribution.
- Regnier & W.S. (1997,1998) – exact and approximate occurrences (memoryless and Markov models).
- P. Nicodéme, Salvy, & P. Flajolet (1999) – regular expressions.
- E. Bender and F. Kochman (1993) – general pattern matching.

Languages and Generating Functions

A language $\mathcal{L}$ is a collection of words satisfying some properties.

For any language $\mathcal{L}$ we define its generating function $L(z)$ as

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

where $P(w)$ is the stationary probability u occurrence, $|u|$ is the length of w.

For Markov sources we define $\mathcal{W}$-conditional generating function:

$$L_{\mathcal{W}}(z) = \sum_{u \in \mathcal{L}} P(u | u_{-m} = w_1 \cdots u_{-1} = w_m) z^{|u|}$$

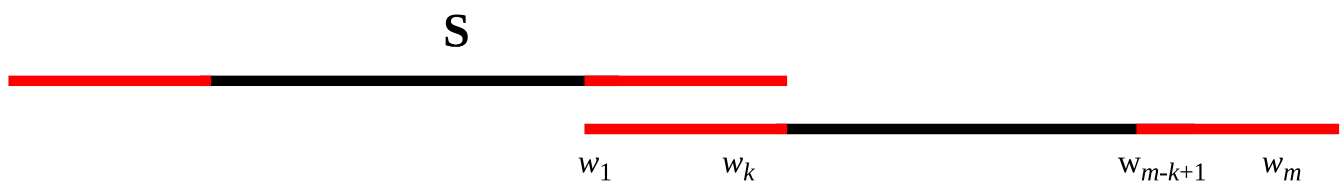
where u_{-i} stands for a symbol preceding the first character of u at distance i.

Autocorrelation Set and Polynomial

Given a pattern $\mathcal{W}$, we define the autocorrelation set $\mathcal{S}$ as:

$$\mathcal{S} = \{w_{k+1}^m : \ w_1^k = w_{m-k+1}^m\}, \quad w_1^k = w_{m-k+1}^m$$

and $\mathcal{WW}$ is the set of positions k satisfying $w_1^k = w_{m-k+1}^m$.



The generating function of $\mathcal{S}$ is denoted as $S(z)$ and we call it the autocorrelation polynomial.

$$S(z) = \sum_{k \in \mathcal{WW}} P(w_{k+1}^m) z^{m-k}.$$

Its $\mathcal{W}$-conditional generating function is denoted $S_{\mathcal{W}}(z)$. For example, for a Markov model we have

$$S_{\mathcal{W}}(z) = \sum_{k \in \mathcal{WW}} P(w_{k+1}^m | w_k^k) z^{m-k} .$$

Example

Example:

Let $\mathcal{W} = bab$ over alphabet $\mathcal{A} = \{a, b\}$.

$$\mathcal{W}\mathcal{W} = \{1, 3\} \quad \text{and} \quad \mathcal{S} = \{\epsilon, ab\},$$

where ϵ is the empty word, since

b a b
  b a b

For the unbiased memoryless source

$$S(z) = 1 + P(ab)z^2 = 1 + \frac{z^2}{4}.$$

For the Markovian model of order one

$$S_{bab}(z) = 1 + P(ab|b)z^2 = 1 + p_{ba}p_{ab}z^2.$$

Language $\mathcal{T}_r$

We are interested in the following language:

$\mathcal{T}_r$ – set of words that contains exactly $r \geq 1$ occurrences of $\mathcal{W}$,

and its generating functions

$$\begin{aligned} O_r(z) &= \sum_{n \geq 0} \Pr\{O_n(\mathcal{W}) = r\} z^n, \quad r \geq 1, \\ O(z,u) &= \sum_{r=1}^{\infty} T_r(z) u^r = \sum_{r=1}^{\infty} \sum_{n=0}^{\infty} \Pr\{O_n(\mathcal{W}) = r\} z^n u^r \end{aligned}$$

for $|z| \leq 1$ and $|u| \leq 1$.

More Languages

(i) Let $\mathcal{T}$ be a language of words containing at least one occurrence of $\mathcal{W}$.

(ii) We define $\mathcal{R}$ as the set of words containing only one occurrence of $\mathcal{W}$, located at the right end. For example, for $\mathcal{W} = aba$

$$cc\,aba \in \mathcal{R}.$$

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

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

that is, a word $u \in \mathcal{U}$ if $\mathcal{W} \cdot u$ has exactly one occurrence of $\mathcal{W}$ at the left end of $\mathcal{W} \cdot u$,

$$c\,ba \in \mathcal{U}, \qquad ba \notin \mathcal{U}.$$

(iv) Let $\mathcal{M}$ be the language:

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

that is, $\mathcal{M}$ is a language such that $\mathcal{W}\mathcal{M}$ has exactly two occurrences of $\mathcal{W}$ at the left and right end of a word from $\mathcal{M}$.

$$ba \in \mathcal{M}.$$

Basic Lemma

Lemma 1. *The language* $\mathcal{T}$ *satisfies the fundamental equation:*

$$\mathcal{T} = \mathcal{R} \cdot \mathcal{M}^* \cdot \mathcal{U} \ .$$

Notably, the language $\mathcal{T}_r$ *can be represented for any* $r \geq 1$ *as follows:*

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

and

$$\mathcal{T}_0 \cdot \mathcal{W} = \mathcal{R} \cdot \mathcal{S} \ .$$

Here, by definition $\mathcal{M}^0 := \{\epsilon\}$ *and* $\mathcal{M}^* := \bigcup_{r=0}^{\infty} M^r$.

Example: Let $\mathcal{W} = TAT$. The following string belongs $\mathcal{T}_3$:

$$\overbrace{CCTAT}^{\mathcal{R}}\underbrace{AT}_{\mathcal{M}}\underbrace{GATAT}_{\mathcal{M}}\overbrace{GGA}^{\mathcal{U}} .$$

More Results

Theorem 1. (i) *The languages $\mathcal{M}, \mathcal{U}$ and $\mathcal{R}$ satisfy:*

$$\begin{aligned}
\bigcup_{k\geq 1} \mathcal{M}^k &= \mathcal{A}^* \cdot \mathcal{W} + \mathcal{S} - \{\epsilon\} , \\
\mathcal{U} \cdot \mathcal{A} &= \mathcal{M} + \mathcal{U} - \{\epsilon\}, \\
\mathcal{W} \cdot \mathcal{M} &= \mathcal{A} \cdot \mathcal{R} - (\mathcal{R} - \mathcal{W}) ,
\end{aligned}$$

where $\mathcal{A}^*$ is the set of all words, $+$ and $-$ are disjoint union and subtraction of languages.

(ii) *The generating functions associated with languages $\mathcal{M}, \mathcal{U}$ and $\mathcal{R}$ satisfy for memoryless sources*

$$\begin{aligned}
\frac{1}{1 - M(z)} &= S_{\mathcal{W}}(z) + P(\mathcal{W}) \frac{z^m}{1 - z}, \\
U_{\mathcal{W}}(z) &= \frac{M(z) - 1}{z - 1}, \\
R(z) &= P(\mathcal{W}) z^m \cdot U_{\mathcal{W}}(z)
\end{aligned}$$

(Extension to Markov sources possible; cf. Regnier & WS.)

Main Results: Exact

Theorem 2. *The generating functions* $T_r(z)$ *and* $T(z,u)$ *are*

$$\begin{aligned} O_r(z) &= R(z)M_{\mathcal{W}}^{r-1}(z)U_{\mathcal{W}}(z)\,, \quad r \geq 1 \\ O(z,u) &= R(z)\frac{u}{1-uM(z)}U_{\mathcal{W}}(z) \\ O_0(z)P(\mathcal{W}) &= R(z)S_{\mathcal{W}}(z) \end{aligned}$$

where

$$\begin{aligned} M(z) &= 1+\frac{z-1}{D_{\mathcal{W}}(z)}\,, \\ U_{\mathcal{W}}(z) &= \frac{1}{D_{\mathcal{W}}(z)}\,, \\ R(z) &= z^m P(\mathcal{W})\frac{1}{D_{\mathcal{W}}(z)}\,. \end{aligned}$$

with

$$D_{\mathcal{W}}(z) = (1-z)S_{\mathcal{W}}(z) + z^m P(\mathcal{W}).$$

Main Results: Asymptotics

Theorem 3. (i) *Moments. The expectation satisfies, for* $n \geq m$*:*

$$\mathbf{E}[O_n(\mathcal{W})] = P(\mathcal{W})(n - m + 1) \ ,$$

while the variance is

$$\mathbf{Var}[O_n(\mathcal{W})] = nc_1 + c_2.$$

with

$$\begin{aligned} c_1 &= P(\mathcal{W})(2S(1) - 1 - (2m-1)P(\mathcal{W}) \ , \\ c_2 &= P(\mathcal{W})((m-1)(3m-1)P(\mathcal{W}) \\ &- (m-1)(2S(1) - 1) - 2S'(1)). \end{aligned}$$

(ii) *Case* $r = O(1)$. *Let* $\rho_{\mathcal{W}}$ *be the smallest root of*

$$D_{\mathcal{W}}(z) = (1-z)S_{\mathcal{W}}(z) + z^m P(\mathcal{W}) = 0.$$

Then

$$\Pr\{O_n(\mathcal{W}) = r\} \quad \sim \quad \sum_{j=1}^{r+1} (-1)^j a_j \binom{n}{j-1} \rho_{\mathcal{W}}^{-(n+j)}$$

where

$$a_{r+1} = \frac{\rho_{\mathcal{W}}^m P(\mathcal{W})\left(\rho_{\mathcal{W}} - 1\right)^{r-1}}{\left(D'_{\mathcal{W}}(\rho_{\mathcal{W}})\right)^{r+1}},$$

and the remaining coefficients can be easily computed, too.

Central Limit and Large Deviations

(iii) *CLT: Case* $r = EO_n + x\sqrt{\mathbf{Var}O_n}$ *for* $x = O(1)$. *Then:*

$$\Pr\{O_n(\mathcal{W}) = r\} = \frac{1}{\sqrt{2\pi c_1 n}} e^{-\frac{1}{2}x^2} \left(1 + O\left(\frac{1}{\sqrt{n}}\right)\right) .$$

(iv) *Large Deviations: Case* $r = (1+\delta)EO_n$. *Let* $a = (1+\delta)P(\mathcal{W})$ *with* $\delta \neq 0$. *For complex* t, *define* $\rho(t)$ *to be the root of*

$$1 - e^t M_{\mathcal{W}}(e^\rho) = 0 ,$$

while ω_a *and* σ_a *are defined as*

$$\begin{aligned} -\rho'(\omega_a) &= a \\ -\rho''(\omega_a) &= \sigma_a^2 \end{aligned}$$

Then

$$\Pr\{O_n(\mathcal{W}) \sim (1+\delta)EO_n\} = \frac{e^{-(n-m+1)I(a)+\delta_a}}{\sigma_a\sqrt{2\pi(n-m+1)}}$$

where $I(a) = a\omega_a + \rho(\omega_a)$ *and* δ_a *is a constant.*

Analysis of a Random Suffix Tree

Recall that:

- For any $i \leq n$ $D_n(i)$ is the largest k such that T_i^{i+k-1} occurs at least twice in the text T_1^n.
- Typical depth D_n is defined as

$$\Pr(D_n = \ell) = \frac{1}{n}\sum_{i=1}^{n}\Pr(D_n(i) = \ell)$$

  for any $1 \leq \ell \leq n$.
- For $w \in \mathcal{A}^k$ let $O_n(w)$ be, as before, the number of times w occurs in the text T_1^n.
- The following is true

$$\Pr(D_n(i) \geq k \;\&\; T_i^{i+k-1} = w) = \Pr(O_n(w) \geq 2 \;\&\; T_i^{i+k-1} = w),$$

  and

$$\sum_{i=1}^{n}\Pr(O_n(w) = r \;\&\; T_i^{i+k-1} = w) = r\Pr(O_n(w) = r).$$

Basic Relationship Between D_n and O_n

Recalling that $O_n(u) = \mathbf{E}[u^{O_n(w)}]$, we have

$$
\begin{aligned}
\Pr(D_n \geq k) &= \frac{1}{n}\sum_{i=1}^{n}\Pr(D_n(i) \geq k) \\
&= \sum_{w\in\mathcal{A}^k}\frac{1}{n}\sum_{i=1}^{n}\Pr(D_n(i) \geq k \;\&\; T_i^{i+k-1} = w) \\
&= \frac{1}{n}\sum_{w\in\mathcal{A}^k}\sum_{r\geq 2} r\Pr(O_n(w) = r) \\
&= \sum_{w\in\mathcal{A}^k}\left(\Pr(w) - \frac{1}{n}O'_{n,w}(0)\right) \\
&= 1 - \frac{1}{n}\sum_{w\in\mathcal{A}^k}\frac{d}{du}\left(\mathbf{E}[u^{O_n(w)}]\right)|_{u=0}
\end{aligned}
$$

where $O'_{n,w}(0)$ denotes the derivative of $O_n(u)$ at $u = 0$

Generating Function of D_n

From previous slide we conclude that the probability generating function

$$D_n(u) = \mathbf{E}[u^{D_n}] = \sum_k \Pr(D_n = k)u^k$$

becomes

$$D_n(u) = \frac{1}{n}\frac{(1-u)}{u}\sum_{w\in\mathcal{A}^*} u^{|w|} O'_{n,w}(0),$$

and the bivariate generating function

$$D(z,u) = \sum_n nD_n(u)z^n$$

is

$$D(z,u) = \frac{1-u}{u}\sum_{w\in\mathcal{A}^*} u^{|w|}\frac{\partial}{\partial u}O_w(z,0)$$

where $O_w(z,u) = \sum_{n=0}^{\infty}\sum_{r=0}^{\infty} \Pr(O_n(w) = r)z^n u^r$.

Pattern Matching and Suffix Tree

In Theorem 1 of pattern matching we derived

$$O_w(z,u) = \frac{z^{|w|}\Pr(w)}{D_w^2(z)} \frac{u}{1-uM_w(z)} + \frac{S_w(z)}{D_w(z)},$$

where

$$\begin{aligned} M_w(z) - 1 &= \frac{z-1}{D_w(z)} \\ D_w(z) &= (1-z)S_w(z) + z^{|w|}\Pr(w) \end{aligned}$$

and $S_w(z)$ is the autocorrelation polynomial for w.

Lemma 2. *The bivariate generating function for D_n is*

$$D(z,u) = \frac{1-u}{u} \sum_{w \in \mathcal{A}^*} (zu)^{|w|} \frac{\Pr(w)}{((1-z)S_w(z) + z^{|w|}\Pr(w))^2}$$

for $|u| < 1$ and $|z| < 1$, where $S_w(z)$ is the autocorrelation polynomial for w.

Main Result on Suffix Tree

Define h to be the entropy and $h_2 = \sum_{i=1}^{V} p_i \log^2 p_i$.

Theorem 4. (i) *For a biased memoryless source (i.e., $p_i \neq p_j$ for some $i \neq j$) and any $\varepsilon > 0$*

$$\begin{aligned} \mathbf{E}D_n &= \frac{1}{h}\log n + \frac{\gamma}{h} + \frac{h_2}{h^2} + P_1(\log n) + O(n^{-\varepsilon}), \\ \mathbf{Var}(D_n) &= \frac{h_2 - h^2}{h^3}\log n + O(1) \end{aligned}$$

where $P_1(\cdot)$ is a periodic function with small amplitude when the tuple $(\log p_1, \ldots,$
$\log p_V)$, is collinear with a rational tuple (i.e., $\log p_j / \log p_1 = r/s$ for some integers r and s) and converges to zero otherwise.
Furthermore, $(D_n - \mathbf{E}[D_n])/\mathbf{Var}(D_n)$ is asymptotically normal with mean zero and variance one that is, for fixed $x \in R$

$$\lim_{n\to\infty} \Pr\{D_n \le \mathbf{E}[D_n] + x\sqrt{\mathbf{Var}(D_n)}\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}dt \ ,$$

and for all integer m

$$\lim_{n\to\infty} \mathbf{E}\left[\frac{D_n - \mathbf{E}[D_n]}{\sqrt{\mathbf{Var}\, D_n}}\right]^m = \begin{cases} 0 & \text{when } m \text{ is odd} \\ \frac{m!}{2^{m/2}(\frac{m}{2})!} & \text{when } m \text{ is even.} \end{cases}$$

(ii) *For the unbiased source (i.e.,* $p_1 = \cdots = p_V = 1/V$*),* $h_2 = h^2$*, the expected value* $\mathbf{E}[D_n]$ *is given above, and for any* $\varepsilon > 0$

$$\mathbf{Var}(D_n) = \frac{\pi^2}{6\log^2 V} + \frac{1}{12} + P_2(\log n) + O(n^{-\varepsilon})$$

where $P_2(\log n)$ *is a periodic function with small amplitude The limiting distribution of* D_n *does not exist, but one finds*

$$\lim_{n\to\infty} \sup_x \mid \Pr(D_n \le x) - \exp(-nV^{-x}) \mid = 0$$

for any fixed real x*.*

Sketch of Proof

1. Consider a trie built from n **independently** generated texts. Let D_n^T be the typical depth in such a trie. Since

$$\Pr(D_n^T(i) < k) = \sum_{w \in \mathcal{A}^k} \Pr(w)(1 - \Pr(w))^{n-1}.$$

we obtain the following formulas for the probability generating function $D_n^T(u) = \mathbf{E}[u^{D_n^T}]$ and the bivariate generating function $D(z, u)$

$$\begin{aligned} D_n^T(u) &= \frac{1-u}{u} \sum_{w \in \mathcal{A}^*} u^{|w|} \Pr(w)(1 - \Pr(w))^{n-1}, \\ D^T(z, u) &= \frac{1-u}{u} \sum_{w \in \mathcal{A}^*} u^{|w|} \frac{z \Pr(w)}{(1 - z + \Pr(w) z)^2} \end{aligned}$$

for all $|u| \le 1$ and $|z| < 1$.

Result on Tries

2. This is known (for independent tries).

Lemma 3 (P. Jacquet, M. Regnier, W.S.). *There exists* $\varepsilon > 0$ *such that*

$$D_n^T(u) = (1-u)n^{\kappa(u)}(\Gamma(\kappa(u)) + P(\log n, u))) + O(n^{\varepsilon}),$$

where Γ *is the Euler gamma function*

$$u \sum_{i=1}^{V} p_i^{1-\kappa(u)} = 1$$

and $P(\log n, u)$ *is periodic function with small amplitude when the vector* $(\log p_1, \ldots, \log p_V)$ *is collinear with a rational tuple, and converges to zero when* $n \to \infty$ *otherwise.*

This lemma implies

$$e^{-tc_1 \log n/\sqrt{c_2 \log n}} D_n^T\left(e^{t/\sqrt{c_2 \log n}}\right) \to e^{t^2/2}$$

where $c_1 = 1/h$ and $c_2 = (h_2 - h^2)/h^3$.

That is,

$$\frac{D_n^T - c_1 \log n}{\mathbf{Var}[c_2 \log n]} \to N(0,1).$$

Our Goal is ...

Our goal now is to prove that $D_n(u)$ and $D_n^T(u)$ are asymptotically close as $n \to \infty$.

3. We accomplish this by proving that for some $\varepsilon > 0$ and all $|u| < \beta$ for $\beta > 1$

$$D_n^T(u) - D_n(u) = (1-u)O(n^{-\varepsilon}),$$

that is,

$$|\Pr(D_n \leq k) - \Pr(D_n^T \leq k)| = O(n^{-\varepsilon}\beta^{-k}).$$

for all positive integer k.

Autocorrelation Polynomial

4. For almost all words w the autocorrelation polynomial

$$S_w(z) \approx 1.$$

Lemma 4. *There exist* $\delta < 1, \theta > 0$ *and* $\rho > 1$ *such that* $\rho\delta < 1$ *and*

$$\sum_{w \in \mathcal{A}^k} [\![|S_w(\rho) - 1| \leq (\rho\delta)^k \theta]\!] \Pr(w) \geq 1 - \theta\delta^k.$$

Based on the following observation: in order for $w \in \mathcal{A}$ to have the same prefix of length $k - i$ (for some i) as the suffix of length $k - i$, that is,

$$w_1^{k-i} = w_{k-i+1}^k$$

pattern w must have a periodic structure, that is, for some $u \in \mathcal{A}^i$ we have

$$w_1^k = \underbrace{u \cdot u \cdots u}_{l} \bar{u}$$

where $\bar{u} \leq i$.

5. The bivariate generating function $D(z, u)$ can be analytically continued to a larger disk.

Lemma 5. *The generating function $D(z, u)$ can be analytically continued for all $|u| < \delta^{-1}$ and $|z| < 1$ where $\delta < 1$.*

We prove this lemma by showing

$$uD(z, u) - \frac{(1-u)}{(1-uz)(1-z)^2} = O\left(\frac{u-1}{1-\delta|u|}\right)$$

for $\delta < 1$ and $|z| < 1$

Finishing Up ...

6. Define

$$Q_n(u) = \frac{u}{1-u}\Big(D_n(u) - D_n^T(u)\Big)\,,$$

and

$$Q(z,u) = \sum_{n=0}^{\infty} n\, Q_n(u) z^n = \frac{u}{1-u}\Big(D(z,u) - D^T(z,u)\Big)\,.$$

That is,

$$Q(z,u) = \sum_w u^{|w|}\mathrm{Pr}(w)\left(\frac{z^{|w|}}{D_w(z)^2} - \frac{z}{(1-z+\mathrm{Pr}(w)z)^2}\right)\,.$$

Lemma 6. *For all $1 < \beta < \delta^{-1}$, there exists $\varepsilon > 0$ such that*

$$Q_n(u) = (1-u)O(n^{-\varepsilon})$$

uniformly for $|u| \le \beta$.