# On the Average Complexity of the Membership Problem for a Generalized Dyck Language

Markus E. Nebel
Johann Wolfgang Goethe-Universität, Frankfurt
Fachbereich Informatik
D-60054 Frankfurt am Main
Germany
e-mail: nebel@sads.informatik.uni-frankfurt.de

**Abstract**

A general concept for analysing the average complexity of the membership problem for any formal language is used in order to examine a generalization of the Dyck language. Our investigation is motivated by the fact that the Dyck language has a distinguished behaviour concerning that parameter. Surprisingly, that behaviour is lost even by small variations without utilising any opposite controls, e.g. adapting probabilities. This observation supports the significance of the Dyck language in computer science.

## 1 Introduction and Fundamental Definitions

In [Kem96] a general concept for calculating the average complexity of the membership problem for any formal language $\mathcal{L} \subseteq T^\star$ has been introduced. In order to decide whether or not a given word $w$ belongs to $\mathcal{L}$ the following assumptions are made:
Assume $l(w)$ is the length of input $w$. Scan $w$ from left to right letter by letter until a prefix $v$ is read which has no extension rightwards to any word of length $l(w)$ belonging to $\mathcal{L}$. If such a $v$ exists we have $w \notin \mathcal{L}$ and only $l(v) \leq l(w)$ symbols were read.
Assume $P_{\mathcal{L}}^{(n)}(v)$ for $(v,n) \in T^\star \times I\!N$ is a predicate with the value true if there is a $u \in T^\star$ with $vu \in \mathcal{L}_n := \mathcal{L} \cap T^n$ and the value false if such a suffix $u$ does not exist. Then we obtain the following formal recognition procedure MEMBER:

Input: $n \in I\!N_0$; $w = a_1 a_2 \cdots a_n$, $a_i \in T$, $1 \leq i \leq n$; $P_{\mathcal{L}}^{(n)}$
Output: $w \in \mathcal{L}$ or $w \notin \mathcal{L}$
Method: $i := 0$; $v := \varepsilon$; /* $\varepsilon$ denotes the emty word with $l(\varepsilon) = 0$ */
**while** $(i < n)$ **and** $(P_{\mathcal{L}}^{(n)}(v)$=true$)$ **do begin**
$i := i + 1$;
$v := va_i$;
**end**;
**if** $(l(v) = n)$ **and** $(P_{\mathcal{L}}^{(n)}(v)$=true$)$ **then** $w \in \mathcal{L}$ **else** $w \notin \mathcal{L}$;

We assume $T$ to be minimal with respect to $\mathcal{L}$, i.e. $(\forall \hat{T} \subset T) : (\mathcal{L} \not\subseteq \hat{T}^\star)$. To any symbol $a \in T := \{a_1, a_2, \ldots\}$ we associate the probability $p_a$ with $\sum_{a \in T} p_a = 1$. Further, we let $\vec{\rho} := (p_{a_1}, p_{a_2}, \ldots)$ be the corresponding probability distribution. For any word $v \in T^\star$ we denote by $\#_a(v)$ the number of appearances of $a$ in $v$. Now, if $\mathsf{INIT}(\mathcal{L}) := \{v \in T^\star | (\exists u \in T^\star) : (vu \in \mathcal{L})\}$ and $\mathsf{INIT}_k(\mathcal{L}) := \mathsf{INIT}(\mathcal{L}) \cap T^k$ the following theorem can be found in [Kem96]:

**Theorem 1** *Let $Y_{\mathsf{pref}}(\mathcal{L}_n)$ be the random variable describing the length of the shortest prefix which the procedure* MEMBER *has to read in order to decide whether or not an input word $w \in T^n$ belongs to the given language $\mathcal{L} \subseteq T^\star$. Then, the $s$-th moment about the origin of the random variable $Y_{\mathsf{pref}}(\mathcal{L}_n)$ is:*

$$\mathbb{E}[Y^s_{\mathsf{pref}}(\mathcal{L}_n)] = \sum_{0 \le k < n} [(k+1)^s - k^s] \sum_{v \in \mathsf{INIT}_k(\mathcal{L}_n)} \prod_{a \in T} p_a^{\#_a(v)}.$$

*If all symbols are equally likely we have*

$$\mathbb{E}[Y^s_{\mathsf{pref}}(\mathcal{L}_n)] = \sum_{0 \le k < n} [(k+1)^s - k^s] |\mathsf{INIT}_k(\mathcal{L}_n)| \, |T|^{-k}.$$

□

Note, that in the original work a more general model was considered in which the input word $w$ might be in $\Sigma^\star$, $\Sigma \neq T$. But since the interesting cases[1] can all be reduced to the case $\Sigma = T$ (by adapting $\vec{p}$) we restrict ourselves to the model as given above.
It was shown that for any $\mathcal{L} \in T^\star$, $d_{\vec{p}}(\mathcal{L}) := \lim_{n\to\infty} \sum_{w \in \mathcal{L}_n} \prod_{a \in T} p_a^{\#_a(w)} \neq 0$ implies the $s$-th moment to be $\Theta(n^s)$.
One of the applications of the general Theorem which R. Kemp considered was the *semi Dyck language* $D \subset \{[,]\}^\star$ for one type of brackets (e.g.[Har78] pp.312). He was able to show that for $(p,q) := (p_[, p_])$ the average length of the shortest prefix has the asymptotical behaviour

$$\mathbb{E}[Y_{\mathsf{pref}}(D_{2n})] \quad \sim \quad \begin{cases} (1-2p)^{-1} & : \quad p < \frac{1}{2} \\ 4\pi^{-\frac{1}{2}} n^{\frac{1}{2}} & : \quad p = \frac{1}{2} \\ (2p-1)p^{-2}n & : \quad p > \frac{1}{2} \end{cases} . \qquad (1)$$

It should be mentioned that the language $D$ is the only known context-free language having a non-linear and non-constant average behaviour of the membership problem provided that all words of length $2n$ are equally likely (second alternative in (1)). In [Kem97] the minimal prefix length of the following generalization $D^{\Re}$ of $D$ was considered. Let $T_[$ and $T_]$ be two disjoint alphabets and $\Re \subseteq T_[ \times T_]$ be a relation. A tuple $(x,y) \in \Re$ identifies $xy$ to be a pair of corresponding brackets. Since there is no restriction on $\Re$ one bracket may correspond to several others contrary to the ordinary Dyck language. The language $D^{\Re}$ consists of all words $w \in (T_[ \cup T_])^\star$ which are equivalent to the empty word under the congruence $\delta$ defined by $xy \equiv \varepsilon \bmod \delta$ for all $(x,y) \in \Re$, i.e. all words correctly bracketed in the above sense. If $T := T_[ \cup T_]$ and $\Re_1 := \{x \in T_[ | (\exists y \in T_])((x,y) \in \Re)\}$ then for $p := |\Re_1||T|^{-1}$ and $q := |\Re|(|\Re_1||T|)^{-1}$ the behaviour of (1) was rediscovered for the special case $p + q = 1$.

The author of this paper therefore expected to observe interesting effects by considering the following generalization of the language $D$.

**Definition 1** *Let $m, n$ be two natural numbers and let $\phi^{(m,n)}$ be the monoidhomomorphism induced by $\phi^{(m,n)}([) := [^m$ and $\phi^{(m,n)}(]) :=]^n$. The* generalized semi Dyck language *$D^{(m,n)}$ is defined as*

$$D^{(m,n)} := \{\phi^{(m,n)}(w) | w \in D\}.$$

It is obvious that $D^{(1,1)} = D$ holds. Further, we have $d_{\vec{\rho}}(D^{(m,n)}) = 0$ for any choice of $m$ and $n$ and any probability distribution $\vec{p}$. As we will see $|D^{(m,n)}_{(m+n)\ell}| =$

[1] There are cases in which it is known a priori that all moments are bounded by a constant.

$\frac{1}{\ell+1}\binom{2\ell}{\ell}$. Thus, we have $\lim_{\ell\to\infty}\sum_{w\in D^{(m,n)}_{(m+n)\ell}} p_[^{n\ell}p_]^{m\ell} = \lim_{\ell\to\infty}\frac{1}{\ell+1}\binom{2\ell}{\ell}p_[^{n\ell}(1-p_[)^{m\ell} \leq \lim_{\ell\to\infty}\frac{1}{\ell+1}\binom{2\ell}{\ell}(\frac{n}{m+n})^{n\ell}(\frac{m}{m+n})^{m\ell} \leq \lim_{\ell\to\infty}\frac{1}{\ell+1}\binom{2\ell}{\ell}2^{-2\ell} = 0$ by Stirling's formula. The larger we choose $m$ and $n$ the faster that limit converges to zero.

# 2 The Average Complexity of the Membership Problem

There is a well known one-to-one correspondence (e.g. [Kem84] p.173) between Dyck words in $D$ of length $2\ell$ and the walks from $(0,0)$ to $(2\ell,0)$ consisting of steps $\nearrow$ (representing a [ ) and $\searrow$ (representing a ] ) only, (see Figure 1). The number of prefixes $v \in \mathsf{INIT}(D)$ consisting of $i$ opening and $j$ closing brackets is equal to the number of different paths from $(0,0)$ to $(i+j, i-j)$ in the corresponding walk. It is well known (see [CRS71] and [Rio79] p.130) that this number is equal to the ballot number $w_{i,j}$ defined by the recurrence:

$$w_{i,j} := \begin{cases} w_{i,j-1} & : \quad i = j \\ w_{i,j-1} + w_{i-1,j} & : \quad i > j \\ 1 & : \quad i = 0,\ j = 0 \end{cases} . \tag{2}$$

Explicitly, we have $w_{i,j} := \binom{i+j}{j} - \binom{i+j}{j-1} = \frac{i+1-j}{i+1}\binom{i+j}{j}$. Counting opening brackets by $x$ and closing brackets by $y$ and summing over all positions of the walk for $D_{2\ell}$ yields the following generating function:

$$D(\ell, x, y) := \sum_{i=0}^{\ell}\sum_{j=0}^{i} w_{i,j}x^i y^j. \tag{3}$$

Recall that for $(p,q) := (p_[, p_])$ Theorem 1 implies

$$\mathbb{E}[Y_{\mathsf{pref}}(D_{2\ell})] = \sum_{\substack{0\leq k<2\ell \\ v\in \mathsf{INIT}_k(D_{2\ell})}} \prod_{a\in T} p_a^{\#_a(v)} = D(\ell,p,q) - p^\ell q^\ell w_{\ell,\ell}.$$

Now, examine the language $D^{(m,n)}$. In Figure 2 the diagram for $D^{(2,3)}_{(2+3)5}$ is drawn. To consider the structure of $D^{(m,n)}$ a $\nearrow$ (resp. $\searrow$) may only appear in sequence

Figure 2: In the generalized version the number of Dyck words of length $(m+n)\ell$ corresponds to the number of paths of the given structure from $(0,0)$ to $((m+n)\ell, (m-n)\ell)$, here $(m,n,\ell) = (2,3,5)$.

with $m-1$ (resp. $n-1$) other ones. Let $\hat{w}^{(m,n)}_{i,j}$ denote the number of prefixes of $D^{(m,n)}$ consisting of $i$ opening and $j$ closing brackets. It is obvious that $\hat{w}^{(m,n)}_{im,jn} = w_{i,j}$ holds. In Figure 2 the corresponding points are marked by a $\bullet$; these points represent a stretched diagram for $D_{10}$ (generally $D_{2\ell}$) where the coordinates were transformed with respect to $m$ and $n$. Those points without a $\bullet$ have exactly one predecessor from which they inherit the number of related prefixes, i.e. $\hat{w}^{(m,n)}_{im+\alpha,jn} = \hat{w}^{(m,n)}_{im,jn+\beta} = \hat{w}^{(m,n)}_{im,jn}$, $1 \leq \alpha < m$, $1 \leq \beta < n$, since the number of paths from $(0,0)$ to such a point is equal to that one from $(0,0)$ to its predecessor $\bullet$. This observation enables us to construct the walks (and thus the generating function) of $D^{(m,n)}$ by multiple overlayed and scaled walks for $D$. In Figure 3 one can see how this is done in the case of $D^{(2,3)}_{25}$ (generally $D^{(m,n)}_{(m+n)\ell}$). Points marked by the same symbol (filled or unfilled) belong to the same (stretched) diagram for $D$, which is stretched by setting $x$ to $x^m$ and $y$ to $y^n$ in the generating function. Note, that there are positions marked by $\blacktriangle$ and $\blacksquare$ not lying on the dotted structure of the $D^{(2,3)}_{25}$-walk. However, in order to generate the right number of different paths we have to consider walks of size 10 (generally $2\ell$) for them which include the positions marked by $\blacktriangle$ and $\blacksquare$.

Figure 3: How to construct the walk for $D^{(m,n)}$ by overlayed walks for $D$, $(m,n) = (2,3)$.

These positions have to be considered by an error term. This is not necessary for $\diamond$ where a walk consisting only of dotted positions is sufficient to obtain the desired result. In order to place the (stretched) walks ($\diamond$, $\triangle$, $\square$) at the right positions they have to be moved in the $\nearrow$ ($\searrow$) direction. This movement corresponds to a multiplication of the generating function by $x^i$ ($y^i$) if $i$ is the distance between $(0,0)$ and the starting-point of the desired walk. Now let $D^{(m,n)}(\ell,x,y)$ be the generating function counting the elements in $\mathsf{INIT}(D^{(m,n)}_{(m+n)\ell})$. By translating our observations into generating functions we get:

$$\begin{aligned} D^{(m,n)}(\ell,x,y) &= \underbrace{D(\ell,x^m,y^n)}_{=\bullet} + \underbrace{\sum_{k=1}^{m-1} x^k D(\ell-1,x^m,y^n) +}_{=\diamond} \\ &\quad + \underbrace{\sum_{k=1}^{n-1} y^k [D(\ell,x^m,y^n)}_{=\triangle,\square,\blacktriangle,\blacksquare} - \underbrace{\sum_{j=0}^{\ell} w_{j,j} x^{jm} y^{jn}]}_{=\blacktriangle,\blacksquare}. \end{aligned} \tag{4}$$

Note, that this is a general method for constructing the generating function for a stretched diagram if that one for the unstretched one is known.

Now, it is not hard to see that $D^{(m,n)}(\ell,p,q) = \mathbb{E}[Y_{\mathsf{pref}}(D^{(m,n)}_{(m+n)\ell})] + p^{m\ell} q^{n\ell} \hat{w}^{(m,n)}_{\ell m,\ell n}$ holds. If $[z^n]f(z)$ denotes the coefficient of $z^n$ in the series expansion of $f(z)$ at $z=0$, the number of $|\mathsf{INIT}_k(D^{(m,n)}_{(m+n)\ell})|$is given by $[z^k]D^{(m,n)}(\ell,z,z)$ and we get by (4) the explicit form

$$\begin{aligned} &|\mathsf{INIT}_k(D^{(m,n)}_{(m+n)\ell})| = \\ &\qquad \sum_{\substack{i \geq \lceil k/(m+n) \rceil \\ n \mid (k-mi)}}^{\ell} w_{i,(k-mi)/n} + \sum_{v=1}^{m-1} \sum_{\substack{i \geq \lceil (k-v)/(m+n) \rceil \\ n \mid (k-v-mi)}}^{\ell-1} w_{i,(k-v-mi)/n} + \sum_{v=1}^{n-1} \sum_{\substack{i > \lfloor (k-v)/(m+n) \rfloor \\ n \mid (k-v-mi)}}^{\ell} w_{i,(k-v-im)/n}. \end{aligned}$$

This expression is not quite handy for further calculations (i.e. for inserting it into the second formula of Theorem 1). So, we return to formula (4). By simple algebraic manipulations and the application of (3) it can be transformed into:

$$\begin{aligned} \sum_{j=0}^{\ell} y^{nj} \sum_{i=j}^{\ell} w_{i,j} x^{mi} &+ \left( \frac{x^m - 1}{x-1} - 1 \right) \sum_{j=0}^{\ell-1} y^{nj} \sum_{i=j}^{\ell-1} w_{i,j} x^{mi} \\ &+ \left( \frac{y^n - 1}{y-1} - 1 \right) \sum_{j=0}^{\ell} y^{nj} \sum_{i=j+1}^{\ell} w_{i,j} x^{mi}. \end{aligned}$$

As stated above this generating function differs from our expected value only by $x^{m\ell} y^{n\ell} w_{\ell,\ell}$ for $x := p_[$ and $y := p_]$. Subtracting this term and changing the order of summation yields the following theorem:

**Theorem 2** *Let $\delta$ denote Kronecker's symbol, then*

$$\mathbb{E}[Y_{\mathsf{pref}}(D^{(m,n)}_{(m+n)\ell})] = \hat{D}^{(m,n)}(\ell,x,y)\Big|_{(x,y):=(p_[,p_])},$$

*where*

$$\hat{D}^{(m,n)}(\ell,x,y) := \sum_{j=0}^{\ell-1} y^{nj} \sum_{i=j}^{\ell} x^{mi} \left( \left[ \sum_{k=0}^{m-1} x^k \right]^{(1-\delta_{\ell,i})} + (1-\delta_{i,j}) \sum_{k=1}^{n-1} y^k \right) w_{i,j}.$$

□

In order to get more information about the mean of $Y_{\mathsf{pref}}$ we need a closed form representation of that generating function. A first step is to regard the difference $\Delta^{(m,n)}(\ell,x,y) := \hat{D}^{(m,n)}(\ell+1,x,y) - \hat{D}^{(m,n)}(\ell,x,y)$. We have

$$\begin{aligned}
&\hat{D}^{(m,n)}(\ell+1,x,y) = \\
&\quad \sum_{j=0}^{\ell} y^{nj} \sum_{i=j}^{\ell+1} x^{mi} \left( \left[ \sum_{k=0}^{m-1} x^k \right]^{(1-\delta_{\ell+1,i})} + (1-\delta_{i,j}) \sum_{k=1}^{n-1} y^k \right) w_{i,j} \\
= \quad & \sum_{j=0}^{\ell-1} y^{nj} \sum_{i=j}^{\ell+1} x^{mi} \left( \left[ \sum_{k=0}^{m-1} x^k \right]^{(1-\delta_{\ell+1,i})} + (1-\delta_{i,j}) \sum_{k=1}^{n-1} y^k \right) w_{i,j} \\
&\quad + y^{n\ell} \left[ x^{m\ell} \sum_{k=0}^{m-1} x^k w_{\ell,\ell} + x^{m(\ell+1)} \sum_{k=0}^{n-1} y^k w_{\ell+1,\ell} \right] \\
= \quad & \sum_{j=0}^{\ell-1} y^{nj} \sum_{i=j}^{\ell} x^{mi} \left( \left[ \sum_{k=0}^{m-1} x^k \right]^{(1-\delta_{\ell,i})} + (1-\delta_{i,j}) \sum_{k=1}^{n-1} y^k \right) w_{i,j} \\
&\quad + \sum_{j=0}^{\ell-1} y^{nj} x^{m\ell} \sum_{k=1}^{m-1} x^k w_{\ell,j} + \sum_{j=0}^{\ell-1} y^{nj} x^{m(\ell+1)} \sum_{k=0}^{n-1} y^k w_{\ell+1,j} \\
&\quad + y^{n\ell} \left[ x^{m\ell} \sum_{k=0}^{m-1} x^k w_{\ell,\ell} + x^{m(\ell+1)} \sum_{k=0}^{n-1} y^k w_{\ell+1,\ell} \right].
\end{aligned}$$

Thus, $\Delta^{(m,n)}(\ell,x,y)$ turns out to be:

$$\frac{x^m-1}{x-1} x^{m\ell} \sum_{j=0}^{\ell} y^{nj} w_{\ell,j} + \frac{y^n-1}{y-1} x^{m(\ell+1)} \sum_{j=0}^{\ell} y^{nj} w_{\ell+1,j} - x^{m\ell} \sum_{j=0}^{\ell-1} y^{nj} w_{\ell,j}.$$

So, we have two different but similar kinds of sums $\sum_{0\le i\le \ell} c^i w_{\ell,i}$ and $\sum_{0\le i\le \ell} c^i w_{\ell+1,i}$ ($c$ constant) which we will examine in detail.

**Lemma 1** *Let $t(\ell,c) := \sum_{0\le i\le \ell} c^i w_{\ell,i}$ and $s(\ell,c) := \sum_{0\le i\le \ell} c^i w_{\ell+1,i}$. The following recurrence relations hold:*

$$\begin{aligned}
t(0,c) &= 1, \\
t(\ell,c) &= \frac{t(\ell-1,c)}{1-c} - \frac{c^{\ell+1} w_{\ell,\ell}}{1-c}, \\
\\
s(0,c) &= 1, \\
s(\ell,c) &= \frac{s(\ell-1,c)}{1-c} + \frac{1}{1-c} \left[ c^\ell \frac{1}{\ell+1} \binom{2\ell}{\ell} - c^{\ell+1} w_{\ell+1,\ell} \right].
\end{aligned}$$

**Proof:** It is obvious that $t(0,c)=1$ holds. For $t(\ell,c)$ we consider $(1-c)t(\ell,c) - t(\ell-1,c)$. The application of (2) proves $\sum c^i w_{\ell,i} - (w_{\ell,i-1} + w_{\ell-1,i})$ to be zero and the whole expression evaluates to $-c^{\ell+1} w_{\ell,\ell}$. The proof for $s(\ell,c)$ can be performed in an analogous way. □

Now, defining $T_c(z) := \sum_{i\ge 0} t(i,c) z^i$ yields

$$T_c(z) = \frac{1}{1-c} \sum_{i\ge 1} \left[ t(i-1,c) - c^{i+1} \underbrace{\frac{1}{i+1} \binom{2i}{i}}_{w_{i,i}} \right] z^i + 1$$

$$
\begin{aligned}
&= \frac{1}{1-c}\sum_{i\geq 1} t(i-1,c)z^i - \frac{c}{1-c}\sum_{i\geq 1}\frac{1}{i+1}\binom{2i}{i}(cz)^i + 1\\
&= \frac{z}{1-c}T_c(z) - \frac{c}{1-c}\left[\frac{1-\sqrt{1-4cz}}{2cz} - 1\right] + 1
\end{aligned}
$$

By defining $S_c(z) := \sum_{i\geq 0} s(i,c)z^i$ we find after a similar computation the following lemma.

**Lemma 2**

$$
\begin{aligned}
T_c(z) &= \frac{1-\sqrt{1-4cz}}{2z(c-1)(1-z/(1-c))} + \frac{1-c/(c-1)}{1-z/(1-c)},\\
S_c(z) &= \frac{1}{1-z/(1-c)}\left[\frac{1-z^{-1}}{1-c}\frac{1-\sqrt{1-4cz}}{2cz} + \frac{z^{-1}}{1-c}\right].
\end{aligned}
$$

□

By Lemma 2 and $\mathcal{D}^{(m,n)}(z,x,y) := \sum_{i\geq 0}\Delta^{(m,n)}(i,x,y)z^i$ we have a closed form expression:

$$
\mathcal{D}^{(m,n)}(z,x,y) = \frac{x^m-1}{x-1}T_{y^n}(zx^m) + \frac{y^n-1}{y-1}x^m S_{y^n}(zx^m) - zx^m S_{y^n}(zx^m). \qquad (5)
$$

Note that the variable $z$ is related to the length $\ell$, i.e. $[z^\ell]\mathcal{D}^{(m,n)}(z,x,y) = \hat{D}^{(m,n)}(\ell+1,x,y) - \hat{D}^{(m,n)}(\ell,x,y) = \Delta^{(m,n)}(\ell,x,y)$.
In order to get information about the parameter in question, i.e. the average prefix length necessary for deciding the membership problem, we have to recombine the generating function for the difference into one for the entire problem. A moment's reflection shows that the multiplication of (5) by $z/(1-z)$ solves this problem since

$$
\begin{aligned}
&\sum_{i\geq 0}\Delta^{(m,n)}(i,x,y)z^i =\\
&\quad \left[\hat{D}^{(m,n)}(1,x,y) - \hat{D}^{(m,n)}(0,x,y)\right]z^0 + \left[\hat{D}^{(m,n)}(2,x,y) - \hat{D}^{(m,n)}(1,x,y)\right]z^1 +\\
&\quad + \left[\hat{D}^{(m,n)}(3,x,y) - \hat{D}^{(m,n)}(2,x,y)\right]z^2 + \ldots\\
&= -\underbrace{\hat{D}^{(m,n)}(0,x,y)}_{=0} + \hat{D}^{(m,n)}(1,x,y)(1-z) + \hat{D}^{(m,n)}(2,x,y)z(1-z) + \ldots\\
&= \frac{1-z}{z}\sum_{i\geq 1}\hat{D}^{(m,n)}(i,x,y)z^i = \frac{1-z}{z}\sum_{i\geq 0}\hat{D}^{(m,n)}(i,x,y)z^i.
\end{aligned}
$$

Now, everything is prepared to establish our main theorem.

**Theorem 3** *The average minimal prefix length $\mathbb{E}[Y_{\mathsf{pref}}(D^{(m,n)}_{(m+n)\ell})]$ needed to decide the membership problem by means of the procedure* MEMBER *for input words $w \in D^{(m,n)}$ of length $(m+n)\ell$ is equal to the coefficient of $z^\ell$ in the expansion of*

$$
\frac{z}{1-z}\left[\frac{p^m-1}{p-1}T_{q^n}(zp^m) + \left(\frac{q^n-1}{q-1} - z\right)p^m S_{q^n}(zp^m)\right].
$$

□

*Here, closed-form expressions for $T_c(z)$ and $S_c(z)$ are stated in Lemma 5.*

We conclude this section by determining asymptotics from the coefficients of the generating function given in Theorem 3. To do this, we have to consider two

Figure 4: The different expansions of our generating function.

Figure 5: Exact [in roman] and asymptotical values [in italics] for $\mathbb{E}[Y_{\mathsf{pref}}(D^{(m,n)}_{(m+n)\ell})]$, $q := 1 - p$.

different cases. The first case is that of $m > 1 \vee n > 1$, i.e. any of our generalizations. We divide our function into part $B(z) := \frac{1}{1-z}$ and part $A(z) := z\left[\frac{p^m-1}{p-1}T_{q^n}(zp^m) + \left(\frac{q^n-1}{q-1} - z\right)p^m S_{q^n}(zp^m)\right]$. Obviously, the radius of convergence $\rho(B)$ is equal to 1. Since the absolute value of all non-zero singularities of $A(z)$ is greater than 1 and further $A(\lim_{n\to\infty}(b_{n-1}/b_n)) = A(1) \neq 0$ we meet the conditions of Theorem 4.8 of [Kem84] or [Odl95] which proves that our number in question is $\sim A(1)$, $\ell \to \infty$.
For the case $m = n = 1$, i.e. the ordinary Dyck language, it is possible to use Darboux's Theorem (see [Kem84],[Odl95] or [GrKn82] for a more detailed discussion). In Figure 4 the corresponding expansions of our generating function can be found. The repeated application of that theorem to our expansions and using the relation $\Gamma(s + \frac{1}{2}) = \pi^{\frac{1}{2}}(2s)!4^{-s}s!^{-1}$ satisfied by the complete gamma function (e.g. [AbSt70]) rediscovers (1). Altogether we have the

**Theorem 4** *The average minimal prefix length needed to decide the membership problem for the language $D^{(m,n)}$ is asymptotically given by*

$$\mathbb{E}[Y_{\mathsf{pref}}(D^{(m,n)}_{(m+n)\ell})] \sim \left\{ \begin{array}{lcll} (1-2p)^{-1} & : & p < \frac{1}{2} & \\ 4\pi^{-\frac{1}{2}}\ell^{\frac{1}{2}} + 2\pi^{-\frac{1}{2}}\ell^{-\frac{1}{2}} - 2 & : & p = \frac{1}{2} & \left.\begin{array}{l} \\ \\ \end{array}\right\} \; m = 1, \\ (2p-1)p^{-2}\ell + (1-p)^2(2p-1)^{-1}p^{-2} & : & p > \frac{1}{2} & \left.\begin{array}{l} \\ \\ \end{array}\right\} \; n = 1 \\ \frac{p^m-1}{p-1}T_{q^n}(p^m) + \left(\frac{q^n-1}{q-1} - 1\right)p^m S_{q^n}(p^m) & : & m > 1 \vee n > 1, & \end{array} \right.$$

*for $\ell \to \infty$.*
*Here, $T_c(z)$ and $S_c(z)$ are established in Lemma 5.* □

Some exact values of our parameter and their asymptotical equivalents for different values of $p, q, m$ and $n$ are given in Figure 5.

# 3 Comments on the Result

Our asymptotical result shows that the behaviour of the membership problem for any of the cases $m > 1$ or $n > 1$ is always constant. So, the interesting behaviour of $D$ was lost. This is an engrossing fact since it also happens for only small variations, e.g. for $m = 1$ and $n = 2$. In those cases the author would have expected a possibility to get at least a linear or sublinear behaviour by means of the probabilities $p_[$ and $p_]$. The lack of any opposite controls shows how sensitive the structure of Dyck words reacts concerning our parameter. Empirical studies have shown that $m < 1$ or $n < 1$[†] leads to an exponentially growing average prefix length (implied by singularities $< 1$). This might explain why $D$ still is the only known context-free languages with a sublinear minimal prefix length. However, our results support the outstanding significance of the Dyck language in computer science.
Further investigations might regard a generalization in which several values for the number of consecutive ['s and ]'s are allowed. In one way this would be a unary coding of multiple types of brackets, e.g. one might interpret [[[[ as ( and [[[[[ as

[†] without any language-theoretic interpretation.

{. However, the unary coding differs from a really extended bracket-alphabet as considered in [Kem97] since there is an influence on the membership problem by the different length of the codes whereas every bracket of an extended bracket-alphabet has the same length, namely 1.

# References


[AbSt70] M. Abramowitz and A. Stegun, *Handbook of Mathematical Functions*, Dover (1970)

[CRS71] L.Carlitz, D. P. Roselle and R. A. Scoville, *Some Remarks on Ballot-Type Sequences of Positive Integers*, J. Comb. Theory (A) **11** (1971) 258-271

[GrKn82] D. H. Greene and D. E. Knuth, *Mathematics for the Analysis of Algorithms*, Birkhäuser (1982)

[Har78] M. A. Harrison, *Introduction to Formal Languages*, Addison-Wesley (1978)

[Kem84] R. Kemp, *Fundamentals of the Average Case Analysis of Particular Algorithms,* Wiley-Teubner Series in Computer Science, Wiley (1984)

[Kem96] R. Kemp, *On Prefixes of Formal Languages and Their Relation to the Average-Case Complexity of the Membership Problem*, Journal of Automata, Languages and Combinatorics **1** (4) (1996) 259-303

[Kem97] R. Kemp, *On the Average Minimal Prefix-Length of the Generalized Semi-Dycklanguage*, RAIRO Theoretical Informatics and Applications, to appear

[Odl95] A. Odlyzko, Asymptotic Enumeration Methods, in: *Handbook of Combinatorics*, Chap. 22, Elsevier (1995)

[Rio79] J. Riordan, *Combinatorial Identities*, Robert E. Krieger Publishing Company (1979)