Discrete Mathematics and Theoretical Computer Science DMTCS vol. **22**:3, 2021, #2

Unary profile of lambda terms with restricted De Bruijn indices[*]

Katarzyna Grygiel and Isabella Larcher

Technische Universität Wien, Austria

received 2019-10-14, *revised 2020-08-11*, *accepted 2021-01-13*.

In this paper we present an average-case analysis of closed lambda terms with restricted values of De Bruijn indices in the size model where each occurrence of a variable contributes one to the size. Given a fixed integer k, a lambda term in which all De Bruijn indices are bounded by k has the following shape: It starts with k De Bruijn levels, forming the so-called hat of the term, to which some number of k-colored Motzkin trees are attached. By means of singularity analysis, we show that the size of this hat is constant on average and that the average number of De Bruijn levels of k-colored Motzkin trees of size n is asymptotically $\Theta(\sqrt{n})$. Combining these two facts, we conclude that, for all $k \geq 1$, the maximal non-empty De Bruijn level in a lambda term with De Bruijn indices at most k and of size n is, on average, also of order $\sqrt{n}$. On this basis, we provide the average unary profile of such lambda terms.

Keywords: profile of lambda terms, singularity analysis, lambda terms with restrictions

1 Introduction

The last decade has seen an abundance of studies on a quantitative analysis of objects originating from logic and computability theory. From a combinatorial viewpoint, these objects provide intriguing asymptotic and stochastic problems related to counting them and an average-case analysis of their parameters. One of these are lambda terms, central objects of lambda calculus, which are investigated in this paper. Due to a simple combinatorial specification of lambda terms, no knowledge concerning lambda calculus is required to understand statements and proofs of the presented theorems. For more information on lambda calculus we refer a curious reader to Barendregt (1984).

In the literature one can find several different ways to define the size of a lambda term. In this paper we adapt the definition that is probably the most intuitive for combinatorialists, namely every constructor in a term (*i.e.*, each variable, abstraction, and application) contributes one to the term size. This size model gives rise to a challenging and still open problem on the asymptotics of the sequence of the number of closed terms of a given size (for a thorough discussion see Bodini et al. (2013)). The encountered difficulties lead Bodini et al. (2018) to study restricted classes of terms, namely terms with a bounded number of De Bruijn levels and terms with bounded De Bruijn indices (precise definitions of these notions

[*] This research has been supported by the Austrian Science Fund (FWF) grant SFB F50-03.

are provided in Section 2). Gittenberger and Larcher (2018) studied some statistical properties of lambda terms with a bounded number of De Bruijn levels. Their results, exhibiting a change in the distribution of leaves within terms, shed some light on reasons for the strange behaviour of the counting sequences.

While the problem of counting terms of a given size is also open for the model where variables do not contribute to the size, David et al. (2013) provided some results concerning typical parameters of closed lambda terms with no additional restrictions. The applied methods, however, seem not to work in the case of the size of each variable being one.

Another way of measuring terms consists in taking into account the reference depth of each variable, *i.e.*, the number of abstractions enclosing it. This approach was motivated by Tromp (2007) and further studied by Grygiel and Lescanne (2015); Bendkowski et al. (2016) with the main focus on enumeration of terms and Bendkowski et al. (2019) describing average values of several parameters in terms.

In this paper, in a similar vein to the work by Gittenberger and Larcher (2018), we perform an average-case analysis of lambda terms with bounded De Bruijn indices. This class has a significantly weaker restriction compared to restricting the number of De Bruijn levels. On the basis of an empirical investigation on existing Haskell programs and their lambda calculus counterparts, we claim that imposing a bound on De Bruijn indices seems natural, as their values in the vast majority of programs remain small and rarely exceed 20 (Berger (2019)). Our research is hence motivated by getting a better understanding of the structure of lambda terms belonging to this class as well as explaining the structural discrepancies between terms from the two discussed classes.

In the next section we describe lambda terms as combinatorial objects and introduce basic concepts used throughout the paper. In Section 3 we discuss the shape of lambda terms with De Bruijn indices bounded by k (which we call k-indexed lambda terms for brevity) and their decomposition into smaller structures, that is the hat and some attached k-colored Motzkin trees. Our results concerning average sizes of these substructures are presented in the two following sections: First, in Section 4, we show that the size of the hat is constant on average and then, in Section 5, we prove that the average number of De Bruijn levels in terms of size n is asymptotically of order $\sqrt{n}$. Section 6 contains the main result of this paper, namely the average unary profile of k-indexed terms. Finally, in the last section, we recall the results by Gittenberger and Larcher (2018) about the distribution of the total number of leaves in k-indexed terms and provide a short conclusion and outlook.

2 Preliminaries

Let $\mathcal{V}$ be a countable set of variables. *Lambda terms* are defined by the following grammar:

$$\mathcal{T} ::= \mathcal{V} \mid \big(\lambda\mathcal{V}.\mathcal{T}\big) \mid \big(\mathcal{T}\,\mathcal{T}\big).$$

A term of the form $\big(\lambda x.M\big)$ is called an *abstraction*, while a term of the form $\big(M\,N\big)$ is called an *application*. For the sake of clarity, we omit some parentheses according to the standard convention, *i.e.*, outermost parentheses are dropped, an application is left- and an abstraction right-associative. By $\mathsf{Var}(M)$ we denote the set of all variables occurring in M. The set $\mathsf{FV}(M)$ of *free variables* in a term M is defined recursively as follows:

$$\mathsf{FV}(x) = x, \quad \mathsf{FV}(\lambda x.M) = \mathsf{FV}(M) \setminus \{x\}, \quad \mathsf{FV}(M\,N) = \mathsf{FV}(M) \cup \mathsf{FV}(N).$$

A term M is called *closed* if it contains no free variables, *i.e.*, when $\mathsf{FV}(M) = \emptyset$.

Lambda terms have a natural representation by means of finite *enriched trees*, *i.e.*, rooted trees with additional directed edges (pointers). In order to construct the corresponding enriched tree for a given lambda term, first we construct a Motzkin tree, *i.e.*, a plane rooted tree with each node of out-degree 0, 1, or 2. In this tree each binary node corresponds to an application, each unary node to an abstraction, and each leaf to a variable. Now, for every occurrence of a bound variable x, we add a directed edge from the unary node corresponding to the particular abstraction, labelled with λx, to the variable. Therefore, each unary node of the Motzkin tree carries (zero, one, or more) pointers to leaves taken from the subtree rooted at this unary node; all leaves receiving a pointer from this unary node correspond to the same variable, and each leaf receives at most one pointer. By $\mathtt{tree}(M)$ we denote the Motzkin tree obtained from a lambda term M by removing all pointers (see Grygiel et al. (2013) for a more detailed description).

In what follows, we will be interested only in closed terms. This means that every lambda term we investigate is represented by some Motzkin tree enriched with pointers from unary nodes to leaves in such a way that every leaf receives precisely one pointer. Moreover, terms that are equal up to α-conversion, *i.e.*, up to renaming of bound variables, are considered equivalent. This allows us to apply the De Bruijn notation for lambda terms, which consists in eliminating names of variables and replacing them by positive integers indicating, in the tree representation, the number of unary nodes on the path from a particular variable to its binder. This representation was introduced by De Bruijn (1972), who used the notion of *reference depth*. In other words, instead of having the set $\mathcal{V}$ of variables, we use the set $\{\mathtt{1}, \mathtt{2}, \mathtt{3}, \ldots\}$ of *De Bruijn indices*, where an index $\mathtt{n}$ occurring in a term M indicates that the n-th lambda (*i.e.*, unary node) lying on the path from the corresponding leaf to the root in $\mathtt{tree}(M)$ points at this leaf. For every $k \in \mathbb{N}$, we say that a lambda term is *k-indexed* if all of its De Bruijn indices belong to the set $\{\mathtt{1}, \mathtt{2}, \ldots, \mathtt{k}\}$.

Let M be a lambda term and v be a vertex in $\mathtt{tree}(M)$. The *unary height of v in* $\mathtt{tree}(M)$, denoted by $\mathtt{h}(v)$, is defined as the number of unary nodes on the path connecting v with the root of $\mathtt{tree}(M)$. For every $\ell \in \mathbb{N}$, the *ℓ-th De Bruijn level of* $\mathtt{tree}(M)$ is defined as the set of all vertices v in $\mathtt{tree}(M)$ such that $\mathtt{h}(v) = \ell$. Finally, by the *unary profile* of a lambda term we define the sequence of numbers of variables in each De Bruijn level of the term.

Remark. De Bruijn (1972) introduced the notion of a *level* exclusively for variables. We extend this concept for all nodes in enriched trees as well as for Motzkin trees, so that all internal nodes are also covered. However, to avoid any confusion, we want to point out that the name *De Bruijn level* has also been used in slightly different contexts in previous papers so far, as for example by Lescanne and Rouyer-Degli (1995), who speak about subterms of a given term to be at the ℓ-th level if they have ℓ unary nodes above them and contain only indices at most ℓ.

As an example let us consider the term $\lambda a.\, \big((\lambda b.(\lambda c.b)b)a\big)\, \big(\lambda d.\big((\lambda e.d)d\big)(\lambda f.fa)\big)$ (see Figure 1). Its De Bruijn notation reads as $\lambda\, \big((\lambda(\lambda\mathtt{2})\mathtt{1})\mathtt{1}\big)\, \big(\lambda\big((\lambda\mathtt{2})\mathtt{1}\big)(\lambda\mathtt{13})\big)$, and it has three non-empty De Bruijn levels, of which only level 1 is connected. Notice that the 0-th De Bruijn level is empty, as the term is an abstraction. Moreover, this term is k-indexed for every $k \geq 3$.

In this paper the size of a lambda term is defined as the total number of its variables, abstractions, and applications, *i.e.*, for any variable x and lambda terms M and N the size is recursively defined as follows:

$$\begin{aligned} |x| &= 1, \\ |\lambda x.M| &= 1 + |M|, \\ |M\ N| &= 1 + |M| + |N|. \end{aligned}$$

Therefore, the size of a lambda term is equal to the number of all vertices in the corresponding tree. The

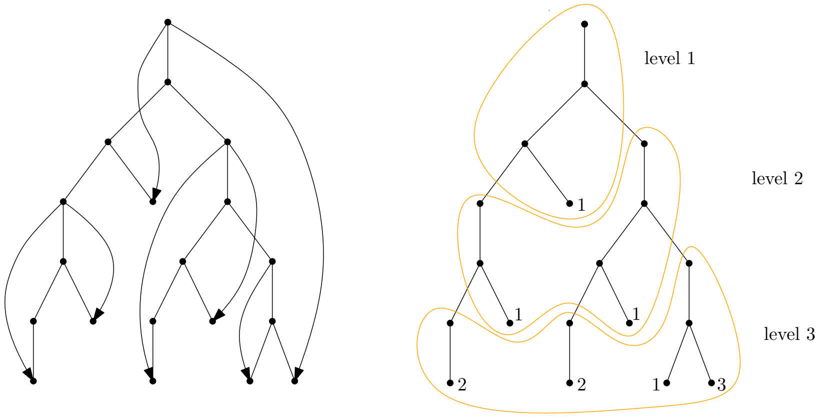

Fig. 1: Left: The enriched tree of $\lambda a.\,\big(\big(\lambda b.(\lambda c.b)b\big)a\big)\,\big(\lambda d.\big((\lambda e.d)d\big)(\lambda f.fa)\big)$. Right: Its counterpart in the De Bruijn notation (where all leaves are labeled with their respective De Bruijn indices) along with its decomposition into De Bruijn levels.

term in Figure 1 is of size 19. It has 7 variables, 6 applications, and 6 abstractions.

Having defined the size of lambda terms, we can now look at lambda terms as at objects from a combinatorial class and hence apply standard combinatorial methodology. Given a combinatorial class $\mathcal{A}$ of objects and its nonempty subclass $\mathcal{A}_n$ containing only the objects of $\mathcal{A}$ that are of size n, we assume the uniform probability distribution over $\mathcal{A}_n$, *i.e.*, every object from $\mathcal{A}_n$ has the equal probability $1/|\mathcal{A}_n|$ to be chosen *at random*. A *parameter* χ on the class $\mathcal{A}$ is any function from $\mathcal{A}$ to $\mathbb{N}$. For example, the size of a hat or the number of non-empty De Bruijn levels of lambda terms are possible instances of parameters. Every parameter χ determines a discrete random variable χ_n defined over the discrete probability space $\mathcal{A}_n$ as follows:

$$\chi_n = \mathbb{P}(\chi = j) = \frac{|\mathcal{A}_{n,j}|}{|\mathcal{A}_n|},$$

where $\mathcal{A}_{n,j}$ denotes the class of objects of size n for which the parameter χ is equal to j. This allows us to speak about the expected value (also called *average* or *mean*) of the parameter χ for the fixed class. It is well-known that the expected value can be computed by means of generating functions via the formula

$$\mathbb{E}(\chi) = \frac{[z^n]\partial_u A(z,u)_{u=1}}{[z^n]A(z,1)},$$

where $A(z,u)$ is the bivariate generating function associated with $\big(|\mathcal{A}_{n,j}|\big)_{n,j\geq 0}$.

3 Structure of lambda terms with bounded De Bruijn indices

In this section we want to illustrate the asymptotic shape of lambda terms with bounded De Bruijn indices in a very general way, while it will be investigated more thoroughly in the subsequent sections.

In order to set up the generating function for closed lambda terms with bounded De Bruijn indices, we proceed analogously to Bodini et al. (2018). For every $k \geq 1$, let $\mathcal{G}_k$ be the class of k-indexed lambda terms and let $G_k(z)$ be the corresponding generating function. In order to write down a formula for $G_k(z)$

we define the following auxiliary functions: For $j \in \{0, \ldots, k\}$, let $\mathcal{G}_{k,j}$ be the class of unary-binary trees such that every leaf v is labelled by a De Bruijn index m with $1 \leq m \leq \min\{\mathtt{h}(v) + j, k\}$. Note that only the terms in $\mathcal{G}_{k,0}$ are necessarily closed, while this does not have to be the case for terms from $\mathcal{G}_{k,j}$ with $j \geq 1$. The classes $\mathcal{G}_{k,j}$ can be recursively specified, starting from a class $\mathcal{Z}$ of atoms, in the following way:

$$\begin{aligned}
\mathcal{G}_{k,k} &= k\,\mathcal{Z} \uplus (\mathcal{Z} \times \mathcal{G}_{k,k} \times \mathcal{G}_{k,k}) \uplus (\mathcal{Z} \times \mathcal{G}_{k,k}), \\
\mathcal{G}_{k,j} &= j\,\mathcal{Z} \uplus (\mathcal{Z} \times \mathcal{G}_{k,j} \times \mathcal{G}_{k,j}) \uplus (\mathcal{Z} \times \mathcal{G}_{k,j+1}), \qquad \text{for } 0 \leq j \leq k-1
\end{aligned}$$

Then the classes $\mathcal{G}_k$ and $\mathcal{G}_{k,0}$ are isomorphic and hence their generating functions coincide. Thus, by translating into generating functions, we directly get (*cf.* Bodini et al. (2018))

$$G_k(z) = \frac{1 - \sqrt{R_{k,k}(z)}}{2z}, \tag{1}$$

where

$$R_{k,j}(z) = \begin{cases} 1 - 2z - (4k-1)z^2, & j = 0, \\ 1 - 2z - 2(2k-3)z^2 + 2z\sqrt{R_{k,0}(z)}, & j = 1, \\ 1 - 2z - 4(k-j)z^2 + 2z\sqrt{R_{k,j-1}(z)}, & j > 1. \end{cases}$$

Bodini et al. (2018, Lemma 5.4) proved that the dominant singularity of $G_k(z)$ comes from the innermost radicand, *i.e.*, $R_{k,0}(z)$, and is equal to $\rho_k = \frac{1}{2\sqrt{k}+1}$. Furthermore, they provide an asymptotic estimate of the n-th coefficient of $G_k(z)$. Before showing this estimate let us define an auxiliary sequence $(c_j)_{j\geq 1}$ via

$$c_1 = 5 \qquad \text{and} \qquad c_j = 4j - 1 + 2\sqrt{c_{j-1}} \qquad \text{for } j \geq 2, \tag{2}$$

and constants $\mathsf{C}_{j,k}$ with $j \geq 1$ and $k \geq j$ via

$$\mathsf{C}_{j,k} = \prod_{s=j}^{k} \frac{1}{\sqrt{c_s}}. \tag{3}$$

These numbers appear both in the announced estimate and throughout this paper. The first values of $(c_k)_{k\geq 1}$ and $(\mathsf{C}_{j,k})_{j\geq k}$ are listed in Table 1.

Lemma 3.1 (Bodini et al. (2018, Theorem 5.6)). *For any fixed $k \geq 1$, let $G_k(z)$ be the generating function of the class of k-indexed lambda terms. Then*

$$[z^n]G_k(z) \sim \frac{\mathsf{C}_{1,k} k^{1/4}}{2\sqrt{\pi \rho_k}} n^{-3/2} \rho_k^{-n},$$

where $\rho_k = \frac{1}{2\sqrt{k}+1}$ and $\mathsf{C}_{1,k}$ is defined as in (3).

Remark. In (Bodini et al., 2018, Lemma 5.7) the authors showed the asymptotic decrease of the constant in Lemma 3.1, reading as

$$\frac{\mathsf{C}_{1,k} k^{1/4}}{2\sqrt{\pi \rho_k}} \sim C \frac{1}{2^k e^{\sqrt{k}}} \sqrt{\frac{(2k + \sqrt{k})k^{1/4}}{(k-1)!}} \left(1 + \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)\right), \qquad \text{as } k \to \infty,$$

with a constant C that does not depend on k.

k	c_k	$\mathsf{C}_{1,k}$	$\mathsf{C}_{2,k}$	$\mathsf{C}_{3,k}$	$\mathsf{C}_{4,k}$
1	5.00000000	0.4472135954			
2	11.47213595	0.1320361509	0.2952418088		
3	17.77410834	0.0313183551	0.0700299709	0.2371953050	
4	23.43187010	0.0064698687	0.0144670662	0.0490007371	0.2065839250
5	28.68129539	0.0012080811	0.0027013514	0.0091496237	0.0385742191
6	33.71098415	0.0002080704	0.0004652596	0.0015758596	0.0066437216
7	38.61223220	0.0000334848	0.0000748743	0.0002536034	0.0010691754
8	43.42774834	0.0000050812	0.0000113619	0.0000384832	0.0001622426
9	48.17994664	0.0000007320	0.0000016369	0.0000055442	0.0000233740
10	52.88235522	0.0000001007	0.0000002251	0.0000007624	0.0000032142

Tab. 1: First values of the sequences $(c_k)_{k\geq 1}$ and $(\mathsf{C}_{j,k})_{k\geq j}$ for $j \in \{1,2,3,4\}$.

The result from Lemma 3.1 already gives us a hint that lambda terms with bounded De Bruijn indices behave somewhat treelike. However, in order to get a better intuition why this is the case and how exactly these terms look like, we set up the generating function $G_k(z)$ in a different manner, which reflects another way of looking at how the corresponding terms are constructed. Instead of interpreting a lambda term belonging to that class as a structure that involves iterated unary-binary trees, we consider it to be built of leaf-labelled binary trees that are glued together via unary nodes (*cf.* Figure 2). Thereby, the labels of the leaves correspond to the respective De Bruijn indices. Obviously, this implies that within the whole tree each of the labels belongs to the set $\{\mathtt{1},\ldots,\mathtt{k}\}$. However, in the first $k-1$ De Bruijn levels (excluding the 0-th level, which contains no variables) we have a stronger restriction. Since we consider only closed terms, no label (*i.e.*, no De Bruijn index) can exceed the De Bruijn level the respective leaf is located in. Thus, with $B(z,w)$ denoting the bivariate generating function of binary trees where z marks the size (*i.e.*, the total number of nodes) and w marks the number of leaves, and with $M_k(z)$ denoting the generating function of Motzkin trees where each leaf can be labelled in k ways (*k-colored Motzkin trees* in short), we get

$$G_k(z) = B\left(z, B\left(z, 1 + B\left(z, 2 + \ldots + B\big(z, k-1+M_k(z)\big)\ldots\right)\right)\right), \tag{4}$$

where

$$B(z,w) = \frac{1-\sqrt{1-4wz^2}}{2z} \qquad \text{and} \qquad M_k(z) = \frac{1-z-\sqrt{(1-z)^2-4kz^2}}{2z}. \tag{5}$$

Now, let us give an interpretation of Equation (4). Each tree representing a lambda term starts with a binary tree, in which all the leaves are replaced by unary nodes to which we add further binary trees. This is necessary for a lambda term to be closed. These newly added binary trees represent the first De Bruijn level. Next, there are two possibilities for each leaf in this level: Either it receives the label $\mathtt{1}$ or, alternatively, it is replaced with a unary node with a new binary tree attached, which belongs to the next De Bruijn level. In this level the leaves can already be labelled with two different labels (namely $\mathtt{1}$ or $\mathtt{2}$), or they can be replaced with unary nodes with new binary trees attached. Starting from the k-th De Bruijn level, the number of possible labellings for the leaves does not increase anymore. Thus, we finally get $B(z, k + B(z, k + B(z, k + \ldots)))$, which is exactly the generating function $M_k(z)$ of k-colored Motzkin trees given in (5).

Therefore, the enriched tree corresponding to a k-indexed lambda term is constructed as follows (*cf.* Figure 2):

- It starts with the *hat* consisting of all De Bruijn levels from 0 to $k-1$ along with all unary nodes from the k-th level;
- To this hat structure we attach k-colored Motzkin trees via unary nodes.

Remark. Note that the glued binary trees in Equation (4) constitute the hat of the structure, to which we attach the (comparatively large) k-colored Motzkin trees.

We emphasize at this point that the generating functions defined in (1) and (4) describe indeed the exact same function. The reason to choose the latter way of representing the function $G_k(z)$ in this paper is that it gives direct access to the De Bruijn levels, which is advantageous for our purposes.

In the subsequent sections we investigate the structure of these terms in more detail. We prove that, for a fixed $k \geq 1$, the hat of a k-indexed lambda term is on average of constant size and that the average number of De Bruijn levels of a term of size n is asymptotically $\sqrt{n}$. Finally, we provide its unary profile.

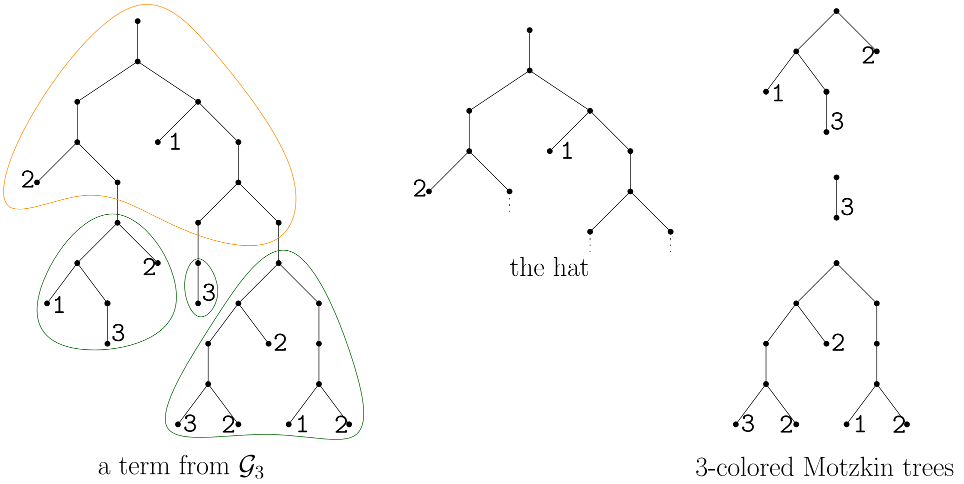

Fig. 2: A lambda term from $\mathcal{G}_3$ decomposed into the hat and three subterms represented by 3-colored Motzkin trees.

4 Average size of a hat

In this section we focus on the size of the hat of k-indexed lambda terms. We prove that the average size of a hat is asymptotically constant, *i.e.*, it does not depend on the size of a term. This implies that on average the number of k-colored Motzkin trees in the decomposition described in the previous section is also constant.

Theorem 4.1. *For $k \geq 1$, let $X_{n,k}$ be the size of the hat of a random k-indexed lambda term of size n. Then, as n tends to infinity,*

$$\mathbb{E}X_{n,k} = 1+\sum_{j=0}^{k-2}\sqrt{c_{k-j}}+4(k+\sqrt{k}-1)\sum_{j=1}^{k}\mathsf{C}_{1,j}+\sum_{j=0}^{k-2}\Big(1+2\sqrt{k}+4j-\sqrt{c_{k-j-1}}\Big)\sum_{m=k-j}^{k}\mathsf{C}_{k-j,m}+o(1)$$

with $(c_j)_{j\geq 1}$ and $\mathsf{C}_{j,k}$ defined in (2) *and* (3).

Proof: To prove this theorem we use the well-known approach of marking the parameter of interest in the generating function with a second variable u and then investigating the bivariate generating function. Remember that throughout this whole paper the quantity k is a fixed constant which is independent of n, and thus when expanding the radicands $R_{k,j}$ the uniformity of the error terms in k is not an issue. Moreover, as a further consequence, the set of singularities of the generating function $G_k(z)$ is finite, since the radicand $R_{k,j}$ is singular if and only if one of the radicands $R_{k,j-1}$, $R_{k,j-2}$, …, $R_{k,0}$, is zero. Thus, there exists a $\delta > 0$ such that ρ_k is the only singularity for $|z| < \rho_k + \delta$ and therefore all requirements for the use of singularity are fulfilled.

So, let $G_k(z,u)$ be the generating function for k-indexed lambda terms with z marking the size of terms and u marking the size of their hats. The average size of a hat is hence given by

$$\mathbb{E}X_{n,k} = \frac{[z^n]\frac{\partial G_k(z,u)}{\partial u}|_{u=1}}{[z^n]G_k(z)}.$$

Since we want to mark by u all the nodes that belong to the hat, we get

$$G_k(z,u) = B\left(zu, B\left(zu, 1 + B\left(zu, 2 + \ldots + B\big(zu, k-1+M_k(z)\big)\ldots\right)\right)\right),$$

where $B(z,w)$ is the function defined in (5). This gives

$$G_k(z,u) = \frac{1-\sqrt{R_{k,k}(z,u)}}{2zu}$$

where

$$R_{k,j}(z,u) = \begin{cases} 1-2z-(4k-1)z^2, & j=0, \\ 1-2zu^2-2(2k-3)z^2u^2+2zu^2\sqrt{R_{k,0}(z,u)}, & j=1, \\ 1-2zu-4(k-j)z^2u^2+2zu\sqrt{R_{k,j-1}(z,u)}, & j>1. \end{cases}$$

Therefore, the derivatives can also be recursively defined via

$$\frac{\partial R_{k,j}(z,u)}{\partial u}\Big|_{u=1} = \begin{cases} 0, & j=0, \\ -4z-4(2k-3)z^2+4z\sqrt{R_{k,0}(z,1)}, & j=1, \\ -2z-8(k-j)z^2+2z\sqrt{R_{k,j-1}(z,1)}+\frac{z}{\sqrt{R_{k,j-1}(z,1)}}\frac{\partial R_{k,j-1}(z,u)}{\partial u}\Big|_{u=1}, & j>1, \end{cases}$$

and we get

$$\begin{aligned}\frac{\partial G_k(z,u)}{\partial u}\Big|_{u=1} &= \frac{\sqrt{R_{k,k}(z,1)}-1}{2z} \;-\; \frac{1}{4z\sqrt{R_{k,k}(z,1)}}\cdot\frac{\partial R_{k,k}(z,u)}{\partial u}\Big|_{u=1}\\ &= \frac{\sqrt{R_{k,k}(z,1)}-1}{2z} \;+\; \sum_{j=0}^{k-2}\frac{z^j(1+4jz-\sqrt{R_{k,k-j-1}(z,1)})}{2\prod_{m=k-j}^{k}\sqrt{R_{k,m}(z,1)}}\\ &\quad+\;\frac{z^{k-1}\big(1+(2k-3)z-\sqrt{R_{k,0}(z,1)}\big)}{\prod_{m=1}^{k}\sqrt{R_{k,m}(z,1)}}.\end{aligned} \tag{6}$$

Expanding the radicands $R_{k,j}$ around $z=\rho_k$ yields (see Bodini et al. (2018))

$$R_{k,j}(\rho_k(1-\varepsilon),1)=\begin{cases}4\rho_k\sqrt{k}\varepsilon+\mathcal{O}(|\varepsilon|^2), & j=0,\\ c_j\rho_k^2+d_{k,j}\sqrt{\varepsilon}+\mathcal{O}(|\varepsilon|), & j>0\end{cases}$$

where $\varepsilon\in\mathbb{C}\setminus\mathbb{R}_-$ and $|\varepsilon|\to 0$ and with $d_{k,j}=4\mathsf{C}_{1,j-1}\rho_k^{3/2}k^{1/4}$.

Hence, we have

$$\sqrt{R_{k,j}\big(\rho_k(1-\varepsilon),1\big)}=\begin{cases}2\sqrt{\rho_k}k^{1/4}\sqrt{\varepsilon}+\mathcal{O}(|\varepsilon|^{3/2}), & j=0,\\ \sqrt{c_j}\rho_k+b_{k,j}\sqrt{\varepsilon}+\mathcal{O}(|\varepsilon|), & j>0\end{cases}$$

with

$$b_{k,j}=\frac{d_{k,j}}{2\rho_k\sqrt{c_j}}=2\mathsf{C}_{1,j-1}\sqrt{\rho_k}k^{1/4}. \tag{7}$$

Plugging this into Equation (6) gives

$$\begin{aligned}\frac{\partial G_k(\rho_k(1-\varepsilon),u)}{\partial u}\Big|_{u=1} &= \frac{\sqrt{c_k}\rho_k+b_{k,k}\sqrt{\varepsilon}-1}{2\rho_k} \;+\; \sum_{j=0}^{k-2}\frac{\rho_k^j\big(1+4j\rho_k-\sqrt{c_{k-j-1}}\rho_k-b_{k-j-1}\sqrt{\varepsilon}\big)}{2\prod_{m=k-j}^{k}\big(\sqrt{c_m}\rho_k+b_{k,m}\sqrt{\varepsilon}\big)}\\ &\quad+\;\frac{\rho_k^{k-1}(1+(2k-3)\rho_k-2\sqrt{\rho_k}k^{1/4}\sqrt{\varepsilon})}{\prod_{m=1}^{k}\big(\sqrt{c_m}\rho_k+b_{k,m}\sqrt{\varepsilon}\big)} \;+\; \mathcal{O}\big(|\varepsilon|\big)\\ &= A_k \;-\; B_k\sqrt{\varepsilon} \;+\; \mathcal{O}\big(|\varepsilon|\big),\end{aligned}$$

where A_k and B_k are constants depending on k with

$$\begin{aligned}B_k \;&= \mathsf{C}_{1,k}\left(k^{1/4}\rho_k^{-1/2}+\frac{1+(2k-3)\rho_k}{\rho_k^2}\sum_{m=1}^{k}\frac{b_{k,m}}{\sqrt{c_m}}\right)\\ &\quad+\frac{1}{2\rho_k^2}\sum_{j=0}^{k-2}\mathsf{C}_{k-j,k}\left(b_{k,k-j}\sqrt{c_{k-j}}\rho_k+(1+4j\rho_k-\rho_k\sqrt{c_{k-j-1}})\sum_{m=k-j}^{k}\frac{b_{k,m}}{\sqrt{c_m}}\right).\end{aligned}$$

Since A_k is not important for the result, we skip computing its exact value. By inserting (7) into the formula for B_k and after some simplifications, we obtain

$$\begin{aligned} B_k = \frac{\mathsf{C}_{1,k}k^{1/4}}{\sqrt{\rho_k}} \Bigg(1 \,+\, & \sum_{j=0}^{k-2} \sqrt{c_{k-j}} + 4(k+\sqrt{k}-1) \sum_{j=1}^{k} \mathsf{C}_{1,j} \\ + & \sum_{j=0}^{k-2} \Big(1+2\sqrt{k}+4j-\sqrt{c_{k-j-1}}\Big) \sum_{m=k-j}^{k} \mathsf{C}_{k-j,m} \Bigg). \end{aligned} \tag{8}$$

By singularity analysis applied to

$$\frac{\partial G_k(z,u)}{\partial u}\Big|_{u=1} = A_k - B_k\sqrt{1-\frac{z}{\rho_k}} + \mathcal{O}\left(\left|1-\frac{z}{\rho_k}\right|\right),$$

we immediately obtain, as n tends to infinity,

$$[z^n]\frac{\partial G_k(z,u)}{\partial u}\Big|_{u=1} \sim \frac{B_k}{2\sqrt{\pi}}\rho_k^{-n}n^{-3/2}.$$

Finally, by (8) and Lemma 3.1, we get

$$\mathbb{E}X_{n,k} = 1+\sum_{j=0}^{k-2}\sqrt{c_{k-j}}+4(k+\sqrt{k}-1)\sum_{j=1}^{k}\mathsf{C}_{1,j}+\sum_{j=0}^{k-2}\Big(1+2\sqrt{k}+4j-\sqrt{c_{k-j-1}}\Big)\sum_{m=k-j}^{k}\mathsf{C}_{k-j,m}+o(1).$$

□

In Figure 3 we list average sizes of hats of k-indexed lambda terms for k up to 10.

1	2	3	4	5	6	7	8	9	10
2.7888	5.3187	7.6761	9.9443	12.1554	14.3260	16.4660	18.5816	20.6774	22.7565

Fig. 3: Average size of the hat (bottom row) for $k \in \{1,\dots,10\}$ (top row).

Since the hat of a k-indexed lambda term, for any $k \geq 1$, is constant on average, such a term has on average a finite number of unary nodes in the k-th De Bruijn level. Therefore, we arrive at the following conclusion.

Corollary 4.2. *For every $k \geq 1$, the average number of k-colored Motzkin trees in the decomposition (see the description in the end of Section 3) of k-indexed lambda terms is constant.*

5 Average number of De Bruijn levels

In order to determine the average number of De Bruijn levels of k-indexed lambda terms, we first compute the average number of De Bruijn levels of k-colored Motzkin trees. To this end, we use the following result by Drmota et al. (2014). The notation $A(z) \preceq B(z)$ used therein means that $[z^n]A(z) \leq [z^n]B(z)$ for every $n \geq 0$.

Lemma 5.1 (Drmota et al. (2014, Lemma 1.4))**.** *Suppose that $F(z,t)$ is an analytic function at $(z,t) = (0,0)$ such that the equation $T(z) = F(z,T(z))$ has a solution $T(z)$ that is analytic at $z = 0$ and has non-negative Taylor coefficients. Suppose that $T(z)$ has a square-root singularity at $z = z_0$ and can be continued to a region $\{z \in \mathbb{C} \colon |z| < z_0 + \varepsilon\} \setminus [z_0, \infty)$ for some $\varepsilon > 0$, such that $F_t(z_0,t_0) = 1$, $F_z(z_0,t_0) \neq 0$, and $F_{tt}(z_0,t_0) \neq 0$, where $t_0 = T(z_0)$.*

Let $T^{[0]}(z)$ be a power series with $0 \preceq T^{[0]}(z) \preceq T(z)$ such that $T^{[0]}(z)$ is analytic at $z = z_0$, and let $T^{[k]}(z)$, $k \geq 1$ be iteratively defined by

$$T^{[k]}(z) = F(z, T^{[k-1]}(z)).$$

Assume that $T^{[k-1]}(z) \preceq T^{[k]}(z) \preceq T(z)$.

Let H_n be an integer valued random variable that is defined by

$$\mathbb{P}\{H_n \leq k\} = \frac{[z^n]T^{[k]}(z)}{[z^n]T(z)}$$

for those n with $[z^n]T(z) > 0$. Then

$$\mathbb{E}H_n \sim \sqrt{\frac{2\pi n}{z_0 F_z(z_0,t_0) F_{tt}(z_0,t_0)}}.$$

Lemma 5.2. *The average number of De Bruijn levels of a k-colored Motzkin tree of size n is asymptotically equal to*

$$\sqrt{\frac{\pi n}{2k + \sqrt{k}}}.$$

Proof: For $k \geq 1$ and $h \geq 0$, the generating function $M_k^{[h]}(z)$ of k-colored Motzkin trees with at most h De Bruijn levels fulfills

$$M_k^{[h+1]}(z) = kz + zM_k^{[h]}(z) + z\big(M_k^{[h+1]}(z)\big)^2,$$

and hence

$$M_k^{[h+1]}(z) = \frac{1 - \sqrt{1 - 4kz^2 - 4z^2 M_k^{[h]}(z)}}{2z}.$$

Let us fix $k \geq 1$ and define $F_k(z,t) := \frac{1-\sqrt{1-4kz^2-4z^2t}}{2z}$. Let us notice that $F_k(z,t)$ satisfies the assumptions of Lemma 5.1. Indeed, the function $M_k(z)$, with a square-root singularity at $z = \rho_k = \frac{1}{1+2\sqrt{k}}$, is a solution of $F_k(z, M_k(z)) = M_k(z)$ fulfilling all necessary conditions. Furthermore, the function $M_k^{[0]}(z)$ enumerates all k-colored Motzkin trees with only one (the 0-th) De Bruijn level. These trees are binary trees with k possible labels for each node, thus $M_k^{[0]}(z) = \frac{1-\sqrt{1-4kz^2}}{2z}$. As $M_k^{[0]}(z)$ has its dominant singularity at $z = \frac{1}{2\sqrt{k}}$, it is analytic at $z = \frac{1}{1+2\sqrt{k}}$. Moreover, by a purely combinatorial argument, $M_k^{[h]}(z) \preceq M_k^{[h+1]}(z) \preceq M_k(z)$ for every $h \geq 0$. Finally, since $F\big(z, M_k^{[h]}(z)\big) = M_k^{[h+1]}(z)$, we can apply Lemma 5.1. We have $M_k(\rho_k) = \sqrt{k}$ and

$$\frac{\partial F_k(z,t)}{\partial z}\Big|_{(z,t)=(\rho_k,\sqrt{k})} = \big(1 + 2\sqrt{k}\big)^2 \sqrt{k} \qquad \text{and} \qquad \frac{\partial^2 F_k(z,t)}{\partial t^2}\Big|_{(z,t)=(\rho_k,\sqrt{k})} = 2,$$

thus the average number of De Brujin levels of k-colored Motzkin trees is asymptotically equal to

$$\sqrt{\frac{2\pi n}{\frac{1}{1+2\sqrt{k}}\cdot\left(1+2\sqrt{k}\right)^2\sqrt{k}\cdot 2}}=\sqrt{\frac{\pi n}{2k+\sqrt{k}}}.$$

□

Corollary 5.3. *For every $k \geq 1$, the average number of De Bruijn levels of a k-indexed lambda term of size n is $\Theta(\sqrt{n})$.*

Proof: By Corollary 4.2, the number of k-colored Motzkin trees in the decomposition of lambda terms is constant on average. Therefore, the size of a largest such tree in the decomposition of a k-indexed lambda term of size n is asymptotically $\Theta(n)$. Since the average number of De Bruijn levels of k-colored Motzkin trees of size asymptotic to n is $\Theta(\sqrt{n})$, the same is true for k-indexed lambda terms, which have just k levels more than a longest (in terms of De Bruijn levels) k-colored Motzkin tree in their decomposition. □

6 Unary profile

In this section, we determine the mean unary profile of a random lambda term of some large size, *i.e.*, the asymptotic number of variables in each De Bruijn level of the term.

In the forthcoming proof, we will make use of the following technical results.

Lemma 6.1 (Gittenberger (1999, Lemma 3.4))**.** *Let γ be a Hankel contour truncated at $K \in \mathbb{R}_+$, i.e., $\Re t \leq K$ for all $t \in \gamma$. Then we have, for $\alpha, \beta > 0$,*

$$\frac{1}{2\pi i}\int_\gamma e^{-\alpha\sqrt{-t}-\beta t}dt=\frac{\alpha\beta^{\frac{-3}{2}}}{2\sqrt{\pi}}\exp\left(-\frac{\alpha^2}{4\beta}\right)+\mathcal{O}\left(\frac{1}{\beta}e^{-K\beta}\right).$$

Lemma 6.2. *Let $\varepsilon > 0$ and $\gamma = \left\{\rho_k\left(1+\frac{t+i}{n}\right) : t \in [\log^2 n, n\varepsilon)\right\}$ with $\rho_k = \frac{1}{2\sqrt{k+1}}$. Then*

$$\max_{z\in\gamma}\left|\frac{\sqrt{1-2z-(4k-1)z^2}}{z}\right|=\mathcal{O}\left(\frac{\log n}{\sqrt{n}}\right).$$

Proof: The closer to ρ_k an argument of the function $\gamma \ni z \mapsto \frac{\sqrt{1-2z-(4k-1)z^2}}{z}$ is, the greater its modulus gets. Thus, we set $z = \rho_k\left(1+\frac{\log^2 n}{n}+\frac{i}{n}\right)$, which is the closest point to ρ_k on γ, and we get

$$\frac{\sqrt{1-2z-(4k-1)z^2}}{z}=\frac{\sqrt{a+ib}}{\rho_k\left(1+\frac{\log^2 n}{n}+\frac{i}{n}\right)}$$

with $a \sim -4\sqrt{k}\rho_k\frac{\log^2 n}{n}$ and $b \sim -4\sqrt{k}\rho_k\frac{1}{n}$. Plugging in the asymptotic formulas for a and b directly yields the desired result. □

Now we are in the position to prove the main theorem of this section.

Theorem 6.3. *Let $\alpha > 0$ be a fixed real number. The expected number of variables at De Bruijn level $\lfloor \alpha\sqrt{n} \rfloor$ in a k-indexed lambda term of size n is asymptotically equal to*

$$2\alpha \exp\left(-\alpha^2(2k+\sqrt{k})\right)\sqrt{n}.$$

Proof: To prove this theorem we use again the approach that we mark the parameter of interest and investigate the thereby obtained bivariate generating function. However, in this case we have to face a different problem that arises due to the fact that we want to mark the number of leaves in a level ℓ proportional to $\sqrt{n}$, while n tends to infinity, leading to infinitely many nested radicals. Thus, we cannot apply the transfer theorems directly in order to get the asymptotics for the coefficients of the generating functions, but instead we write the coefficients by means of Cauchy's integral formula and use a suitable integration contour. Fortunately, we will be able to show that solely finitely many of the occurring radicands are different, while the great majority describes the exact same function, which then provides simplifications that allow us to calculate the integral along the chosen curve asymptotically.

So, first let $U_{k,\ell}(z,u)$ be the bivariate generating function for k-indexed terms with z marking the size and u marking the number of leaves in the $(k+\ell)$-th De Bruijn level, where $\ell \geq 1$. Then we have

$$U_{k,\ell}(z,u) = B\Big(z, B\Big(z, 1{+}B\Big(z, 2{+}\ldots{+}\underbrace{B\big(z, k + B(z, k + B(\ldots B}_{\ell \text{ occurrences of } B}(z, k{+}B(z, ku{+}M_k(z)))))\big)\ldots\Big)\Big)\Big).$$

Applying formulas for $B(z,w)$ and $M_k(z)$ given in (5) yields

$$U_{k,\ell}(z,u) = \frac{1-\sqrt{Q_{k,k+\ell}(z,u)}}{2z}$$

where

$$Q_{k,i}(z,u) = \begin{cases} 1-2z-(4k-1)z^2, & j=0, \\ 1-2z-(4ku-2)z^2+2z\sqrt{Q_{k,0}(z,u)}, & j=1, \\ 1-2z-4kz^2+2z\sqrt{Q_{k,j-1}(z,u)}, & j\in\{2,\ldots,\ell\}, \\ 1-2z-4(k-j+\ell)z^2+2z\sqrt{Q_{k,j-1}(z,u)}, & j\in\{\ell+1,\ldots,\ell+k\}. \end{cases}$$

Furthermore, we have

$$\frac{\partial Q_{k,j}(z,u)}{\partial u} = \begin{cases} 0, & j=0, \\ -4kz^2, & j=1, \\ \frac{-4kz^{j+1}}{\prod_{m=1}^{j-1}\sqrt{Q_{k,m}(z,u)}}, & j>1, \end{cases}$$

and hence

$$\frac{\partial U_{k,\ell}(z,u)}{\partial u} = \frac{kz^{k+\ell}}{\prod_{m=1}^{k+\ell}\sqrt{Q_{k,m}(z,u)}}.$$

Given the De Bruijn level $\ell = \lfloor \alpha\sqrt{n} \rfloor$ with $\alpha > 0$, we are interested in estimating

$$\frac{[z^n]\frac{\partial U_{k,\ell}(z,u)}{\partial u}\Big|_{u=1}}{[z^n]G_k(z)}.$$

In order to make further computations easier, let us notice that $\left|\sqrt{Q_{k,j}(z,1)}\right| = \left|z + \sqrt{Q_{k,0}(z,1)}\right|$ for $j \in \{1,\dots,\ell\}$, *i.e.*, all these radicands describe the same function. Indeed, let us first notice that the above holds for $j = 1$, since

$$Q_{k,1}(z,1) = Q_{k,0}(z,1) + z^2 + 2z\sqrt{Q_{k,0}(z,1)} = \left(\sqrt{Q_{k,0}(z,1)} + z\right)^2.$$

Next, by (5), we can notice that $x = M_k(z)$ is a solution of the equation $x = B(z, k + x)$. Therefore, in particular,

$$B(z, k + B(z, k + M_k(z)))) = B(z, k + M_k(z)),$$

which gives us $\sqrt{Q_{k,1}(z,1)} = \sqrt{Q_{k,2}(z,1)}$. By iteration we obtain the result for $j \in \{3,\dots,\ell\}$.

For $z = \rho_k\left(1 + \frac{t}{n}\right)$ we get the expansions

$$\sqrt{Q_{k,j}(z,1)} = \begin{cases} 2k^{1/4}\rho_k^{1/2}\sqrt{-t/n} + \mathcal{O}(|t|/n), & j = 0, \\ \rho_k + 2k^{1/4}\rho_k^{1/2}\sqrt{-t/n} + \mathcal{O}(|t|/n), & j \in \{1,\dots,\ell\}, \\ \sqrt{c_{j-\ell}}\rho_k + b_{j-\ell}\sqrt{-t/n} + \mathcal{O}(|t|/n), & j \in \{\ell+1,\dots,\ell+k\}, \end{cases} \tag{9}$$

where $(c_j)_{j\geq 1}$ and $(b_{k,j})_{k,j\geq 1}$ are as before (see (2) and (7)).

Let $\varepsilon > 0$. We have

$$[z^n]\frac{\partial U_{k,\ell}(z,u)}{\partial u}\Big|_{u=1} = \frac{k}{2\pi i}\int_\gamma \frac{z^{k+\ell-n-1}}{\prod_{j=1}^{k+\ell}\sqrt{Q_{k,j}(z,1)}}dz,$$

where as an integration path we choose a truncated Hankel contour $\gamma_1 \cup \gamma_2 \cup \gamma_3$ encircling the dominant singularity ρ_k and a circular arc γ_4:

$$\begin{aligned} \gamma_1 &= \left\{z = \rho_k\Big(1 + \frac{t}{n}\Big) : t = e^{-i\theta}, \theta \in [-\pi/2, \pi/2]\right\} \cup \left\{z = \rho_k\Big(1 + \frac{t \pm i}{n}\Big) : t \in (0, \log^2 n)\right\}, \\ \gamma_2 &= \left\{z = \rho_k\Big(1 + \frac{t+i}{n}\Big) : t \in [\log^2 n, n\varepsilon)\right\}, \\ \gamma_3 &= \left\{z = \rho_k\Big(1 + \frac{t-i}{n}\Big) : t \in [\log^2 n, n\varepsilon)\right\}, \\ \gamma_4 &= \left\{z \colon |z| = \rho_k\Big|1 + \varepsilon + \frac{i}{n}\Big|, \Re(z) \leq \rho_k\Big(1 + \varepsilon\Big)\right\}. \end{aligned}$$

We start by estimating the integral along γ_1. To this end, we apply the substitution $z = \rho_k(1 + t/n)$,

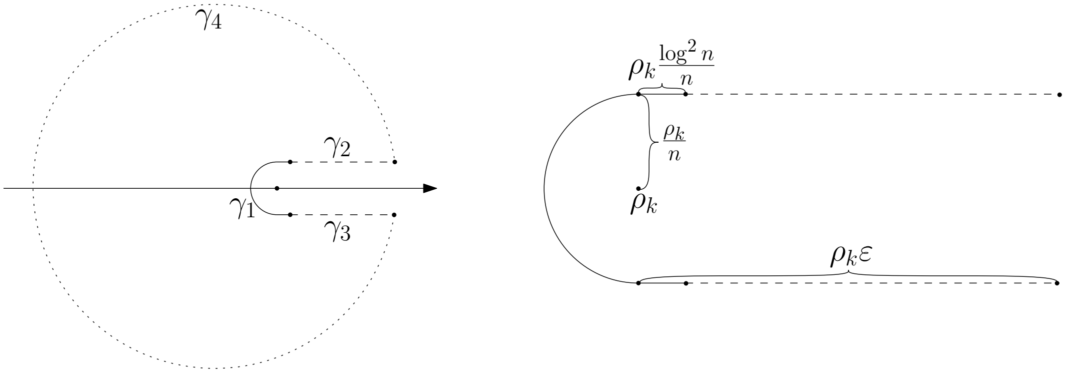

Fig. 4: Contour of integration: γ_1 plotted with a solid line, γ_2 and γ_3 with dashed lines, and γ_4 with a dotted line.

where $\widetilde{\gamma_1}$ denotes the transformed curve and we use the expansions given in (9):

$$\begin{aligned}
&\int_{\gamma_1} \frac{z^{k+\ell-n-1}}{\prod_{j=1}^{k+\ell}\sqrt{Q_{k,j}(z,1)}}dz\\
&= \frac{\rho_k^{k+\ell-n}}{n}\int_{\widetilde{\gamma_1}} \Big(1+\frac{t}{n}\Big)^{-n+k+\ell}\left(\frac{1}{\rho_k+2k^{1/4}\rho_k^{1/2}\sqrt{-t/n}+\mathcal{O}(|t|/n)}\right)^{\ell}\prod_{j=1}^{k}\frac{1}{\sqrt{c_j}\rho_k+b_{k,j}\sqrt{-t/n}+\mathcal{O}(|t|/n)}dt\\
&= \frac{\rho_k^{k+\ell-n}}{n}\int_{\widetilde{\gamma_1}} e^{-t}\Big(1+\frac{t}{n}\Big)^{k+\ell}\left(\frac{1}{\rho_k+2k^{1/4}\rho_k^{1/2}\sqrt{-t/n}}\right)^{\ell}\prod_{j=1}^{k}\frac{1}{\sqrt{c_j}\rho_k+b_{k,j}\sqrt{-t/n}}\left(1+\mathcal{O}\Big(\frac{|t|}{n}\Big)\right)dt\\
&= \frac{\rho_k^{k-n}}{n}\int_{\widetilde{\gamma_1}} e^{-t}\Big(1+\frac{t}{n}\Big)^{k+\alpha\sqrt{n}}\left(\frac{1}{1+2k^{1/4}\rho_k^{-1/2}\sqrt{-t/n}}\right)^{\alpha\sqrt{n}}\prod_{j=1}^{k}\frac{1}{\sqrt{c_j}\rho_k+b_{k,j}\sqrt{-t/n}}\left(1+\mathcal{O}\Big(\frac{|t|}{n}\Big)\right)dt\\
&= \frac{\rho_k^{k-n}}{n}\int_{\widetilde{\gamma_1}} e^{-t-\frac{2\alpha k^{1/4}}{\sqrt{\rho_k}}\sqrt{-t}}\Big(1+\frac{\alpha t}{\sqrt{n}}\Big)\left(1-\frac{2\alpha\sqrt{k}t}{\rho_k\sqrt{n}}\right)\prod_{j=1}^{k}\frac{1}{\sqrt{c_j}\rho_k+b_{k,j}\sqrt{-t/n}}\left(1+\mathcal{O}\Big(\frac{|t|}{n}\Big)\right)dt\\
&= \frac{\mathsf{C}_{1,k}}{n}\rho_k^{-n}\int_{\widetilde{\gamma_1}} e^{-t-\frac{2\alpha k^{1/4}}{\sqrt{\rho_k}}\sqrt{-t}}\Big(1+\frac{\alpha t}{\sqrt{n}}\Big)\left(1-\frac{2\alpha\sqrt{k}t}{\rho_k\sqrt{n}}\right)\left(1-\frac{\sqrt{-t}}{\rho_k\sqrt{n}}\sum_{j=1}^{k}\frac{b_{k,j}}{\sqrt{c_j}}\right)\left(1+\mathcal{O}\Big(\frac{|t|}{n}\Big)\right)dt\\
&= \frac{\mathsf{C}_{1,k}}{n}\rho_k^{-n}\int_{\widetilde{\gamma_1}} e^{-t-\frac{2\alpha k^{1/4}}{\sqrt{\rho_k}}\sqrt{-t}}\left(1+\frac{1}{\rho_k\sqrt{n}}\left(\alpha t(\rho_k-2\sqrt{k})-\sqrt{-t}\sum_{j=1}^{k}\frac{b_{k,j}}{\sqrt{c_j}}\right)\right)\left(1+\mathcal{O}\Big(\frac{|t|}{n}\Big)\right)dt\\
&= \frac{\mathsf{C}_{1,k}}{n}\rho_k^{-n}\int_{\widetilde{\gamma_1}} e^{-t-\frac{2\alpha k^{1/4}}{\sqrt{\rho_k}}\sqrt{-t}}\left(1+\mathcal{O}\Big(\frac{|t|}{\sqrt{n}}\Big)\right)dt.
\end{aligned}$$

Now, by applying Lemma 6.1, we get that the integral above can be further estimated to result in

$$\begin{aligned}\int_{\gamma_1} \frac{z^{k+\ell-n-1}}{\prod_{j=1}^{k+\ell}\sqrt{Q_{k,j}(z,1)}}dz &= \frac{\mathsf{C}_{1,k}}{n}\rho_k^{-n}\int_{\widehat{\gamma_1}} e^{-t-\frac{2\alpha k^{1/4}}{\sqrt{\rho_k}}\sqrt{-t}}dt + \mathcal{O}\Big(\frac{\rho_k^{-n}}{n^{3/2}}\Big)\\ &= \frac{\alpha k^{1/4}\mathsf{C}_{1,k}}{\sqrt{\pi\rho_k}n}\rho_k^{-n}\exp\Big(-\alpha^2(2k+\sqrt{k})\Big) + \mathcal{O}\Big(\frac{\rho_k^{-n}}{n^{3/2}}\Big).\end{aligned} \tag{10}$$

What is left to show is that the integrals along γ_j for $j \in \{2,3,4\}$ are all of order $o\left(\rho_k^{-n}n^{-3/2}\right)$ and hence the whole asymptotic contribution comes from integration along γ_1. In order to do so, we estimate the integrand with its maximum along the respective curve for all cases. For the sake of conciseness we present the precise calculations in the appendix.

Now, by (10), we get

$$[z^n]\frac{\partial U_{k,\ell}(z,u)}{\partial u}\Big|_{u=1} = \frac{\alpha k^{5/4}\mathsf{C}_{1,k}}{\sqrt{\pi\rho_k}n}\rho_k^{-n}\exp\Big(-\alpha^2(2k+\sqrt{k})\Big) + \mathcal{O}\Big(\frac{\rho_k^{-n}}{n^{3/2}}\Big).$$

Finally, combining this result and the asymptotic behavior of the sequence enumerating all k-indexed terms, we obtain that the expected number of leaves at the level $\lfloor\alpha\sqrt{n}\rfloor$ is given by

$$\frac{[z^n]\frac{\partial U_{k,\ell}(z,u)}{\partial u}\Big|_{u=1}}{[z^n]G_k(z)} \sim \frac{\frac{\alpha k^{5/4}\mathsf{C}_{1,k}}{\sqrt{\pi\rho_k}n}\rho_k^{-n}\exp\big(-\alpha^2(2k+\sqrt{k})\big)}{\frac{k^{1/4}\mathsf{C}_{1,k}}{2\sqrt{\pi\rho_k}}n^{-3/2}\rho_k^{-n}} = 2k\alpha\exp\Big(-\alpha^2(2k+\sqrt{k})\Big)\sqrt{n}.$$

□

7 Final remarks

The methods used to obtain the asymptotic mean of the size of the hat, as well as the asymptotic unary profile, in principle also serve to calculate the variances of the respective random variables. However, since calculations get rather involved when taking into account second derivatives, while the results will be not very surprising due to the tree-like structure of the studied terms, we omitted them for the sake of conciseness. Furthermore, by Theorem 6.3, we can observe that the expected unary profile of k-indexed lambda terms looks like the density of a Rayleigh distribution. As this is typical for trees, we also decided not to give a rigorous proof showing the distribution of the unary profile, since it would entail many pages of technical calculations.

It seems that so far the distribution of the number of leaves in each De Bruijn level has not been investigated, however, the total number of leaves within these terms, as well as their distribution, have been studied asymptotically by Gittenberger and Larcher (2018).

Theorem 7.1 (Gittenberger and Larcher (2018), Theorem 1)**.** *Let X_n be the total number of variables in a random closed lambda term of size n where the De Bruijn index of each variable is at most k. Then X_n is asymptotically normally distributed with*

$$\mathbb{E}X_n \sim \frac{\sqrt{k}}{1+2\sqrt{k}}n \qquad \textit{and} \qquad \mathbb{V}X_n \sim \frac{\sqrt{k}}{2(1+2\sqrt{k})^2}n \qquad \textit{as } n\to\infty.$$

Since the number of binary nodes differs only by 1 from the number of leaves, and the remaining nodes (that are neither binary nodes nor leaves) have to be unary nodes, we can state the following corollary.

Corollary 7.2. *Let Y_n be the total number of binary nodes in a random closed lambda term of size n with De Bruijn index at most k, and let Z_n be the total number of unary nodes, respectively. Then*

$$\mathbb{E}Y_n = \mathbb{E}X_n \sim \sqrt{k}\rho_k n \qquad \textit{and} \qquad \mathbb{E}Z_n \sim \rho_k n \qquad \textit{as } \ n \to \infty,$$

with X_n being defined as in Theorem 7.1.

Remark. Thus, it is an immediate observation that on average each lambda binds $\sqrt{k}$ leaves in lambda terms with De Bruijn indices being at most k.

By calculating the asymptotic number of individual constructors that occur in k-colored Motzkin trees, we get exactly the same results as in Theorem 7.1 (and therefore also as in Corollary 7.2). Furthermore, the height and the profile of these k-colored Motzkin trees are also very similar to that of lambda terms with De Bruijn indices at most k. Thus, k-indexed lambda terms are very much alike k-colored Motzkin trees. However, their counting sequences differ significantly (by a factor $\mathsf{C}_{1,k}/2$) due to the restrictions on labelling leaves in hats of the terms. So, there are way more k-colored Motzkin trees than k-indexed lambda terms. Nevertheless the great majority of them exhibits the same structural properties.

This leads to the conjecture that the problem of generating random lambda terms could be solved by means of generating random k-colored Motzkin trees and finding a suitable algorithm for *repairing their hats*. The resulting generation would not be perfectly uniform, but potentially very close to the uniform one and it would definitively be an interesting future topic to investigate.

Acknowledgements

We would like to thank Bernhard Gittenberger for fruitful discussions and suggesting the research topic. We also thank Michael Drmota for pointing out Lemma 5.1.

References

H. P. Barendregt. *The Lambda Calculus: Its Syntax and Semantics*, volume 103. North Holland, revised edition, 1984.

M. Bendkowski, K. Grygiel, P. Lescanne, and M. Zaionc. A natural counting of lambda terms. In R. M. Freivalds, G. Engels, and B. Catania, editors, *SOFSEM 2016: Theory and Practice of Computer Science*, pages 183–194, Berlin, Heidelberg, 2016. Springer Berlin Heidelberg. doi: 10.1017/S0956796815000271.

M. Bendkowski, O. Bodini, and S. Dovgal. Statistical properties of lambda terms. *The Electronic Journal of Combinatorics*, 26:P70, 2019. doi: 10.37236/8491.

M. Berger. Private communication, July 2019.

O. Bodini, D. Gardy, B. Gittenberger, and A. Jacquot. Enumeration of generalized BCI lambda-terms. *Electronic Journal of Combinatorics*, 20:P30.23, 2013.

O. Bodini, D. Gardy, B. Gittenberger, and Z. Gołębiewski. On the number of unary-binary tree-like structures with restrictions on the unary height. *Annals of Combinatorics*, 22(1):45–91, March 2018. doi: 10.1007/s00026-018-0371-7.

R. David, K. Grygiel, J. Kozik, C. Raffalli, G. Theyssier, and M. Zaionc. Asymptotically almost all λ-terms are strongly normalizing. *Logical Methods in Computer Science*, 9(1), 2013. doi: 10.2168/LMCS-9(1:2)2013.

N. G. de Bruijn. Lambda calculus notation with nameless dummies, a tool for automatic formula manipulation, with application to the Church-Rosser theorem. *Indagationes Mathematicae (Proceedings)*, 75: 381–392, December 1972. doi: 10.1016/S0049-237X(08)70216-7.

M. Drmota, A. de Mier, and M. Noy. Extremal statistics on non-crossing configurations. *Discrete Mathematics*, 327:103–117, 2014. ISSN 0012-365X. doi: 10.1016/j.disc.2014.03.016.

B. Gittenberger. On the contour of random trees. *SIAM Journal on Discrete Mathematics*, 12:434–458, 1999. doi: 10.1137/S0895480195289928.

B. Gittenberger and I. Larcher. On the number of variables in special classes of random lambda-terms. In J. A. Fill and M. D. Ward, editors, *29th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms (AofA 2018)*, volume 110 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 25:1–25:14, 2018. doi: 10.4230/LIPIcs.AofA.2018.25.

K. Grygiel and P. Lescanne. Counting and generating terms in the binary lambda calculus. *Journal of Functional Programming*, 25:e24, 2015. doi: 10.1017/S0956796815000271.

K. Grygiel, P. M. Idziak, and M. Zaionc. How big is BCI fragment of BCK logic. *Journal of Logic and Computation*, 23(3):673–691, 2013. doi: 10.1093/logcom/exs017.

P. Lescanne and J. Rouyer-Degli. Explicit substitutions with de Bruijn's levels. In *5th International Conference Rewriting Techniques and Applications (Proceedings)*, volume 914 of *Lecture Notes in Computer Science*, pages 294–308, April 1995. doi: 10.1007/3-540-59200-8_65.

J. Tromp. Binary lambda calculus and combinatory logic. In *Randomness and Complexity, From Leibniz to Chaitin*, pages 237–260, January 2007. doi: 10.1142/9789812770837_0014.

A Proof of Theorem 6.3

Here we show that the integrals along γ_j for $j \in \{2,3,4\}$ are all of order $o\left(\rho_k^{-n} n^{-3/2}\right)$ and hence the whole asymptotic contribution comes from integration along γ_1.

First, let us consider the integral along γ_4:

$$\begin{aligned}\left|\int_{\gamma_4} \frac{z^{k+\ell-n-1}}{\prod_{j=1}^{k+\ell}\sqrt{Q_{k,j}(z,1)}}dz\right| &\le \left(\rho_k(1+\varepsilon)\right)^{k+\lfloor\alpha\sqrt{n}\rfloor-n-1}|\gamma_4|\max_{z\in\gamma_4}\left|\frac{1}{\prod_{j=1}^{k+\lfloor\alpha\sqrt{n}\rfloor}\sqrt{Q_{k,j}(z,1)}}\right| \\ &\le C\rho_k^{-n}(1+\varepsilon)^{-n}\left(\rho_k(1+\varepsilon)\right)^{\lfloor\alpha\sqrt{n}\rfloor}\min_{z\in\gamma_4}\left|\sqrt{Q_{k,1}(z,1)}\right|^{-\lfloor\alpha\sqrt{n}\rfloor},\end{aligned}$$

where C is some positive constant. Here, $(1+\varepsilon)^{-n}$ contributes an exponential factor e^{-Dn} with a positive constant D, which compensates the factor $\min_{z\in\gamma_4}\left|\sqrt{Q_{k,1}(z,1)}\right|^{-\lfloor\alpha\sqrt{n}\rfloor}$ and thus guarantees

$$\int_{\gamma_4}\frac{z^{k+\ell-n-1}}{\prod_{j=1}^{k+\ell}\sqrt{Q_{k,j}(z,1)}}dz = \mathcal{O}\left(\left(\rho_k(1-\varepsilon)^{-n}\right)\right) = o\left(\rho_k^{-n}n^{-3/2}\right).$$

Now, we estimate the integral along γ_2. For some constant $C>0$, we have

$$\begin{aligned}\left|\int_{\gamma_2}\frac{z^{k+\ell-n-1}}{\prod_{j=1}^{k+\ell}\sqrt{Q_{k,j}(z,1)}}dz\right| &\leq C\left|\int_{\log^2 n}^{\varepsilon n}\frac{\rho_k^{\lfloor\alpha\sqrt{n}\rfloor-n}\left(1+\frac{t}{n}+\frac{i}{n}\right)^{\lfloor\alpha\sqrt{n}\rfloor-n}}{\sqrt{Q_{k,1}\left(\rho_k\left(1+\frac{t}{n}+\frac{i}{n}\right),1\right)}^{\lfloor\alpha\sqrt{n}\rfloor}}\frac{1}{n}dt\right|\\ &\leq C\rho_k^{-n}\frac{1}{n}\rho_k^{\lfloor\alpha\sqrt{n}\rfloor}\max_{\gamma_2}\left|\frac{z}{\sqrt{Q_{k,1}(z,1)}}\right|^{\lfloor\alpha\sqrt{n}\rfloor}\int_{\log^2 n}^{\varepsilon n}\left(1+\frac{t}{n}+\frac{i}{n}\right)^{-n}dt.\end{aligned}$$

Using the fact that $\left|\sqrt{Q_{k,1}(z,1)}\right| = \left|z+\sqrt{Q_{k,0}(z,1)}\right|$ and by Lemma 6.2, we get that the maximum contributes a factor

$$\max_{z\in\gamma_2}\left|\frac{z}{\sqrt{Q_{k,1}(z,1)}}\right|^{\lfloor\alpha\sqrt{n}\rfloor} = \max_{z\in\gamma_2}\left|\frac{1}{1+\frac{1}{z}\sqrt{Q_{k,0}(z,1)}}\right|^{\lfloor\alpha\sqrt{n}\rfloor} = \left(1+\widetilde{C}\frac{\log n}{\sqrt{n}}\right)^{\lfloor\alpha\sqrt{n}\rfloor} \sim e^{\overline{C}\log n}$$

for some positive constants $\widetilde{C}$ and $\overline{C}>0$. The remaining integral can be estimated by

$$\int_{\log^2 n}^{\varepsilon n}\left(1+\frac{t}{n}+\frac{i}{n}\right)^{-n}dt = \mathcal{O}\left(e^{-\log^2 n}\right),$$

which finally gives us

$$\left|\int_{\gamma_2}\frac{z^{k+\ell-n-1}}{\prod_{j=1}^{k+\ell}\sqrt{Q_{k,j}(z,1)}}dz\right| = \mathcal{O}\left(\rho_k^{-n}\frac{1}{n}e^{-\log^2 n+\overline{C}\log n}\right) = o\left(\rho_k^{-n}n^{-3/2}\right).$$

The estimate of the integral along γ_3 works analogously.