D I S S E R T A T I O N

PARAMETER STUDIES ON RANDOM TREES AND LAMBDA TERMS

Ausgeführt zum Zwecke der Erlangung des akademischen
Grades einer Doktorin der technischen Wissenschaften unter
der Anleitung von

AO.UNIV.PROF. DIPL.- ING. DR.TECHN. BERNHARD GITTENBERGER
E104, Institut für
Diskrete Mathematik und Geometrie

eingereicht an der Technischen Universität Wien
Fakultät für Mathematik und Geoinformation

von

ISABELLA LARCHER
Matrikelnummer: 0826592
Stolberggasse 41/13
1050 Wien

Datum	Isabella Larcher
	Bernhard Gittenberger
	Olivier Bodini

Declaration

I declare that I have written this thesis on my own, using only specified literature. Furthermore, the thesis has not been submitted, in whole or in parts, in any previous application for a scientific degree. Some parts of the thesis are based on already submitted or published papers of the author.

Vienna, April 2020

Isabella Larcher

Abstract

This thesis is devoted to asymptotic parameter studies of different combinatorial classes. Thereby the key tools are the concept of generating functions, the symbolic method and singularity analysis. The thesis is devided into three parts.

In the first part we briefly introduce the methods that are used within this thesis, give the basic definitions of the combinatorial classes that are investigated and outline their enumeration problems.

The second part is dedicated to the analysis of three tree parameters: First, we derive the asymptotic mean and variance of the protection number of a random tree as well as of a random vertex in simply generated trees, Pólya trees and non-plane binary trees. Then, we compute the average number of non-isomorphic subtree-shapes for two selected classes of increasingly labeled trees. Last, we investigate the average number of embeddings of a given rooted tree into different classes of trees, namely plane and non-plane binary trees and planted plane trees.

The third and last part of this thesis treats the analysis of shape parameters of special subclasses of lambda terms that are restricted by a bounded length of each binding, or a bounded number of nested abstractions, respectively. In particular, we are able to explain the unusual behavior of the counting sequence of the latter class by providing their asymptotic unary profile.

Zusammenfassung

Die vorliegende Dissertation beschäftigt sich mit asymptotischen Parameterstudien verschiedener kombinatorischer Klassen. Als wichtigste Werkzeuge dabei dienen erzeugende Funktionen, die symbolische Methode, sowie die Singularitätenanalyse, welche alle im ersten Teil der Arbeit eingeführt werden. Alle Resultate, die dieser Dissertation entspringen, sind asymptotischer Natur und die Arbeit kann daher klar dem Gebiet der analytischen Kombinatorik zugeordnet werden. Die Dissertation ist in drei Teile aufgeteilt.

Im ersten Teil werden die verwendeten Methoden kurz erklärt, alle benötigten Definitionen der kombinatorischen Klassen, die im weiteren untersucht werden, eingeführt, sowie deren Abzählprobleme in Kürze skizziert.

Der zweite Teil ist der Analyse von drei verschiedenen Parametern von Bäumen gewidmet. Als erstes bestimmen wir asymptotische Werte für den Mittelwert und die Varianz der sogenannten "protection number" (Länge des kürzesten Zweiges) eines zufällig gewählten Baumes bzw. Knotens eines Baumes der Klasse der einfach erzeugten Bäume, der Pólya Bäume, sowie der nicht-planaren Binärbäume.

Des weiteren ermitteln wir asymptotische Schranken für die Anzahl der nicht-isomorphen Teilbaum-Formen in zwei ausgewählten Klassen von aufsteigend markierten Bäumen.

Das letzte Kapitel dieses Teils behandelt die asymptotische Analyse eines Parameters der in verschiedenen so-genannten "optimal stopping"-Problemen eine wichtige Rolle spielt, nämlich die mittlere Anzahl an Einbettungen eines gegebenen gewurzelten Baumes in verschiedene Klassen von gewurzelten Bäumen. Dabei betrachten wir Einbettungen in die Klasse der ebenen und nicht-ebenen Binärbäume, sowie der ebenen Wurzelbäume.

Der dritte und letzte Teil der vorliegenden Dissertation beschäftigt sich mit der Untersuchung einiger Strukturparameter spezieller Klassen von Lambda-Termen. Die zwei Klassen von Lambda-Termen, die wir betrachten sind beschränkt durch eine maximale Länge der einzelnen Abstraktionen (d.h. gebundene Variablen dürfen nicht beliebig weit von dem sie bindenden Quantor entfernt sein), bzw. durch eine maximale Anzahl von ineinander geschachtelten Abstraktionen. Diese Einschränkungen ermöglichen die Anwendung der klassichen Methoden der analytischen Kombinatorik, und sind auch aus praktischer Sicht sinnvoll, da Lambda-Terme, die in der funktionalen Programmierung Anwendung finden, beide der Eigenschaften aufweisen.

Die in diesem Teil analysierten Strukturparameter betreffen die insgesamte Anzahl der Variablen in beiden Klassen von Lambda-Termen, sowie die Verteilung der Variablen über die einzelnen Abstraktionslevels der Terme. Insbesondere ermöglichen

uns die im letzten Kapitel dieses Teils erhaltenen Resultate eine Erklärung für das ungewöhnliche Verhalten der Zählfolge der betreffenden Lambda-Terme zu geben, indem wir das unäre Profil der Terme detailliert aufzeigen.

Einige Ergebnisse dieser Dissertation sind bereits in verschiedenen Arbeiten der Autorin publiziert oder zur Veröffentlichung eingereicht worden. Sämtliche Arbeiten wurden von dem FWF Projekt SFB F50-03 finanziert.

Acknowledgments

First of all, I want to dearly thank my supervisor, Bernhard Gittenberger, for his support and the friendly and encouraging atmosphere that he provided over the last years. I enjoyed our discussions and I am very thankful for the time that he took to collaborate with me. It was thanks to him that I started my PhD in the first place and his passion for the field of combinatorics inspired me throughout my whole doctoral studies. Moreover, I am very grateful that he introduced me to many bright and nice mathematicians from abroad, with whom I had the pleasure to work during several research visits from which I benefited both mathematically and personally.

In this spirit, a special thank goes to Olivier Bodini and Marc Noy for accepting to take the time to review my thesis.

The Technical University has been a very warm and welcoming place to work over the last years, due to my dear colleagues, with whom I shared a lot of nice hours working, talking as well as playing games from time to time. I want to thank all of them for giving me such a memorable time and I am sure that a lot of friendships will be kept long after finishing the PhD. I also want to thank the secretaries of our institute for their competence concerning whatever issue and for their constant care and kindness.

Furthermore, I thank my boyfriend, Philipp, who shared with me the joys and sorrows that come along with doing a PhD. Besides listening to all my talks when I practised, he always supported and motivated me and I cannot thank him enough for always being there for me.

Finally, I want to thank my family, to whom I owe everything. They always believe in me and encourage and support me in all my decisions.

Contents

Part I

Preliminaries

Chapter 1

Introduction

Trees and tree-like structures are widely known and commonly studied objects that find applications in various fields and disciplines starting from computer science up to biological and sociological research. Thereby combinatorial parameters of the investigated structures often comprise interesting information on their behavior and thus lead to an increased understanding of the underlying problem.

This thesis is concerned with parameter studies of several different classes of trees, and special plane directed acyclic graphs (PDAGs), called lambda terms. We are exclusively performing asymptotic analyses and the thereby used key tools are the concept of generating functions, the symbolic method and singularity analysis, which are all introduced in Chapter 2. In Chapter 3 we provide the basic definitions of the objects that will be studied within this thesis and briefly outline their enumeration problems.

The second part of the thesis treats the analysis of three interesting parameters for different classes of trees. In Chapter 4 we study the average length of the shortest path from the root to a leaf in a tree, called the protection number of a tree, as well as the protection number of a random vertex, which is defined as the protection number of the fringe (maximal) subtree having this node as a root. The results given in this chapter are based on the paper [55], which has recently been submitted to a journal. So far, there were several results on the number of 2-protected vertices [24, 39, 81, 82, 83], while the asymptotic average protection number of a tree was solely known for planted plane trees [65]. We generalized these results to a larger class of rooted trees, by studying both the average protection number of a tree as well as a random vertex protection number for the family of simply generated trees (introduced by Meir and Moon [85]) and their nonplane counterparts: unlabeled nonplane rooted trees, also called Pólya trees due to their first extensive treatment by Pólya [93], examined further by Otter [89] including numerical results and the binary case. Thus, our work broadens the results from [65], but maintaining the emphasis on as concrete formulas as possible.

Chapter 5 is devoted to the asymptotic investigation of the average number of non-isomorphic subtree-shapes in selected classes of increasing trees, based on the not yet submitted manuscript [14]. This parameter is often studied in the context of so-called compacted trees [49, 50] that can be constructed from any given tree by means of a post-order traversal where repeatedly occurring subtrees are replaced by directed edges pointing at the already traversed appearance of the respective subtree. In this way,

the size of the compacted tree belonging to a given tree T corresponds to the number of non-isomorphic subtrees of the tree T. So far, it was known that for a random simply generated tree of size n the average size of a compacted tree (*i.e.*, the average number of non-isomorphic subtrees) is asymptotically of size $\Theta\left(\frac{n}{\sqrt{\log n}}\right)$. We extended the definition of a compacted tree to special classes of increasingly labeled trees and proved that for these classes the average size of a compacted tree, corresponding to the average number of non-isomorphic subtree-shapes, is asymptotically $\Omega(\sqrt{n})$ and $\mathcal{O}\left(\frac{n}{\log n}\right)$.

The last tree parameter, which is studied in Chapter 6 of this thesis, is the number of embeddings of a given rooted tree in the family of (plane and non-plane) binary trees, as well as planted plane trees. Here, the notion of embedding is wider than just a copy. We assume the investigated structures to be some kind of Hasse diagrams of partially ordered sets (in short: posets) and by saying that there exists an embedding of S in T we understand that a poset S is a subposet of T. In particular, we call an embedding of S in T to be a good embedding if it contains the root of T. Based on the not yet submitted article [54] we present a follow-up and generalization of the results obtained by Kubicki, Lehel and Morayne in [72, 73] and Georgiou in [51], where they derived the number of (good and all) embeddings of a given plane tree S in a complete balanced binary tree. We give the asymptotic mean of the number of good embeddings and the number of all embeddings of a rooted tree S in the family of plane and non-plane binary trees, as well as planted plane trees, on n vertices. We prove that the ratio of the number of good embeddings to the number of all embeddings is asymptotically equivalent to $c/\sqrt{n}$ in all cases and provide the exact constant c. Furthermore, we show monotonicity properties for this ratio and briefly discuss the case of embedding a disconnected structure S, *i.e.*, a forest.

In Chaper 7 we briefly discuss the results obtained in Part II of this thesis that is devoted to the investigation of tree parameters and provide an outlook to some related open problems that might be interesting for future studies.

The third part of this thesis is dedicated to the study of two special restricted classes of lambda terms, where we perform an asymptotic analysis of their shape by determining the so-called unary profile. Lambda terms are objects stemming from lambda calculus and can be seen as combinatorial objects with a simple description. Nevertheless, the enumeration of lambda terms is not well understood. Combinatorially, they can be seen as words (sequences of symbols) or graphs and thus the combinatorially most natural way to define an enumeration problem is to ask for the number of terms with a given number of symbols or vertices, respectively. This problem appears to be very intriguing, as the standard techniques of analytic combinatorics fail. Considering subclasses by imposing certain restrictions can turn the enumeration problem into an accessible one, but for one particular model the enumeration formulas exhibit a very peculiar behavior. Our motivation to perform the present investigation is to shed light on this oddity and to give a combinatorial explanation for this phenomenon.

By considering lambda terms as formulas which contain a quantifier that binds variables (see Section 3.2 for the precise definition), the first subclass of lambda terms, which is studied in Chapter 8, is restricted by a maximal length of its bindings, *i.e.*, a maximal number of symbols between each quantifier and the thereby bounded

variables. We extend the results presented in [11], where the enumeration problem of this class of lambda terms was studied thoroughly, and perform an analysis of the asymptotic shape of a random term belonging to this class. First, we show that the total number of variables is asymptotically normally distributed and give explicit formulas for the mean and the variance, based on the already published paper [57]. Then, we investigate the structure of these terms and provide their unary profile, which has been processed in the already submitted paper [62].

Chapter 9 treats the analysis of a second subclass of lambda terms that is restricted by a maximum number of nested bindings and is based on the paper [57]. In [11] it was shown that the enumeration sequence of this combinatorial class admits a very unusual behavior (which is outlined in Chapter 3). With the aim to understand the reason for the occurring phenomenon, we perform asymptotic analyses concerning the average number of variables and the unary profile of such terms.

Finally, in Chapter 10 we discuss the results that have been in obtained in Part III of this thesis and give a short outlook on some remaining open problems.

Chapter 2

Concepts of analytic combinatorics

In this chapter we introduce some important tools and methods that will be used within this thesis. Our goal is to gain information on various shape parameters of different combinatorial objects, namely several classes of trees and lambda terms. In order to do so the concepts of generating functions and singularity analysis, which are introduced in Sections 2.1 and 2.2, are essential. Furthermore, in Section 2.3 we introduce combinatorial parameters, bivariate generating functions and probabilistic limit theorems that enable us to obtain distributional results concerning random variables related to shape parameters of combinatorial classes.

The results presented in this chapter are strongly based on the book 'Analytic Combinatorics' of Flajolet and Sedgewick [45], to which we also refer the interested reader for gaining further information.

2.1 The symbolic method and generating functions

The symbolic method provides a very simple and efficient systemic treatment of combinatorial constructions. Yet before we can give an accurate explanation, we need to introduce some basic definitions, as they can be found in [45].

Definition 2.1 (combinatorial class, [45, Definition I.1])**.** *A combinatorial class $\mathcal{A}$ is a finite or denumerable set, on which a size function $|\cdot|$ is defined, satisfying the following conditions:*

(i) The size $|a|$ of an element $a \in \mathcal{A}$ is a non-negative integer.

(ii) The number of elements of any given size is finite.

We denote by $\mathcal{A}_n$ the class of elements of $\mathcal{A}$ having size n, *i.e.*, $A_n = \{a \in \mathcal{A} : |a| = n\}$. Moreover, A_n denotes the number of objects in the class $\mathcal{A}_n$, *i.e.*, $A_n = \operatorname{card}(\mathcal{A}_n)$. Two combinatorial classes $\mathcal{A}$ and $\mathcal{B}$ are called (combinatorially) *isomorphic*, if the sequences $(A_n)_{n\geq 0}$ and $(B_n)_{n\geq 0}$ are identical, i.e. $A_n = B_n, \quad \forall n \in \mathbb{N}_0$.

Definition 2.2 (generating function, [45, Definitions I.4 and II.2])**.** *The ordinary generating function (OGF) of a combinatorial class $\mathcal{A}$, or the sequence $(A_n)_{n\geq 0}$ respectively, is the formal power series*

$$A(z) = \sum_{n=0}^{\infty} A_n z^n. \tag{2.1}$$

Similarly, the exponential generating function (EGF) of $\mathcal{A}$ (or $(A_n)_{n\geq 0}$) is given by

$$A(z) = \sum_{n=0}^{\infty} A_n \frac{z^n}{n!}. \tag{2.2}$$

The choice of using ordinary generating functions or exponential generating functions depends on the kind of problem. One usually uses OGFs for unlabeled structures and EGFs for labeled structures.

Remark 2.3. *Equivalently to* (2.1) *and* (2.2)*, the generating function of the class $\mathcal{A}$ admits the representation $A(z) = \sum_{\alpha\in\mathcal{A}} z^{|\alpha|}$, or $A(z) = \sum_{\alpha\in\mathcal{A}} \frac{z^{|\alpha|}}{|\alpha|!}$, respectively.*

By $[z^n]$ we denote the operation of *coefficient extraction*, *i.e.*,

$$[z^n]A(z) = [z^n]\left(\sum_{n\geq 0} A_n z^n\right) = A_n,$$

which returns the number of objects of the corresponding combinatorial class $\mathcal{A}$ of size n. For some combinatorial problems we can derive an explicit formula for A_n. Unfortunately, in many cases this does not seem to be possible. However, with methods from analytic combinatorics we are able to determine the order of magnitude of A_n asymptotically as n tends to infinity. A first step for the investigation of combinatorial counting problems is to set up equations that specify the respective generating functions. In order to do so, the symbolic method has turned out to be a very simple and efficient tool. Within the symbolic method combinatorial classes are built directly in terms of simpler classes by means of a collection of combinatorial constructions, which can easily be translated into generating functions. Table 2.1 summarizes the most important constructions together with their counterpart in relation to generating functions.

Combinatorial Construction		OGF / EGF
Neutral set	$\mathcal{E} = \{\epsilon\}$	$E(z) = 1$
Atomic set	$\mathcal{Z} = \{a\}$	$Z(z) = z$
Disjoint union	$\mathcal{C} = \mathcal{A} \cup \mathcal{B}$	$C(z) = A(z) + B(z)$
Cartesian/partition product	$\mathcal{C} = \mathcal{A} \times \mathcal{B}$	$C(z) = A(z) \cdot B(z)$
Sequence	$\mathcal{C} = \mathcal{S}\mathrm{eq}(\mathcal{A})$	$C(z) = \frac{1}{1-A(z)}$
Set	$\mathcal{C} = \mathcal{S}\mathrm{et}(\mathcal{A})$	$C(z) = e^{A(z)-\frac{1}{2}A(z^2)+\frac{1}{3}A(z^3)\mp\dots}$ / $e^{A(z)}$
Multiset	$\mathcal{C} = \mathcal{M}(\mathcal{A})$	$C(z) = e^{A(z)+\frac{1}{2}A(z^2)+\frac{1}{3}A(z^3)+\dots}$
Substitution	$\mathcal{C} = \mathcal{A}(\mathcal{B})$	$C(z) = A(B(z))$
Pointing	$\mathcal{C} = \Theta\mathcal{A}$	$C(z) = zA'(z)$
Boxed product	$\mathcal{C} = \mathcal{A}^{\square} \times \mathcal{B}$	not applicable / $C'(z) = A'(z) \cdot B(z)$

Table 2.1: A summary of the most important constructions, and their translations into generating functions. The neutral set $\mathcal{E}$ consists of one element of size 0, while the atomic set $\mathcal{Z}$ contains just one element of size 1. In most cases the translations into OGFs and EGFs work analogously, with the exception of the set construction. The boxed product is solely defined for exponential generating functions, since $\mathcal{C} = \mathcal{A}^{\square} \times \mathcal{B}$ corresponds to the subset of the product $\mathcal{A} \times \mathcal{B}$ consisting of labeled elements such that the smallest label belongs to an element from $\mathcal{A}$.

We will now exemplify the symbolic method by means of the very simple and well-known combinatorial class of binary trees.

Example 2.4 (Binary trees)**.** *A binary tree, is a rooted tree (i.e., a graph without cycles that contains one distinguished node called the root), where all nodes have either two children, or no children at all (then it is called as leaf). Let us denote by $\mathcal{B}$ the class of binary trees and let us define the size of a binary tree to be the number of its internal nodes (i.e., leaves do not count to the size). Then $\mathcal{B}$ can be recursively specified by*

$$\mathcal{B} = \mathcal{E} \cup (\mathcal{Z} \times \mathcal{B} \times \mathcal{B}) .$$

Using the translation rules summarized in Table 2.1 we get the equation

$$B(z) = 1 + zB(z)^2,$$

which defines the generating function $B(z)$ of the class of binary trees. In this case we can solve the equation explicitly and directly get

$$B(z) = \frac{1 - \sqrt{1-4z}}{2z}.$$

In the next section we will turn to the analysis of the coefficients of generating functions in order to gain asymptotic information on the number of structures of a certain size.

2.2 Singularity analysis

Singularity analysis relies on the simple principle that some special points of a generating function, called singularities, are reflected in the function's coefficients. Therefore we gain information on the order of magnitude of A_n by determining the singularities of the OGF $A(z) = \sum_{n\geq 0} A_n z^n$. These results can be obtained by no longer considering generating functions as formal power series, but as functions in the complex plane that are analytic around 0. We refer the reader who is not familiar with basic concepts of complex analysis to [45], since we will use some of these concepts in the sequel.

Definition 2.5 (singularity, [45, Def. IV.4])**.** *Given a function f defined in the region interior to the simple closed curve γ, a point z_0 on the boundary (γ) of the region is a singularity, if f is not analytically continuable to z_0.*

In short, singularities are points where a function ceases to be analytic. The singularities which are closest to the origin, are called the *dominant singularities*, and will turn out to be particularly important. Their distance to the origin equals the radius of convergence. The general form of the coefficients of a generating function looks like $[z^n]A(z) = a^n\theta(n)$, where a denotes the exponential growth factor and $\theta(n)$ a subexponential factor, *i.e.*, $\limsup |\theta(n)|^{\frac{1}{n}} = 1$. In [45] Flajolet and Sedgewick introduced the following two principles:

- First Principle of Coefficient Asymptotics:

  The location of a function's singularities dictates the exponential growth of its coefficients, *i.e.*, a^n.

- Second Principle of Coefficient Asymptotics:

  The nature of a function's singularities determines the associate subexponential factor $\theta(n)$.

The first principle is specified by the following theorem:

Theorem 2.6 (Exponential growth formula, [45, Thm. IV.7]). *If $A(z)$ is analytic at 0 and R is the radius of convergence, i.e. $R := sup\{r \geq 0|\ A(z)$ is analytic in $|z| < r\}$, then the coefficient $A_n = [z^n]A(z)$ satisfies*

$$A_n \sim \left(\frac{1}{R}\right)^n.$$

For functions with non-negative coefficients, including all combinatorial generating functions, one can also adopt

$$R := sup\{r \geq 0|\ A(z) \text{ is analytic at all points of } 0 \leq z < r\}.$$

Thus, the exponential factor can easily be determined by computing the radius of convergence. In order to derive the subexponential factor, we have to distinguish between certain kinds of functions according to the type of their singularities: For meromorphic functions, which have only polar singularities, the subexponential factor $\theta(n)$ is of polynomial growth.

Theorem 2.7 (Expansion of meromorphic functions, [45, Thm. IV.10]). *Let $A(z)$ be a function meromorphic at all points of the closed disc $|z| \leq R$, with poles at points $\alpha_1, \alpha_2, \ldots, \alpha_m$. Assume that $A(z)$ is analytic at all points of $|z| = R$ and at $z = 0$. Then there exist m polynomials $\{\prod_j(x)\}_{j=1}^m$ such that*

$$A_n \equiv [z^n]f(z) = \sum_{j=1}^{m} \prod_j (n)\alpha_j^{-n} + O(R^{-n}).$$

Furthermore the degree of $\prod_j$ is equal to the order of the pole of f at α_j minus one.

Now we consider functions whose singularities are of richer nature than poles. Our goal is to translate an expansion of a generating function $A(z)$ near its singularity, called the *Puiseux expansion*, into an asymptotic approximation of its coefficients. The basic property that allows for such an asymptotic transfer is the so-called *Transfer Theorem* [44], which requires that $A(z)$ is analytic in a so-called *Delta-domain* Δ that is depicted in Figure 2.1.

Theorem 2.8 (Transfer Theorems, [34, Lemma 2.12]). *Let $A(z) = \sum_{n\geq 0} A_n z^n$ be analytic in a Delta-domain*

$$\Delta = \Delta(\rho, \eta, \phi) = \{z : |z| < \rho + \eta, \left|\arg\left(\frac{z}{\rho} - 1\right)\right| > \phi\},$$

in which ρ and η are positive real numbers and $0 < \phi < \pi/2$. Furthermore, suppose that there exists a real number α such that

$$A(z) = \mathcal{O}\left((1 - z/\rho)^{-\alpha}\right),$$

for $z \in \Delta$. Then

$$A_n = \mathcal{O}\left(\rho^{-n} n^{\alpha-1}\right).$$

Similarly, if there exists a real number α such that

$$A(z) = o\left((1 - z/\rho)^{-\alpha}\right),$$

for $z \in \Delta$, we have

$$A_n = o\left(\rho^{-n} n^{\alpha-1}\right).$$

Remark 2.9. *We will call the number $-\alpha$ the type of the singularity. Sometimes a singularity of type $\frac{1}{2}$ is interchangeably called a square root singularity.*

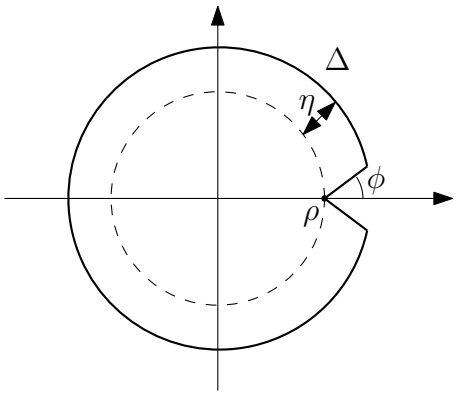


Figure 2.1: Delta-domain $\Delta = \Delta(\rho, \eta, \phi)$.

Furthermore, for functions of the so-called *standard scale*, *i.e.*, functions of the form $(1 - z)^{-\alpha}$ for $\alpha \in \mathbb{C} \setminus \mathbb{Z}_{\leq 0}$, we have the following result.

Theorem 2.10 ([45, Thm. VI.1])**.** *Let $\alpha \in \mathbb{C} \setminus \mathbb{Z}_{\leq 0}$. Then*

$$[z^n](1 - z)^{-\alpha} \sim \frac{n^{\alpha-1}}{\Gamma(\alpha)} \left(1 + \sum_{k=1}^{\infty} \frac{e_k}{n^k}\right),$$

where e_k is a polynomial in α of degree $2k$.

Theorems 2.8 and 2.10 directly yield the following corollary.

Corollary 2.11 ([34, Corollary 2.15])**.** *Suppose that a function is analytic in a Delta-domain and that it has an Puiseux expansion of the form*

$$A(z) = C\left(1 - \frac{z}{\rho}\right)^{-\alpha} + \mathcal{O}\left(\left(1 - \frac{z}{\rho}\right)^{-\beta}\right),$$

for $z \in \Delta$, where $\beta < \mathrm{Re}(\alpha)$. Then we have

$$A_n = [z^n]A(z) = C\frac{n^{\alpha-1}}{\Gamma(\alpha)}\rho^{-n} + \mathcal{O}\left(\rho^{-n} n^{\max\{\mathrm{Re}(\alpha)-2, \beta-1\}}\right).$$

Example 2.12 (Binary trees, continuation)**.** *Let us remember that in Example 2.4 we got*

$$B(z) = \frac{1 - \sqrt{1-4z}}{2z}.$$

In this case we are actually able to obtain an explicit formula for the coefficients B_n of $B(z)$. However, since we want to exemplify the singularity analysis approach, we will not use the explicit solution for B_n. As a first step, we have to calculate the dominant singularity of $B(z)$, which is given by $\rho = \frac{1}{4}$. Since we have a unique isolated singularity, the important requirement of Delta-analyticity is fulfilled. For $z \to \frac{1}{4}$ the generating function $B(z)$ admits the Puiseux expansion

$$B(z) = 2 - 2\sqrt{1 - \frac{z}{\rho}} + \mathcal{O}\left(\left(1 - \frac{z}{\rho}\right)^{3/2}\right),$$

which yields according to the transfer theorems

$$\begin{aligned} B_n = [z^n]B(z) &= -2\frac{n^{-1/2}}{\Gamma(-1/2)}\left(\frac{1}{4}\right)^{-n} + \mathcal{O}\left(\left(\frac{1}{4}\right)^{-n} n^{-5/2}\right) \\ &= \frac{1}{\sqrt{\pi}} n^{-3/2} 4^{-n}\left(1 + \mathcal{O}\left(n^{-1}\right)\right), \end{aligned}$$

where the last equality follows by the use of $\Gamma(-1/2) = -2\sqrt{\pi}$.

2.3 Combinatorial parameters and limit laws

Now we will introduce the concept of combinatorial parameters and explain how studying these parameters by means of bivariate generating functions leads to an understanding of the distribution of the parameter's values. For more thorough information the reader is referred to [45, Chapter III and IX]. Let us start with the formal definition of a combinatorial parameter and bivariate generating functions.

Definition 2.13 (combinatorial parameter, bivariate genrating function, [45, Definition III.2])**.** *Given a combinatorial class $\mathcal{A}$, a (scalar) parameter is a function from $\mathcal{A}$ to $\mathbb{Z}_{\geq 0}$ that associates to any object $a \in \mathcal{A}$ an integer value $\chi(a)$. The sequence*

$$A_{n,k} = \operatorname{card}\left(\{a \in \mathcal{A} : |a| = n, \chi(a) = k\}\right),$$

is called the counting sequence of the pair $(\mathcal{A}, \chi)$. The bivariate generating function (BGF) of $(\mathcal{A}, \chi)$ is then defined as

$$A(z,u) := \sum_{n \geq 0} \sum_{k \geq 0} A_{n,k} \frac{z^n}{\omega_n} u^k,$$

and is ordinary if $\omega_n \equiv 1$ and exponential if $\omega_n \equiv n!$. One says that the variable z marks the size and the variable u marks the parameter χ.

Remark 2.14. *Obviously, $A(z,1)$ reduces to the usual univariate generating function $A(z)$ associated with $\mathcal{A}$.*

At this point we have to distinguish between certain kinds of parameters, namely so-called *inherited parameters*, *recursive parameters* and *extremal parameters*. We omit the formal definitions of these types of parameters and instead try to shortly explain the differences of the three types and how to deal with them.

An inherited parameter fulfils axioms that are in fact a natural extension of the axioms the size itself has to satisfy (see [45, Definition III.5]). Thus, by extending the basic combinatorial constructions to include bivariate generating functions whose second variable carries information about the parameter, the symbolic method is able to directly take such parameters into account. An example for an inherited parameter is for instance the root degree of a plane tree (for the precise definitions see Chapter 3).

Recursive parameters are - as the name indicates - parameters that are defined by recursive rules over structures that are themselves recursively specified. These parameters are typical for trees and tree-like structures and will occur at several points within this thesis (e.g. the number of leaves of a plane tree). The bivariate generating function is set up by means of a marking process where it suffices to distinguish the elements of interest and mark them by the auxiliary variable, see Example 2.15.

Thus, for both cases of inherited and recursive parameters, the symbolic method can be extended to bivariate generating functions, and thus it serves not just as a tool to count combinatorial objects but also to quantify their properties.

Example 2.15 (Binary trees, continuation)**.** *Let us recall the specification for binary trees given in Example 2.4 by*

$$\mathcal{B} = \mathcal{E} \cup (\mathcal{Z} \times \mathcal{B} \times \mathcal{B}) .$$

Now, we want to set up the bivariate generating function $B(z, u)$, where z marks the size (i.e., the number of internal nodes) and u marks the number of leaves. So, the parameter we are interested in is now the total number of leaves. Then we get via the marking process and the symbolic method

$$B(z, u) = u + zB(z, u)^2.$$

Every leaf is now marked with a u, while the internal nodes are solely marked by a z.

The last type of parameters are the so-called extremal parameters, which are defined by a maximum rule. A typical extremal parameter is for example the height of a tree. In this case, the non-linearity of the maximum function prevents a suitable use of bivariate generating functions. The standard technique is to introduce a collection of univariate generating functions defined by imposing a bound on the parameter of interest. For setting up the generating functions of these restricted combinatorial classes, we use again the symbolic method in its univariate version.

Now, that we have an idea how to set up a bivariate generating function $A(z, u)$, where z marks the size and u marks the parameter of interest, we introduce some important approaches that uses $A(z, u)$ in order to get some information on the distribution of the parameter.

In general, given a combinatorial class $\mathcal{A}$, we will always assume an *uniform probability distribution* over $\mathcal{A}_n$, *i.e.* we assume that all $a \in \mathcal{A}_n$ appear with the equal

likelihood of $\frac{1}{A_n}$. Every parameter χ determines a discrete random variable χ_n defined over the discrete probability space $\mathcal{A}_n$ via

$$\chi_n = \mathbb{P}_{\mathcal{A}_n}(\chi = k) = \frac{A_{n,k}}{A_n} = \frac{A_{n,k}}{\sum_k A_{n,k}}.$$

The probability generating function of χ over $\mathcal{A}_n$ is then given by

$$p(u) = \sum_k \mathbb{P}_{\mathcal{A}_n}(\chi = k) = \frac{[z^n]A(z,u)}{[z^n]A(z,1)},$$

where $A(z,u)$ is the bivariate generating function of $(\mathcal{A}, \chi)$. Important information about a random variable can be obtained by calculating its *moments*. Given a discrete random variable X, the moments are defined via

$$\mathbb{E}(X^r) := \sum_k \mathbb{P}(X = k)k^r.$$

Then the expectation (or average, mean) of X, its variance, and its standard deviation, respectively, are expressed as

$$\mathbb{E}(X), \quad \mathbb{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2, \quad \sigma(X) = \sqrt{\mathbb{V}(X)}.$$

Proposition 2.16. *The expected value of a parameter χ is determined from the BGF $A(z,u)$ via*

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

In order to get some information on the distribution of a parameter, one typically investigates the so-called *characteristic function*, which can be obtained from the probability generating function $p(u) = \mathbb{E}(u^X)$ by setting $u = e^{it}$.

Theorem 2.17 (Levy's continuity theorem, [45, Theorem IX.4])**.** *Let Y and Y_n be random variables with characteristic functions $\phi(t)$, $\phi_n(t)$, and assume that y has a continuous distribution function. A necessary and sufficient condition for the convergence in distribution, $Y_n \Rightarrow Y$, is that, pointwise, for each real t,*

$$\lim_{n\to\infty} \phi_n(t) = \phi(t).$$

In particular, the characteristic function of some prominent probability distributions are given in Table 2.2.

Distribution	Characteristic function $\phi(t)$
Normal(μ,σ^2)	$e^{it\mu - \frac{1}{2}\sigma^2 t^2}$
Poisson(λ)	$e^{\lambda(e^{it}-1)}$
Exponential(λ)	$(1 - it\lambda^{-1})^{-1}$

Table 2.2: Characteristic functions of some well-known probability distributions.

Another very powerful theorem, that we use within this thesis in order to prove hat a sequence of random variables is asymptotically normally distributed, is the so-called *Quasi-Powers Theorem*. The idea is that if the characteristic function of a sequence of random variables X_n behaves almost like powers of a function, then the distribution of X_n should be approximated by a corresponding sum of i.i.d. random variables and, thus, one can expect a central limit theorem, [34].

Theorem 2.18 (Quasi-Powers Theorem, [67]). *Let X_n be a sequence of random variables with the property that*

$$\mathbb{E}(u^{X_n}) = A(u)B(u)^{\lambda_n}\left(1 + \mathcal{O}\left(\frac{1}{\phi_n}\right)\right)$$

holds uniformly in a complex neighbourhood of $u = 1$, where $\lambda_n \to \infty$ and $\phi_n \to \infty$, and $A(u)$ and $B(u)$ are analytic functions in a neighbourhood of $u = 1$ with $A(1) = B(1) = 1$. Set $\mu = B'(1)$ and $\sigma^2 = B''(1) + B'(1) - B'(1)^2$. If $\sigma^2 \neq 0$, then

$$\frac{X_n - \mathbb{E}(X_n)}{\sqrt{\mathbb{V}(X_n)}} \to \mathcal{N}(0, 1),$$

with $\mathbb{E}(X_n) = \mu\lambda_n + A'(1) + \mathcal{O}(1/\phi_n))$ and $\mathbb{V}(X_n) = \sigma^2\lambda_n + A''(1) + A'(1) - A'(1)^2 + \mathcal{O}(1/\phi_n))$.

Now, as we have introduced the most important methods that are used within this thesis, we turn to the definition of the basic structures that we are going to investigate, namely trees and lambda terms and outline their enumeration problems.

Chapter 3

Enumeration of trees and lambda terms

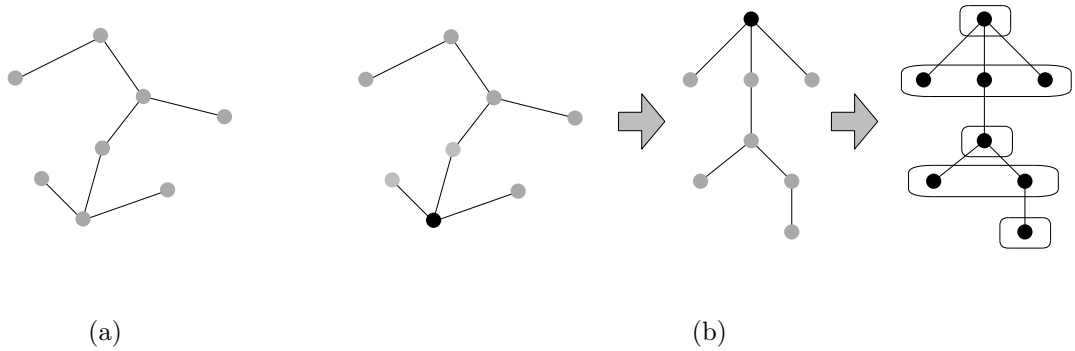
This chapter is split into two parts. In the first part, the basic concepts related to trees, as well as some important tree classes are introduced together with their generating functions. Furthermore, asymptotic results concerning their counting sequences are presented. The second part is devoted to the definition and the counting problems of special classes of lambda terms.

3.1 Trees

In this section we introduce some fundamentals about combinatorial trees, basic definitions and important results concerning the enumeration problem of selected classes of trees. For more detailed information see [34].

3.1.1 Basic definitions

In general, trees are connected graphs that do not contain any cycles. In particular we distinguish between rooted and unrooted, plane and non-plane, and labeled and unlabeled trees. Within this thesis we will exclusively deal with rooted trees, *i.e.*, trees that include one distinguished node, called the *root* of the tree. In analogy to biological trees all nodes with degree 1 are called *leaves* (except for the root, which is solely called a leaf when it has degree 0) and the path connecting the root and a leaf is called a *branch*. The *size* $|T|$ of a tree T is defined as its total number of vertices. The *height* $h(v)$ *of a node* v is defined as the length of the path connecting the root with v, while the *height* $h(T)$ *of a tree* T is the length of the longest path from the root to a leaf, *i.e.*, the length of the longest branch. When visualizing a tree we will always use the convention that the root is drawn as the topmost node. In this way we can speak about *levels* in trees in an intuitive way, thus the level in which a certain node is located coincides with its respective height, see Figure 3.1. Furthermore, for every fixed vertex v we call all nodes that are connected to v and that are located in the next level (*i.e.*, their height is $h(v) + 1$), the *child-nodes* or *children* of v, while v is called their *parent-node*.



(a) (b)

Figure 3.1: Unrooted tree (a), and a rooted tree along with its levels encircled (b).

Rooted trees are planar graphs in the sense that we can embed them into the plane without crossings. However, when we speak of *plane trees*, we mean that we distinguish between all possible different embeddings into the plane. Thus, the trees in Figure 3.2 are assumed to be different plane trees, whereas they represent the same *non-plane tree*. Obviously, this is an important issue when it comes to the problem of counting trees.

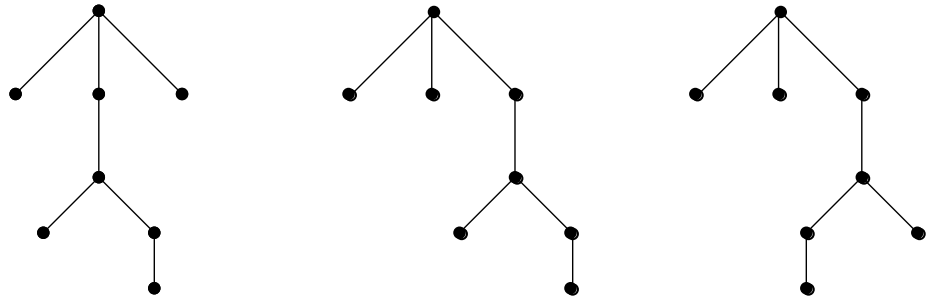


Figure 3.2: Different embeddings that represent the same non-plane tree.

Furthermore, we distinguish between *labeled trees*, where the nodes are labeled with different numbers from 1 to the total number of nodes, and *unlabeled trees* that do not contain any labels. Again, this is particularly important for the counting problem, since there are always more labeled than unlabeled trees with a given number of nodes (except for trees with only 1 node). There are different ways to label a tree. One model that will be used within this thesis is to choose the labels in an increasing manner, *i.e.*, the label of the parent-node has to be always smaller than the label of its child-nodes, see Figure 3.3. This concept leads to the class of *increasing trees*, which will be introduced more thoroughly in the remainder of this subsection.

Now we introduce some important classes of trees that will be considered thereinafter in this thesis and we outline their enumeration problems.

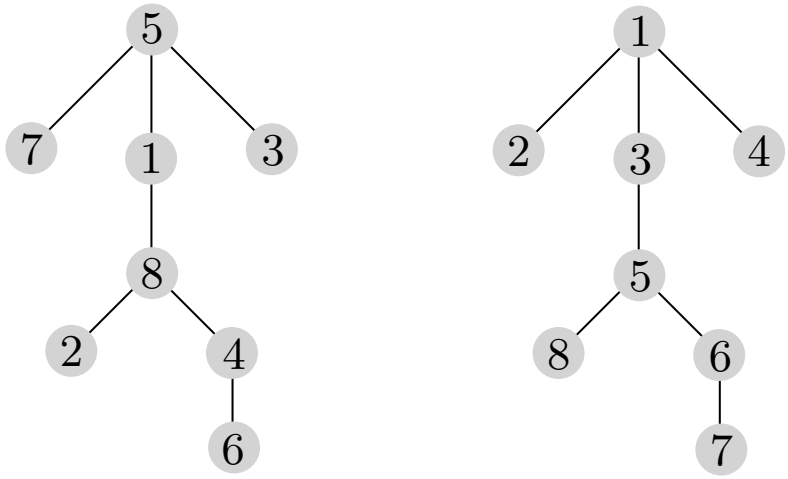


Figure 3.3: Two different labelings of a rooted tree, where the right one is in particular an increasing labeling.

3.1.2 Counting trees

A particularly important class of trees is the class of *simply generated trees*, which comprises many famous representatives, such as binary trees or planted plane trees.

Simply generated trees. Simply generated trees are unlabeled and weighted rooted trees that were introduced by Meir and Moon [85]. Their generating function $S(z)$ is implicitly defined via

$$S(z) = z\Phi(S(z)), \tag{3.1}$$

where $\Phi(z) = \phi_0 + \phi_1 z + \phi_2 z^2 + \ldots = \sum_{j\geq 0} \phi_j z^j$ and $(\phi_j)_{j\geq 0}$ is called the *weight sequence*. Usually one assumes $\phi_0 > 0$ and $\phi_j > 0$ for some $j \geq 2$. The *weight* $w(T)$ of a tree T is then defined as

$$w(T) = \prod_{v\in V(T)} \phi_{d^+(v)},$$

with $V(T)$ denoting the vertex set of T and $d^+(v)$ denoting the out-degree of the vertex v, *i.e.*, the number of children of v.
The number $S_n = [z^n]S(z)$ of simply generated trees of size n is then given by

$$S_n = \sum_{|T|=n} w(T).$$

For some subclasses of simply generated trees an explicit formula for the coefficients S_n can be obtained (mostly by the use of the Lagrange inversion formula), while in most cases it is not possible to get a closed formula for the precise numbers. However, there is a general asymptotic result which relies on the fact that (under certain conditions) the generating function $S(z)$ has a dominant singularity ρ of square root type at a finite radius of convergence and is given in the following theorem, see [34, Theorem 3.6].

Theorem 3.1 ([34, Theorem 3.6])**.** *Let $S_n = [z^n]S(z)$ be the number of simply generated trees of size n, and let R denote the radius of convergence of $\Phi(z)$, with $S(z)$*

and $\Phi(z)$ as in (3.1). Suppose that there exists τ with $0 < \tau < R$ that satisfies $\tau\Phi'(\tau) = \Phi(\tau)$. Set $d = \gcd\{j > 0 : \phi_j > 0\}$. Then

$$S_n = \begin{cases} d\sqrt{\frac{\Phi(\tau)}{2\pi\Phi''(\tau)}}\frac{\Phi'(\tau)^n}{n^{3/2}}\left(1+\mathcal{O}(n^{-1})\right) & \text{if } n \equiv 1 \mod d \\ 0 & \text{if } n \not\equiv 1 \mod d \end{cases},$$

for $n \to \infty$.

Remark 3.2. *The quantity d in Theorem 3.1 is a measure for the periodicity of the function $\Phi(z)$, which is directly related to the number of dominant singularities. Thus, for $d = 1$ we are in the non-periodic case, where there exist trees of arbitrary sizes and where we have a single dominant singularity. For $d > 1$ there are solely trees of size $1 \mod d$ and $\Phi(z)$ has d dominant singularities that all contribute to the asymptotics of the coefficients S_n.*

In the subsequent examples, we introduce some special subclasses of simply generated trees that are particularly important.

Example 3.3 (Planted plane trees)**.** *Planted plane trees (or Catalan trees) are rooted plane trees, where each node has an arbitrary number of children, see Figure 3.4. Their generating function $C(z)$ is obtained by setting $\phi_j = 1$ for all $j \geq 0$. Then all trees T have weight $w(T) = 1$ and $\Phi(z) = \frac{1}{1-z}$. Thus, $C(z)$ satisfies the relation $C(z) = z\frac{1}{1-C(z)}$ (see [34]). The number C_n of planted plane trees of size n is known to be explicitly given by*

$$C_n = \frac{1}{n}\binom{2n-2}{n-1}.$$

The asymptotic behavior of the coefficients C_n can be derived by means of Theorem 3.1 (or directly by use of Stirling's formula) and it reads as

$$C_n = \frac{1}{4\sqrt{\pi}}\frac{4^n}{n^{3/2}}\left(1+\mathcal{O}(n^{-1})\right) \qquad \text{for } n \to \infty.$$

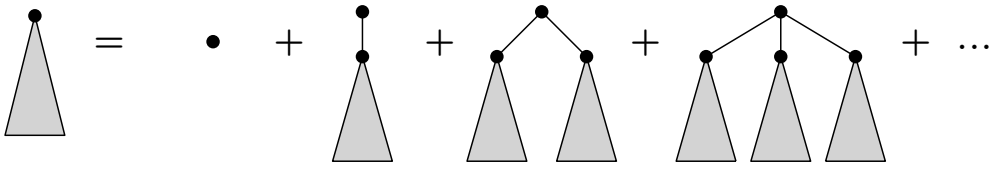


Figure 3.4: Recursive structure of planted plane trees.

Notation 3.4. *As the numbers C_n are just the shifted Catalan numbers, planted plane trees are often called Catalan trees. Within this thesis we will always denote the n-th Catalan number by $\boldsymbol{C}_n$. Thus, $C_n = \boldsymbol{C}_{n-1}$.*

Example 3.5 (Binary trees)**.** *Binary trees are rooted plane trees, where each node has either two children (one to the left, one to the right), or no children at all (then it is a leaf), see Figure 3.5. Their generating function $B(z)$ is obtained by setting $\phi_0 = 1$, $\phi_2 = 1$ and $\phi_j = 0$ for $j = 1$ and for all $j \geq 3$. Then we get $\Phi(z) = 1 + z^2$*

and thus $B(z) = z(1 + B(z)^2)$. *Now we are in a periodic case, since* $d = \gcd\{j > 0 : \phi_j > 0\} = 2$. *This is reflected by the fact that there are no binary trees of even size (when the size is defined as the total number of nodes). By Theorem 3.1 we have for* $n \to \infty$

$$B_n = [z^n]B(z) = \begin{cases} \sqrt{\frac{2}{\pi}} \frac{2^n}{n^{3/2}} \left(1 + \mathcal{O}(n^{-1})\right) & \text{if } n \text{ is odd} \\ 0 & \text{if } n \text{ is even} \end{cases}.$$

This result can also be obtained by means of Stirling's formula, since we know that the exact number of binary trees of size n *is given by the* $\frac{n-1}{2}$*-th Catalan number, i.e.,*

$$B_n = [z^n]B(z) = \begin{cases} \frac{2}{n+1}\binom{n-1}{\frac{n-1}{2}} & \text{if } n \text{ is odd} \\ 0 & \text{if } n \text{ is even} \end{cases}.$$

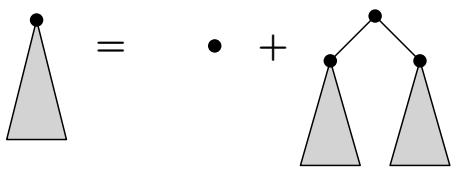


Figure 3.5: Recursive structure of binary trees.

In order to avoid that the counting sequence $(B_n)_{n\geq 0}$ *is zero for every even number* n*, the size of a binary tree is often rather defined as the number of its internal nodes, instead of its total number of nodes. In this case the generating function* $\tilde{B}(z)$ *of binary trees is defined via* $\tilde{B}(z) = 1 + z\tilde{B}(z)$ *and the number* $\tilde{B}_n = [z^n]\tilde{B}(z)$ *is given by the* n*-th Catalan number,* $\tilde{B}_n = \frac{1}{n+1}\binom{2n}{n}$.

In a strict sense, the combinatorial class $\tilde{\mathcal{B}}$ *of binary trees that are counted with respect to their internal nodes does not belong to the class of simply generated trees. However, the functional equation is of the form* $\tilde{B}(z) - 1 = z\Phi(\tilde{B}(z) - 1)$*, with* $\Phi(z) = (1 + z)^2$ *and thus, when considering the shifted generating function this class falls into the framework of simply generated trees as well. Furthermore, the two generating functions* $B(z)$ *and* $\tilde{B}(z)$ *are connected via the identity* $B(z) = z\tilde{B}(z^2)$*, which can easily be verified when observing the explicit solutions for* $B(z)$ *and* $\tilde{B}(z)$*, i.e.,*

$$B(z) = \frac{1 - \sqrt{1 - 4z^2}}{2z} \quad \text{and} \quad \tilde{B}(z) = \frac{1 - \sqrt{1 - 4z}}{2z}.$$

By either applying Stirling's formula to the explicit expression for $\tilde{B}_n$*, i.e., the* n*-th Catalan number, or by applying a tranfer theorem to the Puiseux expansion of* $\tilde{B}(z)$*, or by use of Theorem 3.1 (with* $\Phi(z) = (1 + z)^2$*) the asymptotics for the coefficients* $\tilde{B}_n$ *can be obtained very easily, reading as*

$$\tilde{B}_n = \frac{1}{\sqrt{\pi}} 4^n n^{-3/2} \left(1 + \mathcal{O}(n^{-1})\right) \quad \text{for } n \to \infty.$$

The same asymptotics will also appear in the subsequent example of incomplete binary trees (counted with respect to their total number of vertices), which are in bijection to the number of binary trees counted with respect to their internal nodes.

Example 3.6 (Incomplete binary trees)**.** *Incomplete binary trees are rooted plane trees, where each node has either two children (one to the left, one to the right), or just one child (either to left or to the right), or no children at all (then it is a leaf), see Figure 3.6. Their generating function $I(z)$ is obtained by setting $\phi_0 = 1$, $\phi_1 = 2$, $\phi_2 = 1$ and $\phi_j = 0$ for all $j \geq 3$. Then $\Phi(z) = (1+z)^2$, which yields $I(z) = z(1+I(z))^2$. According to Theorem 3.1 their number behaves asymptotically as*

$$I_n = \frac{1}{\sqrt{\pi}} \frac{4^n}{n^{3/2}} (1 + \mathcal{O}(n^{-1})) \qquad \text{for } n \to \infty.$$

As mentioned earlier, incomplete binary trees are in bijection to ordinary binary trees when the size of the latter ones is defined as the number of internal nodes, i.e., leaves are disregarded. This is reflected by the fact that for both generating functions $\tilde{B}(z)$ and $I(z)$ the function $\Phi(z)$ is defined as $\Phi(z) = (1+z)^2$.

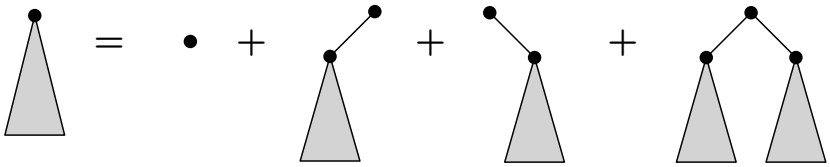


Figure 3.6: Recursive structure of incomplete binary trees.

Example 3.7 (Motzkin trees)**.** *Motzkin trees are rooted plane trees, where each node has either two children (one to the left and one to the right), or one child (centered) or no child at all (then it is a leaf), see Figure 3.7. Their generating function $M(z)$ is obtained by setting $\phi_0 = \phi_1 = \phi_2 = 1$ and $\phi_j = 0$ for $j \geq 3$. Thus, $\Phi(z) = 1 + z + z^2$, and $M(z) = z(1 + M(z) + M(z)^2)$. According to Theorem 3.1 their number behaves asymptotically as*

$$M_n = \sqrt{\frac{3}{4\pi}} \frac{3^n}{n^{3/2}} (1 + \mathcal{O}(n^{-1})) \qquad \text{for } n \to \infty.$$

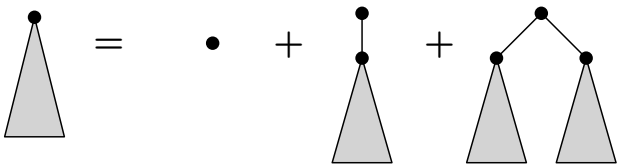


Figure 3.7: Recursive structure of Motzkin trees.

Example 3.8 (Cayley-trees)**.** *Cayley trees are labeled non-plane rooted trees. Their exponential generating function is described through $L(z) = ze^{L(z)}$, thus $\Phi(z) = e^z$. In a strict sense, Cayley trees do not belong to the class of simply generated trees (cf. the discussions in [56] and [69]), since the defining equation* (3.1) *is just valid for the exponential generating function. However, usually they are listed as an example for that class due to their close relation, see [91] for a thorough analysis and [56] for an*

analysis of the differences. By Lagrange inversion the number L_n of Cayley trees can be determined explicitly, which reads as

$$L_n = n![z^n]L(z) = n^{n-1}.$$

Of course we can interpret $L(z)$ also as the ordinary generating function of simply generated trees defined via $L(z) = z\Phi(L(z))$ with $\Phi(z) = e^z$. This has the advantage that this class is easy to investigate (as a special class of simply generated trees) and gives direct access to structural properties of the class of Cayley trees, which is more interesting as it is a very prominent class of trees that occurs in various applications. In order emphasize the difference between these two classes we will subsequently call the respective class of simply generated trees "Cayley-like trees".

The following table summarizes the explicit generating functions of the tree classes that we just introduced in the previous examples together with their dominant singularities.

Tree class	Generating function	Dominant singularity
Planted plane trees	$C(z) = \frac{1-\sqrt{1-4z}}{2}$	$\rho = \frac{1}{4}$
Binary trees	$B(z) = \frac{1-\sqrt{1-4z^2}}{2z}$	$\rho_1 = \frac{1}{2}, \rho_2 = -\frac{1}{2}$
Binary trees (w.r.t. internal nodes)	$\tilde{B}(z) = \frac{1-\sqrt{1-4z}}{2z}$	$\rho = \frac{1}{4}$
Incomplete binary tres	$I(z) = \frac{1-2z-\sqrt{1-4z}}{2z}$	$\rho = \frac{1}{4}$
Motzkin trees	$M(z) = \frac{1-z-\sqrt{1-2z-3z^2}}{2z}$	$\rho = \frac{1}{3}$
Cayley-like trees	$L(z) = -W(-z)$	$\rho = \frac{1}{e}$

Table 3.1: Summary of the explicit closed formulas for the generating functions of selected tree classes falling into the framework of simply generated trees. In a strict sense, for the generating function of Cayley-like trees there is no closed formula known. However, it is possible to express it by means of a special function, namely the Lambert W function (see [27]).

Now, we introduce the unlabeled version of Cayley trees, or the non-plane version of planted plane trees respectively, namely the class of so-called *Pólya trees*.

Pólya trees Pólya trees are unlabeled non-plane rooted trees. By considering them as being constructed of a root to which we attach a multiset of Pólya trees (see ref for the multiset construction), their generating function $P(z)$ satisfies the relation

$$P(z) = ze^{P(z)} \exp\left(\sum_{i\geq 2} \frac{P(z^i)}{i}\right).$$

From the classical results of Pólya [93] we know that $P(z)$ has a unique dominant singularity ρ of type $1/2$ and admits the Puiseux series expansion for $z \to \rho$

$$P(z) \sim 1 - b\sqrt{1-\frac{z}{\rho}} + \frac{b^2}{3}\left(1-\frac{z}{\rho}\right) + d\left(1-\frac{z}{\rho}\right)^{3/2} + \cdots, \tag{3.2}$$

which yields

$$P_n \sim \frac{b}{2\sqrt{\pi}} n^{-3/2} \rho^{-n} \qquad \text{for } n \to \infty.$$

Numerical approximations for the constants were first computed by Otter [89]. This was also topic in Finch [42, Section 5.6] and [45, p. 477] where we find approximations up to 25 digits:

$$\rho \approx 0.33832185689920769519611 26 \quad \text{and} \quad b \approx 1.55949002037464088554226.$$

Another class of non-plane trees that is investigated within this thesis is the class of *non-plane binary trees*.

Non-plane binary trees Like in the plane case, there do not exist any non-plane binary trees of even size. Thus, in order to avoid periodicities in the generating function, the size of a non-plane binary tree is often defined as the number of its internal vertices. Within this thesis, we will use both size models and thus, we will shortly introduce the two cases: Let $N(z)$ denote the generating function of non-plane binary trees where the size is defined as the number of its internal vertices, and let $V(z)$ denote the respective generating function, when the size is defined as the total number of nodes. In analogy to the specification for the generating function of Pólya trees, the generating function $N(z)$ satisfies

$$N(z) = 1 + z\left(\frac{1}{2}N(z)^2 + \frac{1}{2}N(z^2)\right). \tag{3.3}$$

In the binary case this equation can easily be interpreted: The term $N(z)^2$ represents the left and the right subtree. It has to be divided by 2, since we do not distinguish between their left-and-right order. However, in case both subtrees are isomorphic (which corresponds to $N(z^2)$), we have to add the term $\frac{1}{2}N(z^2)$ in order to compensate for the (in this case unnecessary) division by 2. The asymptotic expansion of $N(z)$ is given by

$$N(z) \sim \frac{1}{\sigma} - a\sqrt{1 - \frac{z}{\sigma}}. \tag{3.4}$$

In [45, p. 477] we find the numerical values of the constants σ and a. (*Caveat*: The scaling is different, so [45, p. 477] in fact lists $a \cdot \sigma$, not a.) We have

$$\sigma \approx 0.40269750367144129096904 53 \quad \text{and} \quad a \approx 2.80616022224205389437228 24.$$

The asymptotics of $V(z)$ can easily be obtained from (3.4), by considering that $V(z) = zN(z^2)$. First of all, we immediately know that there are two dominant singularities of $V(z) = zN(z^2)$ at $z = \pm\sqrt{\sigma}$ and we get

$$V(z) = zN(z^2) \sim \pm\sqrt{\sigma}\left(\frac{1}{\sigma} - a\sqrt{2}\sqrt{1 \mp \frac{z}{\sqrt{\sigma}}}\right), \qquad \text{for } z \to \pm\sqrt{\sigma}.$$

Finally, by setting $\rho = \sqrt{\sigma} \approx 0.6346$ and $b = a\sqrt{2\sigma} \approx 2.5184$ we have

$$V(z) \sim \frac{1}{\rho} - b\sqrt{1 \mp \frac{z}{\rho}} \qquad \text{for } z \to \pm\rho.$$

Thus, by means of singularity analysis (see Corollary 2.11) the asymptotics of the coeffcients V_n of $V(z)$ read as

$$V_n = [z^n]V(z) \sim \frac{b}{\sqrt{\pi}}\rho^{-n}n^{-3/2}.$$

Finally, we introduce two important classes of increasing trees, namely *recursive trees* and *increasing binary trees*.

Recursive trees Recursive trees are increasingly labeled non-plane rooted trees. They can be considered as being the result of a growth process, in which one successively picks nodes labeled with growing numbers starting from `1` and attaches them to the thereby growing tree. In this way, the root will always receive the label `1` and for the i-th node ($i > 1$) there are $i-1$ possibilities as to how one can attach it to the tree, see Figure 3.8. Considering that every recursive tree of size n is obtained by a unique growth process, it follows immediately that there are exactly $(n-1)!$ possible trees of size n. Recursive trees can be specified by

$$\mathcal{R} = \{\circ\}^{\square} \times \mathcal{S}\mathrm{et}(\mathcal{R}). \tag{3.5}$$

Since they are labeled, we investigate their asymptotic number by means of an exponential generating function $R(z)$. Translating Equation (3.5) into exponential generating functions yields

$$R'(z) = e^{R(z)}, \tag{3.6}$$

which shows again

$$R(z) = \sum_{n\geq 0}(n-1)!\frac{z^n}{n!} = \ln\frac{1}{1-z}. \tag{3.7}$$

The natural probability distribution on recursive trees of a given size n is to assume that each of the $(n-1)!$ trees occurs with equal possibility. However, by introducing weights of the trees in a similar manner as is has been done for simply generated trees, the class of recursive trees can be generalized to the class of so-called *increasing trees*. They were first introduced by Bergeron, Flajolet, and Salvy [9] and further information can also be found in [34]. Within this thesis we will solely consider two classes of increasing trees, namely the above mentioned recursive trees and the class of plane binary increasing trees, which is defined as follows.

Plane binary increasing trees Plane binary increasing trees can be specified by

$$\mathcal{A} = \{\circ\}^{\square} \times (1+\mathcal{A})^2,$$

which translates as

$$A'(z) = (1+A(z))^2 \tag{3.8}$$

into an equation for the exponential generating function $A(z)$. Solving Equation (3.8) and using the initial condition $A(0) = 0$, the exponential generating function $A(z)$ is given by

$$A(z) = \frac{z}{1-z}.$$

Thus, we get

$$A_n = n![z^n]A(z) = n!,$$

where it is to be emphasized that this result is valid just for odd sizes n, since there are no binary trees of even size, unless one considers incomplete binary trees. In such a case $n!$ is the number of plane increasing incomplete binary trees for arbitrary n.

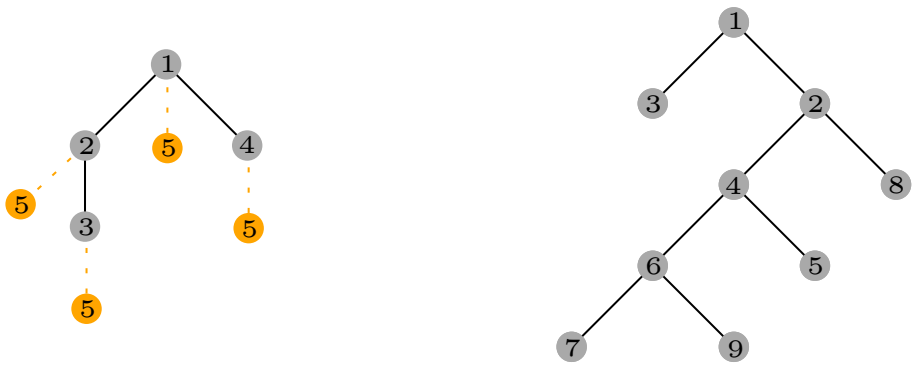


Figure 3.8: Left: Possibilities to add the 5-th node into a given recursive tree of size 4. Right: A (plane) binary increasing tree.

3.2 Lambda terms

This section introduces basic definitions and notations concerning lambda terms in general, as well as two particular subclasses of lambda terms that are studied in Part III. Moreover, the difficulties that arise when trying to solve the enumeration problem of general lambda terms are outlined and the counting problem of the aforementioned subclasses of lambda terms is sketched briefly.

However, before giving all these definitions we shortly want to provide some background information on the lambda calculus and give a brief overview on previous work concerning combinatorics of lambda terms, as it has been done in [57].

3.2.1 Background and previous work

The lambda-calculus was invented by Church and Kleene in the 1930ies as a tool for the investigation of decision problems. Today it still plays an important role in computability theory and for automatic proof systems. Furthermore, it represents the basis for some programming languages, such as LISP or Haskell. In fact, the generation of random lambda terms served for optimising the Glasgow Haskell Compiler [90] and for finding bug in a C-compiler [99] As mentioned at the beginning, recently, rising interest in the number and structural properties of lambda terms can

be observed. This is triggered on the one hand by the fact that random lambda terms have practical application and the understanding of structural properties enables their tuning when generating random terms, see [4], on the other hand, they turned out to be a source of interesting, albeit in part very intricate, combinatorial enumeration problems. Finally, we mention that there is a direct relationship between these random structures acting as computer programs and mathematical proofs (see [28]), but this relationship essentially concerns only typed lambda terms and not general ones.

For a thorough introduction to lambda calculus we refer to [3]. This paper does not require any preliminary knowledge of lambda calculus in order to follow the proofs. Instead we will study the basic objects of lambda calculus, namely lambda terms, by considering them as combinatorial objects, or more precisely as a special class of plane directed acyclic graphs (PDAGs).

To our knowledge, the first appearance of enumeration problems in the sense of enumerative combinatorics which are linked to lambda-calculus is found in [64], where certain models of lambda-calculus are analyzed which have representations as formal power series. More recently, we observe rising interest in the quantitative properties of large random lambda terms. The first work in this direction seems to be [87]. Later David, Grygiel, Kozik, Raffalli, Theyssier,and Zaionc [29] investigated the proportion of normalising terms, which was also the topic of [7] in a different context. Other papers dealing with certain structural properties of lambda terms are for instance [19, 60, 94].

Since studying quantitative aspects of lambda terms using combinatorial methods relies heavily on their enumeration, many papers are devoted to their enumeration, which itself very much depends on the particular class of terms and the definition of the term size. The enumeration may be done by constructing bijections to certain classes of maps, see e.g. [13, 100, 101] or the use of the methodology from analytic combinatorics [45], see e.g. [5, 10, 11, 12, 17, 63, 78].

Another approach to gain structural insight is by random generation. Solving the enumeration problems is the basis for an efficient algorithm for this purpose, namely Boltzmann sampling [40, 43]. The method is extendible to a multivariate setting allowing for a fine tuning according to specified structural properties of the sampled objects, as was demonstrated in [4, 18]. The generation of lambda terms was treated in [6, 13, 63, 90, 96, 98].

3.2.2 Basic definitions

Let us start with the definition of lambda terms.

Definition 3.9 (Lambda terms, [61, Def. 3])**.** *Let $\mathcal{V}$ be a countable set of variables. The set Λ of lambda terms is defined by the following grammar:*

1. *Every variable in $\mathcal{V}$ is a lambda term.*
2. *If T and S are lambda terms then TS is a lambda term.*
3. *If T is a lambda term and x is a variable then $\lambda x.T$ is a lambda term.*

We call TS an application *and $\lambda x.T$ an* abstraction.

The name application arises since a lambda term of the form TS can be regarded as a function $T(S)$, where the function T is applied to S, which in turn is a function itself. An abstraction can be considered as a quantifier that binds the respective variable in the sub-lambda term within its scope. Both application and repeated abstraction are not commutative, *i.e.*, in general the lambda terms TS and ST, as well as $\lambda x.\lambda y.M$ and $\lambda y.\lambda x.M$, are different (with the exceptions of $T = S$ and none of the variables x or y occurring in M).

Definition 3.10 (Bound/free variables, [3, Def. 2.1.6], open/closed lambda term)**.** *A variable x occurs free in a lambda term if it is not in the scope of a λx. Otherwise we call it a bound variable. A lambda term is closed if it contains no free variables; otherwise it is called open.*

Each lambda binds exactly one variable (which may occur several times in the term or even not at all), and each variable can be bound by at most one abstraction. Since we will exclusively deal with closed lambda terms within this thesis, each variable occurrence will always be bound by exactly one lambda.
Throughout this thesis the following notational conventions are used (*cf.* [3, Not. 2.1.3]):

(i) $x, y, z, \ldots$ denote arbitrary variables.

(ii) $S, T, \ldots$ denote arbitrary lambda terms.

(iii) The lambda term $\lambda x_1 \ldots \lambda x_n.S$ is read as $\lambda x_1.(\lambda x_2.(\ldots(\lambda x_n.S))\ldots)$, whereas $ST_1 \ldots T_n$ is an abbreviation for $(\ldots((ST_1)T_2)\ldots T_n)$.

(iv) The symbol $\equiv$ denotes syntactic equality.

Furthermore, we consider lambda terms modulo α-equivalence (*cf.* [3, Def. 2.1.12]), *i.e.*, we identify terms that are equal up to a renaming of their bound variables (by fresh variables that do not occur in the term at all). For example $\lambda x.xz \equiv \lambda y.yz \not\equiv \lambda z.zz$.

In 1972 De Bruijn [23] introduced a representation for lambda terms that completely avoids the use of variables by substituting them by natural numbers that indicate the number of abstractions between the variable and its binding lambda (the binding lambda is counted as well), *i.e.*, $\lambda x.(\lambda y.(xy)) = \lambda(\lambda 21)$.

Definition 3.11 (De Bruijn index, De Bruijn level)**.** *The natural numbers that represent the variables in the De Bruijn representation of a lambda term are called De Bruijn indices. The number of nested lambdas starting from the outermost one specifies the De Bruijn level in which a variable (or De Bruijn index, respectively) is located.*

For example in the lambda term $\lambda x.x(\lambda y.(xy)) = \lambda 1(\lambda 21)$ the first occurrence of the variable x (*i.e.*, the leftmost 1 in the De Bruijn representation) is in the first De Bruijn level, while the other variables are in the second De Bruijn level. In general, free variables are indicated by De Bruijn indices that exceed the De Bruijn level the variable is located in. However, as stated before, we will solely deal with closed lambda terms, and thus we do not have to cope with the modality of free variables.

There is also a combinatorial interpretation of lambda terms that considers them as plane directed acyclic graphs (PDAGs) and thereby naturally identifies two α-equivalent terms to be equal. Combinatorially, lambda terms can be seen as rooted unary-binary trees containing additional directed edges. Note that in general the resulting structures are not trees in the sense of graph theory, but due to their close relation to trees (see Definition 3.12) some authors call them lambda trees or enriched trees. We will call them lambda-PDAGs in order to emphasise that these structures are in fact PDAGs, if we consider the undirected edges of the underlying tree to be directed away from its root.

Definition 3.12 (Lambda-PDAG, [61, Def. 5])**.** *For every lambda term T, the corresponding lambda-PDAG $G(T)$ can be constructed in the following way:*

- *If x is a variable then $G(x)$ is a single node labeled with x. Note that x is free.*
- *$G(PQ)$ is a lambda-PDAG with a binary node as the root, having the two lambda-PDAGs $G(P)$ (to the left) and $G(Q)$ (to the right) as subgraphs.*
- *The PDAG $G(\lambda x.P)$ is obtained from $G(P)$ in four steps:*
  1. *Add a unary node as the new root.*
  2. *Connect the new root by an undirected edge with the root of $G(P)$.*
  3. *Connect all leaves of $G(P)$ labeled with x by directed edges with the new root.*
  4. *Remove all labels x from $G(P)$.*

Obviously, applications correspond to binary nodes and abstractions correspond to unary nodes of the underlying Motzkin tree that is obtained by removing all directed edges. Of course, in the lambda-PDAG some of the vertices that were former unary nodes might have gained out-going edges, so they are no unary nodes in the lambda-PDAG anymore. However, when we speak of unary nodes, we mean the unary nodes of the underlying unary-binary tree that forms the skeleton of the lambda-PDAG. Since the skeleton of a lambda-PDAG is a tree, we sometimes call the variables leaves (*i.e.*, the nodes with out-degree zero), and the path connecting the root with a leaf (consisting of undirected edges) is called a branch.

In the lambda-PDAG, the De Bruijn indices and levels can be easily depicted: The De Bruijn index of a variable v is the number of unary nodes we find on the path from v to its binding lambda in the skeleton of the lambda-PDAG, where the last unary node on the path has to be counted as well. The De Bruijn level of v is the number of unary nodes on the path from v to the root. Figure 3.9 shows two representations of the lambda-PDAG corresponding to the term $(\lambda x.(\lambda y.xy)x)(\lambda z.z)$. The left one presents the PDAG obtained by the algorithm given in Definition 3.12, while the right one is in fact a tree, since the pointers are omitted and instead the leaves are labeled with their respective De Bruijn indices. The latter representation is less common, but will turn out to be very useful for our purposes in Part III of the thesis. Moreover, Figure 3.9 shows the different De Bruijn levels of the lambda-PDAG, extending their definition for leaves to all types of nodes (unary and binary nodes).

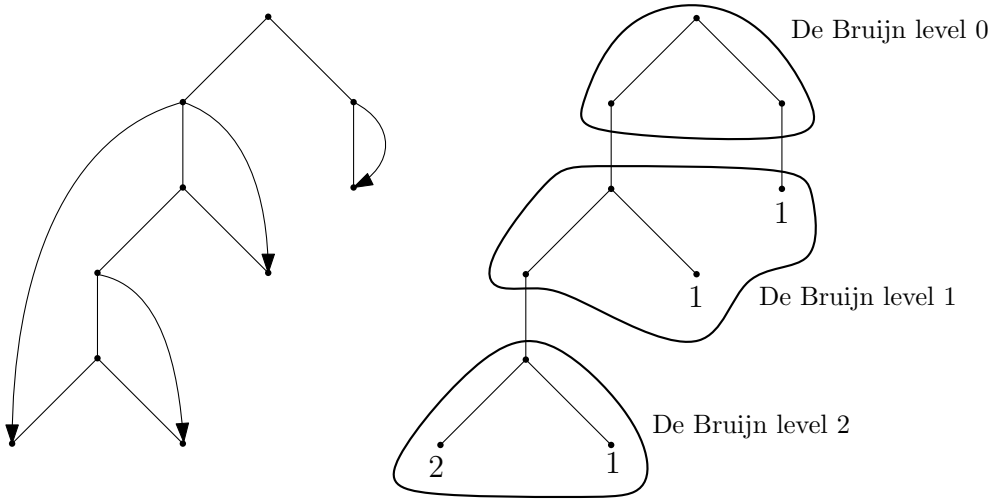


Figure 3.9: Two representations of the lambda-PDAG corresponding to the lambda term $(\lambda x.(\lambda y.xy)x)(\lambda z.z)$. Left: Lambda-PDAG according to Definition 3.12. Right: Lambda-PDAG without pointers (*i.e.*, lambda tree) in which the leaves are labeled with the respective De Brujin indices and with encircled De Bruijn levels.

There are different approaches as to how one can define the size $|t|$ of a lambda term t (see [11, 29, 78]), but within this thesis the size will be defined as the number of nodes in the corresponding lambda-PDAG, *i.e.*,

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

This is combinatorially the most natural definition, and it is equivalent to Barendregt's definition [3].

Now that we introduced all the necessary definitions, we provide a short overview of problems related to counting lambda terms.

3.2.3 Counting lambda terms

At first sight lambda terms appear to be very simple structures, in the sense that their construction can easily be described, but no one has yet accomplished to derive their asymptotic number. However, the asymptotic equivalent of the logarithm of this number can be determined up to the second-order term (see [12]).

One of the difficulties of counting closed lambda terms arises due to the fact that their number increases superexponentially with increasing size, while their specification (as unlabeled objects) requires the use of ordinary generating functions. This rapid growth is caused by the various possibilities of connecting the unary nodes with certain leaves. If we cancel all those pointers, we get ordinary unary-binary trees, which are counted by the large Schröder numbers (OEIS A006318 [95]). These are asymptotically equivalent to $\left(3+2\sqrt{2}\right)^n \frac{1}{\sqrt{\pi}n^{3/2}}$ [29].

However, due to the many degrees of freedom to choose the bindings (see Figure 3.10) a translation of the counting problem into generating functions yields a

generating function that has a radius of convergence equal to zero, which makes the common methods of analytic combinatorics inapplicable.

Consequently, lately some simpler subclasses of lambda terms, which reduce these multiple binding possibilities, have been studied, for example lambda terms with prescribed number of unary nodes [11], or lambda terms in which every lambda binds a prescribed [13, 12, 61] or a bounded [13, 16, 61] number of leaves.

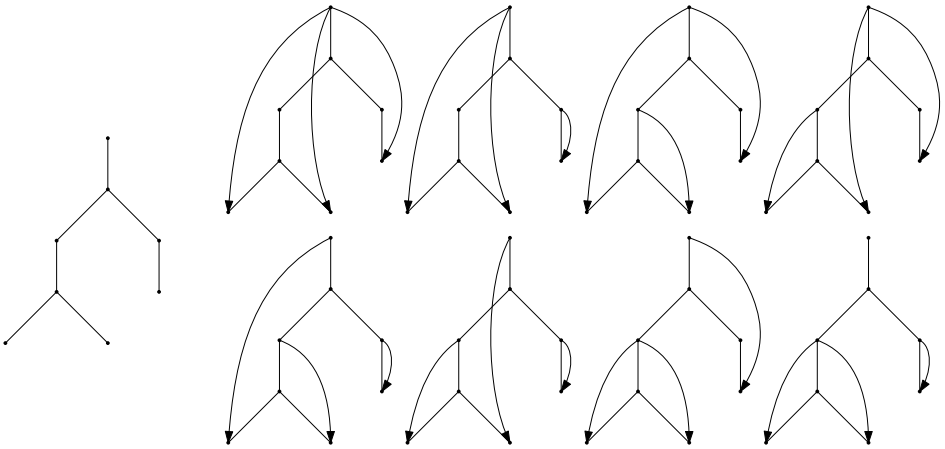


Figure 3.10: All possible variable bindings in order to obtain a closed lambda term from the given unary-binary tree of size 8.

The classes of lambda terms that are the objects of interest within this thesis were introduced in [10] and [11]. The first of these papers studies lambda terms with a bounded number of abstractions between each leaf and its binding lambda, which corresponds to a restriction on the value of De Bruijn indices, while the second one investigates lambda terms with a bounded number of nested levels of abstractions, *i.e.*, lambda terms with a bounded number of De Bruijn levels. From a practical point of view these restrictions appear to be very natural, since the number of abstractions in lambda terms which are used for computer programming is in general assumed to be very small compared to their size (*cf.* [99]).

In what follows we will give an introduction to the counting problems of the two aforementioned classes of lambda terms, since these results will be needed in the remainder of this thesis.

Lambda terms with bounded De Bruijn indices

Now we present the results on the enumeration of lambda terms with bounded De Bruijn indices that have been studied in [11]. Let $\mathcal{G}_k$ denote the class of closed lambda terms where all De Bruijn indices are less than or equal to k.

By the use of the symbolic method one can set up an equation specifying $\mathcal{G}_k$. Therefore we introduce further combinatorial classes as it has been done in [11]: Let $\mathcal{Z}$ denote the class of atoms, $\mathcal{A}$ the class of application nodes (*i.e.*, binary nodes), $\mathcal{U}$ the class of abstraction nodes (*i.e.*, unary nodes), and $\hat{\mathcal{P}}^{(i,k)}$ the class of unary-binary trees such that every leaf e can be labeled in $\min\{\ell(e)+i, k\}$ ways, where $\ell(e)$ denotes the De Bruijn level of e. The objects in $\hat{\mathcal{P}}^{(i,k)}$ may be seen as lambda-PDAGs where the binding of each variable x may come from a unary node at most k

De Bruijn levels above x, even if this means up to i De Bruijn levels above the root (which would indicate that the variable x is free). Thus, the class we are interested in is $\hat{\mathcal{P}}^{(0,k)}$, which is isomorphic to the class $\mathcal{G}_k$. In general, the classes $\hat{\mathcal{P}}^{(i,k)}$ can be recursively specified via

$$\hat{\mathcal{P}}^{(k,k)} = k\mathcal{Z} + (\mathcal{A} \times \hat{\mathcal{P}}^{(k,k)} \times \hat{\mathcal{P}}^{(k,k)}) + (\mathcal{U} \times \hat{\mathcal{P}}^{(k,k)}), \tag{3.9}$$

and

$$\hat{\mathcal{P}}^{(i,k)} = i\mathcal{Z} + (\mathcal{A} \times \hat{\mathcal{P}}^{(i,k)} \times \hat{\mathcal{P}}^{(i,k)}) + (\mathcal{U} \times \hat{\mathcal{P}}^{(i+1,k)}) \qquad \text{for } \; i < k. \tag{3.10}$$

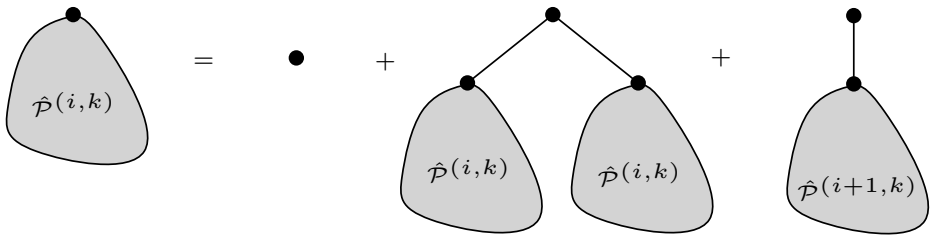


Figure 3.11: Scheme for the specification of the combinatorial classes $\hat{\mathcal{P}}^{(i,k)}$ for the case $i < k$. The sketch is also true for the case $i = k$ when considering that $\hat{\mathcal{P}}^{(k+1,k)} = \hat{\mathcal{P}}^{(k,k)}$.

Translating Equations (3.9) and (3.10) into generating functions and solving for $\hat{P}^{(k,k)}(z)$ and $\hat{P}^{(i,k)}(z)$, we obtain

$$\hat{P}^{(k,k)}(z) = \frac{1 - z - \sqrt{(1-z)^2 - 4kz^2}}{2z}, \tag{3.11}$$

and

$$\hat{P}^{(i,k)}(z) = \frac{1 - \sqrt{1 - 4iz^2 - 4z^2\hat{P}^{(i+1,k)}(z)}}{2z} \qquad \text{for } \; i < k. \tag{3.12}$$

This can be written in the form

$$\hat{P}^{(i,k)}(z) = \frac{1}{2z}\left(1 - \mathbf{1}_{[i=k]}z - \sqrt{\hat{R}_{k-i+1,k}(z)}\right),$$

where $\mathbf{1}_{[i=k]}$ denotes the indicator function

$$\mathbf{1}_{[i=k]} = \begin{cases} 1 & i = k \\ 0 & i \neq k \end{cases},$$

and

$$\begin{aligned}
\hat{R}_{1,k}(z) &= (1-z)^2 - 4kz^2, \\
\hat{R}_{2,k}(z) &= 1 - 4(k-1)z^2 - 2z + 2z^2 + 2z\sqrt{\hat{R}_{1,k}(z)}, \\
\hat{R}_{i,k}(z) &= 1 - 4(k-i+1)z^2 - 2z + 2z\sqrt{\hat{R}_{i-1,k}(z)}, \quad \text{for } \; 3 \leq i \leq k+1.
\end{aligned}$$

Due to the aforementioned isomorphism between the two classes $\hat{\mathcal{P}}^{(0,k)}$ and $\mathcal{G}_k$ we have

$$G_k(z) = \hat{P}^{(0,k)}(z) = \frac{1 - \sqrt{\hat{R}_{k+1,k}(z)}}{2z}. \tag{3.13}$$

Note that the generating function $G_k(z)$ consists of $k+1$ nested radicals, which in fact have a very descriptive combinatorial interpretation that can be seen when setting up the generating function $G_k(z)$ in a different way [62]: Instead of interpreting a lambda term belonging to $\mathcal{G}_k$ as a structure that involves iterated unary-binary trees, we can consider it to be built of leaf-labeled binary trees that are glued together via unary nodes (*cf.* Figure 3.9). 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 labeled in k ways (*k-colored Motzkin trees* in short), we get

$$G_k(z) = B\left(z, B\left(z, 1 + B\left(z, 2 + \ldots + B\big(z, k - 1 + M_k(z)\big)\ldots\right)\right)\right), \tag{3.14}$$

where

$$B(z, w) = \frac{1 - \sqrt{1 - 4wz^2}}{2z} \qquad \text{and} \qquad M_k(z) = \frac{1 - z - \sqrt{(1-z)^2 - 4kz^2}}{2z}. \tag{3.15}$$

Equation (3.14) can be interpreted as follows: Each tree representing a lambda term belonging to $\mathcal{G}_k$ starts with a binary tree, in which all the leaves are replaced by unary nodes to which we add further binary trees, *i.e.* $B(z, B(\ldots))$. 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, *i.e.* $B(z, B(z, 1 + B(\ldots)))$. In this level the leaves can already be labeled with two different labels (namely $\mathtt{1}$ or $\mathtt{2}$), or they can be replaced with unary nodes having new binary trees attached. Starting from the k-th De Bruijn level, the number of possible labelings for the leaves does not increase anymore. Thus, we finally get $\ldots + B(z, k + B(z, k + B(z, k + \ldots)))$, which is exactly the generating function $M_k(z)$ of k-colored Motzkin trees given in (3.15).

Thus, the k outermost radicands, $\hat{R}_{k+1,k}(z), \ldots, \hat{R}_{2,k}(z)$, represent the first k De Bruijn levels of the lambda term, *i.e.*, level 0 to $k-1$, while the innermost radicand, $\hat{R}_{1,k}(z)$, accounts for all the upper De Bruijn levels, starting with level k.

In [11] the authors showed that the dominant singularity of the generating function $G_k(z)$ comes from the innermost radicand for arbitrary fixed k.

Lemma 3.13 ([11, Lemma 5.4])**.** *Let $\hat{\rho}_k$ be the dominant singularity of the function $G_k(z)$, defined by* (3.13)*, or equivalently by* (3.14)*. Then $\hat{\rho}_k = \frac{1}{1+2\sqrt{k}}$ comes from the innermost radicand and is of type $\frac{1}{2}$.*

Furthermore, they provide an asymptotic estimate of the n-th coefficient of $G_k(z)$. For convenience, we will first introduce the auxiliary sequence $(c_i)_{i\geq 1}$, defined via

$$c_1 = 1 \qquad \text{and} \qquad c_i = 4i - 5 + 2\sqrt{c_{i-1}} \text{ for } i \geq 2. \tag{3.16}$$

which appear both in the announced estimate, as well as in the remainder of this thesis.

Theorem 3.14 ([11, Theorem 5.6])**.** *For any fixed $k \geq 1$, let $G_k(z)$ be the generating function of the class of lambda terms where all De Bruijn indices are at most k, which has its dominant singularity at $z = \hat{\rho}_k$. Then*

$$[z^n]G_k(z) \sim \sqrt{\frac{2k+\sqrt{k}}{4\pi \prod_{j=2}^{k+1} c_j}} n^{-3/2} \hat{\rho}_k^{-n}, \qquad \text{as } n \to \infty,$$

where c_i is defined as in (3.16)*.*

Lambda terms with bounded number of De Bruijn levels

Now, we present the counting problem of lambda terms with a bounded number of De Bruijn levels. Let us denote by $\mathcal{H}_k$ the class of closed lambda terms with at most k De Brujin levels. A specification for this class can be set up as in [11] using the classes $\mathcal{P}^{(i,k)}$ of unary-binary trees that contain at most $k-i$ De Bruijn levels and each leaf e can be colored with one out of $i + \ell(e)$ colors, where $\ell(e)$ denotes the De Bruijn level in which the respective leaf is located. By denoting again $\mathcal{Z}$ the class of atoms, $\mathcal{A}$ the class of applications/binary nodes and $\mathcal{U}$ the class of abstractions/unary nodes, the classes $\mathcal{P}^{(i,k)}$ can be recursively specified by

$$\mathcal{P}^{(k,k)} = k\mathcal{Z} + (\mathcal{A} \times \mathcal{P}^{(k,k)} \times \mathcal{P}^{(k,k)}), \tag{3.17}$$

and

$$\mathcal{P}^{(i,k)} = i\mathcal{Z} + (\mathcal{A} \times \mathcal{P}^{(i,k)} \times \mathcal{P}^{(i,k)}) + (\mathcal{U} \times \mathcal{P}^{(i+1,k)}) \qquad \text{for } \; i < k. \tag{3.18}$$

Remark 3.15. *One can see that there is a great similarity in the specification for the classes $\hat{\mathcal{P}}^{(i,k)}$ and $\mathcal{P}^{(i,k)}$. While the recursions* (3.10) *and* (3.18) *for $i > k$ are identical, the only difference lies in the definitions of $\mathcal{P}^{(k,k)}$ and $\hat{\mathcal{P}}^{(k,k)}$. Since the classes $\mathcal{P}^{(i,k)}$ have to fulfill the additional condition to contain at most $k-i$ De Bruijn levels, there cannot be any unary nodes in the class $\mathcal{P}^{(k,k)}$, cf.* (3.17)*. However, the class $\hat{\mathcal{P}}^{(k,k)}$ is not restricted by this condition, and can therefore contain arbitrarily many unary nodes, as long as all the De Bruijn indices are small enough, cf.* (3.9)*.*

By translating (3.17) and (3.18) into generating functions we get

$$P^{(k,k)}(z) = kz + zP^{(k,k)}(z)^2,$$

and

$$P^{(i,k)}(z) = iz + zP^{(i,k)}(z)^2 + zP^{(i+1,k)}(z) \qquad \text{for } \; i < k.$$

Solving yields

$$P^{(k,k)}(z) = \frac{1 - \sqrt{1 - 4kz^2}}{2z},$$

and

$$P^{(i,k)}(z) = \frac{1 - \sqrt{1 - 4iz^2 - 4z^2 P^{(i+1,k)}}}{2z} \quad \text{for } \; i < k.$$

Analogously to the previous section, we are interested in the class $\mathcal{P}^{(0,k)}$, since it is isomorphic to the class $\mathcal{H}_k$. Thus, the corresponding generating function $H_k(z)$ reads as

$$H_k(z) = P^{(0,k)}(z) = \frac{1 - \sqrt{R_{k+1,k}(z)}}{2z},$$

where the radicands $R_{i,k}(z)$ are defined as

$$R_{1,k}(z) = 1 - 4kz^2, \tag{3.19}$$

and

$$R_{i,k}(z) = 1 - 4(k - i + 1)z^2 - 2z + 2z\sqrt{R_{i-1,k}(z)}, \quad \text{for } \; 2 \leq i \leq k + 1. \tag{3.20}$$

Again, this generating function consists of $k + 1$ nested radicands, which can be interpreted in a similar way as the $\hat{R}_{i,k}(z)$ in the previous section. Thereby each radicand indicates one further De Bruijn level of the lambda term and the generating function $H_k(z)$ can be written as

$$H_k(z) = B(z, B(1 + B(z, 2 + \ldots + B(z, k - 1 + B(z, k)) \ldots))), \tag{3.21}$$

where $B(z, w)$ denotes again the bivariate generating function of binary trees given in (3.15), with z marking the size and w marking the number of leaves. This generating function can once more be explained by considering a lambda term belonging to $\mathcal{H}_k$ as a structure consisting of nested binary trees, matching the $k+1$ De Bruijn levels, that are attached to each other via unary nodes. The essential difference in the construction of the generating functions $G_k(z)$ and $H_k(z)$ lies in the choice of the trees that are attached to the $(k - 1)$-th De Bruijn level, *i.e.*, the last structure that we attach, which makes up the innermost radicand. In the case of lambda terms with bounded De Bruijn indices, *i.e.*, $\mathcal{G}_k$, we attached leaf-labeled Motzkin trees, since those terms can have arbitrarily many De Bruijn levels, and starting from De Bruijn level k all leaves can take labels from $\{\mathtt{1}, \ldots, \mathtt{k}\}$. However, for lambda terms with at most k De Bruijn levels, *i.e.*, $\mathcal{H}_k$, we have to attach leaf-labeled binary trees, since in the last (*i.e.*, the k-th) De Bruijn level no more unary nodes are allowed.

Remark 3.16. *Note that the class $\mathcal{H}_k$ is a proper subclass of $\mathcal{G}_k$, since the restriction for all De Bruijn indices to be at most k, has to be fulfilled by terms belonging to both classes.*

In [11] the authors showed a very interesting phenomenon concerning the generating function $H_k(z)$. The asymptotic behavior of its coefficients differs depending on whether the imposed bound k is an element of a certain sequence $(N_i)_{i \geq 0}$, which will be given in Definition 3.17, or not. The remainder of this subsection is devoted to presenting the precise results obtained in [11].

Definition 3.17 (Auxiliary sequences $(u_i)_{i\geq 0}$ and $(N_i)_{i\geq 0}$, [11, Def. 6.1])**.** *Let $(u_i)_{i\geq 0}$ be the integer sequence defined by*

$$u_0 = 0, \quad u_{i+1} = u_i^2 + i + 1 \quad \text{for } i \geq 0,$$

and $(N_i)_{i\geq 0}$ by

$$N_i = u_i^2 - u_i + i, \quad \text{for all } i \geq 0.$$

j	1	2	3	4	5	6
N_j	1	8	135	21760	479982377	23040411505837408
u_j	1	3	12	148	21909	480004287

Table 3.2: The first values of the sequences $(N_j)_{j\geq 0}$ and $(u_j)_{j\geq 0}$ for $j = 1, \ldots, 6$.

As indicated before, the generating function $H_k(z)$ shows a very unusual behavior. The type of its dominant singularity changes when the imposed bound equals N_j. Thus, the subexponential term in the asymptotics of the counting sequence changes.

Theorem 3.18 ([11, Theorem 6.4])**.** *Let $(N_i)_{i\geq 0}$ be the sequence defined in Definition 3.17 and let k be an integer. Define j as the integer such that $k \in [N_j, N_{j+1})$. If $k \neq N_j$, then the dominant radicand of $H_k(z)$ is the $(j+1)$-th radicand (counted from the innermost one outwards), and the dominant singularity ρ_k is of type $\frac{1}{2}$. Otherwise, the j-th and the $(j+1)$-th radicand vanish simultaneously at the dominant singularity of $H_k(z)$, which is equal to $1/(2u_j)$ and of type $\frac{1}{4}$.*

The next theorem contains the asymptotic behavior of the number of lambda terms with at most k De Bruijn levels.

Theorem 3.19 ([11, Theorem 6.23])**.** *Let $(u_i)_{i\geq 0}$ and $(N_i)_{i\geq 0}$ be the integer sequences defined in Definition 3.17 and let $H_k(z)$ be the generating function of the class of closed lambda terms with at most k De Bruijn levels, which has its dominant singularity at $z = \rho_k$. Then the following asymptotic relations hold:*

(i) If there exists $j \geq 0$ such that $N_j < k < N_{j+1}$, then there exists a constant h_k such that

$$[z^n]H_k(z) \sim h_k n^{-3/2} \rho_k^{-n}, \text{ as } n \to \infty.$$

(ii) If there exists $j \geq 0$ such that $k = N_j$, then there exists a constant h_{N_j} such that

$$[z^n]H_k(z) \sim h_{N_j} n^{-5/4} \rho_k^{-n} = h_{N_j} n^{-5/4} (2u_j)^n, \text{ as } n \to \infty.$$

The constants in Theorem 3.19 can be expressed by means of the sequences $(a_i)_{i\geq j+2}$, $(b_i)_{i\geq j+2}$ and $(d_i)_{i\geq j+2}$ given by

$$\begin{aligned} a_{j+2} &= 1 - 4(k-j-1)\rho_k^2 - 2\rho_k^2, \\ a_{i+1} &= 1 - 4(k-i+1)\rho_k^2 - 2\rho_k^2 + 2\rho_k\sqrt{a_i} \qquad \text{for } i \geq j+1, \end{aligned} \tag{3.22}$$

$$\begin{aligned} b_{j+2} &= 2\rho_k\sqrt{2\rho_k}\sqrt[4]{\gamma_j}, \\ b_{i+1} &= \frac{\rho_k}{\sqrt{a_i}} b_i \qquad \text{for } i \geq j+1, \end{aligned} \tag{3.23}$$

$$\begin{aligned} d_{j+2} &= 2\rho_k\sqrt{\gamma_{j+1}}, \\ d_{i+1} &= \frac{\rho_k}{\sqrt{a_i}} d_i \qquad \text{for } i \geq j+1, \end{aligned} \tag{3.24}$$

where γ_i is defined as $\gamma_i := -\frac{d}{dz} R_{i,k}(\rho_k)$.

The constants h_k and h_{N_j} then read as

$$h_k = -\frac{d_{k+1}\sqrt{\rho_k}}{4\rho_k \Gamma(-1/2)\sqrt{a_{k+1}}},$$

and

$$h_{N_j} = -\frac{b_{k+1}\sqrt[4]{\rho_k}}{4\rho_k \Gamma(-1/4)\sqrt{a_{k+1}}}.$$

In the case $k = N_j$ one can easily show that $1 - 4(k-j)\rho_k^2 - 2\rho_k^2 = 0$. This equation arises immediately by evaluating $R_{j+1,k}(z)$ at $z = \rho_k$ by means of the recursive definition (3.20) and by considering the fact that $R_{j+1,k}$ and $R_{j,k}$ both vanish at ρ_k. By the use of this identity, the sequence $(a_i)_{i \geq j+2}$ simplifies to

$$a_{j+l} = 4\rho_k^2 \lambda_{l-1},$$

with $\lambda_0 = 0$ and $\lambda_{i+1} = i + 1 + \sqrt{\lambda_i}$. The advantage of this representation is that the asymptotic behavior of the sequence $(\lambda_i)_{i \geq 0}$ is very simple to derive by bootstrapping and given by

$$\lambda_i = i + \sqrt{i} + \frac{1}{2} - \frac{3}{8\sqrt{i}} - \frac{1}{4i} + \mathcal{O}\left(\frac{1}{i\sqrt{i}}\right), \qquad \text{for } i \to \infty, \tag{3.25}$$

see [11, Lemma 6.32]. This simplified formula for a_{j+l} will be used at some point in Part III and is thus given here for the sake of completeness. However, it is emphasized that the simplification is only true in the case when the bound k is an element of the sequence $(N_j)_{j \geq 0}$.

Part II

Parameters of trees

Chapter 4

Protection number

This chapter is based on joint work with Bernhard Gittenberger, Zbigniew Gołębiewski and Małgorzata Sulkowska, which lead to the article *Protection numbers in simply generated trees and Pólya trees* that has already been submitted to a journal, [55].

The protection number of a tree is the length of the shortest path from the root to a leaf, *i.e.*, the length of the shortest branch of a tree. It is interchangeably called *the protection number of a root*. We define *the protection number of a vertex* v in a tree T as the protection number of the fringe (*i.e.*, maximal) subtree of T having v as a root. We say that a vertex is *k-protected* if k does not exceed its protection number, *cf.* Figure 4.1.

The protection number of a root is closely related to parameters called *minimal fill-up level* and *saturation level*. These were studied previously by, among others, Devroye [31] and Drmota [34, 35].

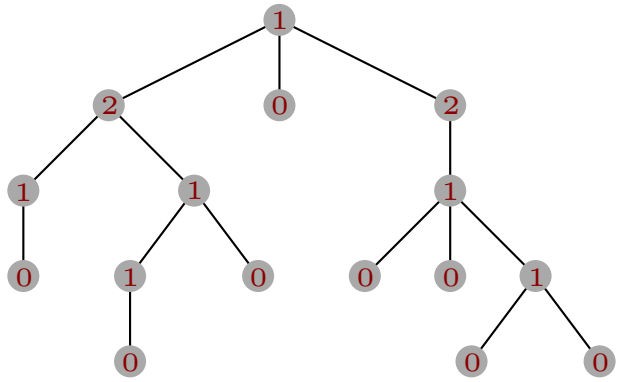


Figure 4.1: Tree with vertices holding their protection numbers. The protection number of the tree is 1, *i.e.*, the tree is 1-protected, as well as 0-protected.

Cheon and Shapiro [24] were the first ones to investigate the number of 2-protected nodes in trees. They stated the results for unlabeled ordered trees and Motzkin trees. Later on Mansour [83] complemented their work by solving k-ary tree case. Over the next several years these results were followed by a series of papers examining the number of k-protected nodes (usually for small values of k) in various models of

random trees. To mention just a few, Du and Prodinger [39] analysed the average number of 2-protected nodes in random digital search trees, Mahmoud and Ward [81] presented a central limit theorem as well as exact moments of all orders for the number of 2-protected nodes in binary search trees and three years later they found the number of 2-protected nodes in recursive trees (consult [82]). The family of binary search trees was investigated also by Bóna and Pittel [20] who showed that the number of its k-protected nodes decays exponentially in k. In 2015 Holmgren and Janson [66] went for more general results. Using probabilistic methods, they derived a normal limit law for the number of k-protected nodes in a binary search tree and a random recursive tree.

Soon after, two particular parameters attracted attention of the algorithmic community. These were the protection number of a root and the protection number of a random vertex.

In 2017 Copenhaver [26] found that in a random unlabeled plane tree the expected value of the protection number of the root and the expected value of the protection number of a random vertex approach 1.62297 and 0.727649, respectively, as the size of the tree tends to infinity. These results were extended by Heuberger and Prodinger [65]. They showed the exact formulas for the first terms of the expectation, the variance and the probability of the respective protection numbers.

The aim of this work is to generalize the protection number results to a larger class of rooted trees, namely simply generated trees (see page 19) and their nonplane counterparts, Pólya trees (see page 23).

For simply generated trees, a general theory of asymptotics of certain functional was developed recently in [30], but this theory does not cover local functionals as the number of protected nodes. Devroye and Janson [32] presented a unified approach to obtaining the number of k-protected nodes in various classes or random trees by putting them in the general context of fringe subtrees introduced by Aldous in [2]. We have obtained analogous results for simply generated trees, but employing a different methodology. This allows an efficient numerical treatment and may serve as a basis for random generation in the framework of Boltzmann sampling [40]. Parts of our investigations fall into the general framework of additive functionals treated in [97], but our focus on concrete expressions allows an easy access to numerical evaluation of the considered parameters.

In Section 4.1 we calculate the asymptotic mean and variance of the protection number of the root and the protection number of a random vertex for a random simply generated tree. In Section 4.2 these parameters are studied for the class of Pólya trees and in Section 4.3 we extend the results to non-plane binary trees. The results for the asymptotic expected values are summarized in Table 4.1. Note that all the obtained values are constants, *i.e.* they do not depend on the size of the respective trees.

4.1 Protection number of simply generated trees

4.1.1 Protection number of the root

Let $T(z)$ denote the generating function of the class of simply generated trees, where z marks the total number of nodes, *i.e.*, $T(z) = z\phi(T(z))$, *cf.* (3.1), which has a unique dominant singularity of square root type at $z = \rho$. Then we denote by $T_k(z)$

Tree model	$\lim_{n\to\infty} \mathbb{E}(X_n)$	$\lim_{n\to\infty} \mathbb{E}(Y_n)$
Simply generated trees		
Plane trees	1.62297	0.72765
Motzkin trees	2.54638	1.30760
Incomplete binary trees	3.53647	1.99182
Cayley trees	2.28620	1.18652
Complete binary trees	1.56298	1.26568
Non-plane trees		
Pólya trees	2.15489	0.99532
Non-plane binary trees	1.70760	1.31241

Table 4.1: Summary of the obtained mean values for the protection number of the root (X_n) and the protection number of a random vertex (Y_n).

the generating function of the class of simply generated trees that have protection number at least k. Furthermore, we assume $\phi(T)$ to be non-periodic. Then, $T_k(z)$ can be defined by

$$T_k(z) = z\left(\phi(T_{k-1}(z)) - \phi_0\right). \tag{4.1}$$

Note that $T_0(z) = T(z)$.

Lemma 4.1. *All generating functions $T_k(z)$ have the same dominant singularity as $T(z)$, and it is a square root singularity.*

Proof. First let us consider that the generating function $T_k(z)$ reads as

$$T_k(z) = \Omega^k(T(z))$$

where $\Omega(t) = z\phi(t) - z\phi_0$ and $\Omega^k(\cdot)$ denotes the k-fold composition. Since $\Omega(t)$ is analytic at $T(\rho)$, inserting a function admitting a Puiseux expansion $t(z) = \alpha_0 + \alpha_1\sqrt{1-\frac{z}{\rho}} + \dots$ results in

$$\Omega(t(z)) = \Omega(\alpha_0) + \Omega'(\alpha_0)\alpha_1\sqrt{1-\frac{z}{\rho}} + \dots,$$

again being a Puiseux expansion at $z = \rho$. It is well known that $T(z)$ admits a Puiseux expansion $\tau_0 + \tau_1\sqrt{1-\frac{z}{\rho}} + \dots$ with nonzero numbers τ_0 and τ_1 (*cf.* Section 3.1.2). Moreover, we always insert one of the functions $T_k(z)$, thus α_0 attains the positive values $T_k(\rho)$, $k = 0, 1, 2, \dots$, implying that $\Omega'(\alpha_0)$ is always positive, as $\Omega(t)$ is a power series with only non-negative coefficients. By induction it is guaranteed that α_1 is always negative and thus all the functions $T_k(z)$ have a unique dominant singularity of square root type at $z = \rho$. □

In order to derive the expected value of the protection number X_n of a random simply generated tree of size n (*i.e.*, with n nodes) asymptotically, we use the well known formula

$$\mathbb{E}(X_n) = \sum_{k\geq 1} \mathbb{P}(X_n \geq k). \tag{4.2}$$

Thus, we need to calculate the probability $\mathbb{P}(X_n \geq k)$, which is given by

$$\mathbb{P}(X_n \geq k) = \frac{[z^n]T_k(z)}{[z^n]T(z)}.$$

Theorem 4.2. *Let X_n be the protection number of a random simply generated tree of size n. Then the expected value $\mathbb{E}(X_n)$ and the variance $\mathbb{V}(X_n)$ satisfy*

$$\lim_{n\to\infty} \mathbb{E}(X_n) = \sum_{k\geq 1} \rho^{k-1} \prod_{i=1}^{k-1} \phi'(T_i(\rho)),$$

and

$$\lim_{n\to\infty} \mathbb{V}(X_n) = \sum_{k\geq 1} (2k-1)\rho^{k-1} \prod_{i=1}^{k-1} \phi'(T_i(\rho)) - \left(\lim_{n\to\infty} \mathbb{E}(X_n)\right)^2.$$

with ρ denoting the dominant singularity of the generating function $T(z) = z\phi(T(z))$ of the class of simply generated trees.

Proof. Using the singularity analysis approach introduced in Chapter 2, we can translate the Puiseux expansion $T(z) = \tau_0 + \tau_1\sqrt{1-\frac{z}{\rho}} + \tau_2\left(1-\frac{z}{\rho}\right) + \ldots$, of the generating function $T(z)$ into asymptotics for its coeffcients, resulting in

$$[z^n]T(z) \sim -\tau_1 \frac{n^{-3/2}}{\Gamma(-1/2)} \rho^{-n}, \qquad \text{as } n \to \infty. \tag{4.3}$$

In order to derive the asymptotic behavior of the n-th coefficient of $T_k(z)$, let us recall that from Lemma 4.1 we know that all generating functions $T_i(z)$ have the same dominant singularity ρ of type $\frac{1}{2}$. Setting $\eta = \sqrt{1-\frac{z}{\rho}}$, the Puiseux expansions of $T_k(z)$ and $T_{k-1}(z)$ read as

$$T_k(z) = \tau_{0,k} + \tau_{1,k}\eta + \tau_{2,k}\eta^2 + \ldots,$$

and

$$T_{k-1}(z) = \tau_{0,k-1} + \tau_{1,k-1}\eta + \tau_{2,k-1}\eta^2 + \ldots.$$

Plugging these expansions into (4.1) and using $z = \rho(1-\eta^2)$ we get

$$\tau_{0,k} + \tau_{1,k}\eta + \tau_{2,k}\eta^2 + \ldots = \rho(1-\eta^2)\left(\sum_{j\geq 0} \phi_j \left(\tau_{0,k-1} + \tau_{1,k-1}\eta + \tau_{2,k-1}\eta^2 + \ldots\right)^j - \phi_0\right).$$

Expanding and comparing coefficients of η^0 and η^1 yields

$$\begin{aligned}
[\eta^0] &: \tau_{0,k} = \rho\phi(\tau_{0.k-1}) - \rho\phi_0, \\
[\eta^1] &: \tau_{1,k} = \rho \sum_{j\geq 0} \phi_j j \tau_{1,k-1} \tau_{0,k-1}^{j-1}.
\end{aligned}$$

Obviously, the $\tau_{0,i} = T_i(\rho),\ \forall i \geq 0$, as the $\tau_{0,i}$'s are the constant terms in the Puiseux expansions of the functions $T_i(z)$, with $0 \leq i \leq k$. Thus, the equation for $\tau_{1,k}$ can be

rewritten as $\tau_{1,k} = \rho\tau_{1,k-1}\phi'(T_{k-1}(\rho))$.

As $\tau_{1,0} = \tau_1$, we get

$$\tau_{1,k} = \tau_1 \rho^{k-1} \prod_{i=1}^{k-1} \phi'(T_i(\rho)).$$

Applying a transfer lemma (see Theorems 2.8 and 2.10) directly gives the asymptotics of the coefficients of $T_k(z)$ and plugging them in conjunction with (4.3) into Equation (4.2) yields the asymptotic value for the mean. In order to derive the formula for the asymptotic variance we use the equations

$$\mathbb{V}(X_n) = \mathbb{E}(X_n^2) - (\mathbb{E}X_n)^2 \quad \text{and} \quad \mathbb{E}(X_n^2) = \sum_{k\geq 1}(2k-1)\mathbb{P}(Y_n \geq k)$$

and immediately get the asserted result. □

It is easy to see that the sequence $(T_i(\rho))_{i\geq 0}$ is monotonically decreasing, since the number of trees with protection number at least i is always greater than the number of trees that have an $(i+1)$-protected root, *i.e.*, protection number at least $i+1$. Since ϕ' is monotonically increasing on the positive real axis, this implies that $\rho\phi'(T_i(\rho)) \leq \rho\phi'(T_1(\rho)) < \rho\phi'(T(\rho)) = 1$. Thus, we can estimate the sum for the expected value by

$$\lim_{n\to\infty} \mathbb{E}(X_n) = \sum_{k\geq 1} \prod_{i=1}^{k-1} (\rho\phi'(T_i(\rho))) < \sum_{k\geq 1} (\rho\phi'(T_1(\rho)))^{k-1},$$

which converges, since $\rho\phi'(T_1(\rho)) < 1$. As the last sum is a convergent geometric series and the inequality even holds term-wise, we can calculate efficiently the asymptotic mean and variance for all classes of simply generated trees with arbitrary accuracy. We will now exemplify this by calculating the limits of mean and variance of the protection number of some prominent classes of simply generated trees that have been introduced in Section 3.1.2.

Example 4.3 (Planted plane trees)**.** *In Chapter 3 we derived that the dominant singularity of the class of planted plane trees is $\rho = \frac{1}{4}$, and thus $C(\rho) = \frac{1}{2}$ (see Table 3.1). Therefore the recursion for the $T_i(\rho)$'s reads as*

$$T_1(\rho) = \frac{1}{4}, \qquad T_i(\rho) = \frac{1}{4 - 4T_{i-1}(\rho)} - \frac{1}{4}.$$

This recursion can be solved explicitly, leading to

$$T_i(\rho) = \frac{3}{2(4^i+2)}.$$

The limits of expected value and variance are therefore given by

$$\lim_{n\to\infty} \mathbb{E}(X_n) = \sum_{k\geq 1} \frac{1}{4^{k-1}} \prod_{i=1}^{k-1} \frac{1}{\left(1 - \frac{3}{2(4^i+2)}\right)^2} \approx 1.622971384715353,$$

and

$$\lim_{n\to\infty}\mathbb{V}(X_n)=\sum_{k\geq 1}\frac{2k-1}{4^{k-1}}\prod_{i=1}^{k-1}\frac{1}{\left(1-\frac{3}{2(4^i+2)}\right)^2}-\Big(\lim_{n\to\infty}\mathbb{E}(X_n)\Big)^2\approx 0.7156950717833327,$$

which has already been calculated by Heuberger and Prodinger in [65].

Example 4.4 (Motzkin trees)**.** *According to Table 3.1 the dominant singulaity of Motzkin trees is* $\rho=\frac{1}{3}$ *and thus* $M(\rho)=1$*. The recursion for the* $T_i(\rho)$*'s reads as*

$$T_1(\rho)=\frac{2}{3},\qquad T_i(\rho)=\frac{1}{3}\left(T_{i-1}(\rho)^2+T_{i-1}(\rho)\right)$$

This recursion can be transformed into another one for the numerators of the rational numbers $T_i(\rho)$*: Indeed, if we write* $T_i(\rho)=A_i\cdot 3^{-2^i+1}$*, then* $A_1=2$ *and* $A_i=A_{i-1}^2+3^{2^{i-1}-1}\cdot A_{i-1}$*, for* $i\geq 2$*. The recurrence for the* A_i*'s does not fall into the scheme of Aho and Sloane [1] and we are not aware of any method to solve it explicitly. But as stated before, the sequence* $(T_i(\rho))_{i\geq 1}$ *is exponentially decreasing and estimates are easily obtained. Thus we can calculate the limits of mean and variance for the protection number numerically with arbitrary accuracy:*

$$\lim_{n\to\infty}\mathbb{E}(X_n)\approx 2.546378248338912,\quad \textit{and}\quad \lim_{n\to\infty}\mathbb{V}(X_n)\approx 1.679348871220563.$$

Example 4.5 (Incomplete binary trees)**.** *The dominant singularity of incomplete binary trees is given by* $\rho=\frac{1}{4}$ *and thus* $I(\rho)=1$ *(cf. Table 3.1). Therefore the recursion for the* $T_i(\rho)$*'s reads as*

$$T_1(\rho)=\frac{3}{4},\qquad T_i(\rho)=\frac{1}{4}(T_{i-1}(\rho)^2+2T_{i-1}(\rho)).$$

As in the previous example we are not aware of a method to explicitly solve this recursion, but the numerical values can be easily computed: They are

$$\lim_{n\to\infty}\mathbb{E}(X_n)\approx 3.536472483525321,\quad \textit{and}\quad \lim_{n\to\infty}\mathbb{V}(X_n)\approx 3.763883442795153.$$

Example 4.6 (Cayley trees)**.** *The exponential generating function* $L(z)$ *of Cayley trees has its dominant singularity at* $\rho=\frac{1}{e}$ *(cf. Table 3.1). Moreover, we have* $L(\rho)=1$*, and the recursion for the* $T_i(\rho)$*'s reads as*

$$T_1(\rho)=1-\frac{1}{e},\qquad T_i(\rho)=\frac{1}{e}(e^{T_{i-1}(\rho)}-1).$$

As in the two previous examples we are not able to solve the recursion for the $T_i(\rho)$*'s explicitly, but the numerical values are*

$$\lim_{n\to\infty}\mathbb{E}(X_n)\approx 2.286198316708012,\quad \textit{and}\quad \lim_{n\to\infty}\mathbb{V}(X_n)\approx 1.598472890455086.$$

This example actually covers two classes of trees, namely the class of simply generated trees whose ordinary generating function is specified via $L(z)=ze^{L(z)}$*, as well as the class of Cayley trees (labeled plane trees) that - in a strict sense - does not belong to the class of simply generated trees, as we already discussed in Example 3.8.*

Example 4.7 (Binary trees counted with respect to the number of internal nodes)**.** *The dominant singularity $\tilde{B}(z)$ of the generating function of binary trees counted with respect to the number of internal nodes is given by $\rho = \frac{1}{4}$ (see Table (3.1)). Although this class does not strictly fall into the simply generated framework, the methodology presented above works here as well, due to their close relation to simply generates trees (cf. Example 3.5). We get $T_0(z) = \tilde{B}(z)$ and $T_k(z) = zT_{k-1}(z)^2$. Since $\rho = 1/4$ we have $T_k(\rho) = 2^{2-2^k}$, for all $k \geq 0$, and then finally $\mathbb{P}(X_n \geq k) \to 2^{k+1-2^k}$, as n tends to infinity. Thus we obtain*

$$\lim_{n\to\infty} \mathbb{E}(X_n) \approx 1.562988296151161, \quad \textit{and} \quad \lim_{n\to\infty} \mathbb{V}(X_n) \approx 0.372985688954940.$$

4.1.2 Protection number of a random vertex

In the first part of this section we studied the average protection number of a simply generated tree, that is the protection number of the root of the simply generated tree. Now we are interested in the average protection number of a randomly chosen vertex in a simply generated tree of size n. We denote this sequence of random variables by Y_n.

As in the previous section we calculate the mean via $\mathbb{E}(Y_n) = \sum_{k\geq 1} \mathbb{P}(Y_n \geq k)$. In order to do so we proceed analogously to Heuberger and Prodinger in [65] and define $S_k(z)$ to be the generating function of the sequence $(s_{n,k})_{n\geq 0}$ of k-protected vertices summed over all trees of size n. As in [65] this generating function can be calculated by

$$S_k(z) = z^{-1} T_k(z) \frac{\partial}{\partial u} T(z, 1), \tag{4.4}$$

by means of the bivariate generating function $T(z, u)$ of simply generated trees, where z marks the size and u the number of leaves, and the generating function $T_k(z)$ of simply generated trees with protection number at least k. The formula for $S_k(z)$ arises from considering a k-protected vertex in the following way: First point at a leaf in a simply generated tree (which yields the factor $\frac{\partial}{\partial u} T(z, 1)$), then remove this leaf (which explains the z^{-1}) and finally attach a tree with protection number at least k (giving the factor $T_k(z)$).

Remark 4.8. *The procedure works also for (complete) binary trees, where only internal vertices contribute to the tree size. The only difference is that for these binary trees the factor z^{-1} in (4.4) must be removed, because removing a leaf does not change the size.*

Using the generating function $S_k(z)$ we can express the probability $\mathbb{P}(Y_n \geq k)$ by

$$\mathbb{P}(Y_n \geq k) = \frac{[z^n] S_k(z)}{n[z^n] T(z)}, \tag{4.5}$$

since $n[z^n]T(z)$ is the total number of vertices summed over all trees of size n.

Theorem 4.9. *Let Y_n be the protection number of a randomly chosen vertex in a random simply generated tree of size n. Then,*

$$\lim_{n\to\infty} \mathbb{E}(Y_n) = \frac{\phi_0}{T(\rho)} \sum_{k\geq 1} T_k(\rho),$$

and

$$\lim_{n\to\infty} \mathbb{V}(Y_n) = \frac{\phi_0}{T(\rho)} \sum_{k\geq 1} (2k-1) T_k(\rho) - \Big(\lim_{n\to\infty} \mathbb{E}(Y_n)\Big)^2.$$

Proof. First we need to determine the n-th coefficient of $S_k(z)$. We have

$$\frac{\partial}{\partial u} T(z,1) = \frac{z\phi_0}{1 - z\phi'(T(z))}. \tag{4.6}$$

Using $T'(z) = z\phi'(T(z))T'(z) + \phi(T(z))$ and $\phi(T(z)) = \frac{T(z)}{z}$ we get

$$z\phi'(T(z)) = \frac{T'(z) - \frac{T(z)}{z}}{T'(z)}.$$

Therefore (4.6) transforms to

$$\frac{\partial}{\partial u} T(z,1) = \frac{T'(z) z^2 \phi_0}{T(z)}.$$

Thus, altogether we have

$$[z^n] S_k(z) = [z^n] z^{-1} T_k(z) \frac{T'(z) z^2 \phi_0}{T(z)},$$

which gives

$$[z^n] S_k(z) \sim \frac{-\tau_{0,k} \tau_1 \phi_0}{2\tau_0} \frac{n^{-1/2}}{\Gamma(1/2)} \rho^{-n}.$$

Finally, we get

$$\mathbb{E}(Y_n) = \sum_{k\geq 1} \mathbb{P}(Y_n \geq k) = \sum_{k\geq 1} \frac{[z^n] S_k(z)}{n [z^n] T(z)} \overset{n\to\infty}{\to} \sum_{k\geq 1} \frac{T_k(\rho)\phi_0}{T(\rho)}.$$

For the variance we use again the formula $\mathbb{V}(Y_n) = \sum_{k\geq 1} (2k-1)\mathbb{P}(Y_n \geq k) - \mathbb{E}(Y_n)^2$ and (4.5). □

Table 4.2 summarizes the values of the asymptotic mean and variance of the protection number of a random vertex for selected classes of simply generated trees.

	$\lim_{n\to\infty} \mathbb{E}(Y_n)$	$\lim_{n\to\infty} \mathbb{V}(Y_n)$
Plane trees	0.7276492769137261	0.8168993794836289
Motzkin trees	1.307604625963334	1.730614214799486
Incomplete binary trees	1.991819588602741	3.638259051495130
Cayley trees	1.186522661652180	1.632206223956926
Binary trees (w.r.t. internal nodes)	1.265686036087572	0.226591112528581

Table 4.2: The approximate values for the limits of mean and variance of the protection number of a random vertex in different classes of simply generated trees.

4.2 Protection number of Pólya trees

4.2.1 Protection number of the root

Let $T(z)$ be the generating function of Pólya trees (*cf.* page 23), which reads as

$$T(z) = ze^{T(z)} \exp\left(\sum_{i\geq 2} \frac{T(z^i)}{i}\right),$$

and in correspondence to the previous section let us denote by $T_k(z)$ the generating function of the class of Pólya trees that have protection number at least k. This generating function can be specified by

$$T_k(z) = ze^{T_{k-1}(z)} \exp\left(\sum_{i\geq 2} \frac{T_{k-1}(z^i)}{i}\right) - z, \tag{4.7}$$

with $T_0(z) = T(z)$.

Lemma 4.10. *All the generating functions $T_k(z)$ have their (unique) dominant singularity at ρ, and the singularity is a square root singularity.*

Proof. First let us recall that $T_0(z) = T(z)$. Thus, for $k = 0$ the lemma is trivial, when considering (3.2). For $k \geq 1$ we proceed by induction. Therefore let us assume that $T_{k-1}(z)$ has the dominant singularity ρ which is of type $\frac{1}{2}$. Then the dominant singularity of $T_k(z)$, satisfying the recurrence relation (4.7), comes from $e^{T_{k-1}(z)}$, since $\exp\left(\sum_{i\geq 2} \frac{T_{k-1}(z^i)}{i}\right)$ is analytic in $|z| < \rho + \epsilon$ with $\epsilon > 0$ sufficiently small. Applying the exponential function to a function having an algebraic singularity does neither change the location nor the type of the singularity, which proves the assertion. □

The goal of this section is to derive an asymptotic value for the average protection number of Pólya trees. We use again the formula $\mathbb{E}(X_n) = \sum_{k\geq 1} \mathbb{P}(X_n \geq k)$, but rewrite this equation as

$$\mathbb{E}(X_n) = \sum_{k\geq 1} \prod_{i=1}^{k} \mathbb{P}(X_n \geq i | X_n \geq i-1),$$

where the conditional probabilities can be obtained by

$$\mathbb{P}(X_n \geq k | X_n \geq k-1) = \frac{[z^n]T_k(z)}{[z^n]T_{k-1}(z)}. \tag{4.8}$$

Lemma 4.11. *The asymptotic expansions of the n-th coefficients of $T_k(z)$ and $T_{k-1}(z)$ read as*

$$[z^n]T_{k-1}(z) = \frac{\gamma_k \rho^{-n} n^{-\frac{3}{2}}}{\Gamma(-1/2)} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right),$$
$$[z^n]T_k(z) = \frac{(T_k(\rho) + \rho)\gamma_k \rho^{-n} n^{-\frac{3}{2}}}{\Gamma(-1/2)} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right),$$

as $n \to \infty$, with a constant $\gamma_k > 0$.

Proof. Let the Puiseux expansion of $T_{k-1}(z)$ be given by

$$T_{k-1}(z) = T_{k-1}(\rho) - \gamma_k \sqrt{1 - \frac{z}{\rho}} + \dots.$$

Then $T_k(z)$ behaves asymptotically as

$$T_k(z) \sim \rho e^{T_{k-1}(\rho)} Q_{k-1}(\rho) e^{-\gamma_k \sqrt{1-\frac{z}{\rho}}},$$

where $Q_{k-1}(\rho) = \exp\left(\sum_{i\geq 2} \frac{T_{k-1}(\rho^i)}{i}\right)$. Applying the asymptotic relation $e^{-\gamma_k\sqrt{1-\frac{z}{\rho}}} \sim 1 - \gamma_k\sqrt{1-\frac{z}{\rho}}$ and using the equation $\rho e^{T_{k-1}(\rho)} Q_{k-1}(\rho) = T_k(\rho) + \rho$ completes the proof. □

Plugging the expansions obtained in Lemma 4.11 into Equation (4.8) gives

$$\mathbb{P}(X_n \geq k | X_n \geq k-1) = T_k(\rho) + \rho,$$

which directly yields the following theorem.

Theorem 4.12. *Let X_n be the protection number of a random Pólya tree of size n. Then*

$$\lim_{n\to\infty} \mathbb{E}(X_n) = \sum_{k\geq 1} \prod_{i=1}^{k} (T_i(\rho) + \rho) \approx 2.154889671973873, \tag{4.9}$$

and

$$\lim_{n\to\infty} \mathbb{V}(X_n) \approx 1.369993017502652. \tag{4.10}$$

Proof. The proof for the asymptotic mean follows directly by Lemma 4.11. In order to determine the variance we use the representation $\lim_{n\to\infty} \mathbb{V}(X_n) = \sum_{k\geq 1}(2k-1)\prod_{i=1}^{k}(T_k(\rho)+\rho) - \mathbb{E}(X_n)^2$. □

Remark 4.13. *Note that in order to get accurate numerical values, we do not compute $T_k(\rho)$ by insertion into a (truncated) series expansion for $T_k(z)$. The reason is that ρ lies on the circle of convergence and thus the convergence is very slow at $z = \rho$. Instead, $T_k(\rho)$ can be directly computed using the recurrence relation* (4.7)*. The values $T_k(\rho^i)$ for $i \geq 2$, which appear in that recurrence relation, can be computed with the help of the series expansion of $T_k(z)$, because ρ^i then lies in the interior of the region of convergence where the series converges at an exponential rate.*

Remark 4.14. *We could also have used the same approach as for simply generated trees in order to get the asymptotic mean. Then the resulting formula looks like*

$$\lim_{n\to\infty} \mathbb{E}(X_n) = \sum_{k\geq 1} \rho^{k-1} \prod_{i=1}^{k-1} C_i e^{T_i(\rho)}, \tag{4.11}$$

where $C_j = \exp\left(\sum_{i\geq 2} \frac{T_j(\rho^i)}{i}\right)$. One can show that C_i tends to 1 and and $T_i(\rho)$ tends to 0 exponentially fast and get the constant given in Theorem 4.12. However, since this approach requires more technical calculations, we decided to switch to the more direct strategy using the conditional probabilities. Moreover note that the equivalence of (4.9) *and* (4.11) *is immediate from* (4.7)*.*

4.2.2 Protection number of a random vertex

The method of marking a leaf and replacing it by a tree with protection number k does not work here. Due to possible symmetries in non-plane trees, this would result in a wrong counting: Indeed, if there are k-protected vertices $x_1, \dots, x_\ell$ which can be mapped to each other by some automorphisms of the tree (*i.e.*, they lie in the same vertex class), then only one of them is counted. Though this is counterbalanced by trees having ℓ leaves in the same vertex class one of which is replaced by a tree with protection number k (the root of this tree is then counted ℓ times), there are further overcounts: As all leaves are marked, trees having several leaves in the same vertex class are counted several times, and so are their k-protected vertices. Thus, to overcome this problem, we appeal to the proof of [97, Theorem 3.1] here: For a tree T let

$$f(T) = \begin{cases} 1 & \text{if } T \text{ has protection number at least } k, \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, we define $F(T)$ to be the number of k-protected nodes in T. Then the generating function $R_k(z,u) = \sum_T z^{|T|} u^{F(T)}$ satisfies (*cf.* [97, Equation (3.1)])

$$z \exp\left(\sum_{i\geq 1} \frac{R_k(z^i,u^i)}{i}\right) = \sum_{n\geq 1} z^n \sum_{T:|T|=n} u^{F(T)-f(T)}. \tag{4.12}$$

As in Section 4.1.2 we utilize the formula $\mathbb{E}(Y_n) = \sum_{k\geq 1} \mathbb{P}(Y_n \geq k)$ and express the occurring probabilities as $\mathbb{P}(Y_n \geq k) = [z^n]S_k(z)/(n[z^n]T(z))$ with $S_k(z)$ being the generating function whose n-th coefficient is the cumulative number of k-protected nodes in all trees of size n. Obviously, $((\partial/\partial u)R_k)(z,1) = S_k(z)$ and thus by differentiating (4.12) with respect to u and inserting $u = 1$ we obtain

$$T(z) \sum_{i\geq 1} S_k(z^i) = S_k(z) - T_k(z). \tag{4.13}$$

This implies

$$S_k(z) = \frac{T(z)\sum_{i\geq 2} S_k(z^i) + T_k(z)}{1-T(z)} \sim \frac{\sum_{i\geq 2} S_k(\rho^i) + T_k(\rho)}{b\sqrt{1-\frac{z}{\rho}}} \tag{4.14}$$

where b is the constant appearing in (3.2). Standard transfer theorems (see Theorems 2.8 and 2.10)) applied to (3.2) give

$$[z^n]T(z) \sim \frac{-bn^{-3/2}\rho^{-n}}{\Gamma(-1/2)} = \frac{bn^{-3/2}\rho^{-n}}{2\sqrt{\pi}},$$

and from (4.14) we get

$$[z^n]S_k(z) \sim \frac{\left(\sum_{i\geq 2} S_k(\rho^i) + T_k(\rho)\right) n^{-1/2}\rho^{-n}}{b\sqrt{\pi}}$$

and thus

$$\mathbb{P}(Y_n \geq k) \sim \frac{2}{b^2}\left(\sum_{i\geq 2} S_k(\rho^i) + T_k(\rho)\right). \tag{4.15}$$

Since $T_k(\rho)$ decreases exponentially (*cf.* remark after Theorem 4.12), and so does $\sum_{i\geq 2} S_k(\rho^i)$, these probabilities decrease exponentially and thus the series for $\mathbb{E}(Y_n)$, namely

$$\mathbb{E}(Y_n) = \sum_{k\geq 1} \mathbb{P}(Y_n \geq k),$$

converges rapidly. But (4.15) still bears a secret, because we do not have an explicit expression for $S_k(z)$ and we cannot solve the functional equation (4.13).

For numerical purposes, however, it is not necessary to have an explicit expression for $S_k(z)$. If we write $S_k(z) = \Psi(S_k(z))$ with Ψ being the operator on the ring of formal power series defined by

$$\Psi(f(z)) = \frac{T(z)\sum_{i\geq 2} f(z^i) + T_k(z)}{1 - T(z)},$$

then Ψ is a contraction on the metric space $\mathbb{R}[[z]]$ equipped with the formal topology (*cf.* [45, Appendix A.5]). Indeed, if $f(z)$ and $g(z)$ coincide up to their ℓ-th coefficient, then the first $2\ell + 2$ coefficients of $\Psi(f(z))$ and$\Psi(g(z))$ coincide.

As there is exactly one tree with $k+1$ vertices which possesses k-protected vertices at all (namely the path of length k has a k-protected root) whereas all smaller trees do not possess any k-protected vertices, we know that the (one-term) series z^{k+1} coincides with $S_k(z) = z^{k+1} + \cdots$ in its first $k+2$ coefficients. Applying Ψ to z^{k+1} a few times, with each application more than doubling the number of known coefficients of $S_k(z)$, gives quickly a fairly accurate expression for $S_k(z)$. We obtain the following theorem:

Theorem 4.15. *Let Y_n be the protection number of a random vertex in a random Pólya tree of size n. Then*

$$\lim_{n\to\infty} \mathbb{E}(Y_n) = \sum_{k\geq 1} \frac{2}{b^2} \left(\sum_{i\geq 2} S_k(\rho^i) + T_k(\rho) \right) \approx 0.9953254987,$$

and

$$\lim_{n\to\infty} \mathbb{V}(Y_n) \approx 1.3818769746.$$

4.3 Protection number of non-plane binary trees

4.3.1 Protection number of the root

We denote by $T(z)$ the generating function of non-plane binary trees (*cf.* page 24), where z marks the number of internal nodes. Then $T(z)$ satisfies

$$T(z) = 1 + z\left(\frac{1}{2}T(z)^2 + \frac{1}{2}T(z^2)\right).$$

The generating function $T_k(z)$ of non-plane binary trees with protection number at least k fulfills

$$T_k(z) = z\left(\frac{1}{2}T_{k-1}(z)^2 + \frac{1}{2}T_{k-1}(z^2)\right),$$

and $T_0(z) = T(z)$.

In order to obtain the asymptotic mean and variance for the protection number of a random non-plane binary tree of size n we proceed analogously as in the previous section for Pólya trees. Thus, we use

$$\mathbb{E}(X_n) = \sum_{k\geq 1}\prod_{i=1}^{k}\mathbb{P}(X_n \geq i | X_n \geq i-1) = \sum_{k\geq 1}\prod_{i=1}^{k}\frac{[z^n]T_i(z)}{[z^n]T_{i-1}(z)}.$$

Theorem 4.16. *Let X_n be the protection number of a random non-plane binary tree of size n. Then*

$$\lim_{n\to\infty}\mathbb{E}(X_n) = \sum_{k\geq 1}\prod_{i=1}^{k-1}(\rho T_i(\rho)) \approx 1.707603060723366,$$

and

$$\lim_{n\to\infty}\mathbb{V}(X_n) \approx 0.431102549825064.$$

Proof. The Puiseux expansions of $T_k(z)$ and $T_{k+1}(z)$ read as

$$T_{k-1}(z) = T_{k-1}(\rho) - \gamma_k\sqrt{1-\frac{z}{\rho}} + \mathcal{O}\left(1-\frac{z}{\rho}\right),$$

and

$$T_k(z) = \rho\left(\frac{1}{2}T_{k-1}(\rho)^2 + \frac{1}{2}T_{k-1}(\rho^2)\right) + \rho T_{k-1}(\rho)\gamma_k\sqrt{1-\frac{z}{\rho}} + \mathcal{O}\left(1-\frac{z}{\rho}\right)$$

Using singularity analysis (see Section 2.2) yields the desired result for the mean. For the variance we use again the formula $\mathbb{V}(X_n) = \sum_{k\geq 1}(2k-1)\mathbb{P}(X_n \geq k) - \mathbb{E}(X_n)^2$. □

4.3.2 Protection number of a random internal vertex

The asymptotic mean and variance for the protection number of a randomly chosen internal vertex in a random non-plane binary tree can be obtained in the same way as in the previous section for Pólya trees. Thus, we again set up an equation for the generating function $R_k(z,u)$ where the coefficients $[z^n u^\ell]R_k(z,u)$ count the number of non-plane binary trees of size n with ℓ k-protected vertices, reading as

$$\frac{z}{2}\left(R_k(z,u)^2 + R_k(z^2,u^2)\right) = \sum_{n\geq 1} z^n \sum_{T:|T|=n} u^{F(T)-f(T)}.$$

Differentiating this equation with respect to u and setting $u=1$ yields

$$zT(z)S_k(z) + zS_k(z^2) = S_k(z) - T_k(z).$$

Using the asymptotic expansion of the generating function $T(z)$ of non-plane binary trees given in (3.4) we get

$$\mathbb{P}(Y_n \geq k) = \frac{[z^n]S_k(z)}{n[z^n]T(z)} \sim \frac{2}{a^2\rho}\left(\rho S_k(\rho^2) + T_k(\rho)\right).$$

By denoting $\Psi(f(z)) = \frac{zf(z^2)+T_k(z)}{1-zT(z)}$ we can use the same arguments as in the Pólya case to efficiently obtain numerical values for the probabilities $\mathbb{P}(Y_n \geq k)$. Finally, we are able to calculate the asymptotic mean and variance for the protection number of a random node in non-plane binary trees.

Theorem 4.17. *Let Y_n be the protection number of a random internal vertex in a random non-plane binary tree of size n. Then*

$$\lim_{n\to\infty} \mathbb{E}(Y_n) = \frac{2}{a^2\rho} \sum_{k\geq 1} \left(\rho S_k(\rho^2) + T_k(\rho)\right) \approx 1.3124128299,$$

and

$$\lim_{n\to\infty} \mathbb{V}(Y_n) \approx 0.2676338724.$$

Chapter 5

Non-isomorphic subtree-shapes

This chapter is based on the not yet submitted manuscript [14], which was joint work with Olivier Bodini, Antoine Genitrini, Bernhard Gittenberger and Mehdi Naima. We prove asymptotic results on the average number of non-isomorphic fringe subtree-shapes for two special classes of trees. This parameter is often studied in the context of so-called *compacted trees* [21, 49, 50], which in fact are no trees, but directed acyclic graphs that can be constructed from every tree in a unique way via a post-order traversal, such that repeatedly occurring subtrees in the original tree are represented by pointers to already existing nodes representing the root of the respective subtree. In this way every node represents a distinct subtree and hence the size of the compacted tree, *i.e.*, the number of its nodes, corresponds to the number of non-isomorphic fringe subtrees of the original tree, see Figure 5.1.

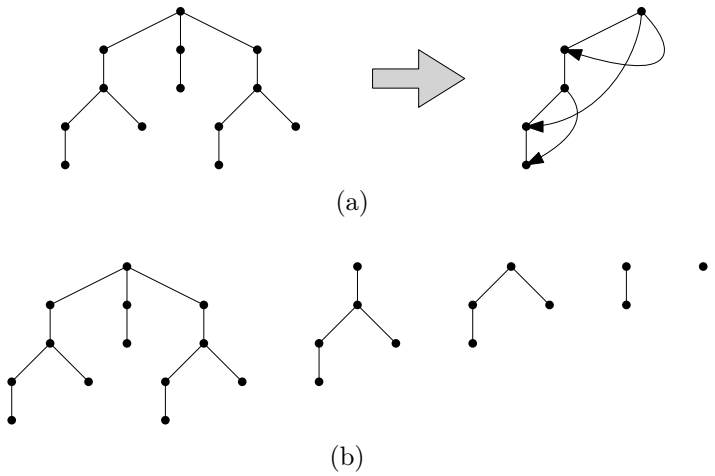


Figure 5.1: A plane tree and its corresponding compacted tree (a), and all 5 of its non-isomorphic fringe subtrees (b).

The two classes of trees that are studied in this chapter are recursive trees (*cf.* page 25) and plane increasing binary trees (*cf.* page 25), which are both comprised of increasingly labeled trees. Thus, we have to extend the definition of a compacted tree to classes of labeled trees. For our purposes, the *compactification of a labeled tree* T is defined as the compacted tree belonging to the tree T after removing its labels. Thus, several different labeled trees can have the same compacted tree. In this way, the size of the compacted tree belonging to a labeled tree T gives exactly the number of non-isomorphic fringe subtree-shapes.

Tree-shape data structures are omnipresent in computer science, for example as syntax structures of programs, symbolic expressions in computer algebra systems or XML data structures. However, in order to reduce redundancy in the storage, usually an algorithmic step called the *common subexpression recognition* is run to identify identical fringe subtrees so that only one occurrence is stored and all other are replaced by pointers to the first one. In the context of tree compaction several studies attempt to quantitatively analyse the process of compaction. The first one, in the context of analytic combinatorics, is presented by Flajolet and his coauthors in [46]. In this paper the authors study the compaction ratio of binary trees and prove that the compacted tree belonging to a large binary tree of size n is on average of size $c\frac{n}{\sqrt{\log n}}$ with a computable constant c. Moreover, the authors state that their analysis is fully adaptable to all families of simply generated trees. In [21] Bousquet-Mélou, Lohrey, Maneth and Noeth present the complete proof for the compaction quantitative analysis of simply generated trees and apply it experimentally on XML-trees.

Within this thesis we extend these results to other classes of trees that do not fall into the framework of simply generated trees. The first family that we study in Section 5.1 is the class of recursive trees, while in Section 5.2 we analyse the class of plane increasing binary trees. Both these families have been extensively studied in the last two decades in both probability studies [22, 33, 37, 80] and combinatorics [9, 71, 92]. In the two subsequent sections we will obtain asymptotic lower and upper bounds for the expected values of the size of the compaction of these two classes of increasingly labeled trees, thereby showing that they can be compacted in a more efficient way than simply generated trees. In [14] a new data structure is introduced, which is based on the compaction of plane increasing binary trees and thus allows for an efficient storage of the involved information.

In Figure 5.2 we have depicted an increasing binary tree of size 500 (after removing its labels), where we highlighted its corresponding compacted tree in red, consisting of 172 remaining nodes.

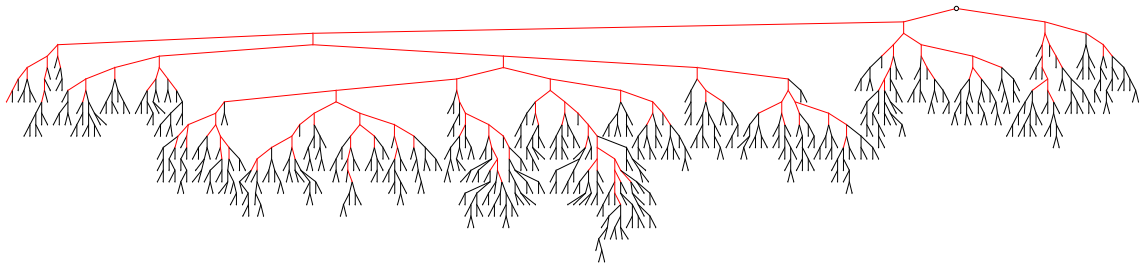


Figure 5.2: A uniformly sampled increasing binary tree structure with 500 internal nodes. After compaction the black fringe subtrees are removed, which yields a compacted structure of size 172.

5.1 Number of non-isomorphic subtree-shapes in recursive trees

Let $\mathcal{T}_n$ be the class of recursive trees (*cf.* page 25) of size n. The size of a tree τ is defined as its total number of vertices and denoted by $|\tau|$. Let X_n be the size of the compacted tree corresponding to a random recursive tree τ of size n. In other words, X_n is the number of distinct fringe subtree-shapes in τ. We define $\mathcal{P}$ as the class of Pólya trees, which corresponds to the set of all possible shapes of recursive trees, once the labels have been removed. For every Pólya tree t, we write $t \in \tau$ if t occurs as a fringe subtree-shape of τ. Otherwise we write $t \notin \tau$. Then we have

$$\mathbb{E}(X_n) = \sum_{t \in \mathcal{P}_{\leq n}} \mathbb{P}(t \in \tau) = \sum_{t \in \mathcal{P}_{\leq n}} \left(1 - \mathbb{P}(t \notin \tau)\right), \tag{5.1}$$

where $\mathcal{P}_{\leq n}$ is the set of all Pólya trees with size at most n.

Remark 5.1. *Recall that the tree t represents a tree-shape, thus it is unlabeled, while τ is a recursive tree and therefore increasingly labeled.*

Let us recall that the exponential generating function $T(z)$ of the class of recursive trees, which was introduced in Chapter 3 (see page 25), reads as

$$T(z) = \ln \frac{1}{1-z},$$

and has its unique dominant singularity at $\rho = 1$.

Now, for a given Pólya tree $t \in \mathcal{P}$ let us consider a perturbed combinatorial class $\mathcal{S}_t$, that contains all recursive trees except for those that contain a t-shape as a (fringe) subtree-shape. The corresponding exponential generating function $S_t(z)$ satisfies

$$S_t'(z) = e^{S_t(z)} - P_t'(z), \tag{5.2}$$

where $P_t(z) = \frac{z^{|t|}}{|t|!}\ell(t)$, with $\ell(t)$ denoting the number of ways to increasingly label the tree-shape t. The expression (5.1) for the expected value of the number of non-isomorphic subtree-shapes can now be rewritten in terms of the generating functions $T(z)$ and $S_t(z)$, reading as

$$\mathbb{E}\left(X_n\right) = \sum_{t \in \mathcal{P}_{\leq n}} \left(1 - \mathbb{P}(t \notin \tau)\right) = \sum_{t \in \mathcal{P}_{\leq n}} \left(1 - \frac{[z^n]S_t(z)}{[z^n]T(z)}\right). \tag{5.3}$$

So, the problem is now essentially reduced to the analysis of the asymptotic behavior of $[z^n]S_t(z)$.

Solving Equation (5.2) we obtain the exponential generating function

$$S_t(z) = \ln\left(\frac{1}{1 - \int_0^z e^{-P_t(v)}\,\mathrm{d}v}\right) - P_t(z). \tag{5.4}$$

Thus, for the dominant singularity $\tilde{\rho}$ of $S_t(z)$ the following equation must hold:

$$\int_0^{\tilde{\rho}} e^{-P_t(v)}\,\mathrm{d}v = 1. \tag{5.5}$$

As $e^{-P_t(v)} < 1$ for positive v, the dominant singularity $\tilde{\rho}$ must be greater than 1. Therefore we write $\tilde{\rho} = \rho(1+\epsilon)$ with suitable $\epsilon > 0$.

Remark 5.2. *Note that the identity $\tilde{\rho} = \rho(1+\epsilon)$ can also be explained combinatorially: Since the class $\mathcal{S}_t$ is a proper subset of the class $\mathcal{T}$ it follows that $[z^n]S_t(z) \leq [z^n]T(z)$ for all n. Thus, the exponential part of the asymptotics of the coefficients of $S_t(z)$ has to be smaller than that of the coefficients of $T(z)$, which implies that $\tilde{\rho}$ has to be a little bit larger than ρ.*

Subsequently, we will use the following notation.

Notation 5.3. *For the size and the weight of a Pólya tree t we use*

$$k := |t| \quad \text{and} \quad w(t) := \frac{\ell(t)}{|t|},$$

respectively. Moreover, let

$$G(z) := \int_0^z e^{-P_t(v)}\,\mathrm{d}v = \int_0^z e^{-w(t)v^k}\,\mathrm{d}v,$$

if $z \geq 0$ and its complex continuation if z is not a nonnegative real number. With this notation (5.5) reads as $G(1+\epsilon) = 1$. By expanding the integrand, we obtain

$$G(z) = \sum_{\ell \geq 0} (-w(t))^\ell \frac{z^{\ell k+1}}{(\ell k+1)\cdot \ell!},$$

which shows that $G(z)$ is an entire function.

We will show the following result concerning the asymptotic mean of the number of non-isomorphic fringe subtree-shapes of a random recursive tree.

Theorem 5.4. *Let X_n be the number of non-isomorphic subtree-shapes of a random recursive tree of size n. Then there exist constants C_1 and C_2 such that*

$$C_1\sqrt{n} \leq \mathbb{E}(X_n) \leq C_2 \frac{n}{\log n}, \qquad \text{for } n \to \infty.$$

The remainder of this section is devoted to the proof of Theorem 5.4, where we proceed as follows: First, in Lemma 5.5, we compute an upper bound for the dominant sigularity $\tilde{\rho}$ of $S_t(z)$, which directly yields the asymptotic behavior of $\tilde{\rho}$, as $k \to \infty$ (see Corollary 5.6). Then, in Lemma 5.8, we provide asymptotics for the n-th coefficient of the generating function $S_t(z)$ when n tends to infinity, thereby showing that the error term is uniform in the size k of the "forbidden" tree t. The average size of a compacted tree corresponding to a random recursive tree is expressed as a sum over the forbidden trees (see Equation (5.3)). Thereby the two cases, whether the size k of the forbidden tree t is smaller or larger than $\log_{\frac{1}{\sigma}} n$ are treated differently in Proposition 5.10 (small trees) and Proposition 5.11 (large trees) in order to obtain an upper bound for the size of the compacted tree. Finally, in Proposition 5.12 a (crude) lower bound for the size of the compacted tree is given.

Lemma 5.5. *Let $S_t(z)$ be the generating function of the perturbed combinatorial class of recursive trees that do not contain the shape t as a subtree (cf. Equation (5.2)). The dominant singularity $\tilde{\rho}$ of $S_t(z)$ is bounded by*

$$\tilde{\rho} = 1 + \epsilon < 1 + \frac{2w(t)}{k},$$

where $w(t) = \frac{\ell(t)}{k!}$ and $\ell(t)$ denotes the number of possible increasing labelings of the Pólya tree t of size k.

Proof. First observe that the number of increasing labelings of the Pólya tree t is bounded by $(k-1)!$, which gives the very crude bound $w(t) \le 1/k$.

Next, as $\tilde{\rho}$ satisfies $G(1+\epsilon) = 1$, it suffices to show the inequality $G\left(1+\frac{2w(t)}{k}\right) > G(1+\epsilon)$. We show the the equivalent inequality $G\left(1+\frac{2w(t)}{k}\right) - G(1) > G(1+\epsilon) - G(1)$.

Then we have

$$G(1+\epsilon) - G(1) = 1 - \int_0^1 e^{-w(t)v^k}\,\mathrm{d}v \le 1 - \int_0^1 (1 - w(t)v^k)\,\mathrm{d}v = \frac{w(t)}{k+1}.$$

On the other hand, if $k \ge 3$, then we have the lower bound

$$\begin{aligned} G\left(1+\frac{2w(t)}{k}\right) - G(1) &\ge \frac{2w(t)}{k}\exp\left(-w(t)\left(1+\frac{2w(t)}{k}\right)^k\right) \\ &\ge \frac{2w(t)}{k}\exp\left(-w(t)\left(1+\frac{2}{k^2}\right)^k\right) > \frac{2w(t)}{k}e^{-2w(t)} \\ &= \frac{w(t)}{k}\cdot 2e^{-2w(t)} > \frac{w(t)}{k+1} \end{aligned}$$

which implies the assertion. In the course of this chain of inequalities we used $w(t) < 1/k$ and then $\left(1+\frac{2}{k^2}\right)^k < 2$ (for $k \ge 3$) in the second line, then again $w(t) < 1/k$, and finally $k \ge 3$ and $2e^{-2/3} > 1$.

If $k = 2$, then t is a path of length one and therefore $w(t) = 1/2$. This gives explicitely $\int_1^{3/2} e^{-v^2/2}\,\mathrm{d}v > 1/6$ which is easily verified. □

Corollary 5.6. *With the notations of Lemma 5.5 we have the following asymptotic relation:*

$$\tilde{\rho} = 1 + \epsilon \sim 1 + \frac{w(t)}{k},\ \textit{as } k \to \infty.$$

Proof. Let us write $G(z)$ as $G(z) = z + R(z)$ with

$$R(z) = \sum_{\ell \ge 1} (-w(t))^\ell \frac{z^{\ell k+1}}{(\ell k+1)\cdot \ell!} \tag{5.6}$$

As $\tilde{\rho} = 1+\epsilon$ is the smallest positive solution of $G(z) = 1$, it is the smallest positive zero of $z - 1 + R(z)$. From Lemma 5.5 we know that $\epsilon = \mathcal{O}(1/k^2)$ and thus $\tilde{\rho}^k \sim 1$, as k tends to infinity, and $R(\tilde{\rho}) = w(t)\tilde{\rho}^{k+1}/(k+1) + \mathcal{O}(1/k^3)$. This implies

$$\epsilon \sim \frac{w(t)}{k+1}\tilde{\rho}^{k+1} \sim \frac{w(t)}{k}, \tag{5.7}$$

as desired. □

Remark 5.7. *Using more terms of the expansion of $G(z)$, it is possible to derive a more accurate asymptotic expression for ϵ (in principle up to arbitrary order). As an example, we state*

$$\tilde{\rho} = 1 + \frac{w(t)}{k+1} + \frac{w(t)^2(3k+1)}{(k+1)(4k+2)} + \frac{w(t)^3(29k^3+32k^2+10k+1)}{6(k+1)^2(2k+1)(3k+1)} + \mathcal{O}\left(\frac{w(t)^4}{k}\right).$$

Now we are able to derive a uniform asymptotic expression for the coefficients of $S_t(t)$.

Lemma 5.8. *Let $S_t(z)$ be the generating function of the perturbed class of recursive trees defined in* (5.4)*. Then for sufficiently small $\delta > 0$ we have*

$$[z^n]S_t(z) = \frac{\tilde{\rho}^{-n}}{n}(1 + \mathcal{O}\left(n^{-\delta}\right)), \text{ as } n \to \infty,$$

which holds uniformly for $D \leq k \leq n$, where $D > 0$ is independent of n and sufficiently large.

Proof. Recall that by (5.4) we have

$$S_t(z) = \ln\left(\frac{1}{1 - G(z)}\right) - P_t(z). \tag{5.8}$$

Since $G(z)$ is an entire function, the singularities of S_t are exactly the zeros of $G(z)-1$. Therefore, consider z_0 such that $G(z_0) = 1$ and write $G(z) = z + R(z)$ with $R(z)$ as in (5.6). Then

$$\begin{aligned} |R(z_0)| &\leq \frac{1}{k+1}\sum_{\ell \geq 1}\frac{|w(t)|^\ell |z_0|^{k\ell+1}}{\ell!} \\ &< \frac{1}{k}(e^{|w(t)||z_0|^k} - 1) \end{aligned} \tag{5.9}$$

Assume first that $|z_0| \leq 1 + \frac{e-1}{k}$. As the dominant singularity of $S_t(z)$ is $\tilde{\rho}$ and $\tilde{\rho} > 1$, we must have $|z_0| > 1$. Thus, the upper bound on $|z_0|$ and (5.9) imply

$$1 - z_0 = R(z_0) = \mathcal{O}\left(1/k^2\right). \tag{5.10}$$

On the other hand, $R(z) \sim -\frac{w(t)}{k}z_0^k$ and $1 - z_0 \sim -w(t)/k$ because of Corollary 5.6. Thus z_0 is asymptotically equal to a k-th root of unity. But then $z_0 = \tilde{\rho}$, because the distnce between the other k-th roots of unity and 1 is greater than $1/k$, which contradicts (5.10).

Now assume that $|z_0| = 1 + \eta$ with $1/k < \eta < (e-1)\ln(k)/k$. Then $w(t)|z_0|^k \leq 1$ and so by (5.9) we have then $|R(z_0)| \leq (e-1)/k$. But we assumed $|z_0 - 1| > (e-1)/k$. Summarizing what we have so far, we obtain that either $z : 0 = \tilde{\rho}$ or $|z_0| > 1 + \frac{\ln k}{k}$.

Notice that the asymptotic relation $R(z) \sim \frac{w(t)}{k}z_0^k$ from the first case discussed above (namely $|z_0| - 1| \leq 1/k$) implies also that $\tilde{\rho}$ is a simple zero of $G(z) - 1$. Thus $G(z) - 1 = (z - \tilde{\rho})\tilde{G}(z)$ where $\tilde{G}(z)$ is analytic in the domain $|z| \leq 1 + \frac{\ln k}{k}$ and does not have any zeros there. Thus,

$$\begin{aligned} S_t(z) &= \ln\left(\frac{1}{1 - G(z)}\right) - P_t(z) \\ &= -\ln\left(1 - \frac{z}{\tilde{\rho}}\right) - \ln(\tilde{\rho}\tilde{G}(z)) - P_t(z), \end{aligned}$$

where, apart from the first summand, there are no singularities in $|z| \leq 1 + \frac{\ln k}{k}$. Applying singularity analysis (*cf.* Corollary 2.11) gives

$$[z^n]S_t(z) = \frac{\tilde{\rho}^{-n}}{n}\left(1 + \mathcal{O}\left(n\left(\frac{\gamma}{\tilde{\rho}}\right)^{-n}\right)\right) \tag{5.11}$$

with $\gamma = 1 + \frac{(1+2\delta)\ln k}{k}$ for sufficiently small $\delta > 0$. Finally, notice that

$$\frac{\gamma}{\tilde{\rho}} \sim 1 + \frac{(1+2\delta)\ln k}{k} \geq 1 + \frac{(1+2\delta)\ln n}{n},$$

as k tends to infinity staying not larger than n. Therefore, for sufficiently large k the inequality

$$\frac{\gamma}{\tilde{\rho}} \geq 1 + \frac{(1+\delta)\ln n}{n}$$

holds. Plugging this estimate into (5.11) yields the desired result after all. □

Remark 5.9. *Within this section many logarithms that occur are to the base $\frac{1}{\sigma}$, where $\sigma \approx 0.338\ldots$ denotes the dominant singularity of Pólya trees. To ensure a simpler reading we omit this base subsequently and instead just write $\log n$. For differentiation purposes the natural logarithm will always be denoted by $\ln n$.*

Now we decompose the sum (5.3) into

$$\mathbb{E}\,(X_n) = \sum_{\substack{t\in\mathcal{P}_{\leq n}\\ k<\log n}} \left(1 - \frac{[z^n]S_t(z)}{[z^n]T(z)}\right) + \sum_{\substack{t\in\mathcal{P}_{\leq n}\\ k\geq\log n}} \left(1 - \frac{[z^n]S_t(z)}{[z^n]T(z)}\right), \tag{5.12}$$

and investigate the two sums individually, starting with the leftmost one, whose summands can be estimated by 1.

Proposition 5.10. *Let $T(z)$ be the generating function of recursive trees, $S_t(z)$ the generating function of the perturbed class of recursive trees that do not contain the tree-shape t as a fringe subtree-shape, and let $\mathcal{P}_{\leq n}$ denote the class of Pólya trees with size at most n. Then*

$$\sum_{\substack{t\in\mathcal{P}_{\leq n}\\ k<\log n}} \left(1 - \frac{[z^n]S_t(z)}{[z^n]T(z)}\right) = \mathcal{O}\left(\frac{n}{\sqrt{(\log n)^3}}\right), \qquad \text{for } n\to\infty,$$

where the logarithm $\log n$ is to the base $\frac{1}{\sigma}$ with σ denoting the dominant singularity of the generating function of Pólya trees.

Proof. Remember that we denote $k := |t|$. Furthermore, we denote by $P(z)$ be the generating function of Pólya trees and by σ its dominant singularity. Then

$$\begin{aligned}\sum_{\substack{t\in\mathcal{P}_{\leq n}\\ k<\log n}} \left(1 - \frac{[z^n]S_t(z)}{[z^n]T(z)}\right) &\leq \sum_{\substack{t\in\mathcal{P}_{\leq n}\\ k<\log n}} 1 = \sum_{k<\log n} [z^k]P(z)\\ &\sim \frac{1}{1-\sigma}[z^{\lfloor\log n\rfloor}]P(z) = \mathcal{O}\left(\frac{\sigma^{-\lfloor\log n\rfloor}}{\sqrt{(\log n)^3}}\right).\end{aligned}$$

Since $\log n$ has the base $\frac{1}{\sigma}$, we estimate $\sigma^{-\lfloor\log n\rfloor} \leq n$, which completes the proof. □

Now we are able to estimate the asymptotic behavior of the second sum in (5.12).

Proposition 5.11. *Let $T(z)$ be the generating function of recursive trees, $S_t(z)$ the generating function of the perturbed class of recursive trees that do not contain the tree-shape t as a fringe subtree-shape, and let $\mathcal{P}_{\leq n}$ denote the class of Pólya trees with size at most n. Then*

$$\sum_{\substack{t\in\mathcal{P}_{\leq n}\\ k\geq\log n}} \left(1-\frac{[z^n]S_t(z)}{[z^n]T(z)}\right) = \mathcal{O}\left(\frac{n}{\log n}\right), \qquad \text{for } n\to\infty,$$

where the logarithm $\log n$ is to the base $\frac{1}{\sigma}$ with σ denoting the dominant singularity of the generating function of Pólya trees.

Proof. Using Lemma 5.8 we get

$$\frac{[z^n]S_t(z)}{[z^n]T(z)} \sim \tilde{\rho}^{-n} = (1+\epsilon)^{-n}, \qquad \text{for } \to\infty,$$

uniformly in $|t|=k$. Thus,

$$\sum_{\substack{t\in\mathcal{P}_{\leq n}\\ k\geq\log n}} \left(1-\frac{[z^n]S_t(z)}{[z^n]T(z)}\right) \sim \sum_{\substack{t\in\mathcal{P}_{\leq n}\\ k\geq\log n}} \left(1-(1+\epsilon)^{-n}\right).$$

By means of the Bernoulli inequality we get

$$\sum_{\substack{t\in\mathcal{P}_{\leq n}\\ k\geq\log n}} 1-(1+\epsilon)^{-n} \leq \sum_{\substack{t\in\mathcal{P}_{\leq n}\\ k\geq\log n}} n\cdot\epsilon,$$

which by use of (5.5) can be further simplified to

$$\sum_{\substack{t\in\mathcal{P}_{\leq n}\\ k\geq\log n}} n\cdot\epsilon \sim \sum_{k=\log n}^{n}\sum_{\substack{t\in\mathcal{P}_{\leq n}\\ |t|=k}} n\cdot\frac{w(t)}{k} = \sum_{k=\log n}^{n}\frac{n}{k}\sum_{\substack{t\in\mathcal{P}_{\leq n}\\ |t|=k}} w(t).$$

Using the fact that

$$\sum_{\substack{t\in\mathcal{P}_{\leq n}\\ |t|=k}} w(t) = [z^k]T(z) = \frac{1}{k},$$

we further get

$$\sum_{k=\log n}^{n}\frac{n}{k}\sum_{\substack{t\in\mathcal{P}_{\leq n}\\ |t|=k}} w(t) = \sum_{k=\log n}^{n}\frac{n}{k^2} = \Theta\left(n\int_{\log n}^{\infty}\frac{1}{x^2}\,\mathrm{d}x\right) = \Theta\left(\frac{n}{\log n}\right).$$

Thus, the statement is proved. □

Finally, we now prove a lower bound for the average size of the compacted tree based on a random recursive tree of size n and thereby finish the proof of Theorem 5.4.

Proposition 5.12. *Let $T(z)$ be the generating function of recursive trees, $S_t(z)$ the generating function of the perturbed class of recursive trees that do not contain the tree-shape t as a fringe subtree-shape, and let $\mathcal{P}_{\leq n}$ denote the class of Pólya trees with size at most n. Then*

$$\sum_{\substack{t\in\mathcal{P}_{\leq n}\\ k\geq \log n}} \left(1-\frac{[z^n]S_t(z)}{[z^n]T(z)}\right) = \Omega\left(\sqrt{n}\right), \qquad \text{for } n\to\infty,$$

where the logarithm $\log n$ is to the base $\frac{1}{\sigma}$ with σ denoting the dominant singularity of the generating function of Pólya trees.

Proof. First, we use the inequality

$$(1+\epsilon)^{-n} \leq e^{-n\epsilon+\frac{n\epsilon^2}{2}}$$

in order to estimate

$$\sum_{\substack{t\in\mathcal{P}_{\leq n}\\ k\geq \log n}} \left(1-\frac{[z^n]S_t(z)}{[z^n]T(z)}\right) = \sum_{\substack{t\in\mathcal{P}_{\leq n}\\ k\geq \log n}} \left(1-(1+\epsilon)^{-n}\right) \geq \sum_{k\geq\log n}\sum_{\substack{t\in\mathcal{P}_{\leq n}\\ |t|=k}} (1-e^{-n\epsilon+\frac{n\epsilon^2}{2}}). \quad (5.13)$$

For the sake of simplicity we will use the abbreviation

$$\sum_t := \sum_{\substack{t\in\mathcal{B}_{\leq n}\\ |t|=k}}$$

in the remainder of this proof. Since $x\mapsto 1-\exp\left(-nx+\frac{nx^2}{2}\right)$, $x\geq 0$, is a concave nonnegative function with a zero in the origin and $w(t)>0$ for all t, we can estimate the inner sum in (5.13), which yields

$$\sum_{k\geq\log n}\sum_t (1-e^{-n\epsilon+\frac{n\epsilon^2}{2}}) \geq \sum_{k\geq\log n} (1-e^{-n\sum_t\epsilon+\frac{n}{2}(\sum_t\epsilon)^2}).$$

Note that ϵ depends on t, and that

$$\sum_t \epsilon \sim \sum_t \frac{w(t)}{k} = \frac{1}{k}\sum_t w(t) = \frac{1}{k^2} \qquad \text{as } n\to\infty.$$

Thus, we get

$$\sum_{\substack{t\in\mathcal{P}_{\leq n}\\ k\geq \log n}} \left(1-\frac{[z^n]S_t(z)}{[z^n]T(z)}\right) \geq \sum_{k\geq\log n} (1-e^{-\frac{n}{k^2}+\frac{n}{2k^4}}) \sim \int_{\log n}^{\infty} (1-e^{-\frac{n}{x^2}+\frac{n}{2x^4}})\,\mathrm{d}x$$

$$= \sqrt{n}\int_{\sqrt{n}\log n}^{\infty} (1-e^{-\frac{1}{y^2}+\frac{1}{2ny^4}})\,\mathrm{d}y.$$

Since the integral is convergent this gives a lower bound that is $\Theta(\sqrt{n})$. □

On the left hand side, Figure 5.3 shows a recursive tree structure containing 5000 nodes, which has been uniformly sampled among all trees of the same size. The original root of the tree is depicted by a small circle ∘. The right hand side shows the structure that is left after the compaction of the latter tree, which remains only of 663 nodes.

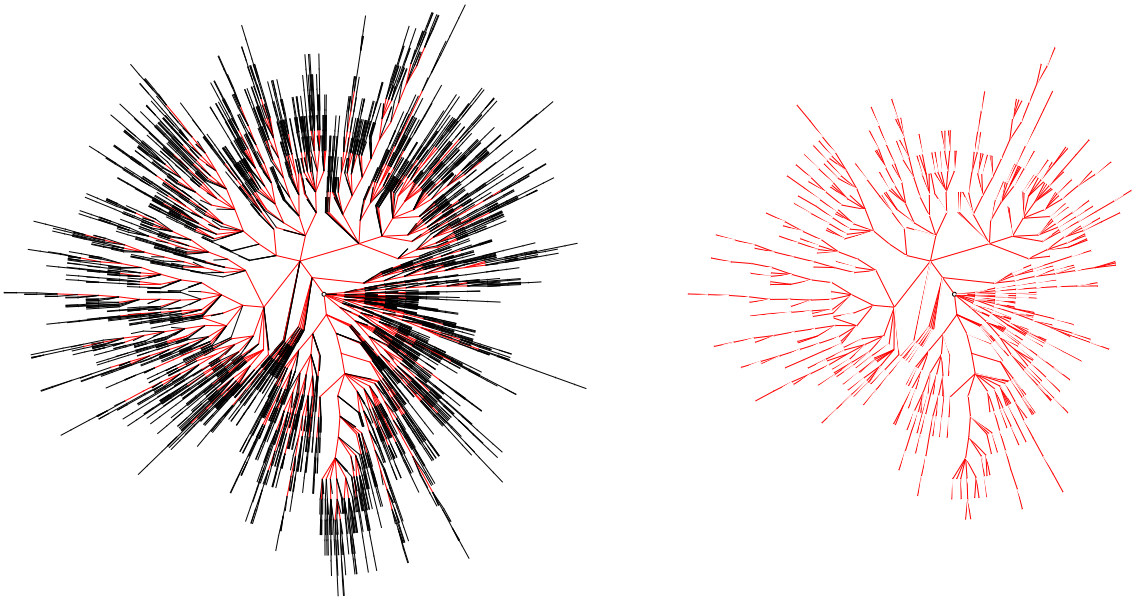


Figure 5.3: A uniformly sampled plane recursive tree of size 5000 before (left) and after (right) compaction. The black fringe subtrees are removed by the compaction; the resulting compacted tree is of size 663.

5.2 Number of non-isomorphic subtree-shapes in increasing binary trees

The exponential generating function $T(z)$ of plane increasing binary trees (*cf.* page 25) is given by

$$T(z) = \frac{z}{1-z},$$

and has its unique dominant singularity at $\rho = 1$.

The exponential generating function $S_t(z)$ of the perturbed class of plane increasing binary trees that do not contain the tree-shape t (where t is now a non-labeled binary tree) as a fringe subtree-shape, satisfies the equation

$$S_t'(z) = (1 + S_t(z))^2 - P_t'(z), \qquad \text{with } S_t(0) = 0, \tag{5.14}$$

where $P_t(z) = \frac{\ell(t)}{|t|!} z^{|t|}$ and $\ell(t)$ denotes the number of ways to increasingly label a plane binary tree t. Since we are in the plane setting now $\ell(t)$ can be calculated by means of the hook-length formula (see for example [70, p.67] or [15]).

We start by establishing an improved upper bound for the weights $w(t)$.

Lemma 5.13. *Let t be a binary tree of size k. By defining the weight of the tree t as $w(t) := \frac{\ell(t)}{k!}$, where $\ell(t)$ denotes the number of ways to increasingly label the tree t, we have*

$$w(t) \leq \frac{1}{2^{k-2}}.$$

Proof. Recall that the hook length equals $|t|!$ devided by the product of the sizes of all fringe subtrees s of t. If we write $s \leq t$ to say that s is a fringe subtree of t, then this means that $w(t) = 1/\prod_{s\,:\,s\leq t} |s|$. Consider now a tree t. If $k = 1$, then t is a single node and hence $w(t) = 1$. Otherwise, the root of t has children being roots of fringe subtrees. If $s \leq t$, the neither $s = t$ and so $|s| = k$ or s is one of the fringe subtrees of one of the subtrees rooted at a child of the root of t. Therefore

$$w(t) = \begin{cases} \frac{1}{k} w(t') & \text{if the root of } t \text{ has one child } t' \\ \frac{1}{k} w(t_\ell) w(t_r) & \text{if the root of } t \text{ has the two children } t_\ell \text{ and } t_r. \end{cases}$$

Now we proceed by induction: Set $w_n := max_{t\,:\,|t|=n} w(t)$. Then we have obviously that $w_n = \max\{w_\ell \cdot w_{n-1-\ell} \mid \ell = 0..n-1\}/n$ with $w_0 = 1$. For the first seven values a direct computation shows

$$(w_1, w_2, \ldots, w_7) = \left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{8}, \frac{1}{15}, \frac{1}{36}, \text{and } \frac{1}{63}\right).$$

As the first seven values of the sequence $1/2^{k-2}$ are

$$2, 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \text{and } \frac{1}{32},$$

we assume thta the result is correct until $k-1$.

Let t be a binary tree of size k. If the root of t has only one child t' of size $k-1$, then by induction we obtain

$$w(t) = \frac{w(t')}{k} \leq \frac{1}{k\ 2^{k-3}} \leq \frac{1}{2^{k-2}}.$$

Otherwise, the root of t has two children. Let us denote the corresponding fringe subtrees by t_ℓ of size ℓ and t_r of size $k-\ell-1$, (with $\ell < k$). By the induction hypothesis, we have $w(t_\ell) \leq 1/2^{\ell-2}$ and $w(t_r) \leq 1/2^{k-\ell-3}$ and thus

$$w(t) = \frac{1}{k} w(t_\ell) w(t_r) \leq \frac{1}{k}\ \frac{1}{2^{k-5}} = \frac{8}{k}\ \frac{1}{2^{k-2}},$$

which is smaller than $1/2^{k-2}$ for $k \geq 8$. □

By the same combinatorial argument as in the previous section we know that $S_t(z)$ has a unique dominant singularity $\tilde{\rho}$, which is greater than the dominant singularity $\rho = 1$ of $T(z)$. Thus, we set again $\tilde{\rho} = \rho(1+\epsilon) = 1+\epsilon$ with $\epsilon > 0$.

Since (5.14) is a Riccati-differential equation (*cf.* [68] for a background on Riccati equations), we use the ansatz $S_t(z) = \frac{-u'(z)}{u(z)}$ to get the transformed equation

$$u''(z) - 2u'(z) + (1 - w(t)kz^{k-1})u(z) = 0, \tag{5.15}$$

where we use the same abbreviations as in the previous section, namely $k := |t|$ and $w(t) := \frac{\ell(t)}{k!}$. Note that the condition $S_t(0) = 0$ implies $u'(0) = 0$ and $u(0) \neq 0$. The singularities of a function $u(z)$ solving a linear differential equation (with polynomial coefficients) are given by the singularities of the coefficient of the highest derivative, *i.e.*, in our case the coefficient of $u''(z)$, which is 1. We refer the reader to Miller [86]

for more details. Thus, we can conclude that $u(z)$ is an entire function. As a direct consequence we know that the singularities of $S_t(z)$ are given by the zeros of $u(z)$ (that are no zeros of $u'(z)$) and are therefore poles. More precisely, the dominant singularity $\tilde{\rho}$ has to be a simple pole for $S_t(z)$, since for $u(z) = (\tilde{\rho} - z)^{\ell} v(z)$ (such that ρ is not a zero of $v(z)$), it follows that $u'(z) = -(\tilde{\rho} - z)^{\ell-1} v(z) + (\tilde{\rho} - z)^{\ell} v'(z)$, which implies

$$S_t(z) = \frac{\ell}{\tilde{\rho} - z} - \frac{v'(z)}{v(z)}. \tag{5.16}$$

Thus,

$$S_t(z) \sim \frac{C}{1 - \frac{z}{\tilde{\rho}}}, \qquad \text{for } z \to \tilde{\rho}.$$

Taking the derivative, we get $S_t'(z) \sim \frac{1}{\tilde{\rho}} \frac{C}{\left(1 - \frac{z}{\tilde{\rho}}\right)^2}$. Plugging the asymptotic expressions for S_t and S_t' into (5.14) gives

$$\frac{1}{\tilde{\rho}} \frac{C}{\left(1 - \frac{z}{\tilde{\rho}}\right)^2} \sim \left(1 + \frac{C}{1 - \frac{z}{\tilde{\rho}}}\right)^2, \qquad \text{for } z \to \tilde{\rho},$$

since the monomial P_t is analytic in $\tilde{\rho}$ (and thus does not contribute to the asymptotics). Comparing the main coefficients yields $C = \frac{1}{\tilde{\rho}}$, and thus

$$S_t(z) \sim \frac{1}{\tilde{\rho} - z}, \qquad \text{for } z \to \tilde{\rho}. \tag{5.17}$$

Remark 5.14. *Note that this directly implies that $\tilde{\rho}$ is a unique zero of the function $u(z)$, when comparing equations* (5.16) *and* (5.17).

With X_n denoting the number of non-isomorphic subtree-shapes of a random plane increasing binary tree of size n, and $\mathcal{B}_{\leq n}$ denoting the class of binary trees with size at most n, we can use again the expression

$$\mathbb{E}(X_n) = \sum_{t \in \mathcal{B}_{\leq n}} (1 - \mathbb{E}(t \notin \tau)) = \sum_{t \in \mathcal{B}_{\leq n}} \left(1 - \frac{[z^n] S_t(z)}{[z^n] T(z)}\right). \tag{5.18}$$

in order to show the following result concerning the asymptotic mean of X_n.

Theorem 5.15. *Let X_n be the number of non-isomorphic subtree-shapes of a random plane increasing tree of size n. Then there exist constants C_1 and C_2 such that*

$$C_1 \sqrt{n} \leq \mathbb{E}(X_n) \leq C_2 \frac{n}{\log n}, \qquad \textit{for } n \to \infty.$$

In order to prove this theorem, we proceed similarly to the recursive tree case: Lemma 5.17 gives an asymptotic expression of the dominant singularity $\tilde{\rho}$ of the generating function $S_t(z)$ that quantifies its dependence on t, when the size k of the "forbidden" tree tends to infinity. As a next step, Lemma 5.18 shows that $S_t(z)$ has a unique dominant singularity $\tilde{\rho}$ on the circle of convergence, which is used in Lemma 5.19 to obtain the asymptotic behavior of the coefficients of the generating

function $S_t(z)$. Again, the average size of a compacted tree can be represented as a sum over the forbidden trees (see (5.18)), where we distinguish between the two cases whether the size of the trees is smaller or larger than $\log_4 n$ in order to get an upper bound (see Proposition 5.10 and Proposition 5.22). Finally, a (crude) lower bound for the size of the compacted tree is given in Proposition 5.24, which uses a better estimate for the weights $w(t)$ (see Lemma 5.13) in order to provide suitable expansions for the summands (see Lemma 5.23).

Starting from equation $u''(z) - 2\,u'(z) + (1 - w(t)kz^{k-1})\,u(z) = 0$, which was derived in (5.15), with the initial conditions $u(0) = \gamma$, and $u'(0) = 0$, we can choose $u(0)$ to attain any non-zero number, since for $u(0) = \gamma$ we have $S_t(z) = \gamma u'(z)/(\gamma u(z))$, and thus, the γ cancels. For simplification reasons we choose $u(0) = -1$.

Lemma 5.16. *The function $u(z)$ defined by the differential equation* (5.15) *with the initial conditions $u(0) = -1$, and $u'(0) = 0$ satisfies*

$$u(z) = ze^z \sum_{m\geq 0} \left(\frac{w(t)k}{(k+1)^2}\right)^m \frac{1}{m!\,(m+\alpha)_m} z^{(k+1)m} - e^z \sum_{m\geq 0} \left(\frac{w(t)k}{(k+1)^2}\right)^m \frac{1}{m!\,(m-\alpha)_m} z^{(k+1)m},$$

where $(x)_m$ denotes the falling factorials $(x)_m = x(x-1)\dots(x-m+1)$ and $\alpha = 1/(k+1)$.

Proof. Solving equation (5.15) gives

$$u(z) = e^z\sqrt{z}\left(\tilde{C}_1 J_{\frac{1}{k+1}}(\beta) + \tilde{C}_2 Y_{\frac{1}{1+k}}(\beta)\right),$$

for constants $\tilde{C}_1, \tilde{C}_2 > 0$, where $\beta = \frac{2\sqrt{-w(t)k}z^{\frac{k+1}{2}}}{k+1}$, and $J_\alpha(x)$ and $Y_\alpha(x)$ denote the Bessel functions that are defined via

$$J_\alpha(x) = \sum_{m=0}^{\infty} \frac{(-1)^m}{m!\Gamma(m+\alpha+1)} \left(\frac{x}{2}\right)^{2m+\alpha},$$

and

$$Y_\alpha(x) = \frac{J_\alpha(x)\cos(\alpha\pi) - J_{-\alpha}(x)}{\sin(\alpha\pi)}.$$

Thus, we can write $u(z)$ as

$$u(z) = C_1 f(z) + C_2 h(z), \tag{5.19}$$

where C_1 and C_2 are constants that depend on k, and $f(z)$ and $h(z)$ are given by

$$f(z) = \sqrt{z}e^z J_\alpha\left(2Xz^{\frac{1}{2\alpha}}\right),$$

and

$$h(z) = \sqrt{z}e^z J_{-\alpha}\left(2Xz^{\frac{1}{2\alpha}}\right),$$

with $X := \frac{\sqrt{-w(t)k}}{k+1}$ and $\alpha := \frac{1}{k+1}$. Note that f and h are analytic functions that can be expanded around 0 as

$$f(z) = X^{\alpha}\frac{1}{\Gamma(1+\alpha)}z + \dots,$$

and

$$h(z) = X^{-\alpha}\frac{1}{\Gamma(1-\alpha)} + X^{-\alpha}\frac{1}{\Gamma(1-\alpha)}z + \dots.$$

Hence, we get

$$u(0) = -1 = C_2 h(0) = C_2 X^{-\alpha}\frac{1}{\Gamma(1-\alpha)}, \tag{5.20}$$

and

$$u'(0) = 0 = \frac{C_1 X^{\alpha}}{\Gamma(1+\alpha)} + \frac{C_2 X^{-\alpha}}{\Gamma(1-\alpha)}, \tag{5.21}$$

which gives

$$C_1 = X^{-\alpha}\Gamma(1+\alpha),$$

and

$$C_2 = -X^{\alpha}\Gamma(1-\alpha).$$

Plugging the expressions for C_1 and C_2 into (5.19) and using $\frac{\Gamma(1+\alpha)}{\Gamma(m+1+\alpha)} = \frac{1}{(m+\alpha)_m}$, where $(x)_m$ denotes the falling factorials $(x)_m = x(x-1)\dots(x-m+1)$, gives

$$\begin{aligned} u(z) = ze^z \sum_{m\geq 0}\left(\frac{w(t)k}{(k+1)^2}\right)^m \frac{1}{m!(m+\alpha)_m} z^{(k+1)m} \\ - e^z \sum_{m\geq 0}\left(\frac{w(t)k}{(k+1)^2}\right)^m \frac{1}{m!(m-\alpha)_m} z^{(k+1)m}. \end{aligned}$$

□

We are now ready to analyze the dominant singularity of $S_t(z)$.

Lemma 5.17. *Let $S_t(z)$ be the generating function of the perturbed combinatorial class of plane increasing binary trees that do not contain the shape t as a subtree. With $\tilde{\rho}$ denoting the dominant singularity of $S_t(z)$, we get*

$$\tilde{\rho} = 1 + \epsilon \sim 1 + \frac{2w(t)}{k^2}, \qquad \text{for } k \to \infty,$$

where $w(t) = \frac{\ell(t)}{k!}$ and $\ell(t)$ denotes the number of ways to increasingly label the tree-shape t.

Proof. For combinatorial reasons we deduced that the equation $u(z) = 0$ must have a solution $\tilde{\rho} > 1$ and no smaller positive solution. When k tends to infinity we expect that $\tilde{\rho} = 1 + \epsilon$ tends to 1, *i.e.* ϵ tends to 0.

First observe that $u(0) = -1$ and

$$u\left(1 + \frac{1}{k^2}\right) = \frac{1}{k^2} + \mathcal{O}\left(\frac{w(t)}{k}\right) > 0,$$

as $w(t)$ decays exponentially due to Lemma 5.13. Thus $\epsilon = \mathcal{O}\left(1/k^2\right)$ and plugging $z = 1 + \epsilon$ into $u(z) = 0$ gives then

$$\epsilon + (1+\epsilon)^{k+1} \frac{w(t)k}{(k+1)^2} \left(\frac{1+\epsilon}{1+\alpha} - \frac{1}{1-\alpha}\right) = \mathcal{O}\left(\frac{w(t)^2}{k^2}\right).$$

This implies $\epsilon - 2\frac{w(t)}{k^2} = \mathcal{O}\left(\frac{w(t)^2}{k^2}\right)$ and hence $\epsilon \sim 2\frac{w(t)}{k^2}$, which finishes the proof. □

So, Lemma 5.17 guarantees that for $|t| = k \to \infty$ the generating function $S_t(z)$ has a dominant singularity at $\tilde{\rho} \sim 1 + \frac{2w(t)}{k^2}$. Now we show that in a circle with radius smaller than $1 + \frac{2\log k}{k}$ there is no other singularity of $S_t(z)$.

Lemma 5.18. *Let $\tilde{\rho}$ be the dominant singularity of $S_t(z)$. Then for $\tilde{\rho} < |z| < 1 + \frac{(2-\delta)\log k}{k}$ the generating function $S_t(z)$ does not have any singularities.*

Proof. First let us remember that the singularities of $S_t(z)$ are given by the zeros of the function $u(z)$ that is defined in Lemma 5.16. Now let us write $\tilde{u}(z) := u(z)\exp(-z)$ and note that $u(z)$ and $\tilde{u}(z)$ have the same zeros. Thus, in the remainder of this proof we investigate $\tilde{u}(z)$, which can be written as $\tilde{u}(z) = zF(z) - G(z)$ with

$$F(z) = \sum_{m\geq 0} \left(\frac{w(t)k}{(k+1)^2}\right)^m \frac{1}{m!} \frac{1}{(m+\alpha)_m} z^{(k+1)m},$$

and

$$G(z) = \sum_{m\geq 0} \left(\frac{w(t)k}{(k+1)^2}\right)^m \frac{1}{m!} \frac{1}{(m-\alpha)_m} z^{(k+1)m},$$

still with $\alpha := 1/(k+1)$. Therefore we get

$$\begin{aligned} |F(z) - G(z)| &= \left| \sum_{m\geq 0} \left(\frac{w(t)k}{(k+1)^2}\right)^m \frac{1}{m!} \left(\frac{1}{(m-\alpha)_m} - \frac{1}{(m+\alpha)_m}\right) z^{(k+1)m} \right| \\ &= \mathcal{O}\left(\frac{w(t)}{k}\ \alpha\ |z|^{k+1}\right) = \mathcal{O}\left(\frac{w(t)}{k^2}\right) |z|^{k+1}. \end{aligned}$$

Now, let us rewrite $\tilde{u}(z)$ as

$$\tilde{u}(z) = (z-1)F(z) + F(z) - G(z), \tag{5.22}$$

set $|z| = 1 + \eta$ and perform a distinction of two cases:

- $\eta = \mathcal{O}\left(\frac{1}{k}\right)$: This implies $|z|^{k+1} = \Theta(1)$ for k tending to infinity. Thus $F(z) \sim 1$, $G(z) \sim 1$, and then $F(z)-G(z)$ tends to 0 when k tends to infinity. Furthermore, equation (5.22) implies $u(z) \sim z - 1$. The equation $\tilde{u}(z) = 0$ therefore yields $z - 1 \sim F(z) - G(z)$, which is $\mathcal{O}\left(\frac{w(t)}{k^2}\right)$. Since we know that $\tilde{\rho} \sim 1 + \frac{2w(t)}{k^2}$ we get $|z - 1| = \Theta(\tilde{\rho} - 1)$.

  But for zeros z_0 of $\tilde{u}(z)$ with $|z_0| = 1 + o\left(\frac{1}{k}\right)$ we know $z_0 - 1 \sim \frac{2w(t)}{k^2} z_0^k \sim \frac{2w(t)}{k^2}$, which is equivalent to $z_0^k \sim 1$. Hence $z_0 \sim \sqrt[k]{1} = \cos\left(\frac{2\pi}{k}\right) + i \sin\left(\frac{2\pi}{k}\right)$ and

$$\tilde{\rho}\sqrt[k]{1} \sim \left(1 + \frac{2w(t)}{k^2}\right)\left(1 - \frac{2\pi^2}{k^2} + i\frac{2\pi}{k}\right) \sim 1 + i\frac{2\pi}{k},$$

  which is a contradiction to $z_0 - 1 \sim \frac{2w(t)}{k^2}$. Thus, the function $\tilde{u}(z)$ has no zeros for $\tilde{\rho} < |z| \leq 1 + \mathcal{O}\left(\frac{1}{k}\right)$.

- $\eta = \frac{C_k}{k}$, with $C_k \leq (2 - \delta) \ln k$, and C_k tends to infinity with k: In this case we have $|z|^{k+1} \sim e^{C_k} = o(k^2)$, and thus $|F(z) - G(z)| = o(w(t))$ and $F \sim 1 + o\left(w(t)k\right) \sim 1$ when k tends to infinity. Using again equation (5.22) yields $\tilde{u}(z) = z-1+o(w(t)) \sim z-1$. Since $|z| = 1+\eta$ we have $|z-1| \geq \frac{c_k}{k}$ and because of $o(w(t)) = o\left(\frac{1}{k}\right)$ we know that $\tilde{u}(z)$ cannot be zero in $\tilde{\rho} < |z| < 1 + \frac{(2-\delta)\ln k}{k}$.

□

Now we are interested in the ratio $[z^n]S_t(z)/[z^n]T(z)$, which corresponds to the probability that a random plane increasing binary tree of size n does not contain the binary tree shape t as a fringe subtree-shape.

Lemma 5.19. *Let $T(z)$ be the generating function of plane increasing trees and $S_t(z)$ the generating function of the perturbed class that has the dominant singularity $\tilde{\rho}$. Then, for any $\eta > 0$ we have*

$$\frac{[z^n]S_t(z)}{[z^n]T(z)} \underset{n\to\infty}{=} \tilde{\rho}^{-n-1}\left(1 + \mathcal{O}\left(\frac{\ln n}{n^{1-\eta}}\right)\right),$$

uniformly for $D \leq k \leq n$, if D is sufficiently large (but independent of n).

Proof. First, let us remember that $\tilde{\rho}$ is a unique zero of the function $u(z)$. Thus, we can write

$$u(z) = \left(1 - \frac{z}{\tilde{\rho}}\right) v(z), \tag{5.23}$$

with $v(\tilde{\rho}) \neq 0$ and by Lemma 5.18 we additionally know that $v(z) \neq 0$ in $\tilde{\rho} < |z| < 1 + \frac{(2-\delta)\ln k}{k}$, provided that k is sufficiently large. Furthermore, we have

$$u'(z) = \left(1 - \frac{z}{\tilde{\rho}}\right) v'(z) - \frac{1}{\tilde{\rho}}\, v(z),$$

which yields

$$S_t(z) = \frac{1}{\tilde{\rho} - z} - \frac{v'(z)}{v(z)}.$$

Thus,

$$[z^n]S_t(z) = \tilde{\rho}^{-n-1} - [z^n]\frac{v'(z)}{v(z)} = \tilde{\rho}^{-n-1} - (n+1)[z^{n+1}]\ln v(z). \tag{5.24}$$

Now, we estimate the second summand in (5.24). First we use a Cauchy coefficient integral to write

$$n[z^n]\ln v(z) = \frac{n}{2\pi i}\int_{\mathcal{C}} \frac{\ln v(t)}{t^{n+1}}\,\mathrm{d}t, \tag{5.25}$$

where the curve $\mathcal{C}$ is described by $|t| = 1 + \frac{(2-\delta)\ln k}{k}$ with some $\delta > 0$. The absolute value of the logarithm of $v(z)$ is given by $\left|\ln v(z)\right| = \left|\ln\left(|v(z)|e^{i\arg v(z)}\right)\right| = |\ln|v(z)| + i\arg(v(z))|$. Furthermore, by (5.23) we have $|v(z)| = |u(z)|/\left|1 - z/\tilde{\rho}\right|$, which can be estimated along $\mathcal{C}$ via

$$|v(z)| \le \frac{|u(z)|k}{(2-\delta)\ln k}.$$

Now, we have to estimate $|u(z)|$. By Lemma 5.16 we get

$$|u(z)| \le \sum_{m\ge 0}\left(\frac{w(t)}{k}\right)^m \frac{1}{m!}\left|\frac{z}{(m+\alpha)_m} - \frac{1}{(m-\alpha)_m}\right||z|^{(k+1)m}.$$

Along $\mathcal{C}$ we have $|z|^{(k+1)m} \le (k^{2-\delta})^m$ and the absolute value $\left|\frac{z}{(m+\alpha)_m} - \frac{1}{(m-\alpha)_m}\right|$ can be estimated by $\left|\frac{z}{(m+\alpha)_m} - \frac{1}{(m-\alpha)_m}\right| \le \frac{2+\mu}{(m-\alpha)_m}$, for some $\mu > 0$ which results in

$$|u(z)| \le \sum_{m\ge 0}(w(t)k^{1-\delta})^m \frac{2+\mu}{m!(m-\alpha)_m} \le K,$$

for a constant K independent of k.

Putting all together, we can estimate the integral (5.25) by

$$\begin{aligned} n[z^n]\ln v(z) = \frac{n}{2\pi i}\int_{\mathcal{C}} \frac{\ln v(t)}{t^{n+1}}\,\mathrm{d}t &\le n(\ln k + \ln K - \ln((2-\delta)\ln k)\left(1 + \frac{(2-\delta)\ln k}{k}\right)^{-n-1} \\ &\le n\ln n\left(1 + \frac{(2-\delta)\ln k}{k}\right)^{-n} \end{aligned}$$

which implies the following asymptotic relation:

$$[z^n]S_t(z) = \tilde{\rho}^{-n-1}\left(1 + \mathcal{O}\left(n\ln n\left(1 + \frac{(2-\delta)\ln k}{k}\right)^{-n}\tilde{\rho}^n\right)\right)$$

Finally, note that for sufficiently large k we have the estimate

$$\begin{aligned} \tilde{\rho}\left(1 + \frac{(2-\delta)\ln k}{k}\right)^{-1} &\le \left(1 + \frac{(2-2\delta)\ln k}{k}\right)^{-1} \\ &\le \left(1 + \frac{(2-2\delta)\ln n}{n}\right)^{-1} \end{aligned}$$

and, as

$$\left(1 + \frac{(2-2\delta)\ln n}{n}\right)^{-n} = \mathcal{O}\left(n^{-2+2\delta}\right),$$

we obtain the assertion by setting $\eta = 2\delta$. □

Now, we separate the sum of interest (5.18) analogously as we did in the previous section for recursive trees.

Remark 5.20. *Since now our underlying class of tree-shapes is the class of plane binary trees instead of Pólya trees, we subsequently use* $\log n$ *as an abbreviation for the logarithm with the base* $\frac{1}{1/4} = 4$.

$$\mathbb{E}\left(X_n\right) = \sum_{\substack{t\in\mathcal{B}_{\leq n}\\ k<\log n}} \left(1 - \frac{[z^n]S_t(z)}{[z^n]T(z)}\right) + \sum_{\substack{t\in\mathcal{B}_{\leq n}\\ k\geq\log n}} \left(1 - \frac{[z^n]S_t(z)}{[z^n]T(z)}\right) \tag{5.26}$$

In order to estimate the first sum, we proceed analogously to Lemma 5.10.

Proposition 5.21. *Let* $\mathcal{B}_{\leq n}$ *be the class of plane binary trees of size at most* n*, and let* $T(z)$ *be the generating function of plane in creasing binary trees, and* $S_t(t)$ *the generating function of the perturbed class of plane increasing binary trees that do not contain the unlabeled binary tree* t *as a fringe subtree-shape. Then we have*

$$\sum_{\substack{t\in\mathcal{B}_{\leq n}\\ k<\log n}} \left(1 - \frac{[z^n]S_t(z)}{[z^n]T(z)}\right) = \mathcal{O}\left(\frac{n}{\sqrt{(\log n)^3}}\right), \qquad \text{as } n\to\infty,$$

where the logarithm $\log n$ *is to the base 4.*

Proof. First let us recall that the dominant singularity of the generating function $B(z)$ of plane binary trees is $\frac{1}{4}$. With the notation $k := |t|$, we get

$$\begin{aligned}\sum_{\substack{t\in\mathcal{B}_{\leq n}\\ k<\log n}} \left(1 - \frac{[z^n]S_t(z)}{[z^n]T(z)}\right) \leq \sum_{\substack{t\in\mathcal{B}_{\leq n}\\ k<\log n}} 1 &= \sum_{k<\log n} [z^k]B(z)\\ &\sim \frac{1}{1-\frac{1}{4}}[z^{\lfloor\log n\rfloor}]B(z) = \mathcal{O}\left(\frac{\left(\frac{1}{4}\right)^{-\lfloor\log n\rfloor}}{\sqrt{(\log n)^3}}\right).\end{aligned}$$

Since $\log n$ has the base 4, we estimate $\left(\frac{1}{4}\right)^{-\lfloor\log n\rfloor} \leq n$, which completes the proof. □

Estimating the second sum in (5.26) works analogously to the proof of Theorem 5.11 in the previous section.

Proposition 5.22. *Let* $\mathcal{B}_{\leq n}$ *be the class of plane binary trees of size at most* n*, and let* $T(z)$ *be the generating function of plane in creasing binary trees, and* $S_t(t)$ *the generating function of the perturbed class of plane increasing binary trees that do not contain the unlabeled binary tree* t *as a fringe subtree-shape. Then we have*

$$\sum_{\substack{t\in\mathcal{B}_{\leq n}\\ k\geq\log n}} \left(1 - \frac{[z^n]S_t(z)}{[z^n]T(z)}\right) = \mathcal{O}\left(\frac{n}{\log n}\right),$$

where the logarithm $\log n$ *is to the base 4.*

Proof. Using Lemma 5.19 we get that for $n \to \infty$

$$\frac{[z^n]S_t(z)}{[z^n]T(z)} \sim \tilde{\rho}^{-n-1} = (1+\epsilon)^{-n-1}.$$

Thus,

$$\sum_{\substack{t\in\mathcal{B}_{\leq n}\\ k\geq\log n}} \left(1 - \frac{[z^n]S_t(z)}{[z^n]T(z)}\right) \sim \sum_{\substack{t\in\mathcal{B}_{\leq n}\\ k\geq\log n}} \left(1-(1+\epsilon)^{-n-1}\right) \qquad \text{for } n\to\infty.$$

Using the Bernoulli inequality gives

$$\sum_{\substack{t\in\mathcal{B}_{\leq n}\\ k\geq\log n}} 1-(1+\epsilon)^{-n-1} \leq \sum_{\substack{t\in\mathcal{B}_{\leq n}\\ k\geq\log n}} (n+1)\cdot\epsilon,$$

which by the use of Lemma 5.17 further simplifies to

$$\sum_{\substack{t\in\mathcal{B}_{\leq n}\\ k\geq\log n}} (n+1)\cdot\epsilon \sim \sum_{k=\log n}^{n} \sum_{\substack{t\in\mathcal{B}_{\leq n}\\ |t|=k}} (n+1)\cdot\frac{2w(t)}{k^2} = \sum_{k=\log n}^{n} \frac{2n}{k^2} \sum_{\substack{t\in\mathcal{B}_{\leq n}\\ |t|=k}} w(t).$$

But since

$$\sum_{\substack{t\in\mathcal{B}_{\leq n}\\ |t|=k}} w(t) = 1,$$

we finally get

$$\sum_{\substack{t\in\mathcal{B}_{\leq n}\\ k\geq\log n}} (n+1)\cdot\epsilon = \sum_{k=\log n}^{n} \frac{2n}{k^2} = \Theta\left(n\int_{\log n}^{\infty}\frac{1}{x^2}\,\mathrm{d}x\right) = \Theta\left(\frac{n}{\log n}\right). \qquad \square$$

Finally, we provide a crude lower bound for the number of non-isomorphic subtree-shapes in a random increasing binary tree. Using the improved upper bound for $w(t)$ that we obtained in Lemma 5.13, we prove the following lemma.

Lemma 5.23. *Let ϵ be defined as in Lemma 5.17, i.e., $\epsilon \sim \frac{2w(t)}{k^2}$. Then*

$$(1+\epsilon)^{-n} \sim e^{-n\epsilon}$$

holds for $n\to\infty$ and for $k \geq n$.

Proof. First of all, let us consider the expansion

$$(1+\epsilon)^{-n} = \exp(-n\log(1+\epsilon)) = e^{-n\epsilon+n\frac{\epsilon^2}{2}\mp\cdots}. \tag{5.27}$$

By Lemma 5.17 we know that $\epsilon \sim \frac{2w(t)}{k^2}$. Using Lemma 5.13 we have $\epsilon^2 \leq \frac{1}{k^4 4^{k-1}}$. Furthermore, $k \geq \log n = \log_4 n$ implies $4^k \geq n$, which gives

$$\frac{1}{k^4 4^{k-1}} \leq \frac{4}{k^4 n} = o\left(\frac{1}{n}\right)$$

for $k \geq \log n$ and $n\to\infty$. Finally, it follows that

$$\frac{n\epsilon^2}{2} = o(1),$$

which completes the proof when considering Equation (5.27). $\square$

Proposition 5.24. *Let $\mathcal{B}_{\leq n}$ be the class of plane binary trees of size at most n, and let $T(z)$ be the generating function of plane in creasing binary trees, and $S_t(t)$ the generating function of the perturbed class of plane increasing binary trees that do not contain the unlabeled binary tree t as a fringe subtree-shape. Then we have*

$$\sum_{\substack{t\in\mathcal{B}_{\leq n}\\ k\geq\log n}}\left(1-\frac{[z^n]S_t(z)}{[z^n]T(z)}\right)=\Omega\left(\sqrt{n}\right),$$

where the logarithm $\log n$ is to the base 4.

Proof. First we use Lemma 5.23 to get

$$\begin{aligned}\sum_{\substack{t\in\mathcal{B}_{\leq n}\\ k\geq\log n}}\left(1-\frac{[z^n]S_t(z)}{[z^n]T(z)}\right)&=\sum_{\substack{t\in\mathcal{B}_{\leq n}\\ k\geq\log n}}\left(1-(1+\epsilon)^{-n}\right)\\ &\sim\sum_{\substack{t\in\mathcal{B}_{\leq n}\\ k\geq\log n}}\left(1-e^{-n\epsilon}\right)=\sum_{k=\log n}^{n}\sum_{\substack{t\in\mathcal{B}_{\leq n}\\ |t|=k}}\left(1-e^{-n\epsilon}\right)\end{aligned}\tag{5.28}$$

For the sake of simplified reading we will use again the abbreviation

$$\sum_t:=\sum_{\substack{t\in\mathcal{B}_{\leq n}\\ |t|=k}}$$

in the remainder of this proof. Since $1-e^{-x}$ is a concave and nonnegative function for $x\geq 0$ and zero for $x=0$, we can estimate the inner sum in (5.28), which yields

$$\sum_t\left(1-e^{-n\epsilon}\right)\geq 1-e^{-n\sum_t\epsilon}\sim 1-e^{-n\sum_t\frac{2w(t)}{k^2}},$$

where the asymptotic equivalence holds due to Lemma 5.17. By means of further simplifications and the identity $\sum_t w(t)=1$, we get

$$1-e^{-n\sum_t\frac{2w(t)}{k^2}}=1-e^{\frac{-2n}{k^2}\sum_t w(t)}=1-e^{\frac{-2n}{k^2}}.$$

Finally, we get

$$\begin{aligned}\sum_{\substack{t\in\mathcal{B}_{\leq n}\\ k\geq\log n}}\left(1-\frac{[z^n]S_t(z)}{[z^n]T(z)}\right)&\geq\sum_{k=\log n}^{n}\left(1-e^{\frac{-2n}{k^2}}\right)\\ &\sim\int_{\log n}^{\infty}\left(1-e^{-\frac{2n}{x^2}}\right)\mathrm{d}x=\sqrt{2n}\int_{\frac{\log n}{\sqrt{2n}}}^{\infty}\left(1-e^{-\frac{1}{v^2}}\right)\mathrm{d}v.\end{aligned}$$

Since the integral is convergent this gives a lower bound that is $\Theta(\sqrt{n})$. □

Remark 5.25. *In order to prove Theorem 5.24 one could proceed analogously to the proof of Theorem 5.12 in the previous section. However, we decided to give the proof that uses the better estimate for $w(t)$, since this result will be needed in order to obtain improved bounds.*

On the left hand side, in Figure 5.4 we have depicted a plane increasing binary tree structure containing 5000 nodes, which has been uniformly sampled among all trees of the same size. The original root of the tree is represented using a small circle $\circ$. The right hand side of Figure 5.4 shows the structure that is left after the compaction of the latter tree, consisting of only 1361 nodes.

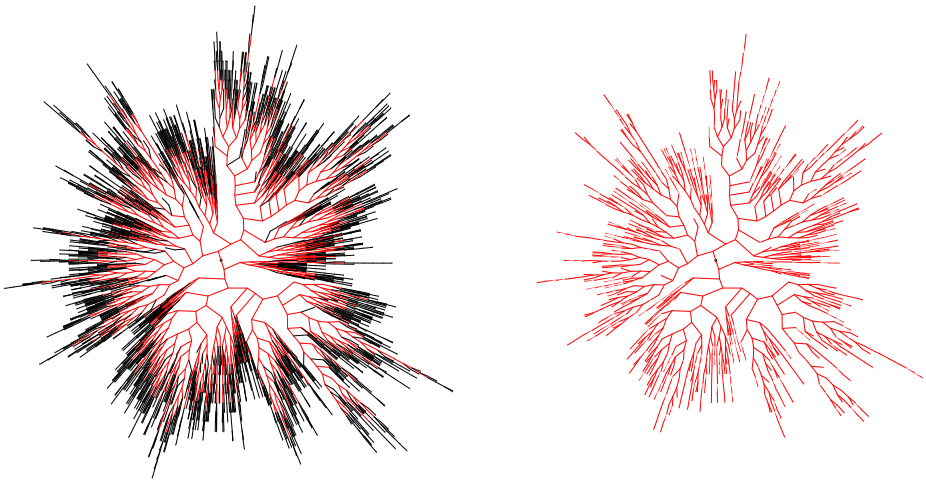


Figure 5.4: A uniformly sampled plane increasing binary tree of size 5000 before (left) and after (right) compaction. The black fringe subtrees are removed by the compaction, so that the resulting compacted tree is of size 1361.

Chapter 6

Tree embeddings

This chapter is based on the not yet submitted manuscript [54], which was joint work with Bernhard Gittenberger, Zbigniew Gołębiewski and Małgorzata Sulkowska, [55]. It is concerned with the enumeration of so-called *tree embeddings* of a given rooted tree into three selected classes of trees, namely plane and non-plane binary trees, and planted plane trees. An embedding of a rooted tree S into another rooted tree T can be seen as a kind of generalized pattern occurrence of S in T, defined as follows, where we distinguish between the plane and the non-plane case.

Definition 6.1 (non-plane embedding)**.** *Let S and T be two non-plane rooted trees. When interpreting T as the cover graph of a partially ordered set (poset), rooted at the root of T, i.e., at the 1-element of the poset, then an embedding of S in T can be defined as any subposet of T isomorphic to S.*

Remark 6.2. *Note that there exists a non-plane embedding of S in T if and only if S is a minor of T.*

Definition 6.3 (plane embedding)**.** *Let S and T be two plane rooted trees. If we interpret T to be a Hasse diagram of a poset, then an embedding of S in T can be defined as any "subposet" of T isomorphic to S in which the order of the children of each node is preserved (thus, a plane version of a subposet).*

Remark 6.4. *So, in the plane case S and T can be interpreted as Hasse diagrams of posets, and whenever S can be embedded in T it follows that S is a subposet of T. However, note that the respective posets can eventually be represented as different Hasse diagrams such that no embedding of the corresponding trees is possible.*

We say that an embedding of S in T is *good* if it contains the root of T, which is subsequently denoted by $\mathbb{1}_T$. Otherwise we call it a *bad embedding*. If there exists at least one embedding of S in T, we write $S \subseteq T$. All embeddings of a cherry (*i.e.*, a tree composed only of a root and its two children) in a given binary tree of size 5 are given in Figure 6.1. Four of them are good and the last one is bad.

Subsequently the size of the tree S will always be denoted by m, while the size of T is consistently denoted by n. Thus, for the asymptotic analysis of the number of embeddings of a tree S into a class of trees of size n, the quantity m is considered to be a constant, while n tends to infinity.

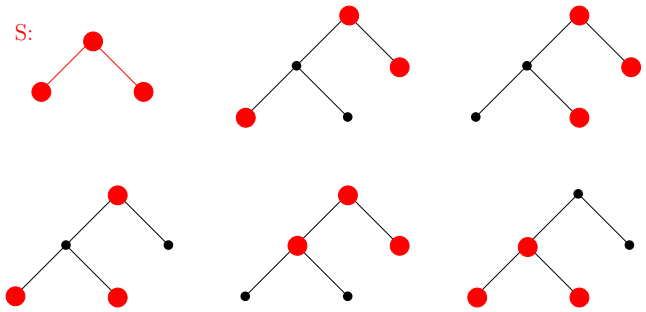


Figure 6.1: All five embeddings of a cherry S in a given plane binary tree of size 5. Or all four embeddings of a cherry S in the a given non-plane binary tree of size 5, since in the non-plane case the two rightmost pictures in the upper row represent the same embedding (they can easily be mapped onto each other via a simple automorphism that changes the order of the two leftmost leaves).

For S, the structure that we embed, we define its *degree distribution sequence* as $d_S = (d_0, d_1, \ldots, d_{m-1})$, where d_i is the number of vertices in S with out-degree equal to i (*i.e.*, the number of vertices in S with i children). Note that d_0 is simply the number of leaves, which will be, interchangeably, denoted by l (*i.e.*, $l = d_0$). Similarly, d_1 is the number of unary nodes, which will be, interchangeably, denoted by u (*i.e.*, $u = d_1$). The number of all embeddings of a structure S in T will be denoted by $a_T(S)$ and the number of its good embeddings in T by $g_T(S)$. Moreover, the number of all embeddings of a structure S in a family $\mathcal{F}_n = \{F_1, \ldots, F_N\}$ will be denoted by $a_{\mathcal{F}_n}(S)$ and understood as $a_{\mathcal{F}_n}(S) = \sum_{i=1}^{N} a_{F_i}(S)$. Analogously, we define the number of good embeddings of S in $\mathcal{F}_n$ by $g_{\mathcal{F}_n}(S) = \sum_{i=1}^{N} g_{F_i}(S)$.

For S being a cherry and $\mathcal{B}_5 = \{T_1, T_2\}$ being the set of plane binary trees of size 5 (see Figure 6.2), we obtain $a_{T_1}(S) = a_{T_2}(S) = 5$, $g_{T_1}(S) = g_{T_2}(S) = 4$, and thus $a_{\mathcal{B}_n}(S) = 10$ and $g_{\mathcal{B}_n}(S) = 8$ (compare Figure 6.1).

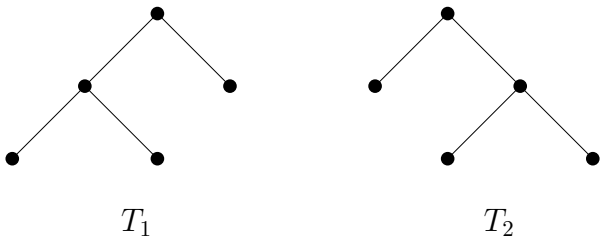


Figure 6.2: The family $\mathcal{B}_5 = \{T_1, T_2\}$ of plane binary trees of size 5, $|\mathcal{B}_5| = \boldsymbol{C}_2 = 2$.

We study the number of embeddings of a given rooted tree in the family of (plane and non-plane) binary trees, as well as planted plane trees. The number of good and bad embeddings of a rooted structure in a complete binary tree was first investigated by Morayne in [88]. His research was motivated by optimal stopping problems. The ratio of the number of good embeddings to the number of all embeddings and its monotonicity properties were used in estimates of conditional probabilities needed to obtain an optimal policy for the best choice problem considered on a complete (balanced) binary tree. This and similar results first served just as tools but soon became interesting questions about the structural features of posets on their own and

resulted in a series of self-standing papers [51, 72, 73]. Counting chains and antichains in trees took a special place in this pool [74, 75, 76].

We present a follow-up and generalization of the results obtained by Kubicki, Lehel and Morayne [72, 73] and Georgiou [51]. We give the asymptotic mean of the number of good and all embeddings of a rooted tree S in the family of plane and non-plane binary trees, as well as planted plane trees, on n vertices. We prove that the ratio of good embeddings to all asymptotically equivalent to $c/\sqrt{n}$ in all cases and provide the exact constant c. Furthermore, we show that this ratio is non-decreasing in S in the plane binary case and asymptotically non-decreasing in the non-plane binary case. We comment also on the case where S is disconnected, *i.e.*, a forest.

Before investigating the number of embeddings of a given tree into the different classes of trees, we want to give some examples of stopping problems in which either the values $a_{\mathcal{F}_n}(S)$ or $g_{\mathcal{F}_n}(S)/a_{\mathcal{F}_n}(S)$ play a crucial role for estimating the conditional probabilities needed to obtain the optimal policy.

6.1 Applications in optimal stopping problems

The most prominent problem in optimal stopping is maybe the so-called "secretary problem" [25, 41, 47], where one assumes a linear order on the applicants for a secretary position concerning their qualifications. The applicants are interviewed in a random order and the decision whether to hire an applicant has to be made immediately after the interview - a rejected applicant cannot be hired at a later point. Thus, if we interview all the candidates, we have to hire the last applicant. The goal is to find the optimal stopping strategy to find the best applicant. Thus, we want to stop at the time with the highest probability that the present applicant is the best one overall, *i.e.*, the maximum element in the linear order. It has been proved (see for example [52]) that for a large number of applicants it is optimal to wait until approximately 37% (more precisely $\frac{100}{e}\%$) of the applicants have been interviewed and then to select the next relatively best one. This problem has been extended and generalized in many different directions.

In the remainder of this section we give examples of stopping problems in which either the value $a_{\mathcal{V}_n}(S)$ or the ratio $g_{\mathcal{V}_n}(S)/a_{\mathcal{V}_n}(S)$ (both investigated in this paper) plays a crucial role in estimating the conditional probabilities needed to obtain the optimal policy. One can consider analogous examples for the families $\mathcal{B}_n$ or $\mathcal{T}_n$ as well.

Let us think about elements of $\mathcal{V}_n$ as of Hasse diagrams of partially ordered sets (in short, posets). Consider the following process. Elements (*i.e.*, nodes) of some T from $\mathcal{V}_n$ appear one by one in a random order (all permutations of elements of T are equiprobable). At time t, *i.e.*, when t elements have already appeared, the selector can see a poset induced on those elements. He knows that the underlying structure is drawn uniformly at random from $\mathcal{V}_n$.

Example 6.5 (Best choice problem for the family of binary trees)**.** *The selector's task is to stop the process maximizing the probability that the element that has just appeared is the root of the underlying structure. He wins only if the chosen element is indeed the root of T. Note that it neither pays off to stop the process when the induced structure is disconnected nor when the currently observed element is not the maximal one in the induced poset. The selector wonders whether to stop only if the emerged*

element at time t is the unique maximal element in the induced structure. In order to decide whether to stop at time t, he needs to know the probability of winning if he stops now. Let W_t denote the event of winning when stopping at time t, S_t the event that at time t he observes a certain structure S with degree distribution sequence d_S and R_i denote the event that T_i has been drawn as the underlying structure. Then the probability of winning if he stops at time t is given by

$$\mathbb{P}[W_t|S_t] = \sum_{i=1}^{M} \mathbb{P}[W_t|S_t \cap R_i]\mathbb{P}[R_i|S_t] = \sum_{i=1}^{M} \frac{g_{T_i}(S)}{a_{T_i}(S)} \frac{\mathbb{P}[S_t|R_i]\mathbb{P}[R_i]}{\mathbb{P}[S_t]}.$$

Since $\mathbb{P}[R_i] = 1/N$, $\mathbb{P}[S_t|R_i] = a_{T_i}(S)/\binom{n}{t}$ and

$$\mathbb{P}[S_t] = \sum_{i=1}^{M} \mathbb{P}[S_t|R_i]\mathbb{P}[R_i] = \sum_{i=1}^{M} \frac{a_{T_i}(S)}{\binom{n}{t}} \frac{1}{N} = \frac{a_{\mathcal{T}_n}(S)}{N\binom{n}{t}},$$

we get

$$\mathbb{P}[W_t|S_t] = \sum_{i=1}^{M} \frac{g_{T_i}(S)}{a_{T_i}(S)} \frac{a_{T_i}(S)}{\binom{n}{t}} \frac{1}{N} \frac{N\binom{n}{t}}{a_{\mathcal{T}_n}(S)} = \frac{g_{\mathcal{T}_n}(S)}{a_{\mathcal{T}_n}(S)}.$$

Example 6.6 (Identifying complete balanced binary trees)**.** *The selector has to identify whether the underlying structure is a complete balanced binary tree or not. The payoff of the game, if he stops the process at time t, is $n - t$ if he guesses correctly and 0 otherwise. He has to maximize the expected payoff. At moment t he observes a structure S, which is not necessarily connected. Again, in order to make a decision whether to stop, he needs to know what is the probability that the currently observed structure is a subposet of a complete balanced binary tree. For a rooted tree S this probability is given by*

$$\frac{a_{T_b}(S)}{a_{\mathcal{V}_n}(S)},$$

where $T_b \in \mathcal{V}_n$ denotes the complete balanced binary tree of size n. If S is a forest, i.e. not connected, than we have to add a factor $\frac{1}{k}$ where k is the number of trees in S.

6.2 Embeddings in plane binary trees

Let $\mathcal{B}_n$ denote the set of plane binary trees of size n. The goal of this section is to derive the expected value for the number of both good and all embeddings of a given rooted plane tree S in a random tree from the class $\mathcal{B}_n$. Furthermore, we study the ratio $\frac{g_{\mathcal{B}_n}(S)}{a_{\mathcal{B}_n}(S)}$ of the number of good to the number of all embeddings and briefly discuss the case when the embedded structure S is disconnected. We start by setting up generating functions for the sequences $a_{\mathcal{B}_n}(S)$ and $g_{\mathcal{B}_n}(S)$.

Theorem 6.7. *Consider a rooted tree S with a degree distribution sequence $d_S = (l, u, d_2, \ldots, d_{m-1})$. The generating function $A_S(z)$ of the sequence $a_{\mathcal{B}_n}(S)$ (counting the number of all embeddings of S into all trees of the family $\mathcal{B}_n$) is given by*

$$A_S(z) = \left(\frac{1}{1 - 2zB(z)}\right)^{3l+u-2} z^{u+l-1}\ B(z)^{l+u}\ 2^u \prod_{i=3}^{m-1} (\boldsymbol{C}_{i-1})^{d_i},$$

where $B(z)$ is the generating function of the class of plane binary trees with z marking the total number of nodes (see Example 3.5).

Remark 6.8. *Note that $A_S(z)$ depends only on the degree distribution sequence d_S, not the particular shape of S. As long as d_{S_1} and d_{S_2} are the same, $A_{S_1}(z)$ and $A_{S_2}(z)$ coincide even if S_1 and S_2 are not isomorphic. However, we use the subscript S to provide a transparent notation. Moreover, note that $A_S(z)$ does also depend on the tree class $\mathcal{B}_n$ in which we embed the tree S. In order to avoid a large number of indices we will omit to indicate this dependence and just emphasize at this point that the generating function $A_S(z)$ is different in each of the Sections 6.2, 6.3 and 6.4 due to the different underlying tree classes.*

Proof. We start with the case where S is a Motzkin tree (*cf.* 3.7), and thereby distinguish between the three cases whether S is a single node, or it starts with a unary node, or a binary node, respectively. The generating function $A_S(z)$ of the number of embeddings of S in the family $\mathcal{B}_n$ can then be recursively defined by

$$A_S(z) = \begin{cases} zB'(z) & \text{if } S = \{\bullet\} \\ \frac{2zB(z)}{1-2zB(z)} A_{\tilde{S}} & \text{if } S = \{\bullet\} \times \tilde{S} \\ \frac{z}{(1-2zB(z))^2} A_{S_L}(z) A_{S_R}(z) & \text{if } S = \{\bullet\} \times \mathcal{S}_L \times S_R \end{cases} . \tag{6.1}$$

The first case, which yields a factor $zB'(z)$, corresponds to marking a node in the underlying tree T (*i.e.*, pointing at a node, *cf.* Table 2.1), because obviously a single vertex can be embedded in every node. Note that instead of counting the number of possible ways to mark a single vertex, we can also interpret it as counting the number of pairs (T, E) where E is an embedding of S in T.

Now we show how an embedding of S in T can be constructed in a recursive way - see Figure 6.3 for a visualization of the used approach: We start with the case that the root of S is a unary node. This root has to be embedded at some point in the tree T. The part of T that is above the embedded root node of S can be expressed as a path of left-or-right trees, which contributes a factor $\frac{1}{1-2zB(z)}$. The embedded root vertex of S itself yields a factor z, since the generating function of an object of size one is given by z. To the embedded root vertex we have to attach an additional tree T in order to create a binary structure, yielding a factor $B(z)$, as well as the remaining tree that contains the embedding of $\tilde{S}$. The factor 2 that appears in the coefficient in the second case of (6.1) indicates that we work with plane trees - the unary vertex may either become the left or the right child of its parent node.

The third case of (6.1), where S starts with a binary node, is very similar to the previous case. Thus, the factor $\frac{1}{(1-2zB(z))^2}$ corresponds to two consecutive paths of left-or-right trees, which are separated by the embedded root which itself gives the additional factor z. At some point the lower path splits into two subtrees containing the embeddings of the subtrees S_L and S_R.

By simple iteration one can see that in case of embedding a Motzkin tree S, the generating function $A_S(z)$ reads as

$$A_S(z) = \left(\frac{z}{(1-2zB(z))^2}\right)^{l-1} \left(\frac{2zB(z)}{1-2zB(z)}\right)^{u} (zB'(z))^{l}, \tag{6.2}$$

where l denotes the number of leaves and u the number of unary nodes in S. The exponent $l-1$ in (6.2) arises from the fact that a Motzkin tree with l leaves has $l-1$

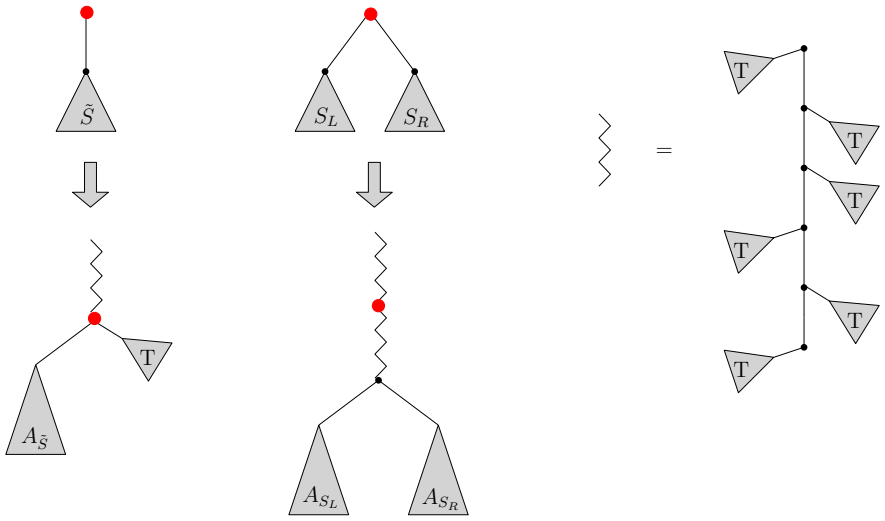


Figure 6.3: Sketch of the recursive construction of the generating function $A_S(z)$, when S is a Motzkin tree consisting of more than one vertex.

binary nodes, and for each of these nodes we get the respective factor.
Finally, we now consider the general case where S is an arbitrary plane tree without any restrictions on the degree distribution sequence. Then we proceed as follows: Every d-ary node with $d \geq 3$ is replaced by a binary tree having d leaves. There are exactly $\boldsymbol{C}_{d-1}$ possible ways to construct such a binary tree. Unary and binary nodes stay unaltered. When doing this for all nodes, the resulting tree is a Motzkin tree, and the number of Motzkin trees that can be constructed in that way is $\prod_{i=3}^{m-1} \boldsymbol{C}_{i-1}^{d_i}$. These Motzkin trees are then embedded with the approach described above, resulting in

$$A_S(z) = \left(\frac{1}{1-2zB(z)}\right)^{2l-2} z^{l-1}(zB'(z))^l \left(\frac{2zB(z)}{1-2zB(z)}\right)^u \cdot \prod_{i=3}^{m-1} (\boldsymbol{C}_{i-1})^{d_i}.$$

Using some basic simplifications and the identity $zB'(z) = \frac{B(z)}{1-2zB(z)}$, which holds for plane binary trees, we get the desired result. See Figure 6.4 for a sketch of the principle of embedding an arbitrary plane tree. □

Corollary 6.9. *Consider a rooted tree S. The generating function $G_S(z)$ of the sequence $g_{\mathcal{B}_n}(S)$ (counting the cumulative number of good embeddings of S into all trees of the family $\mathcal{B}_n$) is given by*

$$G_S(z) = (1-2zB(z))A_S(z).$$

Proof. The corollary follows immediately as the only difference in the case of good embeddings is that the root vertex is always an embedded node and thus, we have to omit the path of left-or-right trees in the beginning of the construction. This corresponds to a multiplication by the factor $(1-2zB(z))$. □

The following theorem provides the asymptotics of $a_{\mathcal{B}_n}(S)$ and $g_{\mathcal{B}_n}(S)$.

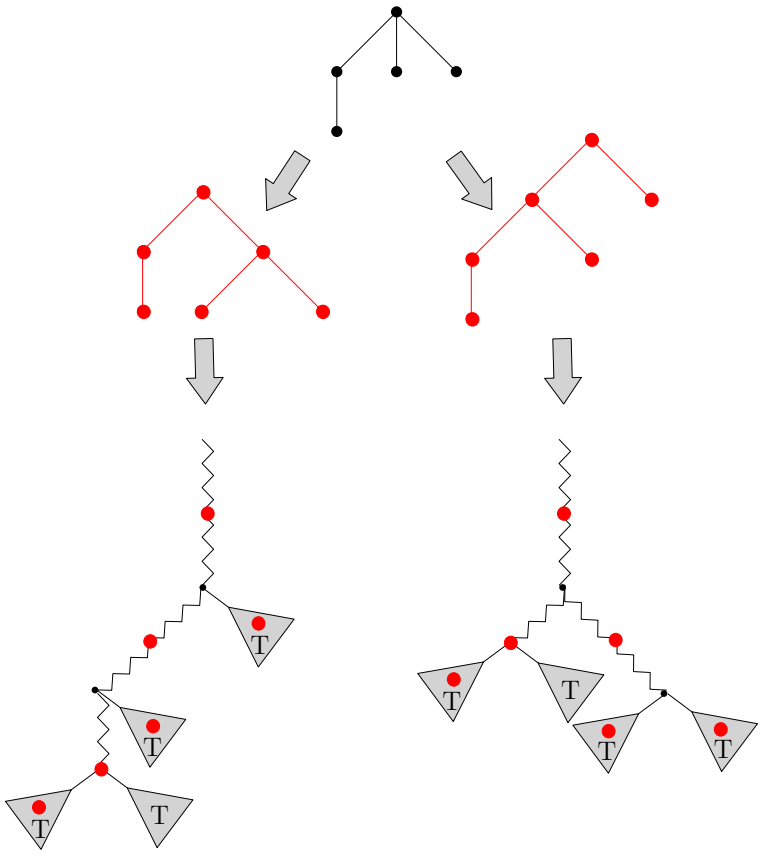


Figure 6.4: Sketch of the principle of embedding an arbitrary plane tree.

Theorem 6.10. *Consider a rooted tree S with degree distribution sequence $d_S = (l, u, d_2, \dots, d_{m-1})$. Let $C = \prod_{i=3}^{m-1} (\boldsymbol{C}_{i-1})^{d_i}$. The asymptotics of the number of all embeddings of S in $\mathcal{B}_n$, is given by*

$$a_{\mathcal{B}_n}(S) \sim \frac{C \cdot 2^{\frac{6-5l-u}{2}}}{\Gamma(\frac{3l+u-2}{2})} \cdot 2^n \cdot n^{\frac{3l+u-4}{2}}, \qquad \text{as } n \to \infty,$$

for n being odd and $a_{\mathcal{B}_n}(S) = 0$ for even n. The number of good embeddings of S in $\mathcal{B}_n$ is asymptotically given by

$$g_{\mathcal{B}_n}(S) \sim \begin{cases} \frac{C \cdot 2^{\frac{7-5l-u}{2}}}{\Gamma(\frac{3l+u-3}{2})} \cdot 2^n \cdot n^{\frac{3l+u-5}{2}} & \text{if } \ 3l+u-3 > 0, \\ \frac{\sqrt{2} \cdot 2^n}{\sqrt{\pi n^3}} & \text{if } \ 3l+u-3 = 0, \end{cases}$$

for n being odd and $g_{\mathcal{B}_n}(S) = 0$ for even n.

Proof. Recall that $a_{\mathcal{B}_n}(S) = [z^n] A_S(z)$. The function $A_S(z)$ has two dominant singularities at $\rho_1 = 1/2$ and $\rho_2 = -1/2$. The Puiseux expansion of $A_S(z)$ for $z \to \rho_1 = 1/2$ reads as

$$A_S(z) = C \cdot 2^{\frac{4-5l-u}{2}} \cdot \left(1 - \frac{z}{\rho_1}\right)^{-\frac{3l+u-2}{2}} \left(1 + O\left(\left(1 - \frac{z}{\rho_1}\right)^{1/2}\right)\right).$$

Note that $3l+u-2 \geq 1$, since always $l \geq 1$ and $u \geq 0$. Expanding $A_S(z)$ in Puiseux series for $z \to \rho_2 = -1/2$ gives

$$A_S(z) = -C \cdot 2^{\frac{4-5l-u}{2}} \cdot \left(1 - \frac{z}{\rho_2}\right)^{-\frac{3l+u-2}{2}} \left(1 + O\left(\left(1 - \frac{z}{\rho_2}\right)^{1/2}\right)\right).$$

Using the transfer theorems (see Theorems 2.8 and 2.10) we get that

$$\begin{aligned}[z^n]A_S(z) &\sim \frac{C \cdot 2^{\frac{4-5l-u}{2}}}{\Gamma(\frac{3l+u-2}{2})} \cdot (\rho_1)^{-n} \cdot n^{\frac{3l+u-4}{2}} - \frac{C \cdot 2^{\frac{4-5l-u}{2}}}{\Gamma(\frac{3l+u-2}{2})} \cdot (\rho_2)^{-n} \cdot n^{\frac{3l+u-4}{2}} \\ &= \begin{cases} \frac{C \cdot 2^{\frac{6-5l-u}{2}}}{\Gamma(\frac{3l+u-2}{2})} \cdot 2^n \cdot n^{\frac{3l+u-4}{2}} & \text{if n is odd,} \\ 0 & \text{if n is even.} \end{cases}\end{aligned}$$

The asymptotic analysis for the number of good embeddings is analogous, since we have again, $g_{\mathcal{B}_n}(S) = [z^n]G_S(z)$ with $G_S(z)$ having two dominant singularities of a square root type at $\rho_1 = 1/2$ and $\rho_2 = -1/2$. For $3l+u-3 > 0$ we obtain

$$[z^n]G_S(z) \sim \begin{cases} \frac{C \cdot 2^{\frac{7-5l-u}{2}}}{\Gamma(\frac{3l+u-3}{2})} \cdot 2^n \cdot n^{\frac{3l+u-5}{2}} & \text{if n is odd,} \\ 0 & \text{if n is even.} \end{cases}$$

The case $3l+u-3 = 0$ needs to be treated separately. Note that $3l+u-3 = 0$ implies that $l = 1$ and $u = 0$. Thus, the structure S that we embed is a single vertex. Therefore the number of good embeddings is just the cardinality of $\mathcal{B}_n$, *i.e.*, $g_{\mathcal{B}_n}(S) = \boldsymbol{C}_{\frac{n-1}{2}} \sim \frac{\sqrt{2} \cdot 2^n}{\sqrt{\pi n^3}}$, see Example 3.5. (Note also that for S being a single vertex $a_{\mathcal{B}_n}(S) = n\boldsymbol{C}_{\frac{n-1}{2}} \sim \frac{\sqrt{2} \cdot 2^n}{\sqrt{\pi n}}$.) □

Now, we can give the average number of embeddings of a given tree S in the class $\mathcal{B}_n$ asymptotically.

Theorem 6.11. *Let S be a rooted tree with degree distribution sequence $d_S = (l, u, d_2, \ldots, d_{m-1})$ and let $C = \prod_{i=3}^{m-1}(\boldsymbol{C}_{i-1})^{d_i}$. For even n the average number of embeddings of S in a random tree $T \in \mathcal{B}_n$ is asymptotically given by*

$$\frac{[z^n]A_S(z)}{[z^n]B(z)} = \frac{a_{\mathcal{B}_n}(S)}{[z^n]B(z)} \sim \frac{C 2^{\frac{5-5l-u}{2}}\sqrt{\pi}}{\Gamma(\frac{3l+u-2}{2})} n^{\frac{3l+u-1}{2}}, \quad \text{as } n \to \infty,$$

and the average number of good embeddings of S in a random tree $T \in \mathcal{B}_n$ is asymptotically given by

$$\frac{[z^n]G_S(z)}{[z^n]B(z)} = \frac{g_{\mathcal{B}_n}(S)}{[z^n]B(z)} \sim \begin{cases} \frac{C 2^{\frac{6-5l-u}{2}}\sqrt{\pi}}{\Gamma(\frac{3l+u-3}{2})} n^{\frac{3l+u-2}{2}} & \text{if } 3l+u-3 > 0, \\ 1 & \text{if } 3l+u-3 = 0. \end{cases}$$

Proof. The proof follows directly by Theorem 6.10 and the asymptotics of the number of plane binary trees (see Example 3.5). Moreover, note that if S is a single node, *i.e.*, $l = 1$ and $d_i = 0\ \forall i \geq 1$ (thus, in particular $u = 0$), the average number of all embeddings of S in a random tree from $\mathcal{B}_n$ is n, which is obvious, since in this case S can be embedded in each of the n nodes. The average number of good embeddings of a single node in a tree T from $\mathcal{B}_n$ is 1, since the only possibility to embed the node is in the root of T. □

Now, we investigate the ratio $\frac{g_{\mathcal{B}_n}(S)}{a_{\mathcal{B}_n}(S)}$, which occurs in optimal stopping problems that were briefly introduced in Section 6.1.

Corollary 6.12. *Consider a rooted tree S with degree distribution sequence $d_S = (l, u, d_2, \ldots, d_{m-1})$. Let $k = \frac{3l+u-3}{2}$ and let n be odd. The asymptotic ratio of the cumulative number of good embeddings of S into $\mathcal{B}_n$ to the number of all embeddings into $\mathcal{B}_n$ is given by*

$$\frac{g_{\mathcal{B}_n}(S)}{a_{\mathcal{B}_n}(S)} \sim \begin{cases} \frac{\Gamma(k+1/2)}{\Gamma(k)}\frac{\sqrt{2}}{\sqrt{n}} & \text{if } k > 0, \\ 1/n & \text{if } k = 0. \end{cases}$$

Proof. We have $\frac{g_{\mathcal{B}_n}(S)}{a_{\mathcal{B}_n}(S)} = \frac{[z^n]G_S(z)}{[z^n]A_S(z)}$. The corollary follows immediately from Theorem 6.10. □

In [72] Kubicki, Lehel and Morayne. proved that if T is a complete balanced binary tree of arbitrary size and S_1, S_2 are rooted trees in which each node has at most 2 descendants and $S_1 \subseteq S_2$, then $\frac{g_T(S_1)}{a_T(S_1)} \leq \frac{g_T(S_2)}{a_T(S_2)}$. They also conjectured that the ratio $\frac{g_T(S)}{a_T(S)}$ is weakly increasing with S for S being any rooted tree. One year later in [73] they also stated an asymptotic result for the ratio $\frac{g_T(S)}{a_T(S)}$ when S is an arbitrary rooted tree and T a complete balanced binary tree of size n. They showed that $\lim_{n\to\infty} \frac{g_T(S)}{a_T(S)} = 2^{l-1} - 1$ where l is the number of leaves in S. Thereby they proved that for any rooted tree S the asymptotic ratio $\frac{g_T(S)}{a_T(S)}$ is non-decreasing with S (the function $2^{l-1} - 1$ increases with l and if $S_1 \subseteq S_2$ then the number of leaves of S_2 equals at least the number of leaves of S_1).

The conjecture from [72] was disproved by Georgiou in [51] who chose specific ternary trees as embedded structures to construct a counterexample. He also generalized the underlying structure to a complete k-ary tree and considered strict-order preserving maps instead of embeddings. In this setting he proved that a correlation inequality (corresponding to $\frac{g_{\mathcal{T}}(S_1)}{a_{\mathcal{T}_n}(S_1)} \leq \frac{g_{\mathcal{T}}(S_2)}{a_{\mathcal{T}_n}(S_2)}$) already holds for S_1, S_2 being arbitrary rooted trees such that $S_1 \subseteq S_2$.

Referring to the asymptotic result from [73], we show below that in our case the asymptotic ratios $\frac{\sqrt{n}\ g_{\mathcal{T}_n}(S)}{a_{\mathcal{T}_n}(S)}$ and $\frac{\sqrt{n}\ g_{\mathcal{V}_n}(S)}{a_{\mathcal{V}_n}(S)}$ are both weakly increasing with S for S being an arbitrary rooted tree. Using this asymptotic result we then show that the ratio $\frac{g_{\mathcal{T}_n}(S)}{a_{\mathcal{T}_n}(S)}$ itself (unlike in the case from [72]) is weakly increasing with S. In order to do so, we use Gautschi's inequality given in the following lemma.

Lemma 6.13 (Gautschi's inequality, [48])**.** *Let x be a positive real number and let $s \in (0, 1)$. Then*

$$x^{1-s} < \frac{\Gamma(x+1)}{\Gamma(x+s)} < (x+1)^{1-s}.$$

Theorem 6.14. *Let S_1, S_2 be rooted trees such that $S_1 \subseteq S_2$. Then*

$$\lim_{n\to\infty} \sqrt{n}\ \frac{g_{\mathcal{B}_n}(S_1)}{a_{\mathcal{B}_n}(S_1)} \leq \lim_{n\to\infty} \sqrt{n}\ \frac{g_{\mathcal{B}_n}(S_2)}{a_{\mathcal{B}_n}(S_2)}.$$

Proof. Let $d_{S_1} = (l_1, u_1, \ldots)$, $d_{S_2} = (l_2, u_2, \ldots)$, $k_1 = \frac{3l_1+u_1-3}{2}$, $k_2 = \frac{3l_2+u_2-3}{2}$ and $k_1 > 0$ (the case when $k_1 = 0$ is trivial). By Corollary 6.12 we have

$$\lim_{n\to\infty} \sqrt{n}\ \frac{g_{\mathcal{B}_n}(S_1)}{a_{\mathcal{B}_n}(S_1)} = \frac{\sqrt{2}\cdot\Gamma(k_1+1/2)}{\Gamma(k_1)} \quad \text{and} \quad \lim_{n\to\infty} \sqrt{n}\ \frac{g_{\mathcal{B}_n}(S_2)}{a_{\mathcal{B}_n}(S_2)} = \frac{\sqrt{2}\cdot\Gamma(k_2+1/2)}{\Gamma(k_2)}.$$

Note that the values k_1, $k_1+1/2$, k_2 and $k_2+1/2$ all belong to the set $\{\frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}, \ldots\}$. First, we are going to show that the function $f(k) = \frac{\Gamma(k+1/2)}{\Gamma(k)}$ is increasing in k for $k \in \{\frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}, \ldots\}$. Indeed, applying twice Gautschi's inequality (Lemma 6.13) we get for $k > 1/2$

$$\frac{f(k+1/2)}{f(k)} = \frac{\Gamma(k+1)}{\Gamma(k+1/2)} \frac{\Gamma(k)}{\Gamma(k+1/2)} > k^{1/2}(k+1/2)^{1/2}.$$

Thus, for $k > \frac{\sqrt{17}-1}{4} \approx 0.78$, we obtain $\frac{f(k+1/2)}{f(k)} > 1$. For $k = 1/2$ we also have $\frac{f(k+1/2)}{f(k)} = \frac{\pi}{2} > 1$.

Now it suffices to show that whenever $S_1 \subseteq S_2$, then $k_1 \leq k_2$ (equivalently $3l_1+u_1 \leq 3l_2 + u_2$). First, observe that if $S_1 \subseteq S_2$, then $l_1 \leq l_2$. (The number of leaves in a tree is the cardinality of its largest antichain, when considering the tree as the Hasse diagram of a poset. If S_1 has l_1 leaves and $S_1 \subseteq S_2$, then S_2 needs to contain an antichain of cardinality l_1 as a subposet, which means that its number of leaves has to satisfy $l_2 \geq l_1$.) Aiming for a contradiction, we assume that $S_1 \subseteq S_2$ and $3l_1 + u_1 > 3l_2 + u_2$. Since $l_2 \geq l_1$, we need to have $u_1 > u_2$. Thus there exists at least one vertex which was unary in S_1 and evolved into a s-ary node (with $s \geq 2$) in S_2. Such a single transformation decreases the number of unary nodes by one but at the same time increases the number of leaves by at least one. This means that in the process of evolving S_1 to any structure in which S_1 can be embedded, the sum of unary nodes and triplicated number of leaves never drops, which is a contradiction to $3l_1 + u_1 > 3l_2 + u_2$. □

Theorem 6.15. *Let S_1, S_2 be rooted trees such that $S_1 \subseteq S_2$. Then for each n*

$$\frac{g_{\mathcal{B}_n}(S_1)}{a_{\mathcal{B}_n}(S_1)} \leq \frac{g_{\mathcal{B}_n}(S_2)}{a_{\mathcal{B}_n}(S_2)}.$$

Proof. Let $d_{S_1} = (l_1, u_1, \ldots)$, $d_{S_2} = (l_2, u_2, \ldots)$, $k_1 = \frac{3l_1+u_1-3}{2}$, $k_2 = \frac{3l_2+u_2-3}{2}$. Aiming for a contradiction, we assume that $S_1 \subseteq S_2$ and $\frac{g_{\mathcal{B}_n}(S_1)}{a_{\mathcal{B}_n}(S_1)} > \frac{g_{\mathcal{B}_n}(S_2)}{a_{\mathcal{B}_n}(S_2)}$. Then by Theorem 6.14 we get

$$\lim_{n\to\infty} \sqrt{n}\, \frac{g_{\mathcal{B}_n}(S_1)}{a_{\mathcal{B}_n}(S_1)} = \lim_{n\to\infty} \sqrt{n}\, \frac{g_{\mathcal{B}_n}(S_2)}{a_{\mathcal{B}_n}(S_2)} = \frac{\sqrt{2}\cdot\Gamma(k_1+1/2)}{\Gamma(k_1)} = \frac{\sqrt{2}\cdot\Gamma(k_2+1/2)}{\Gamma(k_2)}.$$

Recall that the function $f(k) = \frac{\Gamma(k+1/2)}{\Gamma(k)}$ is increasing in k for $k \in \{\frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}, \ldots\}$ thus the above equality implies $k_1 = k_2$, or equivalently $3l_1 + u_1 = 3l_2 + u_2$.

First, assume that $l_1 = l_2$ and $u_1 = u_2$. Observing the generating functions $A_{S_1}(z)$ and $G_{S_1}(z)$, note that the ratio $\frac{g_{\mathcal{B}_n}(S_1)}{a_{\mathcal{B}_n}(S_1)} = \frac{[z^n]G_{S_1}(z)}{[z^n]A_{S_1}(z)}$ depends only on l_1 and u_1, since the constant $\prod_{i=3}^{n-1}(\boldsymbol{C}_{i-1})^{d_i}$ cancels out. Thus, in this case $\frac{g_{\mathcal{B}_n}(S_1)}{a_{\mathcal{B}_n}(S_1)} = \frac{g_{\mathcal{B}_n}(S_2)}{a_{\mathcal{B}_n}(S_2)}$, which is a contradiction.

Now, assume that either $l_1 \neq l_2$ or $u_1 \neq u_2$ by $3l_1 + u_1 = 3l_2 + u_2$. Since $S_1 \subseteq S_2$, we get $l_2 \geq l_1$ (see the proof of Theorem 6.14). This implies $u_1 \geq u_2$. Note also that there are at least $u_1 - u_2$ nodes that were unary in S_1 and evolved into s-ary for $s \geq 2$ in S_2. Each such transformation increases the number of leaves by at least one, thus $l_2 \geq l_1 + (u_1 - u_2)$. Therefore,

$$3l_2 + u_2 \geq 3(l_1 + u_1 - u_2) + u_2 = 3l_1 + u_1 + 2(u_1 - u_2).$$

Since $3l_2 + u_2 = 3l_1 + u_1$, we get $u_1 = u_2$ which implies $l_1 = l_2$. This contradicts the assumption that either $l_1 \neq l_2$ or $u_1 \neq u_2$. □

Now, we briefly discuss the case of embedding disconnected structures in $\mathcal{B}_n$. Note that in this case all the embeddings must be bad, since the underlying structure T has only one maximal element $\mathbb{1}_T$ and if S does not have a single maximal element, its embedding can not contain $\mathbb{1}_T$.

Assume that S is a forest, *i.e.*, a set of rooted trees $S_1, S_2, \ldots, S_r$ ($r \geq 2$) with the degree distribution sequence $d_S = (l, u, d_2, \ldots, d_{m-1})$. The underlying structure T is connected, thus the embedded structures $S_1, S_2, \ldots, S_r$ always have a common parent in T. Let $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_r)$ be a permutation of the set $\{1, 2, \ldots, r\}$. Define $S^{(\sigma)}$ to be a structure constructed as shown in Figure 6.5, *i.e.*, we add an additional vertex $\mathbb{1}_{S^{(\sigma)}}$ to S, which is a common parent of $S_1, S_2, \ldots, S_r$ appearing in the order given by σ. Now, instead of counting the number of embeddings of S in T we can simply count the numbers of good embeddings of $S^{(\sigma)}$ in T for all permutations σ generating non-isomorphic structures $S^{(\sigma)}$ and sum them up. Thus,

$$a_{\mathcal{B}_n}(S) = \sum_{\sigma \in \Sigma} g_{\mathcal{B}_n}(S^{(\sigma)}),$$

where Σ is a set of permutations of $\{1, 2, \ldots, r\}$ such that whenever $\sigma, \tau \in \Sigma$ and $\sigma \neq \tau$ then $S^{(\sigma)}$ and $S^{(\tau)}$ are not isomorphic. Moreover, whenever τ is a permutation of $\{1, 2, \ldots, r\}$ and $\tau \notin \Sigma$ then there exists $\sigma \in \Sigma$ such that $S^{(\sigma)}$ and $S^{(\tau)}$ are isomorphic.

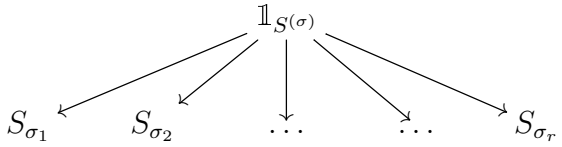


Figure 6.5: The structure of $S^{(\sigma)}$, $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_r)$.

Note that the asymptotics of $g_{\mathcal{B}_n}(S^{(\sigma)})$ is the same for all $\sigma \in \Sigma$ since the degree distribution sequence of $S^{(\sigma)}$ is the same for all $\sigma \in \Sigma$. It is given by $d_{S^{(\sigma)}} = (\tilde{d}_0, \tilde{d}_1, \ldots, \tilde{d}_{m-1}) = (l, u, \ldots, d_{r-1}, d_r + 1, d_{r+1}, \ldots, d_{m-1})$. Therefore, by Theorem 6.10

$$a_{\mathcal{B}_n}(S) \sim \begin{cases} \frac{m!}{a_1! a_2! \ldots a_\ell!} \frac{\tilde{C} \cdot 2^{\frac{7-5l-u}{2}}}{\Gamma(\frac{3l+u-3}{2})} \cdot 2^n \cdot n^{\frac{3l+u-5}{2}} & \text{if n is odd,} \\ 0 & \text{if n is even,} \end{cases}$$

where r is the number of equivalence classes of the set $\{S_1, S_2, \ldots, S_r\}$ with respect to the equivalence relation of being isomorphic and $a_1, a_2, \ldots, a_\ell$ are the cardinalities of those classes. Here $\tilde{C} = \prod_{i=3}^{m-1} (\boldsymbol{C}_{i-1})^{\tilde{d}_i}$. (Note that here we do not consider the case $3l + u - 3 = 0$ from Theorem 6.10, because by $r \geq 2$ we always have $3l + u - 3 > 0$.)

6.3 Embeddings in non-plane binary trees

In this section we explain how to take advantage of the results obtained for the plane case in order to infer about the asymptotics of the number of good and all embeddings of a rooted tree S in the family $\mathcal{V}_n$ of non-plane binary trees.

Theorem 6.16. *Consider a rooted tree S with degree distribution sequence $d_S = (l, u, d_2, \ldots, d_{m-1})$. The generating function $A_S(z)$ of the sequence $a_{\mathcal{V}_n}(S)$ (counting the cumulative number of all embeddings of S into the family $\mathcal{V}_n$) is given by*

$$A_S(z) = \left(\frac{1}{1 - zV(z)}\right)^{2l+u-1} z^{u+l-1}\, V(z)^{l+u}\, C_S\, (1 + o(1)), \qquad \text{for } z \to \pm\rho, \quad (6.3)$$

where C_S is a constant dependent on d_S and $V(z)$ is the generating function of the family of non-plane binary trees, which has two dominant singularities at $\rho \approx \pm 0.6346$ (see page 24).

Proof. Throughout this proof we write $S_1 \cong S_2$ whenever the structures S_1 and S_2 are isomorphic. This time we introduce a bivariate generating function, where z still marks the total number of vertices of a tree, while u is associated with *classes of vertices*. Two vertices v, w are meant to belong to the same class whenever there exists an isomorphism $f : T \to T$ such that $f(v) = w$. From [79] we have

$$V(z, u) = zu + \frac{zu}{2}(V(z, u)^2 - V(z^2, u^2) + 2V(z^2, u)). \quad (6.4)$$

By $V_u(z, u)$ we denote the derivative of $V(z, u)$ with respect to u, *i.e.*, $V_u(z, u) = \frac{\partial V(z,u)}{\partial u}$. We proceed as in the plane case and start with recursively defining the generating function $A_S(z)$ for the number of embeddings of S into the family $\mathcal{V}_n$, when S is a Motzkin tree:

$$A_S(z) = \begin{cases} V_u(z, 1) & \text{if } S = \{\bullet\} \\ \frac{zV(z)}{1-zV(z)} A_{\tilde{S}} & \text{if } S = \{\bullet\} \times \tilde{S} \\ \frac{z}{(1-zV(z))^2} A_{S_L}(z) A_{S_R}(z) & \text{if } S = \{\bullet\} \times S_L \times S_R \quad \text{and} \quad S_L \ncong S_R \\ \frac{z}{(1-zV(z))^2} \frac{1}{2}(A_{S_L}(z)^2 + A_{S_L}(z^2)) & \text{if } \mathcal{S} = \{\bullet\} \times S_L \times S_R \quad \text{and} \quad S_L \cong S_R \end{cases}.$$

The idea of setting up this recursive definition for $A_S(z)$ is similar to the plane case with the following differences: In the first case, corresponding to embedding a single node, we can mark an arbitrary vertex class, instead of an arbitrary vertex, since there might be some non-trivial isomorphisms that would lead to multiple countings of the same embedding. Furthermore, the paths of left-or-right trees from the previous section, yielding a factor $\frac{1}{1-2zB(z)}$, are now replaced by paths of trees where we do not distinguish between the left-or-right order, since we are in the non-plane setting. Thus, these paths give a factor $\frac{1}{1-zV(z)}$. Finally, in the case when the Motzkin tree starts with a binary root, we have to distinguish between the cases whether the two attached trees are isomorphic or not. The non-isomorphic case works analogously to its plane version, while in the isomorphic case we have to eliminate potential double-countings by using the same idea as for Equation (3.3). We do not have to solve the recursion for $A_S(z)$ explicitly, since we are solely interested in the asymptotic behavior of its coefficients and it is easy to see that asymptotically the contribution of the term $A_{S_L}(z^2)$ is negligible: Since $\rho < 1$ the function $A_S(z^2)$ is analytic at ρ^2. Thus, $[z^n]A_S(z^2) < \rho^{-n+\varepsilon}$, which is exponentially smaller than $C\rho^{-n}n^\beta = [z^n]A_S(z)$. Thus, by iterating we obtain

$$A_S(z) \sim \left(\frac{z}{1 - zV(z)}\right)^{l-1} V_u(z, 1)^l \left(\frac{zV(z)}{1 - zV(z)}\right)^u \left(\frac{1}{2}\right)^s C, \qquad \text{as } z \to \rho,$$

where l denotes the number of leaves, u the number of unary nodes and s the number of symmetry nodes in S (a symmetry node is a parent of two isomorphic subtrees). In the general case where S is an arbitrary non-plane tree, *i.e.*, a Pólya tree, we proceed as in the previous section and consider the embeddings of all binary trees with the same number of leaves as S. Thus, we get

$$A_S(z) \sim \left(\frac{z}{1-zV(z)}\right)^{l-1} V_u(z,1)^l \left(\frac{zV(z)}{1-zV(z)}\right)^u C_S, \qquad \text{as } z \to \rho. \tag{6.5}$$

The constant C_S arises from the isomorphisms and reads as

$$C_S = \sum_{\substack{s \in \mathcal{B}_S \\ s \text{ symmetry node}}} \left(\frac{1}{2}\right)^s, \tag{6.6}$$

where $\mathcal{B}_S$ denotes the set of non-plane binary trees that have the same number of leaves as S. Deriving Equation (6.4) with respect to u and plugging $u = 1$ yields

$$V_u(z,1) \sim \frac{V(z)}{1-zV(z)}, \qquad \text{as } z \to \rho$$

Finally, substituting this expression for $V_u(z,u)$ in Equation (6.5) yields the desired result. Note that the asymptotic equivalence (6.5), or (6.3) respectively, is also true for the case when S is a single node, *i.e.*, $l = 1$ and $u = s = 0$. □

Theorem 6.17. *Consider a rooted tree S with degree distribution sequence $d_S = (l, u, d_2, \ldots, d_{m-1})$. For even n the asymptotics of the number $a_{\mathcal{V}_n}(S)$ of all embeddings of S into $\mathcal{V}_n$, is given by*

$$a_{\mathcal{V}_n}(S) \sim \frac{2C_S b^{-2l-u+1} \rho^{-2l-u}}{\Gamma(\frac{2l+u-1}{2})} \cdot \rho^{-n} \cdot n^{\frac{2l+u-3}{2}}, \qquad \text{as } n \to \infty,$$

while the asymptotic behavior of the number of good embeddings of S into $\mathcal{V}_n$ is given by

$$g_{\mathcal{V}_n}(S) \sim \begin{cases} \frac{2C_S b^{-2l-u+2} \rho^{-2l-u+1}}{\Gamma(\frac{2l+u-2}{2})} \cdot \rho^{-n} \cdot n^{\frac{2l+u-4}{2}} & \text{if } \ 2l+u-2 > 0, \\ \frac{b}{\sqrt{\pi}} \cdot \rho^{-n} \cdot n^{-3/2} & \text{if } \ 2l+u-2 = 0, \end{cases}$$

where $b \approx 2.5184$, $\rho \approx 0.6346$ and the constant C_S, given in (6.6)*, depends on the structure of S.*

Proof. First, note that $V(\rho) \sim \frac{1}{\rho}$ (see page 24). Therefore, the dominant part of the asymptotics of the coefficients of $A_S(z)$ comes from the factors $\frac{1}{1-zV(z)}$, which give

$$\frac{1}{1-zV(z)} \sim \frac{1}{\rho b \sqrt{1-\frac{z}{\rho}}} \qquad \text{for } z \to \rho.$$

The result for $a_{\mathcal{V}_n}(S)$ follows immediately by use of the transfer theorems (*cf.* Theorems 2.8 and 2.10). As in the plane case, the generating function $G_S(z)$ for the good embeddings just differs from $A_S(z)$ by a factor $(1-zV(z))$ corresponding to the staring path of binary trees and thus, the asymptotic behavior of its coefficients can be determined analogously. Recall that $2l+u-2 = 0$ represents the case where S is a single vertex. The number of good embeddings is therefore just the cardinality of $\mathcal{V}_n$. □

By means of Theorem 6.17 we can directly give the expected values for the number of embeddings of a given rooted tree S in a random non-plane binary tree.

Theorem 6.18. *Let S be a rooted tree with degree distribution sequence $d_S = (l, u, d_2, \ldots, d_{m-1})$ and let C_S be defined as in (6.6). For even n the average number of embeddings of S in a random tree $T \in \mathcal{V}_n$ is asymptotically given by*

$$\frac{[z^n]A_S(z)}{[z^n]V(z)} = \frac{a_{\mathcal{V}_n}(S)}{[z^n]V(z)} \sim \frac{C_S\Gamma(-1/2)\rho^{-2l-u}b^{-2l-u}}{\Gamma((2l+u-1)/2)} n^{\frac{2l+u}{2}}, \quad \text{as } n \to \infty,$$

and the average number of good embeddings of S in a random tree $T \in \mathcal{V}_n$ is asymptotically given by

$$\frac{[z^n]G_S(z)}{[z^n]V(z)} = \frac{g_{\mathcal{V}_n}(S)}{[z^n]V(z)} \sim \begin{cases} \frac{C_S\Gamma(-1/2)\rho^{-2l-u+1}b^{-2l-u+1}}{\Gamma((2l+u-2)/2)} n^{\frac{2l+u-1}{2}} & \text{if } 2l+u-2 > 0, \\ 1 & \text{if } 2l+u-2 = 0. \end{cases}$$

Proof. The result is obtained immediately by Theorem 6.17 and the asymptotics of the number of non-plane binary trees given on page 24. □

Now we can formulate a corollary analogous to Corollary 6.12 from the plane case.

Corollary 6.19. *Consider a rooted tree S with degree distribution sequence $d_S = (l, u, d_2, \ldots, d_{m-1})$. Let $k = \frac{2l+u-2}{2}$. The asymptotic ratio of the number of good embeddings of S in $\mathcal{V}_n$ to the number of all embeddings, for $n \to \infty$ is given by*

$$\frac{g_{\mathcal{V}_n}(S)}{a_{\mathcal{V}_n}(S)} \sim \begin{cases} \frac{\Gamma(k+1/2)}{\Gamma(k)} \frac{\rho b}{\sqrt{n}} & \text{if } \ k > 0, \\ 1/n & \text{if } \ k = 0. \end{cases}$$

Theorem 6.20. *Let S_1, S_2 be rooted trees such that $S_1 \subseteq S_2$. Then*

$$\lim_{n\to\infty} \sqrt{n}\, \frac{g_{\mathcal{V}_n}(S_1)}{a_{\mathcal{V}_n}(S_1)} \leq \lim_{n\to\infty} \sqrt{n}\, \frac{g_{\mathcal{V}_n}(S_2)}{a_{\mathcal{V}_n}(S_2)}.$$

Proof. By Corollary 6.19 we get that for any S with $d_S = (l, u, d_2, \ldots, d_{m-1})$

$$\lim_{n\to\infty} \sqrt{n}\, \frac{g_{\mathcal{V}_n}(S)}{a_{\mathcal{V}_n}(S)} = \frac{\Gamma(k+1/2)}{\Gamma(k)} \rho b$$

where $k = \frac{2l+u-2}{2} > 0$. The rest of the proof is then analogous to the proof of Theorem 6.14. □

Now, let us briefly comment on embedding disconnected structures in the non-plane case. Let S be a forest, *i.e.*, a set of rooted trees $S_1, S_2, \ldots, S_r$, $r \geq 2$. Again, instead of counting all embeddings of S in $\mathcal{V}_n$, we can count the good embeddings of $\tilde{S}$ in $\mathcal{V}_n$, where $\tilde{S}$ is a forest S with an additional common parent that clips together all S_i's. Note that in the non-plane case the order of S_i's does not matter, thus we simply have

$$a_{\mathcal{V}_n}(S) = g_{\mathcal{V}_n}(\tilde{S}).$$

6.4 Embeddings in planted plane trees

In this section we extend the results from plane binary trees to the class $\mathcal{T}_n$ of planted plane trees (*cf.* Example 3.3). The structures that we embed are as well planted plane trees, and therefore every such tree S is of the form $S = \{\bullet\} \times S_1 \times \ldots \times S_k$, where the S_i's denote the subtrees that are attached to the root. The following Lemma contains the construction of the generating function $A_S(z)$ of all embeddings of the tree S in the family $\mathcal{T}_n$ of planted plane trees of size n.

Lemma 6.21. *The generating function $A_S(z)$ of all embeddings of $S = \{\bullet\} \times S_1 \times \ldots \times S_k$ into the family $\mathcal{T}_n$ of planted plane trees of size n can be recursively specified as*

$$A_S(z) = \begin{cases} zT'(z) = \frac{T(z)(1-T(z))}{1-2T(z)} & \text{if } k = 0 \\ \frac{T(z)}{1-2T(z)} A_{S_1}(z) & \text{if } k = 1 \\ \frac{T(z)^2}{(1-2T(z))^2(1-T(z))} A_{S_1}(z) A_{S_2}(z) & \text{if } k = 2 \\ \frac{T(z)}{(1-2T(z))^2} \left(\frac{1-2T(z)}{1-T(z)} A_{S_1}(z) A_{S_{2,k}}(z) \right. & \\ \quad + \frac{T(z)(1-2T(z))}{(1-T(z))^2} A_{S_{1,k-1}}(z) A_{S_k}(z) & \text{if } k > 2 \\ \quad \left. + \left(\frac{1-2T(z)}{1-T(z)} \right) \left(A_{S_{1,2}}(z) A_{S_{3,k}}(z) + \ldots A_{S_{1,k-2}}(z) A_{S_{k-2,k}}(z) \right) \right) & \end{cases}, \tag{6.7}$$

where $T(z)$ denotes the generating function of the class $\mathcal{T}$ of planted plane trees, and $S_{i,j}$ denotes the tree $S_{i,j} = \{\bullet\} \times S_i \times \ldots \times S_j$ that consists of a root to which the $j - i + 1$ subtrees $S_i, \ldots, S_j$ are attached (in that order).

Proof. The case $k = 0$ is equivalent to the binary cases, and corresponds to marking an arbitrary node in the tree T. Deriving both sides of the specification $T(z) = \frac{z}{1-T(z)}$ of planted plane trees with respect to z and solving for $T'(z)$ yields the equality

$$zT'(z) = \frac{z}{1-2T(z)} = \frac{T(z)(1-T(z))}{1-2T(z)}.$$

Now, let us continue with the proof of the recurrence for the case $k > 2$. In order to do so let us observe Figure 6.6 that visualizes how an embedding of a tree S in a tree T can be constructed: We start with a path of left-or-right plane trees, followed by the embedded root node. Attached to the root node there is another such path, ending with the so-called "splitting node". To the left and the right of this second path there can of course be several planted plane trees attached to the embedded root node, which themselves do not contain any embedded vertices. The two paths that are separated by the embedded root node contribute a factor $\left(\frac{1}{1-\frac{z}{(1-T(z))^2}} \right)^2$, which can be simplified to $\left(\frac{1-T(z)}{1-2T(z)} \right)^2$ by means of the functional equation $T(z) = \frac{z}{1-T(z)}$.

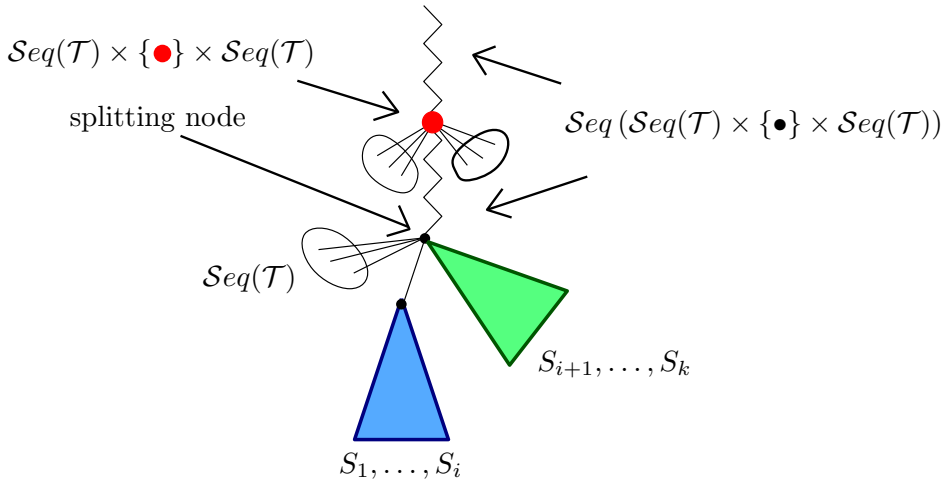


Figure 6.6: Sketch of the principle of embedding a plane tree $S = \{\bullet\} \times S_1 \times \ldots \times S_k$ into the family of plane trees $\mathcal{T}$.

The root node together with the two sequences of planted plane trees that can be attached to the left or to the right of the path give a factor $\frac{z}{(1-T(z))^2} = \frac{T(z)}{1-T(z)}$.

The splitting node can as well have a sequence of plane trees attached, that do not contain any embedded nodes, yielding a factor $\frac{1}{1-T(z)}$, but at some point there has to appear the first plane tree that contains some embedded nodes (pictured in blue in Figure 6.6). All subtrees attached to the splitting node that are to the right of this blue one are comprised in one plane tree (pictured in green in Figure 6.6). Now we have to distinguish between the cases where a different number of the subtrees $S_1, \ldots, S_k$ are embedded in the left (*i.e.*, the blue) subtree, while the remaining ones are embedded in the right (*i.e.*, the green) tree. These case distinctions give rise to the recursion (6.7) for the generating function. The first two summands of the last case in (6.7), *i.e.* the case $k > 2$, represent the cases where one of the S_i's is embedded in a separate subtree:

- Solely S_1 is embedded in the left tree: In this case we count all embeddings of S_1 in the left subtree, giving a factor $A_{S_1}(z)$, while in the right subtree we count exclusively the good embeddings of $S_{2,k} = \{\bullet\} \times S_2 \times \ldots \times S_k$, since the splitting node has to be the embedded root of $S_{2,k}$ in order to prevent multiple embeddings of the root. We already know that the generating function of good embeddings is obtained from the generating function of all embeddings by multiplication with $\frac{1-2T(z)}{1-T(z)}$ (corresponding to 1 divided by the generating function of the starting path) and thus we get the factor $\frac{1-2T(z)}{1-T(z)} A_{S_{2,k}}(z)$, where $A_{S_{2,k}}(z) = A_{\{\bullet\} \times S_2 \times \ldots \times S_k}$.

- Solely S_k is embedded in the right tree: Here we count the good embeddings of $S_{1,k-1}$ in the left tree, as this is general necessary for all cases where we consider more than just one of the S_i's to be embedded in the same subtree. However, in this case we have to count only the bad embeddings of S_k in the right tree, since no node of S can be embedded into the splitting node, except the root of S, but then the embedding of S_k is still a bad embedding into the green tree. Altogether this yields the factor $\frac{1-2T(z)}{1-T(z)} \frac{T(z)}{1-T(z)} A_{S_{1,k-1}}(z) A_{S_k}(z)$.

In all other cases where we embed at least two of the subtrees $S_1, \ldots, S_k$ in both the left and the right (*i.e.*, the blue and the green) subtree, we consider good embeddings for both subtrees, yielding a factor $\left(\frac{1-2T(z)}{1-T(z)}\right)^2$. Together with the factors from the two paths, the embedded root and the sequence of plane trees we get the desired coefficients.

The cases $k = 1$ and $k = 2$ can be treated in the exact same way as we just did for $k > 2$. However, note that in the case $k = 1$ the green (*i.e.*, the right) tree and two of the sequences of planted plane trees are merged together, such that we end up with one path, the embedded root of S together with its two sequences of planted plane trees and finally the attached blue tree that contains the embedding of the only subtree S_1. This yields the factor

$$\frac{1-T(z)}{1-2T(z)}\frac{T(z)}{1-T(z)}A_{S_1}(z).$$

In the case $k = 2$ we have the pre-factor $\frac{T(z)}{(1-2T(z))^2}$ that covers the two paths, the embedded root node with its attached sequences of plane trees and the sequence of plane trees that is attached to the splitting node. Now there is just one splitting option: S_1 has to embedded in the left tree, where we consider all embeddings, and S_2 has to be embedded in the right tree, where we solely count the bad embeddings of S_2, since the splitting node must not be an embedded node. It is easy to verify that this case gives the factor

$$\frac{T(z)^2}{(1-2T(z))^2(1-T(z))}A_{S_1}(z)A_{S_2}(z).$$

□

Remark 6.22. *Note that for the cases $k = 0, 1, 2$ the generating function $A_S(z)$ of all embeddings of $S = \{\bullet\} \times S_1 \times \ldots \times S_k$ into the family $\mathcal{T}_n$ of planted planes trees of size n given in (6.7) is of the form $f(T) \cdot A_{S_1}(z) \ldots A_{S_k}(z)$, where $f(T)$ is a function that is solely depending on $T(z)$. We want to emphasize that, by digging into the structure of S and by recursive application of the formulas given in (6.7), it follows that $A_S(z)$ is in fact of the form*

$$A_S(z) = f(T) \cdot A_{S_1}(z) \ldots A_{S_k}(z),$$

for arbitrary $S = \{\bullet\} \times S_1 \times \ldots \times S_k$.

Now we are in the position to obtain the number of embeddings of a given plane tree S in the family of planed plane trees asymptotically.

Theorem 6.23. *Consider a rooted tree $S\{\bullet\} \times S_1 \times \ldots \times S_k$ with degree distribution sequence $d_S = (l, d_1, d_2, \ldots, d_{m-1})$. The asymptotics of the number of all embeddings of S in the class $\mathcal{T}_n$ of planted plane trees is given by*

$$a_{\mathcal{T}_n}(S) \sim \prod_{i=1}^{k}\left(\frac{\boldsymbol{C}_{i-1}}{2}\right)^{d_i}\left(\frac{1}{4}\right)^{l}\frac{n^{-(l+d-2)/2}}{\Gamma(-(l+d)/2)}4^n,$$

with $d := \sum_{i=1}^{k} id_i$. The asymptotic behavior of the number $g_{\mathcal{T}_n}(S)$ of good embeddings of S in the class $\mathcal{T}_n$ of planted plane trees of size n is given by

$$g_{\mathcal{T}_n}(S) \sim \prod_{i=1}^{k}\left(\frac{\boldsymbol{C}_{i-1}}{2}\right)^{d_i}\left(\frac{1}{4}\right)^{l}\frac{n^{-(l+d-1)/2}}{\Gamma(-(l+d+1)/2)}4^n$$

Proof. Let us set $f_1 = 1$, $f_2 = \frac{T(z)^2}{(1-2T(z))^2(1-T(z))}$ and $f_k = \frac{A_S(z)}{\prod_{i=1}^k A_{S_i}(z)}$ for $k > 2$. Then we get

$$f_k(z) = \frac{T(z)}{(1-T(z))^2}\left(\frac{1}{1-2T(z)} f_1 f_{k-1} + \sum_{j=2}^{k-2} f_j f_{k-j}\right) \qquad \text{for } k \geq 3. \tag{6.8}$$

Now we set

$$g_k(z) = \begin{cases} \frac{1}{2} f_1 & k = 1 \\ (1-2T(z))^k f_k(z) & k > 1 \end{cases}. \tag{6.9}$$

Plugging (6.9) into (6.8) gives

$$g_k(z) = \frac{T(z)}{(1-T(z))^2}\left(2g_1 g_{k-1} + \sum_{j=2}^{k-2} g_j g_{k-j}\right) = \frac{T(z)}{(1-T(z))^2} \sum_{j=1}^{k-1} g_j g_{k-j}.$$

By evaluating this recurrence at $z = \frac{1}{4}$, using $T(\frac{1}{4}) = \frac{1}{2}$ and setting $h_k := g_k\left(\frac{1}{4}\right)$, we get

$$h_k = \sum_{j=1}^{k-1} h_j h_{k-j}, \quad \text{and} \quad h_1 = 1,$$

which is exactly the recurrence for the Catalan numbers, and thus, $h_k = \boldsymbol{C}_{k-1}$.

Hence, for $z \to \frac{1}{4}$ we have

$$f_k(z) \sim \frac{1}{2}\boldsymbol{C}_{k-1}(1-4z)^{-k/2},$$

which implies

$$A_S(z) \sim \frac{\boldsymbol{C}_{k-1}}{2}(1-4z)^{-k/2} A_{S_1} \dots A_{S_k} = \prod_{i=1}^{k}\left(\frac{\boldsymbol{C}_{i-1}}{2}(1-4z)^{-i/2}\right)^{d_i} \left(A_{\{\bullet\}}\right)^l,$$

where $S = \{\bullet\} \times S_1 \times \ldots \times S_k$ and d_i denotes the number of nodes with out-degree i, and l denotes the number of leaves, *i.e.* $l = d_0$. Using the equality $A_{\{\bullet\}} = zT'(z)$ we get for $z \to \frac{1}{4}$

$$A_S(z) = \prod_{i=1}^{k}\left(\frac{\boldsymbol{C}_{i-1}}{2}\right)^{d_i} (1-4z)^{(l+\sum_{i=1}^k i d_i)/2}\left(\frac{1}{4}\right)^l \left(1 + \mathcal{O}(\sqrt{1-4z})\right). \tag{6.10}$$

Finally, by means of singularity analysis (*cf.* Corollary 2.11) we get

$$[z^n]A_S(z) \sim \prod_{i=1}^{k}\left(\frac{\boldsymbol{C}_{i-1}}{2}\right)^{d_i}\left(\frac{1}{4}\right)^l \frac{n^{-(l+d-2)/2}}{\Gamma(-(l+d)/2)} 4^n,$$

with $d := \sum_{i=1}^k i d_i$. □

Finally, by the use of the results from Theorem 6.23 we can calculate the average number of embeddings of a given tree S in a random planted plane tree asymptotically, as well as study the ratio $\frac{g_{\mathcal{T}_n}(S)}{a_{\mathcal{T}_n(S)}}$ of the number of good to all embeddings.

Theorem 6.24. *Let $S = \{\bullet\} \times S_1 \times \ldots \times S_k$ be a rooted tree with degree distribution sequence $d_S = (l, u, d_2, \ldots, d_{m-1})$ and let $d := \sum_{i=1}^{k} i d_i$. Then the average number of embeddings of S in a random tree $T \in \mathcal{T}_n$ is asymptotically given by*

$$\frac{[z^n]A_S(z)}{[z^n]T(z)} = \frac{a_{\mathcal{T}_n}(S)}{[z^n]T(z)} \sim \frac{4\sqrt{\pi}\prod_{i=1}^{k}\left(\frac{C_i}{2}\right)^{d_i}\left(\frac{1}{4}\right)^{l}}{\Gamma(-(d+l)/2)} n^{-(l+d-5)/2}, \quad \text{as } n \to \infty,$$

and the average number of good embeddings of S in a random tree $T \in \mathcal{V}_n$ is asymptotically given by

$$\frac{[z^n]G_S(z)}{[z^n]T(z)} = \frac{g_{\mathcal{T}_n}(S)}{[z^n]T(z)} \sim \frac{4\sqrt{\pi}\prod_{i=1}^{k}\left(\frac{C_i}{2}\right)^{d_i}\left(\frac{1}{4}\right)^{l}}{\Gamma(-(l+d+1)/2)} n^{-(l+d-4)/2}.$$

Proof. The proof follows immediately by Theorem 6.23 and the asymptotics of the number of planted plane trees given in Example 4.3. □

Corollary 6.25. *Consider a rooted tree S. The asymptotic ratio of the number of good embeddings of S into $\mathcal{T}_n$ to the number of all embeddings is given by*

$$\frac{g_{\mathcal{V}_n}(S)}{a_{\mathcal{V}_n}(S)} \sim \frac{2\Gamma(-(l+d)/2)}{\Gamma(-(d+l+1)/2)\sqrt{n}},$$

where $d := \sum_{i=1}^{k} i d_i$.

Chapter 7

Conclusion

In Part II we investigated different types of parameters of various different tree classes.

The protection number of a tree is an extremal parameter (*cf.* Section 2.3), which was already studied by Heuberger and Prodinger in [65] for the class of planted plane trees. We generalized this work to a more general framework and obtained the average protection number for all simply generated trees, as well as for Pólya trees and non-plane binary trees. We did not include Pólya trees with general degree restrictions, since the general expressions will look clumsy and only numerical results for specific classes may be of interest. But it is immediate that the asymptotic mean and variance of the protection number for Pólya-trees with any kind of degree restriction can be calculated in the very same way. As we saw in some of the examples in Chapter 4, there are classes of trees, for which the obtained formulas involve a recurrence that might not be solvable explicitly. However, using the equations that we obtained, it is possible to calculate the asymptotic mean and variance in an arbitrarily accurate way with a very low computational effort.

It is well known that Cayley trees and Pólya trees are very similar, but the latter are not simply generated, as the simple proof presented in [36] shows. A detailed analysis of the structural differences was done in [56, 91]: Roughly speaking, Pólya trees are Cayley trees (more precisely, the simply generated class whose ordinary generating function is the exponential generating function of Cayley trees) with small forests attached to each vertex. Comparing the resulting values from Table 4.1 for Cayley trees and Pólya trees shows the quantitative effect of those forests, which have on average less than one vertex. As expected, these additional forests decrease the protection numbers.

The most striking feature of the obtained results is that the average protection number of all investigated tree classes is asymptotically constant, regardless of any degree restrictions or planarity properties. While this is also the case for (plane and non-plane) recursive trees [59], the average protection number of PATRICIA trees is asymptotically $\log_2 n$ [31] and for d-ary recursive trees it is asymptotically $\alpha_d \log n$ [35, 34]. Another interesting parameter concerning the protection number of a vertex is the maximal protection number of a node. In [59] the authors mention that experiments suggest that the average maximal protection number of a node in plane oriented recursive trees (PORTs), which are the plane counterpart of recursive trees and frequently studied objects [34, 77, 84, 89], and non-plane recursive trees is asymptotically $\Theta(\log \log n)$. However, this has not been proven yet and further

studies concerning this parameter in various classes of trees might deliver interesting structural results and enable nice comparisons between the classes.

The second parameter that we studied was the number of non-isomorphic subtree-shapes in two selected families of increasing trees, namely recursive trees and increasing binary trees. The major novelty in our results is that we extended the compactification process to labeled trees, which can be useful in practice since it enables an efficient search in the compacted data structure [14].

So far there have only been results concerning the number of non-isomorphic subtrees in unlabeled trees. In particular, it was shown in [8] that the expected number of non-isomorphic subtrees in a random simply generated tree of size n is asymptotically $\Theta\left(\frac{n}{\sqrt{\log n}}\right)$. Experiments suggested that for increasingly labeled trees this number is smaller. We proved that the average number of non-isomorphic subtree-shapes in a random recursive or increasing binary tree of size n is $\Omega(\sqrt{n})$ and $\mathcal{O}\left(\frac{n}{\log n}\right)$. Numerical simulations give rise to conjecture that this upper bound is already sharp, *i.e.*, that the size of the compacted tree is $\Theta\left(\frac{n}{\log n}\right)$. However, in order to prove this claim, one has to find the distribution of the weights $w(t)$, which turns out to be a very challenging task, especially in case of non-plane trees due to the appearance of automorphisms. Thus, obtaining the (maximum) number of labelings of non-plane trees of a given size is still work in progress, with the aim to improve the lower bounds such that we can show the Θ-result. Furthermore, we conjecture that the average number of non-isomorphic subtree-shapes is $\Theta\left(\frac{n}{\log n}\right)$ for all classes of increasing trees. The reason why we chose to investigate recursive trees and increasing binary trees was that for these two classes we were able to solve the differential equation defining $S_t(z)$, although in case of increasing binary trees the solution is already more complicated and involves some Bessel functions. However, in case of PORTs, we did not get any explicit solution for $S_t(z)$. Thus, the compaction rate of PORTs is still an open problem.

The last tree parameter that was studied in Part II of this thesis was the number of good/all embeddings of a given rooted tree S into three selected families of trees, namely plane and non-plane binary trees, and planted plane trees. We calculated the asymptotic mean of the number of embeddings and proved that the ratio of the number of good embeddings to the number of all embeddings of a given tree $S = \{\bullet\} \times S_1 \times \ldots \times S_k$ into the above mentioned families of trees of size n is asymptotically of the same order for all the three considered classes of trees. We expect that this result will also hold for Pólya trees, which are the closest counterpart to posets that admit a (rooted) treelike shape, *i.e.*, they have a single maximal element. In principle, the approach that we used within this thesis works for embeddings into Pólya trees as well. However, one would have to consider all possible partitions of $S_1, \ldots, S_k$ indicating potential groups of isomorphisms between the S_i's, which can get rather involved and is therefore omitted in this work.

To our knowledge we were the first ones to use the methods of analytic combinatorics in order to study optimal stopping problems on tree structures. It seems that so far all investigations were based on probabilistic methods. This new approach may provoke several new ideas of what and how could be studied next in the context of optimal stopping. One restriction that comes along with our approach is that we solely get a result on the cumulative number of embeddings in all trees of a given size belonging to the class of interest. Thus, we assume to have no additional preliminary

knowledge of the underlying structure of the occurring poset. However, further studies concerning this topic with the aim to extend our methodology to both restricted tree-shapes, as well as to general posets, might reveal interesting results.

Part III

Parameters of lambda terms

Chapter 8

Lambda terms with bounded De Bruijn indices

This chapter is organized in two parts. First, we will provide asymptotic results on the total number of variables in lambda terms with bounded De Bruijn indices, *i.e.*, with a bounded number of abstractions between each leaf and its binding lambda. This section is based on the article *Distribution of variables in lambda terms with restrictions on De Bruijn indices and De Bruijn levels*, which was joint work with Bernhard Gittenberger and has already been published in the Electronic Journal of Combinatorics, [57]. Then, in the second part, we will investigate the shape of such lambda terms and thereby obtain the so-called unary profile of a random term belonging to this class of lambda terms, based on joint work with Katarzyna Grygiel in the already submitted manuscript [62].

8.1 Total number of variables

In this section we prove that the number of all variables in closed lambda terms with bounded De Bruijn indices is asymptotically normally distributed and we provide expressions for the mean and the variance when the size n tends to infinity. This is done by means of bivariate generating functions and the use of Hwang's Quasi Powers Theorem 2.18.

By translating Equations (3.9) and (3.10) into bivariate generating functions $\hat{P}^{(i,k)}(z,u)$, where z marks the size and u marks the number of leaves, we get

$$\hat{P}^{(k,k)}(z,u) = kzu + z\hat{P}^{(k,k)^2}(z,u) + z\hat{P}^{(k,k)}(z,u),$$

and

$$\hat{P}^{(i,k)}(z,u) = izu + z\hat{P}^{(i,k)^2}(z,u) + z\hat{P}^{(i+1,k)}(z,u),$$

which can be solved and written in the form

$$\hat{P}^{(i,k)}(z,u) = \frac{1 - \mathbf{1}_{[i=k]}z - \sqrt{\hat{R}_{k-i+1,k}(z,u)}}{2z},$$

with

$$\hat{R}_{1,k}(z,u) = (1-z)^2 - 4kuz^2, \tag{8.1}$$

$$\hat{R}_{2,k}(z,u) = 1 - 4(k-1)z^2u - 2z + 2z^2 + 2z\sqrt{\hat{R}_{1,k}(z,u)}, \tag{8.2}$$

and for $3 \leq i \leq k+1$

$$\hat{R}_{i,k}(z,u) = 1 - 4(k-i+1)z^2u - 2z + 2z\sqrt{\hat{R}_{i-1,k}(z,u)}. \tag{8.3}$$

In analogy to (3.13) the bivariate generating function $G_k(z,u)$ then reads as

$$G_k(z,u) = \hat{P}^{(0,k)}(z,u) = \frac{1 - \sqrt{\hat{R}_{k+1,k}(z,u)}}{2z}.$$

Due to continuity arguments and Lemma 3.13 we know that in a sufficiently small neighborhood of $u = 1$ the dominant singularity $\hat{\rho}_k(u)$ of $G_k(z,u)$ comes only from the innermost radicand $\hat{R}_{1,k}(z,u)$ and is of type $\frac{1}{2}$. By calculating the smallest positive root of $\hat{R}_{1,k}(z,u)$ we get $\hat{\rho}_k(u) = \frac{1}{1+2\sqrt{ku}}$. Given the nested structure of $G_k(z,u)$, it follows that $\hat{\rho}_k(u)$ is the dominant singularity of all $\hat{R}_{j,k}(z,u)$, for $j = 2, \ldots, k+1$. By determining the local behavior of $\hat{R}_{1,k}(z,u)$ near $z = \hat{\rho}_k(u)$, we are able to determine Puiseux expansions of all $\hat{R}_{j,k}(z,u)$ for $j = 2, \ldots, k+1$ at $z = \hat{\rho}_k(u)$. This will be done in Proposition 8.1. In particular, this gives us the Puiseux expansion of $G_k(z,u)$ from which we can derive the asymptotic behavior of its coefficients by transfer theorems (see Theorems 2.8 and 2.10). This will be the task of Theorem 8.2 below. It will then turn out that the shape of $\hat{\rho}_k(u)$ near $u = 1$ determines the characteristic function of the random variable "number of leaves", because $\hat{\rho}_k(u)$ depends on u in a nicely regular way. This characteristic function has then the shape of a so-called quasi-power involving the function $\hat{\rho}_k(u)$ so that we can use the Quasi-Powers Theorem (Theorem 2.18) to obtain a central limit theorem.

Proposition 8.1. *Let $\hat{\rho}_k(u)$ be the root of the innermost radicand $\hat{R}_{1,k}(z,u)$, i.e. $\hat{\rho}_k(u) = \frac{1}{1+2\sqrt{ku}}$, where u is in a sufficiently small neighborhood of 1, i.e. $|u-1| < \delta$ for $\delta > 0$ sufficiently small. Then we have for $1 \leq i \leq k+1$*

$$\hat{R}_{i,k}(\hat{\rho}_k(u)(1-\epsilon), u) = \begin{cases} \hat{\rho}_k(u)\big(2 - 2\hat{\rho}_k(u) + 8ku\hat{\rho}_k(u)\big)\epsilon + \mathcal{O}(|\epsilon|^2) & i = 1 \\ c_i(u)\hat{\rho}_k(u)^2 + d_i(u)\sqrt{\epsilon} + \mathcal{O}(|\epsilon|) & i > 1 \end{cases}, \tag{8.4}$$

for $\epsilon \to 0$ so that $\epsilon \in \mathbb{C} \setminus \mathbb{R}^-$, uniformly in u, with

$$c_1(u) = 1, \quad \text{and} \qquad c_i(u) = 4(i-1)u - 1 + 2\sqrt{c_{i-1}(u)}, \quad \text{for } 2 \leq i \leq k+1,$$

and

$$d_i(u) = \frac{2\hat{\rho}_k(u)\sqrt{\hat{\rho}_k(u)\big(2 - 2\hat{\rho}_k(u) + 8ku\hat{\rho}_k(u)\big)}}{\prod_{l=2}^{i}\sqrt{c_l}} \qquad \text{for } 2 \leq i \leq k+1.$$

Proof. Using the Taylor expansion of $\hat{R}_{1,k}(z,u)$ around $\hat{\rho}_k(u)$ we obtain

$$\hat{R}_{1,k}(z,u) = \hat{R}_{1,k}(\hat{\rho}_k(u),u) + (z - \hat{\rho}_k(u))\frac{\partial}{\partial z}\hat{R}_{1,k}(\hat{\rho}_k(u),u) + \mathcal{O}((z-\hat{\rho}_k(u))^2).$$

Per definition, the first summand $\hat{R}_{1,k}(\hat{\rho}_k(u), u)$ is equal to zero. Setting $z = \hat{\rho}_k(u)(1-\epsilon)$ and using (8.1) we obtain the claim of Proposition 8.1 for the case $i = 1$.
The next step is to compute an expansion of $\hat{R}_{j,k}(z, u)$ around $\hat{\rho}_k(u)$ for $2 \leq j \leq k+1$. Using the recursive relation (8.2) for $\hat{R}_{2,k}(z, u)$ and the formula $\hat{\rho}_k(u) = \frac{1}{1+2\sqrt{ku}}$ yields

$$\hat{R}_{2,k}(\hat{\rho}_k(u)(1-\epsilon), u) = (1+4u)\hat{\rho}_k(u)^2 + 2\hat{\rho}_k(u)\sqrt{\hat{\rho}_k(u)(2 - 2\hat{\rho}_k(u) + 8ku\hat{\rho}_k(u))}\sqrt{\epsilon} + \mathcal{O}(|\epsilon|).$$

We set $c_2(u) := 1+4u$ and $d_2(u) := 2\hat{\rho}_k(u)\sqrt{\hat{\rho}_k(u)(2 - 2\hat{\rho}_k(u) + 8ku\hat{\rho}_k(u))}$ and assume that for $2 \leq i \leq k+1$ the equation $\hat{R}_{i,k}(\hat{\rho}_k(u)(1-\epsilon), u) = c_i(u)\hat{\rho}_k^2(u) + d_i(u)\sqrt{\epsilon} + \mathcal{O}(|\epsilon|)$ holds. Now we proceed by induction. Observe that

$$\hat{R}_{i+1,k}(\hat{\rho}_k(u)(1-\epsilon), u) = 1 - 4(k-i)\hat{\rho}_k^2(u)u - 2\hat{\rho}_k(u) + 2\hat{\rho}_k(u)\sqrt{c_i(u)\hat{\rho}_k^2 + d_i(u)\sqrt{\epsilon} + \mathcal{O}(|\epsilon|)}.$$

Expanding, using again $\hat{\rho}_k(u) = \frac{1}{1+2\sqrt{ku}}$ and $\hat{R}_{1,k}(\hat{\rho}_k(u), u) = 1 - 2\hat{\rho}_k(u) + \hat{\rho}_k(u)^2 - 4ku\hat{\rho}_k(u) = 0$ yields

$$\hat{R}_{i+1,k}(\hat{\rho}_k(u)(1-\epsilon), u) = 4iu\hat{\rho}_k(u)^2 - \hat{\rho}_k(u)^2 + 2\hat{\rho}_k(u)^2\sqrt{c_i(u)} + \frac{d_i(u)}{\sqrt{c_i(u)}}\sqrt{\epsilon} + \mathcal{O}(|\epsilon|).$$

Setting $c_{i+1}(u) := 4iu - 1 + 2\sqrt{c_i(u)}$ and $d_{i+1}(u) := \frac{d_i(u)}{\sqrt{c_i(u)}}$ for $2 \leq i \leq k$, we obtain $\hat{R}_{i+1,k}(\hat{\rho}_k(u) - \epsilon, u) = c_{i+1}\hat{\rho}_k(u)^2 + d_{i+1}\sqrt{\epsilon} + \mathcal{O}(|\epsilon|)$. Expanding $d_{i+1}(u)$, using its recursive relation and $d_2(u) = 2\hat{\rho}_k(u)\sqrt{\hat{\rho}_k(u)(2 - 2\hat{\rho}_k(u) + 8ku\hat{\rho}_k(u))}$, we get for $2 \leq i \leq k$

$$d_{i+1}(u) = \frac{2\hat{\rho}_k(u)\sqrt{\hat{\rho}_k(u)(2 - 2\hat{\rho}_k(u) + 8ku\hat{\rho}_k(u))}}{\prod_{l=2}^{i}\sqrt{c_l(u)}}.$$

Finally, we show that the $c_l(u)$'s are greater than zero in a neighborhood of $u = 1$. By induction it can easily be seen that they are always positive for $u = 1$. Since $c_1(1) = 1$ and assuming $c_{i-1}(1) < c_i(1)$ we get

$$c_{i+1}(1) = 4i - 1 + 2\sqrt{c_i(1)} > 4(i-1) + 4 - 1 + 2\sqrt{c_{i-1}(1)} = c_i(1) + 4.$$

Using continuity arguments we can see that the functions $c_l(u)$ have to be positive in a sufficiently small neighborhood of $u = 1$ as well, which completes the proof of (8.4). □

Theorem 8.2. *Let for any fixed k, $G_k(z, u)$ denote the bivariate generating function of the class of closed lambda terms where all De Bruijn indices are at most k. Then the equation*

$$[z^n]G_k(z, u) = \sqrt{\frac{\sqrt{ku} + 2ku}{4\pi\prod_{l=2}^{k+1} c_l(u)}}(1 + 2\sqrt{ku})^n n^{-\frac{3}{2}}\left(1 + O\left(\frac{1}{\sqrt{n}}\right)\right), \quad \text{for } n \to \infty,$$

with $c_1(u) = 1$ and $c_j(u) = 4(j-1)u - 1 + 2\sqrt{c_{j-1}(u)}$, for $2 \leq j \leq k+1$, holds uniformly in u for $|u-1| < \delta$, with $\delta > 0$ sufficiently small.

Proof. Using $G_k(z,u) = \frac{1-\sqrt{\hat{R}_{k+1,k}(z,u)}}{2z}$ and (8.4), we get for $\epsilon \in \mathbb{C} \setminus \mathbb{R}^-$ with $|\epsilon| \to 0$

$$G_k(\hat{\rho}_k(u)(1-\epsilon),u) = \frac{1-\sqrt{c_{k+1}(u)}\hat{\rho}_k(u)}{2\hat{\rho}_k(u)} - \frac{d_{k+1}(u)}{4\hat{\rho}_k(u)^2\sqrt{c_{k+1}(u)}}\sqrt{\epsilon} + \mathcal{O}(|\epsilon|).$$

Hence,

$$[z^n]G_k(z,u) = -\frac{d_{k+1}(u)\sqrt{\hat{\rho}_k(u)}}{4\hat{\rho}_k^2(u)\sqrt{c_{k+1}(u)}}[z^n]\sqrt{1-\frac{z}{\hat{\rho}_k(u)}} + [z^n]\mathcal{O}\left(\left|1-\frac{z}{\hat{\rho}_k(u)}\right|\right). \tag{8.5}$$

The singularity $\hat{\rho}_k(u) = \frac{1}{1+2\sqrt{ku}}$ is of type $\frac{1}{2}$ and if we plug

$$d_{k+1}(u) = \frac{2\hat{\rho}_k(u)\sqrt{\hat{\rho}_k(u)(2-2\hat{\rho}_k(u)+8ku\hat{\rho}_k(u))}}{\prod_{l=2}^{k}\sqrt{c_l(u)}} = \frac{4\hat{\rho}_k(u)^2\left(\sqrt{\sqrt{ku}+2ku}\right)}{\prod_{l=2}^{k}\sqrt{c_l(u)}}$$

into (8.5) and apply the standard transfer theorems (see Theorems 2.8 and 2.10), we obtain the desired result. □

Using Theorems 8.2 and 3.14, we get for $n \to \infty$

$$\mathbb{E}(u^{X_n}) = \frac{[z^n]G_k(z,u)}{[z^n]G_k(z,1)} = \left(\frac{1+2\sqrt{ku}}{1+2\sqrt{k}}\right)^n \sqrt{\frac{\sqrt{ku}+2ku}{2k+\sqrt{k}}\prod_{j=2}^{k+1}\frac{c_j(1)}{c_j(u)}}\left(1+O\left(\frac{1}{n}\right)\right),$$

where $c_1(u) = 1$ and $c_{i+1}(u) = 4iu - 1 + 2\sqrt{c_i(u)}$.

Thus, all assumptions for the Quasi-Powers Theorem (Theorem 2.18) are fulfilled, and we directly obtain the following theorem.

Theorem 8.3. *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*

$$\mathbb{E}(X_n) \sim \frac{k}{\sqrt{k}+2k}n, \quad \text{and} \quad \mathbb{V}(X_n) \sim \frac{k^2}{2\sqrt{k}(\sqrt{k}+2k)^2}n, \quad \text{as } n \to \infty.$$

Remark 8.4. *Note that $\mathbb{E}(X_n) \to \frac{n}{2}$ and $\mathbb{V}(X_n) \to 0$ for $k \to \infty$. Since these values are known for the number of leaves in binary trees, this gives a hint that almost all leaves of a large random unrestricted lambda term are located within an almost purely binary structure. However, one has to be careful, since we have to deal with two limits, $n \to \infty$ and $k \to \infty$, and thus it is necessary to check whether we can choose the order of these two limits arbitrarily.*

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 8.5. *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*

$$\mathbb{E}(Y_n) = \mathbb{E}(X_n) \sim \frac{k}{\sqrt{k}+2k}n \quad \text{and} \quad \mathbb{E}(Z_n) \sim \frac{\sqrt{k}}{\sqrt{k}+2k}n \quad \text{as } n \to \infty,$$

with X_n being defined as in Theorem 8.3.

Remark 8.6. *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.*

8.2 Unary profile

In this section we investigate the expected shape of random lambda term with bounded De Bruijn indices. Considering Equation (3.14) together with its interpretation, it can easily be seen that the enriched tree corresponding to a lambda term from $\mathcal{G}_k$ is constructed as follows (*cf.* Figure 8.1):

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

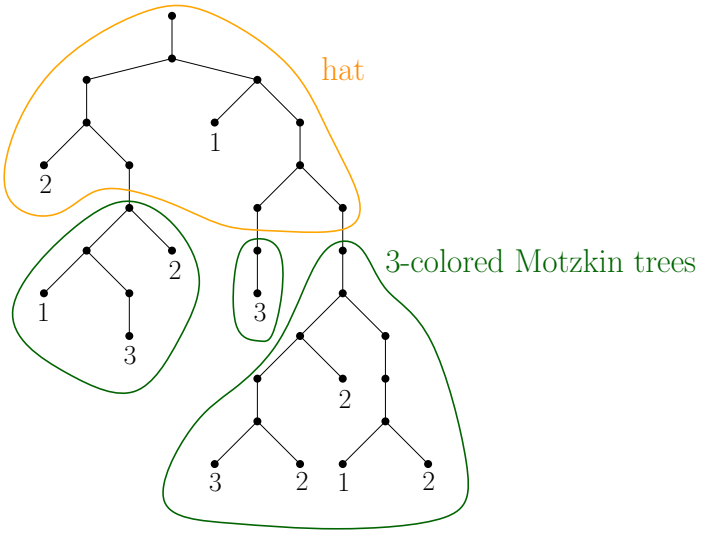


Figure 8.1: A lambda term from $\mathcal{G}_3$ decomposed into the hat (encircled in yellow) and three attached 3-colored Motzkin trees (encircled in green).

Remark 8.7. *Note that the glued binary trees in Equation* (3.14) *constitute the hat of the structure, to which we attach the k-colored Motzkin trees.*

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 lambda term belonging to $\mathcal{G}_k$ is on average of constant size and that the average number of De Bruijn levels of a term of size n is asymptotically of order $\sqrt{n}$. Finally, we provide its unary profile.

8.2.1 Average size of a hat

In this section 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 above is also constant.

Theorem 8.8. *For $k \geq 1$, let X_k be the random variable defined as the size of the hat of a lambda term where all De Bruijn indices are at most k. Then, as $n \to \infty$*

$$\mathbb{E}X_k = \frac{B_k \prod_{l=2}^{k+1} \sqrt{c_l}}{\sqrt{2k+\sqrt{k}}},$$

with $c_i := c_i(1)$, where $c_i(u)$ is defined as in Proposition 8.1, and B_k is given by

$$\begin{aligned} B_k &= \left(\sqrt{\hat{\rho}_k(2-2\hat{\rho}_k+8k\hat{\rho}_k)}\left(1+\frac{1}{\hat{\rho}_k}\right)+\frac{1+(2k-3)\hat{\rho}_k}{\hat{\rho}_k^2}\sum_{l=2}^{k+1} g_l\right)\prod_{j=2}^{k+1}\frac{1}{\sqrt{c_j}} \\ &\quad + \frac{1}{2}\sum_{i=0}^{k-2}\left(\left(\frac{1+4i\hat{\rho}_k-\sqrt{c_{k-i}}\hat{\rho}_k}{\hat{\rho}_k^2}\sum_{l=k+1-i}^{k+1} g_l\prod_{j=k+1-i}^{k+1}\frac{1}{\sqrt{c_j}}\right)-\frac{g_{k-i}}{\hat{\rho}_k}\prod_{j=k+1-i}^{k+1}\frac{1}{\sqrt{c_j}}\right), \end{aligned}$$

with

$$g_i = \frac{\sqrt{\hat{\rho}_k(2-2\hat{\rho}_k+8k\hat{\rho}_k)}}{\prod_{l=2}^{i}\sqrt{c_i}} \qquad \text{for } 2\le i\le k+1.$$

Proof. Let $G_k(z,u)$ be the bivariate generating function of the class $\mathcal{G}_k$ with z marking the size of the terms and u marking the size of their hats. The average size of a hat is hence given by

$$\mathbb{E}X_k = \frac{[z^n]\frac{\partial G_k(z,u)}{\partial u}|_{u=1}}{[z^n]G_k(z)}. \tag{8.6}$$

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 $B(z,w)$ and $M_k(z)$ are the functions defined in (3.15). This gives

$$G_k(z,u) = \frac{1-\sqrt{\hat{R}_{k+1,k}(z,u)}}{2zu},$$

where

$$\hat{R}_{i,k}(z,u) = \begin{cases} 1-2z-(4k-1)z^2, & i=1,\\ 1-2zu^2-(4k-6)z^2u^2+2zu^2\sqrt{\hat{R}_{1,k}(z,u)}, & i=2,\\ 1-2zu-4(k-i+1)z^2u^2+2zu\sqrt{\hat{R}_{i-1,k}(z,u)}, & i>2. \end{cases}$$

Note that the radicands $\hat{R}_{i,k}(z,u)$ defined above are not equal to the radicands defined in (8.1) - (8.3), since the u marks a different variable in this section. However, for $u=1$ the radicands coincide.

The derivatives of the radicands can also be recursively defined via

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

and we get

$$\begin{aligned}\frac{\partial G_k(z,u)}{\partial u}\Big|_{u=1} &= \frac{\sqrt{\hat{R}_{k+1,k}(z,1)}-1}{2z} \;-\; \frac{1}{4z\sqrt{\hat{R}_{k+1,k}(z,1)}}\cdot\frac{\partial \hat{R}_{k+1,k}(z,u)}{\partial u}\Big|_{u=1}\\ &= \frac{\sqrt{\hat{R}_{k+1,k}(z,1)}-1}{2z} \;+\; \sum_{i=0}^{k-2}\frac{z^i\left(1+4iz-\sqrt{\hat{R}_{k-i,k}(z,1)}\right)}{2\prod_{j=k+1-i}^{k+1}\sqrt{\hat{R}_{j,k}(z,1)}}\\ &\quad+\; \frac{z^{k-1}\left(1+(2k-3)z-\sqrt{\hat{R}_{1,k}(z,1)}\right)}{\prod_{j=2}^{k+1}\sqrt{\hat{R}_{j,k}(z,1)}}. \end{aligned} \tag{8.7}$$

Analogously to Proposotion 8.1 we expand the radicands $\hat{R}_{i,k}(z,1)$ around $z=\hat{\rho}_k := \hat{\rho}_k(1)$, which yields

$$\hat{R}_{i,k}(\hat{\rho}_k(1-\varepsilon),1) = \begin{cases} \hat{\rho}_k(2-2\hat{\rho}_k+8k\hat{\rho}_k)\varepsilon+\mathcal{O}(|\varepsilon|^2) & i=1,\\ c_i\hat{\rho}_k^2+d_i\sqrt{\varepsilon}+\mathcal{O}(|\varepsilon|) & i>1,\end{cases}$$

where $\varepsilon\in\mathbb{C}\setminus\mathbb{R}_-$ and $|\varepsilon|\to 0$, and $c_i=c_i(1)$ and $d_i=d_i(1)$ with $c_i(u)$ and $d_i(u)$ defined as in Proposition 8.1. Hence, we have

$$\sqrt{\hat{R}_{i,k}\big(\hat{\rho}_k(1-\varepsilon),1\big)} = \begin{cases} \sqrt{\hat{\rho}_k(2-2\hat{\rho}_k+8k\hat{\rho}_k)}\sqrt{\varepsilon}+\mathcal{O}(|\varepsilon|), & i=1,\\ \sqrt{c_i}\hat{\rho}_k+g_i\sqrt{\varepsilon}+\mathcal{O}(|\varepsilon|), & i>1\end{cases}$$

with

$$g_i = \frac{d_i}{2\hat{\rho}_k\sqrt{c_i}} = \frac{\sqrt{\hat{\rho}_k(2-2\hat{\rho}_k+8k\hat{\rho}_k)}}{\prod_{l=2}^{i}\sqrt{c_i}} \qquad \text{for } 2\le i\le k+1. \tag{8.8}$$

Plugging this into Equation (8.7) gives

$$\begin{aligned}&\frac{\partial G_k(\hat{\rho}_k(1-\varepsilon),u)}{\partial u}\Big|_{u=1} =\\ &\frac{\sqrt{c_{k+1}}\hat{\rho}_k+g_{k+1}\sqrt{\varepsilon}-1}{2\hat{\rho}_k}+\sum_{i=0}^{k-2}\frac{\hat{\rho}_k^i\big(1+4i\hat{\rho}_k-\sqrt{c_{k-i}}\hat{\rho}_k-g_{k-i}\sqrt{\varepsilon}\big)}{2\prod_{j=k-i+1}^{k+1}\big(\sqrt{c_j}\hat{\rho}_k+g_j\sqrt{\varepsilon}\big)}\\ &\quad+\;\frac{\hat{\rho}_k^{k-1}(1+(2k-3)\hat{\rho}_k-\sqrt{\hat{\rho}_k(2-2\hat{\rho}_k+8k\hat{\rho}_k)}\sqrt{\varepsilon})}{\prod_{j=2}^{k+1}\big(\sqrt{c_j}\hat{\rho}_k+g_j\sqrt{\varepsilon}\big)}\;+\;\mathcal{O}\big(|\varepsilon|\big)\\ &= A_k\;-\;B_k\sqrt{\varepsilon}\;+\;\mathcal{O}\big(|\varepsilon|\big),\end{aligned}$$

where A_k and B_k are constants depending on k with

$$\begin{aligned}B_k\;&=\frac{d_{k+1}}{2\hat{\rho}_k\sqrt{c_{k+1}}}+\left(\frac{\sqrt{\hat{\rho}_k(2-2\hat{\rho}_k+8k\hat{\rho}_k)}}{\hat{\rho}_k}+\frac{1+(2k-3)\hat{\rho}_k}{\hat{\rho}_k^2}\sum_{l=2}^{k+1}g_l\right)\prod_{j=2}^{k+1}\frac{1}{\sqrt{c_j}}\\ &\quad+\frac{1}{2}\sum_{i=0}^{k-2}\left(\left(\frac{1+4i\hat{\rho}_k-\sqrt{c_{k-i}}\hat{\rho}_k}{\hat{\rho}_k^2}\sum_{l=k+1-i}^{k+1}g_l\prod_{j=k+1-i}^{k+1}\frac{1}{\sqrt{c_j}}\right)-\frac{g_{k-i}}{\hat{\rho}_k}\prod_{j=k+1-i}^{k+1}\frac{1}{\sqrt{c_j}}\right).\end{aligned}$$

Since A_k is not important for the result, we omit to give its exact value. By singularity analysis (see Corollary 2.11) 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

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

Pluggind this and the asymptotics of $[z^n]G_k(z)$ given in Theorem 3.14 into (8.6) completes the proof. □

Since the hat of a lambda term where all De Bruijn indices are at most k, 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 can conclude the following corollary.

Corollary 8.9. *For every $k \geq 1$, the average number of k-colored Motzkin trees in the decomposition (see page 107) of lambda terms where all De Bruijn indices are at most k, is constant.*

8.2.2 Average number of De Bruijn levels

In order to determine the average number of De Bruijn levels of lambda terms with bounded De Bruijn indices, we first compute the average number of "De Bruijn levels" (*i.e.*, unary levels) of k-colored Motzkin trees. To this end, we use the following result by [38]. 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 8.10 ([38, 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

$$\mathbb{E}H_n \sim \sqrt{\frac{2\pi n}{z_0F_z(z_0,t_0)F_{tt}(z_0,t_0)}}.$$

Lemma 8.11. *The average number of De Bruijn levels of a k-colored Motzkin tree of size n is asymptotically equal to*

$$\sqrt{\frac{\pi n}{2k+\sqrt{k}}} \qquad \text{for } n\to\infty.$$

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 $F_k(z,t) := \frac{1-\sqrt{1-4kz^2-4z^2t}}{2z}$. Then $F_k(z,t)$ satisfies the assumptions of Lemma 8.10. Indeed, the function $M_k(z)$, with a square-root singularity at $z = \hat{\rho}_k = \frac{1}{1+2\sqrt{k}}$, 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 8.10. 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

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

As a corollary we immediately get the main result of this subsection.

Corollary 8.12. *For every $k \geq 1$, the average number of De Bruijn levels of a lambda term from $\mathcal{G}_k$ of size n is $\Theta(\sqrt{n})$.*

Proof. By Corollary 8.9, the number of k-colored Motzkin trees in the decomposition of lambda terms is constant on average. Therefore, the size of a largest such a tree in the decomposition of a lambda term with bounded De Bruijn indices 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 lambda terms from $\mathcal{G}_k$, which have just k levels more than a longest (in terms of De Bruijn levels) k-colored Motzkin tree in their decomposition. □

8.2.3 Unary profile

By the *unary profile* of a lambda term we mean the sequence counting the numbers of variables in each De Bruijn level of the term. In this section, we determine the mean unary profile of a random lambda term from $\mathcal{G}_k$ asymptotically.

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

Lemma 8.13 ([53, Lemma 3.4])**.** *Let γ be a Hankel contour truncated at K. Then we have, for $\alpha, \beta > 0$,*

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

Lemma 8.14. *Let $\varepsilon > 0$ and $\gamma = \left\{\rho_k\left(1 + \frac{t+i}{n}\right) : t \in [\log^2 n, n\varepsilon)\right\}$. 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 subsection.

Theorem 8.15. *Let $\kappa > 0$ be a fixed real number. The expected number of variables in De Bruijn level $\lfloor\kappa\sqrt{n}\rfloor$ in a random lambda term from $\mathcal{G}_k$ of size n is asymptotically equal to*

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

Proof. Let $U_{k,\ell}(z,u)$ be the bivariate generating function for lambda terms where all De Bruijn indices are at most k, 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

$$\begin{aligned} U_{k,\ell}(z,u) = \quad & B\Big(z, B\Big(z, 1 + B\Big(z, 2 + \dots \\ & + \underbrace{B\big(z, k + B(z, k + B(\dots B}_{\ell \text{ occurrences of } B(z,k+\dots)}(z, k + B(z, ku + M_k(z))))))\big)\dots\Big)\Big)\Big). \end{aligned}$$

Applying the formulas for $B(z,w)$ and $M_k(z)$ given in (3.15) yields

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

where

$$\hat{R}_{i,k}(z,u) = \begin{cases} 1-2z-(4k-1)z^2, & i = 1, \\ 1-2z-(4ku-2)z^2 + 2z\sqrt{R_{1,k}(z,u)}, & i = 2, \\ 1-2z-4kz^2 + 2z\sqrt{R_{i-1,k}(z,u)}, & i \in \{3,\dots,\ell\}, \\ 1-2z-4(k-i+\ell)z^2 + 2z\sqrt{R_{i-1,k}(z,u)}, & i \in \{\ell+1,\dots,\ell+k\}. \end{cases}$$

Furthermore, we have

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

and hence

$$\frac{\partial U_{k,\ell}(z,u)}{\partial u} = \frac{z^{k+\ell}}{\prod_{j=1}^{k+\ell}\sqrt{\hat{R}_{j,k}(z,u)}}.$$

Given the De Bruijn level $\ell = \lfloor \kappa\sqrt{n} \rfloor$ with $\kappa > 0$, 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 $\left|\sqrt{\hat{R}_{j,k}(z,1)}\right| = \left|z + \sqrt{R_{1,k}(z,1)}\right|$ for $j \in \{2,\dots,\ell\}$, *i.e.* all these radicands describe the same function. Indeed, let us first notice that the above holds for $j = 2$, since

$$\hat{R}_{2,k}(z,1) = \hat{R}_{1,k}(z,1) + z^2 + 2z\sqrt{\hat{R}_{1,k}(z,1)} = \left(\sqrt{\hat{R}_{1,k}(z,1)} + z\right)^2.$$

Next, by (3.15), 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 $\sqrt{\hat{R}_{2,k}(z,1)} = \sqrt{\hat{R}_{3,k}(z,1)}$. By iteration we obtain the result for $j \in \{4,\dots,\ell\}$. For $z = \hat{\rho}_k\big(1 + \frac{t}{n}\big)$ we get the expansions

$$\sqrt{\hat{R}_{i,k}(z,1)} = \begin{cases} 2k^{1/4}\hat{\rho}_k^{1/2}\sqrt{-t/n} + \mathcal{O}(|t|/n), & i = 1, \\ \hat{\rho}_k + 2k^{1/4}\hat{\rho}_k^{1/2}\sqrt{-t/n} + \mathcal{O}(|t|/n), & i \in \{2,\dots,\ell\}, \\ \sqrt{c_{i-\ell}}\hat{\rho}_k + g_{i-\ell}\sqrt{-t/n} + \mathcal{O}(|t|/n), & i \in \{\ell+1,\dots,\ell+k\}, \end{cases} \tag{8.9}$$

where $(c_i)_{i\geq 1}$ and $(g_i)_{i\geq 2}$ are as before (see (3.16) and (8.8)). Let $\varepsilon > 0$. We have

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

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

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

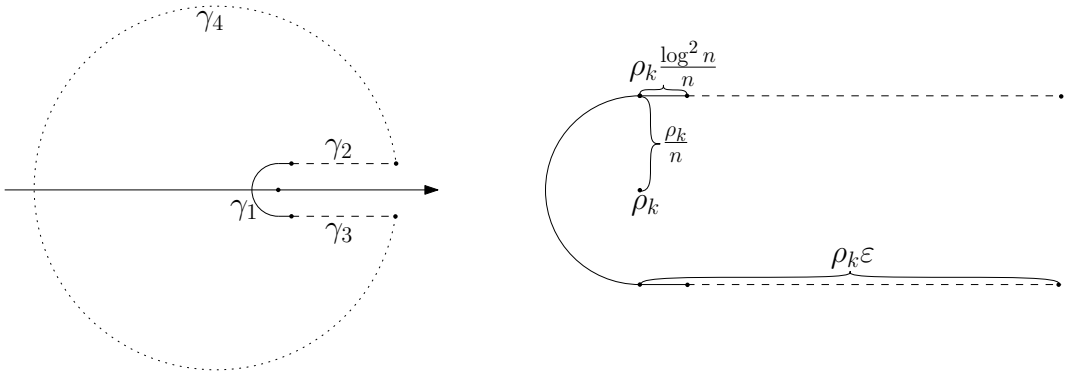


Figure 8.2: Left: Contour of integration: γ_1 plotted with a solid line, γ_2 and γ_3 with dashed lines, and γ_4 with a dotted line. Right: Enlarged truncated Hankel contour.

We start by estimating the integral along γ_1. To this end, we apply the substitution $z = \hat{\rho}_k(1 + t/n)$, where $\widetilde{\gamma_1}$ denotes the transformed curve and we use the expansions given in (8.9):

$$
\begin{aligned}
&\int_{\gamma_1} \frac{z^{k+\ell-n-1}}{\prod_{j=2}^{k+\ell} \sqrt{\hat{R}_{j,k}(z,1)}} \,\mathrm{d}z \\
&= \frac{\hat{\rho}_k^{k+\ell-n}}{n} \int_{\widetilde{\gamma_1}} \Big(1+\frac{t}{n}\Big)^{-n+k+\ell} \left(\frac{1}{\hat{\rho}_k + 2k^{1/4}\hat{\rho}_k^{1/2}\sqrt{-t/n} + \mathcal{O}(|t|/n)}\right)^{\ell} \\
&\qquad\qquad \cdot \prod_{j=2}^{k+1} \frac{1}{\sqrt{c_j}\hat{\rho}_k + g_j\sqrt{-t/n} + \mathcal{O}(|t|/n)} \,\mathrm{d}t \\
&= \frac{\hat{\rho}_k^{k+\ell-n}}{n} \int_{\widetilde{\gamma_1}} e^{-t}\Big(1+\frac{t}{n}\Big)^{k+\ell} \left(\frac{1}{\hat{\rho}_k + 2k^{1/4}\hat{\rho}_k^{1/2}\sqrt{-t/n}}\right)^{\ell} \prod_{j=2}^{k+1} \frac{1}{\sqrt{c_j}\hat{\rho}_k + g_j\sqrt{-t/n}} \Big(1+\mathcal{O}\Big(\frac{|t|}{n}\Big)\Big) \,\mathrm{d}t \\
&= \frac{\hat{\rho}_k^{k-n}}{n} \int_{\widetilde{\gamma_1}} e^{-t}\Big(1+\frac{t}{n}\Big)^{k+\kappa\sqrt{n}} \left(\frac{1}{1 + 2k^{1/4}\hat{\rho}_k^{-1/2}\sqrt{-t/n}}\right)^{\kappa\sqrt{n}} \\
&\qquad\qquad \cdot \prod_{j=2}^{k+1} \frac{1}{\sqrt{c_j}\hat{\rho}_k + g_j\sqrt{-t/n}} \Big(1+\mathcal{O}\Big(\frac{|t|}{n}\Big)\Big) \,\mathrm{d}t \\
&= \frac{\hat{\rho}_k^{k-n}}{n} \int_{\widetilde{\gamma_1}} e^{-t-\frac{2\kappa k^{1/4}}{\sqrt{\hat{\rho}_k}}\sqrt{-t}}\Big(1+\frac{\kappa t}{\sqrt{n}}\Big) \left(1-\frac{2\kappa\sqrt{k}t}{\hat{\rho}_k\sqrt{n}}\right) \prod_{j=2}^{k+1} \frac{1}{\sqrt{c_j}\hat{\rho}_k + g_j\sqrt{-t/n}} \Big(1+\mathcal{O}\Big(\frac{|t|}{n}\Big)\Big) \,\mathrm{d}t \\
&= \frac{\hat{\rho}_k^{-n}}{n\prod_{i=2}^{k+1}\sqrt{c_i}} \int_{\widetilde{\gamma_1}} e^{-t-\frac{2\kappa k^{1/4}}{\sqrt{\hat{\rho}_k}}\sqrt{-t}}\Big(1+\frac{\kappa t}{\sqrt{n}}\Big) \left(1-\frac{2\kappa\sqrt{k}t}{\hat{\rho}_k\sqrt{n}}\right) \left(1-\frac{\sqrt{-t}}{\hat{\rho}_k\sqrt{n}}\sum_{j=1}^{k}\frac{g_j}{\sqrt{c_j}}\right) \Big(1+\mathcal{O}\Big(\frac{|t|}{n}\Big)\Big) \,\mathrm{d}t \\
&= \frac{\hat{\rho}_k^{-n}}{n\prod_{i=2}^{k+1}\sqrt{c_i}} \int_{\widetilde{\gamma_1}} e^{-t-\frac{2\kappa k^{1/4}}{\sqrt{\hat{\rho}_k}}\sqrt{-t}} \left(1+\frac{1}{\hat{\rho}_k\sqrt{n}}\left(\kappa t(\hat{\rho}_k - 2\sqrt{k}) - \sqrt{-t}\sum_{j=1}^{k}\frac{g_j}{\sqrt{c_j}}\right)\right) \Big(1+\mathcal{O}\Big(\frac{|t|}{n}\Big)\Big) \,\mathrm{d}t \\
&= \frac{\hat{\rho}_k^{-n}}{n\prod_{i=2}^{k+1}\sqrt{c_i}} \int_{\widetilde{\gamma_1}} e^{-t-\frac{2\kappa k^{1/4}}{\sqrt{\hat{\rho}_k}}\sqrt{-t}} \Big(1+\mathcal{O}\Big(\frac{|t|}{\sqrt{n}}\Big)\Big) \,\mathrm{d}t.
\end{aligned}
$$

Now, by applying Lemma 8.13, we get that the integral above can be further estimated to result in

$$\begin{aligned}\int_{\gamma_1} \frac{z^{k+\ell-n-1}}{\prod_{j=2}^{k+\ell}\sqrt{\hat{R}_{j,k}(z,1)}}\,\mathrm{d}z &= \frac{1}{n\prod_{i=2}^{k+1}\sqrt{c_i}}\hat{\rho}_k^{-n}\int_{\widehat{\gamma_1}} e^{-t-\frac{2\kappa k^{1/4}}{\sqrt{\hat{\rho}_k}}\sqrt{-t}}\,\mathrm{d}t + \mathcal{O}\Big(\frac{\hat{\rho}_k^{-n}}{n^{3/2}}\Big)\\ &= \frac{\kappa k^{1/4}}{\sqrt{\pi\rho_k}}\frac{1}{n\prod_{i=2}^{k+1}\sqrt{c_i}}\hat{\rho}_k^{-n}\exp\left(-\kappa^2(2k+\sqrt{k})\right) + \mathcal{O}\Big(\frac{\hat{\rho}_k^{-n}}{n^{3/2}}\Big).\end{aligned} \tag{8.10}$$

Next, we show that the integrals along γ_j for $j\in\{2,3,4\}$ are all of order $o\left(\hat{\rho}_k^{-n}n^{-3/2}\right)$ and hence the whole asymptotic contribution comes from integration along γ_1. First, let us consider the integral along γ_4:

$$\begin{aligned}\left|\int_{\gamma_4} \frac{z^{k+\ell-n-1}}{\prod_{j=2}^{k+\ell}\sqrt{\hat{R}_{j,k}(z,1)}}\,\mathrm{d}z\right| &\le \left(\hat{\rho}_k(1+\varepsilon)\right)^{k+\lfloor\kappa\sqrt{n}\rfloor-n-1}|\gamma_4|\max_{z\in\gamma_4}\left|\frac{1}{\prod_{j=2}^{k+\lfloor\kappa\sqrt{n}\rfloor}\sqrt{\hat{R}_{j,k}(z,1)}}\right|\\ &\le C\hat{\rho}_k^{-n}(1+\varepsilon)^{-n}\left(\hat{\rho}_k(1+\varepsilon)\right)^{\lfloor\kappa\sqrt{n}\rfloor}\min_{z\in\gamma_4}\left|\sqrt{\hat{R}_{2,k}(z,1)}\right|^{-\lfloor\kappa\sqrt{n}\rfloor},\end{aligned}$$

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 $\min_{z\in\gamma_4}\left|\sqrt{\hat{R}_{2,k}(z,1)}\right|^{-\lfloor\kappa\sqrt{n}\rfloor}$ and thus guarantees

$$\int_{\gamma_4} \frac{z^{k+\ell-n-1}}{\prod_{j=2}^{k+\ell}\sqrt{\hat{R}_{j,k}(z,1)}}\,\mathrm{d}z = \mathcal{O}\left(\left(\hat{\rho}_k(1-\varepsilon)^{-n}\right)\right) = o\left(\hat{\rho}_k^{-n}n^{-3/2}\right).$$

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

$$\begin{aligned}\left|\int_{\gamma_2} \frac{z^{k+\ell-n-1}}{\prod_{j=2}^{k+\ell}\sqrt{\hat{R}_{j,k}(z,1)}}\,\mathrm{d}z\right| &\le C\left|\int_{\log^2 n}^{\varepsilon n}\frac{\hat{\rho}_k^{\lfloor\kappa\sqrt{n}\rfloor-n}\left(1+\frac{t}{n}+\frac{i}{n}\right)^{\lfloor\kappa\sqrt{n}\rfloor-n}}{\sqrt{\hat{R}_{2,k}\left(\hat{\rho}_k\left(1+\frac{t}{n}+\frac{i}{n}\right),1\right)}^{\lfloor\kappa\sqrt{n}\rfloor}}\frac{1}{n}\,\mathrm{d}t\right|\\ &\le C\hat{\rho}_k^{-n}\frac{1}{n}\hat{\rho}_k^{\lfloor\kappa\sqrt{n}\rfloor}\max_{\gamma_2}\left|\frac{z}{\sqrt{\hat{R}_{2,k}(z,1)}}\right|^{\lfloor\kappa\sqrt{n}\rfloor}\int_{\log^2 n}^{\varepsilon n}\left(1+\frac{t}{n}+\frac{i}{n}\right)^{-n}\mathrm{d}t.\end{aligned}$$

Using the fact that $\left|\sqrt{\hat{R}_{2,k}(z,1)}\right| = \left|z+\sqrt{\hat{R}_{1,k}(z,1)}\right|$ and by Lemma 8.14, we get that the maximum contributes a factor

$$\begin{aligned}\max_{z\in\gamma_2}\left|\frac{z}{\sqrt{\hat{R}_{2,k}(z,1)}}\right|^{\lfloor\kappa\sqrt{n}\rfloor} &= \max_{z\in\gamma_2}\left|\frac{1}{1+\frac{1}{z}\sqrt{\hat{R}_{1,k}(z,1)}}\right|^{\lfloor\kappa\sqrt{n}\rfloor}\\ &= \left(1+\widetilde{C}\frac{\log n}{\sqrt{n}}\right)^{\lfloor\kappa\sqrt{n}\rfloor} \sim e^{\overline{C}\log n},\end{aligned}$$

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

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

which finally gives us

$$\left| \int_{\gamma_2} \frac{z^{k+\ell-n-1}}{\prod_{j=2}^{k+\ell} \sqrt{\hat{R}_{j,k}(z,1)}} \mathrm{d}z \right| = \mathcal{O}\left(\hat{\rho}_k^{-n} \frac{1}{n} e^{-\log^2 n + \overline{C}\log n}\right) = o\left(\hat{\rho}_k^{-n} n^{-3/2}\right).$$

The estimate of the integral along γ_3 works analogously. Therefore, by (8.10), we get

$$[z^n] \frac{\partial U_{k,\ell}(z,u)}{\partial u}\Big|_{u=1} = \frac{\kappa k^{1/4}}{n\sqrt{\pi\hat{\rho}_k} \prod_{i=2}^{k+1} \sqrt{c_i}} \hat{\rho}_k^{-n} \exp\left(-\kappa^2(2k+\sqrt{k})\right) + \mathcal{O}\Big(\frac{\hat{\rho}_k^{-n}}{n^{3/2}}\Big).$$

Combining this result and the asymptotic behavior of the sequence enumerating all terms from $\mathcal{G}_k$, we finally obtain that the expected number of leaves at the level $\lfloor \kappa\sqrt{n} \rfloor$ is given by

$$\begin{aligned} \frac{[z^n] \frac{\partial U_{k,\ell}(z,u)}{\partial u}\Big|_{u=1}}{[z^n] G_k(z)} &\sim \frac{\frac{\kappa k^{1/4}}{\sqrt{\pi\hat{\rho}_k} \prod_{i=2}^{k+1} \sqrt{c_i}} n^{-1} \hat{\rho}_k^{-n} \exp\left(-\kappa^2(2k+\sqrt{k})\right)}{\sqrt{\frac{2k+\sqrt{k}}{4\pi \prod_{j=2}^{k+1} \sqrt{c_j}}} n^{-3/2} \hat{\rho}_k^{-n}} \\ &= 2\kappa \exp\left(-\kappa^2(2k+\sqrt{k})\right) \sqrt{n}. \end{aligned}$$

□

Chapter 9

Lambda terms with bounded number of De Bruijn levels

In this chapter we study parameters of lambda terms with a bounded number of De Bruijn levels, based on the article *Distribution of variables in lambda terms with restrictions on De Bruijn indices and De Bruijn levels*, which was joint work with Bernhard Gittenberger and published in the Electronic Journal of Combinatorics, [57]. A preliminary version of the presented results has been published in the Proceedings of the International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms (AofA2018), [58].

As in the previous chapter, the first part is devoted to proving asymptotic results concerning the total number of variables, while in the second part we investigate the distribution of the variables, as well as applications and abstractions, within the term, thereby providing the asymptotic unary profile.

9.1 Total number of variables

This section is devoted to the enumeration of all variables in closed lambda terms with a bounded number of De Bruijn levels. Analogously to Section 8.1 we translate the specifications (3.17) and (3.18) into bivariate generating functions $P^{(i,k)}(z,u)$, where z marks the size and u the number of leaves. Solving for $P^{(i,k)}(z,u)$ and simplifying yields

$$P^{(i,k)}(z,u) = \frac{1-\sqrt{R_{k-i+1,k}(z,u)}}{2z},$$

where

$$R_{1,k}(z,u) = 1 - 4kz^2u, \tag{9.1}$$

and for $2 \leq i \leq k+1$

$$R_{i,k}(z,u) = 1 - 4(k-i+1)z^2u - 2z + 2z\sqrt{R_{i-1,k}(z,u)}. \tag{9.2}$$

For the bivariate generating function of closed lambda terms with at most k De Bruijn levels we therefore get

$$H_k(z,u) = P^{(0,k)}(z,u) = \frac{1-\sqrt{R_{k+1,k}(z,u)}}{2z}. \tag{9.3}$$

In this section we will prove the following theorem concerning the total number of variables in lambda terms with a bounded number of De Bruijn levels.

Theorem 9.1. *Let $\rho_k(u)$ denote the dominant singularity of the bivariate generating function $H_k(z,u)$ given in (9.3) and let $B(u) := \frac{\rho_k(1)}{\rho_k(u)}$. Furthermore, assume that $B''(1) + B'(1) - B'(1)^2 \neq 0$. Then the total number of variables in a random closed lambda term with at most k De Bruijn levels is asymptotically normally distributed with asymptotic mean μn and asymptotic variance $\sigma^2 n$, where $\mu = B'(1)$ and $\sigma^2 = B''(1) + B'(1) - B'(1)^2$.*

As stated in Theorem 3.18 the type of the dominant singularity of the generating function $H_k(z,1)$ changes when the imposed bound k equals an entry N_j of the sequence $(N_i)_{i\geq 0}$ given in Definition 3.17. Thus, although the result concerning the total number of variables in lambda terms with a bounded number of De Bruijn levels is the same for both cases, whether k is an element of $(N_i)_{i\geq 0}$ or not, the method of proof is different and we will present both approaches in separate subsections.

9.1.1 The case $N_j < k < N_{j+1}$

In this case we can proceed analogously to Section 8.1, since the dominant singularity of the generating function $H_k(z,1)$ comes solely from one radicand, namely from $R_{j+1,k}$ (see Theorem 3.18). Thus, we can again use continuity arguments to guarantee that sufficiently close to $u = 1$ the dominant singularity $\rho_k(u)$ of $H_k(z,u)$ comes from the $(j+1)$-th radicand $R_{j+1,k}(z,u)$ and is of type $\frac{1}{2}$. Now we will determine the expansions of the radicands in a neighborhood of the dominant singularity.

Proposition 9.2. *Let $\rho_k(u)$ be the dominant singularity of $H_k(z,u)$, where u is in a sufficiently small neighbourhood of 1, i.e. $|u-1| < \delta$ for $\delta > 0$ sufficiently small. Then the expansions*

(i) $\forall i < j+1$ *(inner radicands)* : $R_{i,k}(\rho_k(u) - \epsilon, u) = R_{i,k}(\rho_k(u), u) + \mathcal{O}(|\epsilon|)$

(ii) $R_{j+1,k}(\rho_k(u) - \epsilon, u) = \gamma_{j+1}(u)\epsilon + \mathcal{O}(|\epsilon|^2)$, *with* $\gamma_{j+1}(u) = -\frac{\partial}{\partial z} R_{j+1,k}(\rho_k(u), u)$

(iii) $\forall i > j+1$ *(outer radicands)* : $R_{i,k}(\rho_k(u) - \epsilon, u) = a_i(u) + b_i(u)\sqrt{\epsilon} + \mathcal{O}(|\epsilon|)$, *with*

$$\begin{aligned} a_{j+2}(u) &= 1 - 4(k-j-1)\rho_k(u)^2 u - 2\rho_k(u), \\ a_{i+1}(u) &= 1 - 4(k-i)\rho_k(u)^2 u - 2\rho_k(u) + 2\rho_k(u)\sqrt{a_i(u)}, \quad \text{for } j+2 \leq i \leq k, \end{aligned}$$

and

$$\begin{aligned} b_{j+2}(u) &= 2\rho_k(u)\sqrt{\gamma_{j+1}(u)}, \\ b_{i+1}(u) &= \frac{b_i(u)\rho_k(u)}{\sqrt{a_i(u)}} \quad \text{for } j+2 \leq i \leq k, \end{aligned}$$

hold for $\epsilon \to 0$ so that $\epsilon \in \mathbb{C} \setminus \mathbb{R}^-$, uniformly in u.

Proof. The proof works analogously to the proof of Proposition 8.1.

(i) The first statement (for $i < j+1$) follows immediately by Taylor expansion around $\rho_k(u)$ and setting $z = \rho_k(u) - \epsilon$.

(ii) The equation for $i = j + 1$ follows analogously to the first case, knowing that $R_{j+1,k}(z, u)$ cancels for $z = \rho_k(u)$.

(iii) The next step is to expand $R_{i,k}(z, u)$ around $\rho_k(u)$ for $i > j+1$. From the second claim of Proposition 9.2 and from the recurrence relation (9.2) for $R_{i,k}(z, u)$ it results

$$R_{j+2,k}(\rho_k(u)-\epsilon, u) = 1-4(k-j-1)\rho_k(u)^2u-2\rho_k(u)+2\rho_k(u)\sqrt{\gamma_{j+1}(u)}\sqrt{\epsilon}+\mathcal{O}(|\epsilon|).$$

We set $a_{j+2}(u) := 1-4(k-j-1)\rho_k^2(u)u-2\rho_k(u)$ and $b_{j+2}(u) := 2\rho_k(u)\sqrt{\gamma_{j+1}(u)}$. Now we proceed by induction. Assume

$$R_{i,k}(\rho_k(u) - \epsilon, u) = a_i(u) + b_i(u)\sqrt{\epsilon} + \mathcal{O}(|\epsilon|). \tag{9.4}$$

We have just checked that it holds for $i = j + 2$. Now we perform the induction step $i \mapsto i + 1$. Using the recursion (9.2) for $R_{i,k}$ and plugging in the induction hypothesis (9.4) yields

$$\begin{aligned} R_{i+1,k}(\rho_k(u) - \epsilon, u) = 1 - 4(k - i)\rho_k(u)^2u - 2\rho_k(u) \\ + 2\rho_k(u)\sqrt{a_i(u)} + \frac{b_i(u)\rho_k(u)}{\sqrt{a_i(u)}}\sqrt{\epsilon} + \mathcal{O}(|\epsilon|). \end{aligned}$$

Setting $a_{i+1}(u) := 1 - 4(k - i)\rho_k^2(u)u - 2\rho_k(u) + 2\rho_k(u)\sqrt{a_i(u)}$ and $b_{i+1}(u) := \frac{b_i(u)\rho_k(u)}{\sqrt{a_i(u)}}$ for $i \geq j + 2$ we obtain

$$R_{i+1,k}(\rho_k(u) - \epsilon, u) = a_{i+1}(u) + b_{i+1}(u)\sqrt{\epsilon} + \mathcal{O}(|\epsilon|).$$

Expanding $b_i(u)$, using its recursive relation and $b_{j+2}(u) = 2\rho_k(u)\sqrt{\gamma_{j+1}(u)}$ we get for $i > j + 1$

$$b_i(u) = \frac{2\rho_k^{i-j}(u)\sqrt{\gamma_{j+1}(u)}}{\prod_{l=j+1}^{i-1}\sqrt{a_l(u)}}.$$

□

We know that for sufficiently large i the sequence u_i, defined in Definition 3.17, is given by $u_i = \lfloor \chi^{2^i} \rfloor$, with $\chi \approx 1.36660956\ldots$ (see [11, Lemma 18]). Therefore we have $N_j \sim u_j^2 \sim \chi^{2^{j2}}$ and $N_j < k < N_{j+1} = O(N_j^2)$, which gives $j = \Theta(\log\log k)$. Taking a look at values of j corresponding to the initial values of $k = 1, \ldots, 135$, which read as

k	1	2	…	7	8	9	…	134	135
j	1	1	…	1	2	2	…	2	3

we can deduce that $j + 1 < k + 1$, *i.e.*, that the dominant singularity $\rho_k(u)$ cannot come from the outermost radical.

Remark 9.3. *Obviously the same is true for the case $k = N_j$. Thus, the dominant singularity never comes from the outermost radical.*

Theorem 9.4. *Let for any fixed k, $H_k(z,u)$ denote the bivariate generating function of the class of closed lambda terms with at most k De Bruijn levels. Furthermore, let $N_j < k < N_{j+1}$, where N_i is defined in Definition 3.17. Then for $n \to \infty$ the equation*

$$[z^n]H_k(z,u) = h_k(u)\rho_k(u)^{-n}\frac{n^{-\frac{3}{2}}}{\Gamma(-\frac{1}{2})}\left(1+O\left(\frac{1}{\sqrt{n}}\right)\right), \tag{9.5}$$

with

$$h_k(u) = -\frac{b_{k+1}(u)\sqrt{\rho_k(u)}}{4\rho_k(u)\sqrt{a_{k+1}(u)}} \neq 0,$$

where $a_i(u)$ and $b_i(u)$ are defined as in Proposition 9.2, holds uniformly in u for $|u-1| < \delta$, with $\delta > 0$ sufficiently small.

Proof. Using Proposition 9.2 and $H_k(z,u) = \frac{1}{2z}(1-\sqrt{R_{k+1,k}(z,u)})$ we get

$$H_k(\rho_k(u)-\epsilon,u) = \frac{1-\sqrt{a_{k+1}(u)}}{2\rho_k(u)} - \frac{b_{k+1,k}(u)}{4\rho_k(u)\sqrt{a_{k+1}(u)}}\sqrt{\epsilon} + \mathcal{O}(|\epsilon|),$$

which directly yields Equation (9.5) by applying a transfer theorem (see Theorems 2.8 and 2.10).

Now we show that $h_k(u) \neq 0$ in a sufficiently small neighborhood of $u = 1$: First, let us consider that $a_{j+2} := a_{j+2}(1)$ is positive, since

$$a_{j+2} = 1-4(k-j-1)\rho_k(1)^2 - 2\rho_k(1) = 1-4(k-j)\rho_k(1)^2 - 2\rho_k(1) + 4\rho_k^2,$$

and $1-4(k-j)\rho_k(1)^2 - 2\rho_k(1) > 0$ (see [11]). Now we show by induction that the sequence $a_i := a_i(1)$ is monotonically increasing. Let us assume that $a_{i-1} < a_i$, then we get

$$\begin{aligned} a_{i+1} &> 1-4(k-i)\rho_k(1)^2 - 2\rho_k(1) + 2\rho_k(1)\sqrt{a_i} \\ &> 1-4(k-i+1)\rho_k(1)^2 - 2\rho_k(1) + 2\rho_k(1)\sqrt{a_{i-1}} + 4\rho_k(1)^2 > a_i + 4\rho_k(1)^2. \end{aligned}$$

Moreover, it is obvious that if $b_{j+2} := b_{j+2}(1)$ is non-zero, than all the b_i's, which are defined via

$$b_i = \frac{\rho_k(1)b_{i-1}}{a_{i-1}},$$

are non-zero as well. In order to prove that $b_{j+2} = 2\rho_k(1)\sqrt{-\frac{\partial}{\partial z}R_{j+1,k}(\rho_k(1),1)}$ is non-zero, we also proceed by induction. Since

$$R_{1,k}(z,1) = 1-4kz^2,$$

we can see that $\frac{\partial}{\partial z}R_{1,k}(\rho_k(1),1) < 0$. Assuming $\frac{\partial}{\partial z}R_{i,k}(\rho_k(1),1) < 0$ and using

$$\frac{\partial}{\partial z}R_{i+1,k}(z,1) = -8(k-i)z - 2 + 2\sqrt{R_{i,k}(z,1)} + \frac{z}{\sqrt{R_{i,k}(z,1)}}\frac{\partial}{\partial z}R_i(z,1),$$

we proved that all b_i's are non-zero. Thus, by continuity arguments it follows that $h_k(u) \neq 0$ in a sufficiently small neighborhood of $u = 1$. □

Now, using (9.5) and Theorem 3.19 we get for $n \to \infty$

$$\frac{[z^n]H_k(z,u)}{[z^n]H_k(z,1)} = \frac{h_k(u)}{h_k\Gamma(-1/2)} \left(\frac{\rho_k(1)}{\rho_k(u)}\right)^n \left(1 + O\left(\frac{1}{n}\right)\right). \tag{9.6}$$

Assuming that $\sigma^2 := B''(1) + B'(1) - B'(1)^2 \neq 0$ with $B(u) = \frac{\rho_k(1)}{\rho_k(u)}$ we can finally apply the Quasi-Powers Theorem (Theorem 2.18). Unfortunately, the proof of this assumption appears to be quite difficult, since there is only very little known about the function $\rho_k(u)$. However, numerical data supports the conjecture that this condition will be fulfilled for arbitrary k (see Table 9.1), which allows for the use of the Quasi-Powers Theorem (Theorem 2.18) that then directly yields Theorem 9.1 for the case that $k \in (N_j, N_{j+1})$.

bound k	$j+1$	$B''(1) + B'(1) - B'(1)^2$	$B'(1)$
1	**2**	**0**	**0**
2	2	0.0385234386	0.4381229337
3	2	0.0210625856	0.4414407371
4	2	0.0167136805	0.4463973717
5	2	0.0148700270	0.4504258849
6	2	0.0138224393	0.4536185043
7	2	0.0131157948	0.4561987871
8	**3**	**0.0125868052**	**0.4583333333**
9	3	0.0582322465	0.4566104777
10	3	0.0470481360	0.4560418340
11	3	0.0396601986	0.4560810348
12	3	0.0345090124	0.4564489368
⋮	⋮	⋮	⋮
133	3	0.0077469541	0.4821900098
134	3	0.0077234960	0.4822482745
135	**4**	**0.0077002803**	**0.4823059361**
136	4	0.0132855719	0.4823515285
137	4	0.0131816901	0.4823968564
138	4	0.0130800422	0.4824419195
139	4	0.0129805564	0.4824867175

Table 9.1: Table summarizing the coefficients occurring in the variance and the mean for some initial values of k, where the cases $k = N_j$ are bold.

9.1.2 The case $k = N_j$

We know from Theorem 3.18 that in the case $k = N_j$ both radicands $R_{j,k}(z,1)$ and $R_{j+1,k}(z,1)$ vanish simultaneously and the dominant singularity is therefore of type $\frac{1}{4}$. This is not true for the radicands $R_{j,k}(z,u)$ and $R_{j+1,k}(z,u)$ when u is in a neighborhood of 1. Thus, we have a discontinuity at $\rho_k(1)$, which is why we do not get any uniform expansions of the radicands in a neighborhood of $\rho_k(1)$ and cannot use the same approach as in the previous section.
In order to overcome this problem we proceed as follows (see Figure 9.1 for a sketch of the idea of the proof): First, we show that the dominant singularity of the generating function $H_k(z, 1+\epsilon)$ comes solely from the radicand $R_{j,k}(z, 1+\epsilon)$ (*cf.* Lemma 9.5). Then we investigate the expansions of the radicands thoroughly for $u = 1 + \frac{s}{\sqrt{n}}$ in a

neighborhood of $z = \rho_k(1)$ with radius $\frac{t}{n}$, where s and t are both bounded complex numbers (*cf.* Lemma 9.6). This approach of choosing the considered neighborhoods of $z = \rho_k(1)$ and $u = 1$ to be dependent on each other constitutes the main idea of the applied method. By use of Cauchy's coefficient formula we are then able to obtain an asymptotic expression for the n-th coefficient of the generating function $H_k\left(z, 1 + \frac{s}{\sqrt{n}}\right)$ by choosing a suitable integration contour (*cf.* Proposition 9.7). Finally, we show that the characteristic function of the random variable counting the total number of variables in a random lambda term with at most k De Bruijn levels tends to the characteristic function of the normal distribution as the size tends to infinity (*cf.* Lemma 9.8). For convenience, we will subsequently use the abbreviation $\rho_k := \rho_k(1)$.

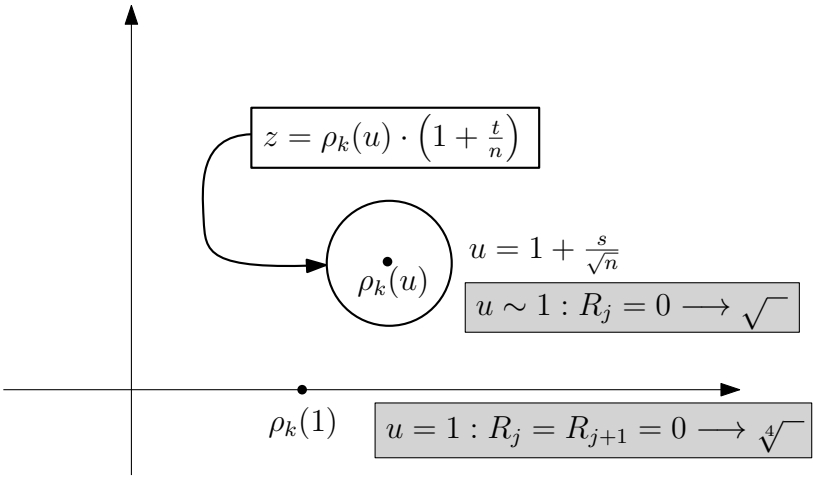


Figure 9.1: Sketch of the idea of the proof.

Lemma 9.5. *For $u = 1 + \epsilon$ with $\epsilon \to 0$ so that $\epsilon \in \mathbb{C} \setminus \mathbb{R}_0^-$, the dominant singularity $\rho_k(u) = \rho_k(1+\epsilon)$ of the bivariate generating function $H_k(z, 1+\epsilon)$ comes from the j-th radicand $R_{j,k}(z, u)$.*

Proof. Setting $u = 1 + \epsilon$, expanding $\rho_k(u)$ around 1 and plugging this expansion into the recursive definition of the radicands yields

$$\begin{aligned} R_{j,k}\left(\rho_k(1+\epsilon), 1+\epsilon\right) =& 1 - 4(k-j+1)(\rho_k^2 + 2\rho_k\rho_k'\epsilon + (\rho_k'^2 + 2\rho_k\rho_k'')\epsilon^2 + \rho_k^2\epsilon) \\ & - (2\rho_k + 2\rho_k'\epsilon + 2\rho_k''\epsilon^2)\left(1 - \sqrt{R_{j-1,k}\left(\rho_k(1+\epsilon), 1+\epsilon\right)}\right) \\ & + \mathcal{O}(|\epsilon|). \end{aligned}$$

Using $\sqrt{R_{j-1,k}\left(\rho_k(1+\epsilon), 1+\epsilon\right)} = \sqrt{R_{j-1,k}\left(\rho_k, 1\right) + \mathcal{O}(|\epsilon|)} = 2\rho_k + \mathcal{O}(|\epsilon|)$ and $1 - 4(k-j)\rho_k^2 - 2\rho_k = 0$, which are both shown in [11], we get

$$\begin{aligned} R_{j,k}\left(\rho_k(1+\epsilon), 1+\epsilon\right) =& -4(k-j+1)(2\rho_k\rho_k'\epsilon + (\rho_k'^2 + 2\rho_k\rho_k'')\epsilon^2) \\ & - (2\rho_k'\epsilon + 2\rho_k''\epsilon^2)\left(1 - 2\rho_k + \mathcal{O}(|\epsilon|)\right) + \mathcal{O}(|\epsilon|). \end{aligned}$$

Thus, $R_{j,k}\left(\rho_k(1+\epsilon), 1+\epsilon\right) = \Theta(|\epsilon|)$. Using this result and again the recursive definition of the radicands results in

$$R_{j+1,k}\left(\rho_k(u), 1+\epsilon\right) = 2\sqrt{R_{j,k}(\rho_k(u), 1+\epsilon)} + \mathcal{O}\left(|\epsilon|\right) = \Theta(\sqrt{|\epsilon|}).$$

Thus, we see that $|R_{j+1,k}(\rho_k(u),u)| \gg |R_{j,k}(\rho_k(u),u)|$ in a neighborhood of $u=1$, which implies that the dominant singularity has to come from the j-th radicand, *i.e.*, $R_{j,k}(\rho_k(u),u)=0$ for u being sufficiently close to 1. □

Lemma 9.6. *Let $z=\rho_k(u)=\rho_k(1+\frac{s}{\sqrt{n}})$ be the dominant singularity of the bivariate generating function $H_k(z,1+\frac{s}{\sqrt{n}})$ with bounded $s\in\mathbb{C}$. Then, as $n\to\infty$,*

(i) $R_{j,k}\Big(\rho_k(u)\left(1+\frac{t}{n}\right),1+\frac{s}{\sqrt{n}}\Big)=\frac{1}{n}p_j(t)+\mathcal{O}\left(\frac{1}{n^{3/2}}\right)$,

with $p_j(t):=-8t(k-j+1)\rho_k^2-2\rho_k t+4\rho_k^2 t+2t\rho_k f(\frac{t}{n})$ where f is an analytic function around 0;

(ii) $R_{j+1,k}\Big(\rho_k(u)\left(1+\frac{t}{n}\right),1+\frac{s}{\sqrt{n}}\Big)=\frac{1}{\sqrt{n}}p_{j+1}(s,t)+\mathcal{O}\left(\frac{1}{n}\right)$,

where $p_{j+1}(s,t)=2\rho_k\sqrt{p_j(t)}-4(k-j)(2\rho_k\rho_k' s+\rho_k^2 s)-2\rho_k' s$;

(iii) $R_{i,k}\Big(\rho_k(u)\left(1+\frac{t}{n}\right),1+\frac{s}{\sqrt{n}}\Big)=\hat{C}_i+\frac{1}{\sqrt[4]{n}}p_i(s,t)+O\left(\frac{1}{\sqrt{n}}\right)\quad$ *for $i\geq j+2$,*

where $\hat{C}_i$ are constants and p_i are analytic functions in the variables s and t.

Proof. We start with setting $u=1+\frac{s}{\sqrt{n}}$ and $z=\rho_k(u)(1+\frac{t}{n})$ with bounded $s,t\in\mathbb{C}$ (*cf.* Figure 9.1), which results in

$$R_{j+1,k}\left(\rho_k(u)\left(1+\frac{t}{n}\right),1+\frac{s}{\sqrt{n}}\right)=$$
$$1-4(k-j)\rho_k(u)^2\left(1+\frac{t}{n}\right)^2\left(1+\frac{s}{\sqrt{n}}\right)-2\rho_k(u)\left(1+\frac{t}{n}\right)\left(1-\sqrt{R_{j,k}}\right),$$

and

$$R_{j,k}\left(\rho_k(u)\left(1+\frac{t}{n}\right),1+\frac{s}{\sqrt{n}}\right)=$$
$$1-4(k-j+1)\rho_k(u)^2\left(1+\frac{t}{n}\right)^2\left(1+\frac{s}{\sqrt{n}}\right)-2\rho_k(u)\left(1+\frac{t}{n}\right)\left(1-\sqrt{R_{j-1,k}}\right),$$

where the radicand in the square root in the last bracket of both equations is of course also evaluated at $(z,u)=\Big(\rho_k(1+\frac{s}{\sqrt{n}})(1+\frac{t}{n}),1+\frac{s}{\sqrt{n}}\Big)$, but we will omit this notation from now on to ensure a simpler reading, *i.e.*, subsequently we will write $R_{i,k}$ instead of $R_{i,k}\Big(\rho_k(1+\frac{s}{\sqrt{n}})\left(1+\frac{t}{n}\right),1+\frac{s}{\sqrt{n}}\Big)$. Expanding $\rho_k(1+\frac{s}{\sqrt{n}})$ around 1 and using the recursive definition for the radicands yields

$$R_{j,k}=1-4(k-j+1)\left(\rho_k^2+2\rho_k\rho_k'\frac{s}{\sqrt{n}}+(\rho_k'^2+2\rho_k\rho_k'')\frac{s^2}{n}+\rho_k^2\frac{s}{\sqrt{n}}+2\rho_k\rho_k'\frac{s^2}{n}+\rho_k^2\frac{2t}{n}\right)$$
$$-2\left(\rho_k+\rho_k'\frac{s}{\sqrt{n}}+\rho_k''\frac{s^2}{2n}+\rho_k\frac{t}{n}\right)\left(1-\sqrt{R_{j-1,k}}\right)+\mathcal{O}\left(\frac{1}{n^{3/2}}\right).\qquad(9.7)$$

From Lemma 9.5 we know that for u in a sufficiently small vicinity of 1 the dominant singularity of $H_k(z,u)$ comes from the j-th radicand, *i.e.* $R_{j,k}\left(\rho_k(u),u\right)=0$.

Expanding $R_{j,k}\left(\rho_k(1+\frac{s}{\sqrt{n}}), 1+\frac{s}{\sqrt{n}}\right)$ this yields

$$1 - 4(k-j+1)\left(\rho_k^2 + 2\rho_k\rho_k'\frac{s}{\sqrt{n}} + (\rho_k'^2 + 2\rho_k\rho_k'')\frac{s^2}{n} + \rho_k^2\frac{s}{\sqrt{n}} + 2\rho_k\rho_k'\frac{s^2}{n}\right)$$
$$- 2\left(\rho_k + \rho_k'\frac{s}{\sqrt{n}} + \rho_k''\frac{s^2}{2n}\right)\left(1 - \sqrt{R_{j-1,k}\left(\rho_k\left(1+\frac{s}{\sqrt{n}}\right), 1+\frac{s}{\sqrt{n}}\right)}\right)$$
$$+ \mathcal{O}\left(\frac{1}{n^{3/2}}\right) = 0.$$

Thus, Equation (9.7) simplifies to

$$R_{j,k} = -4(k-j+1)\rho_k^2\frac{2t}{n} - 2\rho_k\frac{t}{n} + 4\rho_k^2\frac{t}{n} + 2\rho_k\frac{t}{n}f\left(\frac{t}{n}\right) + \mathcal{O}\left(\frac{1}{n^{3/2}}\right), \tag{9.8}$$

where $\frac{t}{n}f\left(\frac{t}{n}\right) = \sqrt{R_{j-1,k}} - \sqrt{R_{j-1,k}\left(\rho_k(1+\frac{s}{\sqrt{n}}), 1+\frac{s}{\sqrt{n}}\right)}$. Notice that $f(x)$ is analytic around $x = 0$. Therefore, the proof of (i) is finished.
Proceeding equivalently for $R_{j+1,k}$ results in

$$R_{j+1,k} = \frac{1}{\sqrt{n}}\Big(-4(k-j)(2\rho_k\rho_k's + \rho_k^2 s) - 2\rho_k' s\Big) + 2\rho_k\sqrt{R_{j,k}} + O\left(\frac{1}{n}\right).$$

Finally, to complete the proof of the second statement of the assertion we simply have to replace $R_{j,k}$ by the right-hand side of (9.8). Going one step further leads to

$$R_{j+2,k} = \hat{C}_{j+2} + \frac{1}{\sqrt[4]{n}}p_{j+2}(s,t) + O\left(\frac{1}{\sqrt{n}}\right),$$

with $\hat{C}_{j+2} := 4\rho_k^2$ and $p_{j+2}(s,t) := 2\rho_k\sqrt{p_{j+1}(s,t)}$, where $p_{j+1}(s,t)$ is defined as in Lemma 9.6. Now we proceed by induction. Therefore we assume that $R_{i,k} = \hat{C}_i + \frac{1}{\sqrt[4]{n}}p_i(s,t) + O\left(\frac{1}{\sqrt{n}}\right)$ with $i \geq j+2$. Thus, we get

$$R_{i+1,k} = 1 - 4(k-i)\left(\rho_k^2 + 2\rho_k\rho_k'\frac{s}{\sqrt{n}} + (\rho_k'^2 + 2\rho_k\rho_k'')\frac{s^2}{n} + \rho_k^2\frac{s}{\sqrt{n}} + 2\rho_k\rho_k'\frac{s^2}{n} + \rho_k^2\frac{2t}{n}\right)$$
$$- 2\left(\rho_k + \rho_k'\frac{s}{\sqrt{n}} + \rho_k''\frac{s^2}{2n} + \rho_k\frac{t}{n}\right)\left(1 - \sqrt{R_{i,k}}\right) + \mathcal{O}\left(\frac{1}{n^{3/2}}\right).$$

Inserting the induction hypothesis and simplifying yields

$$R_{i+1,k} = 4(i-j)\rho_k^2 + 2\rho_k\sqrt{\hat{C}_i} + \frac{1}{\sqrt[4]{n}}\frac{\rho_k p_i(s,t)}{\sqrt{\hat{C}_i}} + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

Setting $\hat{C}_{i+1} := 4(i-j)\rho_k^2 + 2\rho_k\sqrt{\hat{C}_i}$ and $p_{i+1}(s,t) := \frac{\rho_k}{\sqrt{\hat{C}_i}}p_i(s,t)$ completes the proof. □

Proposition 9.7. *Let $H_k(z,u)$ be the bivariate generating function of the class of closed lambda terms with at most k De Bruijn levels. Then the n-th coefficient of $H_k(z, 1+\frac{s}{\sqrt{n}})$ with bounded $s \in \mathbb{C}$ is given by*

$$[z^n]H_k(z, 1+\frac{s}{\sqrt{n}}) = C_k(s)\rho_k^{-n}n^{-\frac{5}{4}}\left(1 + O\left(n^{-\frac{3}{4}}\right)\right), \quad \text{as } n \to \infty,$$

with a constant $C_k(s) \neq 0$.

Proof. Let us remember that $H_k(z, 1+\frac{s}{\sqrt{n}}) = \frac{1-\sqrt{R_{k+1,k}(z,1+\frac{s}{\sqrt{n}})}}{2z}$. Thus, with the well-known Cauchy coefficient formula we get

$$\begin{aligned}[z^n]H_k\left(z, 1+\frac{s}{\sqrt{n}}\right) &= \frac{1}{2i\pi}\int_\gamma \frac{H_k\left(z, 1+\frac{s}{\sqrt{n}}\right)}{z^{n+1}}\mathrm{d}z \\ &= \frac{1}{2i\pi}\int_\gamma \frac{1-\sqrt{R_{k+1,k}\left(z, 1+\frac{s}{\sqrt{n}}\right)}}{2z^{n+2}}\mathrm{d}z,\end{aligned}$$

where γ encircles the dominant singularity $\rho_k(u)$ as depicted in Figure 9.2. We denote the small Hankel-like part of the integration contour γ that contributes the main part of the asymptotics by γ_H (*cf.* Figure 9.2). The curve γ_H encircles $\rho_k(u)$ at a distance $\frac{1}{n}$ and its straight parts (that lead into the direction $\rho_k(u)\cdot\infty$) have the length $\frac{\log^2(n)}{n}$. On $\gamma \setminus \gamma_H$ we have $|z| = |\rho_k(u)|\left|1+\frac{\log^2(n)}{n}+\frac{i}{n}\right|$. This enables use to estimate the contribution of $\gamma \setminus \gamma_H$, which turns out to be exponentially small. Next, we use the transformation $z = \rho(u)\left(1+\frac{t}{n}\right)$, which changes the integration contour γ_H to $\tilde{\gamma}_H$. On the new contour $\tilde{\gamma}_H$ the integrand is now represented in a way that Lemma 9.6 becomes directly applicable. Summarizing all those arguments, we know now that there exisits a $K > 0$ such that

$$\begin{aligned}&[z^n]H_k\left(z, 1+\frac{s}{\sqrt{n}}\right) \\ &\quad= \frac{1}{2i\pi}\int_{\tilde{\gamma}_H} \frac{1-\sqrt{\hat{C}_{k+1}+\frac{1}{\sqrt[4]{n}}p_{k+1}(s,t)+\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)}}{2\rho^{n+1}e^t n}\mathrm{d}t + \mathcal{O}\left(e^{-K\log^2(n)}\right) \\ &\quad= \frac{1}{2i\pi}\int_{\tilde{\gamma}_H} \frac{1-\sqrt{\hat{C}_{k+1}}-\frac{1}{2\sqrt[4]{n}\sqrt{\hat{C}_{k+1}}}p_{k+1}(s,t)+\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)}{2\rho_k^{n+1}e^t n}\mathrm{d}t + \mathcal{O}\left(e^{-K\log^2(n)}\right).\end{aligned} \tag{9.9}$$

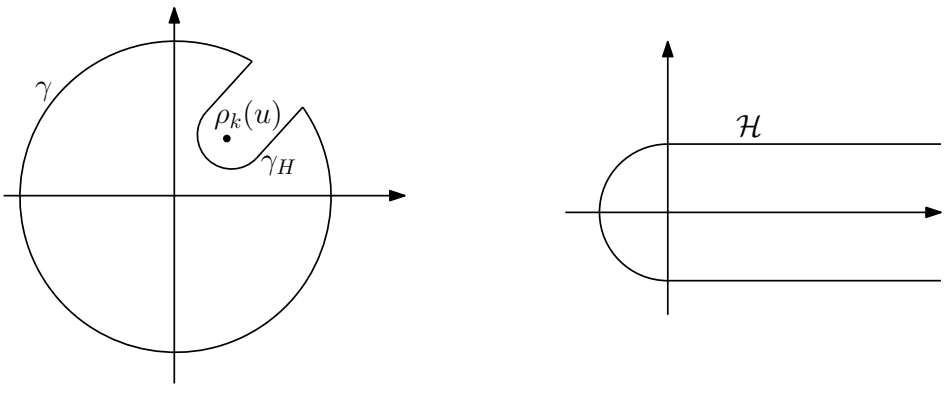


Figure 9.2: The integral contours γ and $\mathcal{H}$.

Now, let us observe how the function $p_{k+1}(s,t)$ looks like by using the recursive definition $p_{i+1}(s,t) = \frac{\rho_k}{\sqrt{\hat{C}_i}}p_i(s,t)$ and $p_{j+2}(s,t) = 2\rho_k\sqrt{2\rho_k\sqrt{p_j(t)}+q(s)}$, with a polynomial $q(s) = s\left(-4(k-j)(2\rho_k\rho_k'+\rho_k^2)-2\rho_k'\right)$ that is linear in s. Thus, $p_{k+1}(s,t) =$

$D \cdot p_{j+2}(s,t)$ with a constant D. Inserting this into (9.9) and splitting the integral yields

$$[z^n]H_k\left(z, 1+\frac{s}{\sqrt{n}}\right) = \frac{\rho_k^{-n}}{4i\pi\rho_k n}\cdot\Bigg(\int_{\tilde{\gamma}_H}\left(1-\sqrt{\hat{C}_{k+1}}\right)e^{-t}\mathrm{d}t \\ - \int_{\tilde{\gamma}_H}\frac{De^{-t}}{2\sqrt[4]{n}\sqrt{\hat{C}_{k+1}}}\sqrt{2\rho_k\sqrt{p_j(t)}+q(s)}\mathrm{d}t \\ + \int_{\tilde{\gamma}_H}\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)e^{-t}\mathrm{d}t\Bigg).$$

The first integral is zero and the third integral contributes $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$. Thus, the main part of the asymptotics results from the second integral: There are some constants $A(s)$ and $B(s)$ such that

$$\begin{aligned}
&- \int_{\tilde{\gamma}_H}\frac{De^{-t}}{\sqrt[4]{n}\sqrt{\hat{C}_{k+1}}}\sqrt{2\rho_k\sqrt{p_j(t)}+q(s)}\mathrm{d}t \\
&\quad = -\int_{\tilde{\gamma}_H}\frac{De^{-t}}{\sqrt[4]{n}\sqrt{\hat{C}_{k+1}}}\sqrt{A(s)t+B(s)+\mathcal{O}\left(\frac{\log^4(n)}{n}\right)}\mathrm{d}t \\
&\quad = -\int_{\tilde{\gamma}_H}\frac{De^{-t}}{\sqrt[4]{n}\sqrt{\hat{C}_{k+1}}}\sqrt{A(s)t+B(s)}\mathrm{d}t + \mathcal{O}\left(\frac{\log^6(n)}{n}\right) \\
&\quad = -\int_{\mathcal{H}}\frac{De^{-t}}{\sqrt[4]{n}\sqrt{\hat{C}_{k+1}}}\sqrt{A(s)t+B(s)}\mathrm{d}t + \mathcal{O}\left(e^{-\tilde{K}\log^2(n)}\right) \\
&\quad \sim \tilde{C}(s)\frac{1}{\sqrt[4]{n}}.
\end{aligned}$$

Here $\tilde{K}$ denotes a suitable positive constant, and $\mathcal{H}$ denotes the classical Hankel curve, *i.e.*, the noose-shaped curve that winds around 0 and starts and ends at $+\infty$ (*cf.* Figure 9.2).

Finally, using this result we get

$$[z^n]H_k\left(z, 1+\frac{s}{\sqrt{n}}\right) = C(s)\rho_k\left(1+\frac{s}{\sqrt{n}}\right)^{-n} n^{-5/4}\left(1+\mathcal{O}\left(\frac{1}{\sqrt[4]{n}}\right)\right), \quad \text{for } n\to\infty,$$

with a constant $C(s)$ that depends on s. □

Now we show that the characteristic function of our standardized sequence of random variables tends to the characteristic function of the normal distribution.

Lemma 9.8. *Let X_n be the total number of variables in a random lambda term with at most k De Bruijn levels. Set $\sigma^2 := -\frac{\rho_k'(1)}{\rho_k(1)} - \frac{\rho_k''(1)}{\rho_k(1)} + \frac{\rho_k'(1)^2}{\rho_k(1)^2}$. If $\sigma^2 \neq 0$, then*

$$Z_n = \frac{X_n - \mathbb{E}(X_n)}{\sqrt{n}} \to \mathcal{N}\left(0, \sigma^2\right).$$

Proof. For the standardised sequence of random variables Z_n we have with $\mu := \frac{\mathbb{E}(X_n)}{n}$

$$Z_n = \frac{X_n - \mathbb{E}(X_n)}{\sqrt{n}} = \frac{X_n}{\sqrt{n}} - \mu\sqrt{n}.$$

Its characteristic function reads as

$$\phi_{Z_n}(s) = \mathbb{E}(e^{isZ_n}) = e^{-is\mu\sqrt{n}}\phi_{X_n}\left(\frac{s}{\sqrt{n}}\right) = e^{-is\mu\sqrt{n}}\mathbb{E}(e^{\frac{isX_n}{\sqrt{n}}}) = e^{-is\mu\sqrt{n}}\frac{[z^n]H_k(z, e^{\frac{is}{\sqrt{n}}})}{[z^n]H_k(z,1)}.$$

From Proposition 9.7 we know

$$\frac{[z^n]H_k(z, 1+\frac{s}{\sqrt{n}})}{[z^n]H_k(z,1)} \sim C(s)\left(\frac{\rho_k(1+\frac{s}{\sqrt{n}})}{\rho_k(1)}\right)^{-n},$$

where the constant $C(s) \sim 1$ for $n \to \infty$. Thus,

$$\begin{aligned}
\phi_{Z_n}(s) &= e^{-is\mu\sqrt{n}}\frac{[z^n]H_k(z, e^{\frac{is}{\sqrt{n}}})}{[z^n]H_k(z,1)} \sim e^{-is\mu\sqrt{n}}\left(\frac{\rho_k\left(1+\frac{si}{\sqrt{n}}-\frac{s^2}{2n}+\mathcal{O}\left(\frac{|s^3|}{n^{3/2}}\right)\right)}{\rho_k(1)}\right)^{-n} \\
&= e^{-is\mu\sqrt{n}}\exp\left(-n\cdot\left(\log\left(1+\frac{\rho_k' is}{\rho_k\sqrt{n}}-\frac{s^2}{2n}\frac{\rho_k'}{\rho_k}+\frac{s^2}{2n}\frac{\rho_k''}{\rho_k}\right)+\mathcal{O}\left(\frac{|s^3|}{n^{3/2}}\right)\right)\right) \\
&\sim e^{-is\mu\sqrt{n}}e^{-is\sqrt{n}\frac{\rho_k'}{\rho_k}}e^{\frac{s^2}{2}\left(\frac{\rho_k'}{\rho_k}+\frac{\rho_k''}{\rho_k}-\frac{\rho_k'^2}{\rho_k^2}\right)}.
\end{aligned}$$

Since we know that the expected value of the standardised random variable is zero, we get $\mu = -\frac{\rho_k'(1)}{\rho_k(1)} + o\left(\frac{1}{\sqrt{n}}\right)$, and thus

$$\phi_{Z_n}(s) \sim e^{-\frac{s^2\sigma^2}{2}},$$

with $\sigma^2 = -\frac{\rho_k'(1)}{\rho_k(1)} - \frac{\rho_k''(1)}{\rho_k(1)} + \frac{\rho_k'(1)^2}{\rho_k(1)^2}$, which completes the proof. □

Thus, we get that the total number of leaves in lambda terms with a bounded number of De Bruijn levels is asymptotically normally distributed and Theorem 9.1 is proved for both cases whether k is an element of the sequence $(N_i)_{i\geq 0}$ or not.

9.2 Unary profile

9.2.1 Location of leaves among the De Bruijn levels

The aim of this section is the investigation of the distribution of the number of leaves in the different De Bruijn levels in closed lambda terms with a bounded number of De Bruijn levels. In order to do so, we make use of the representation (3.21) of the generating function $H_k(z)$ that considers a lambda term from $\mathcal{H}_k$ as a structure obtained by glued binary trees. Now, let ${}_{k-l}\tilde{H}_k(z,u)$ be the bivariate generating function of closed lambda terms with at most k De Bruijn levels, where z marks the size and u marks the number of leaves in the $(k-l)$-th unary level $(0 \leq l \leq k)$, reading as

$${}_{k-l}\tilde{H}_k(z,u) = B(z, B(z, 1+\ldots+B(z,(k-l)\cdot u+\ldots+B(z,(k-1)+B(z,k)))\ldots)\ldots)).$$

This can be written as

$$ {}_{k-l}\tilde{H}_k(z,u) = \frac{1-\sqrt{\tilde{R}_{k+1,k}(z,u)}}{2z}, $$

with

$$ \tilde{R}_{i,k}(z,u) = \begin{cases} 1-4z^2k & i=1 \\ 1-4z^2(k-i+1)-2z+2z\sqrt{\tilde{R}_{i-1,k}(z,u)} & i>1, i\neq l+1 \\ 1-4z^2u(k-l)-2z+2z\sqrt{\tilde{R}_{l-1,k}(z,u)} & i=l+1 \end{cases}. $$

Thus, if we investigate the number of leaves in the $(k-l)$-th De Bruijn level, for $0 \leq l \leq k$, a factor u is inserted in the recursive definition of the $(l+1)$-th radicand (*cf.* Figure 9.3).

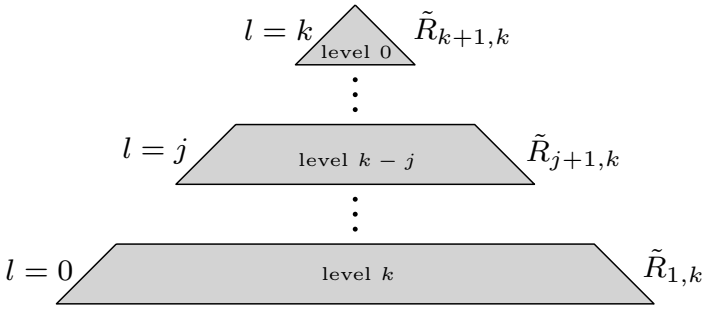


Figure 9.3: A schematic sketch of a lambda term with at most k De Bruijn levels that exemplifies the notation that is used within this section: For $l = j$ we investigate the number of leaves in the $(k-j)$-th De Brujin level, and the u appears in the $(j+1)$-th radicand. The levels that are closer to the root will subsequently be denoted by the "lower levels", since they have a lower number, while the so-called "upper levels" are the ones further away from the root.

Remark 9.9. *Note that the radicands $\tilde{R}_{i,k}$ that are introduced above are very similar to the radicands $R_{i,k}$ that were used in the previous section. The only difference is that now we have a u only in the $(l+1)$-th radicand, while in the previous case u was occurring in all radicands. Thus, from now on we will have further distinctions of cases now depending on the relative position (w.r.t. l) of the radicand(s) where the dominant singularity comes from.*

The remainder of this section is devoted to the proof of the following theorem.

Theorem 9.10. *Let ${}_{k-l}\tilde{H}_k(z,u)$ denote the bivariate generating function of the class of closed lambda terms with at most k De Bruijn levels, where z marks the size and u marks the number of leaves in the $(k-l)$-th De Bruijn level. Additionally, we denote its dominant singularity by $\tilde{\rho}_k(u)$, and $\tilde{B}(u) = \frac{\tilde{\rho}_k(u)}{\tilde{\rho}_k(1)}$. Then the following assertions hold:*

- *If $k \in (N_j, N_{j+1})$, then the average number of leaves in the first $k-j$ De Bruijn levels is $O(1)$, as the size $n \to \infty$, while it is $\Theta(n)$ for each of the last $j+1$ levels.*

In particular, if $\tilde{B}''(1) + \tilde{B}'(1) - \tilde{B}'(1)^2 \neq 0$, the number of leaves in each of the last $j+1$ De Bruijn levels is asymptotically normally distributed with mean and variance proportional to the size n of the lambda term.

- *If $k = N_j$, then the average number of leaves in the first $k-j$ De Bruijn levels is $O(1)$, as $n \to \infty$, while the average number of leaves in the $(k-j)$-th level is $\Theta(\sqrt{n})$. Each of the last j De Bruijn levels have asymptotically $\Theta(n)$ leaves.*

  In particular, if $\tilde{B}''(1) + \tilde{B}'(1) - \tilde{B}'(1)^2 \neq 0$, the number of leaves in each of the last j De Bruijn levels is asymptotically normally distributed with mean and variance proportional to the size n of the lambda term.

This subsection consists of two subsubsections. In the first part we will derive the mean values for the number of leaves in the different De Bruijn levels and the second part deals with the distributions of the number of leaves in these levels.

Mean values

Now we want to determine the mean for the number of leaves in the $(k-l)$-th De Bruijn level for abitrary $0 \leq l \leq k$, *i.e.*

$$\mathbb{E}(X_n) = \frac{[z^n]\left(\frac{\partial}{\partial u}\ {}_{k-l}\tilde{H}_k(z,u)\right)\Big|_{u=1}}{[z^n]_{k-l}\tilde{H}_k(z,1)},$$

where X_n denotes the number of leaves in the $(k-l)$-th De Bruijn level of a random closed lambda term of size n with at most k De Bruijn levels. In order to do so, we start with determining the derivatives of the radicands $\tilde{R}_{i,k}(z,u)$ recursively:

$$\frac{\partial}{\partial u}\tilde{R}_{i,k}(z,u) = \begin{cases} 0 & i < l+1, \\ -4z^2(k-l) & i = l+1, \\ z \cdot \frac{\frac{\partial}{\partial u}\tilde{R}_{i-1,k}(z,u)}{\sqrt{\tilde{R}_{i-1,k}(z,u)}} & i > l+1. \end{cases}. \tag{9.10}$$

Therefore we get

$$\left(\frac{\partial}{\partial u}\ {}_{k-l}\tilde{H}_k(z,u)\right)\Big|_{u=1} = z^{k-l+1}(k-l)\prod_{i=l+1}^{k+1}\frac{1}{\sqrt{\tilde{R}_{i,k}(z,1)}}. \tag{9.11}$$

Again we perform a distinction of cases starting with k not being an element of the sequence $(N_j)_{j\in\mathbb{N}}$.

The case: $N_j < k < N_{j+1}$. Let $\tilde{\rho}_k(u)$ be the dominant singularity of ${}_{k-l}\tilde{H}_k(z,u)$, which we know comes from the $(j+1)$-th radicand $\tilde{R}_{j+1,k}(z,u)$. Obviously, $\tilde{\rho}_k(1) = \rho_k(1)$. Therefore we will again use the abbreviation $\rho_k := \tilde{\rho}_k(1)$. From Proposition 9.2 we get the following expansions of the radicands for $u = 1$ and $\epsilon \to 0$ so that $\epsilon \in \mathbb{C}\backslash\mathbb{R}^-$:

- $\forall i < j+1$ (inner radicands) : $\tilde{R}_{i,k}(\rho_k - \epsilon, 1) = \tilde{R}_{i,k}(\rho_k, 1) + \mathcal{O}(|\epsilon|)$,
- $\tilde{R}_{j+1,k}(\rho_k - \epsilon, 1) = \gamma_{j+1}\epsilon + \mathcal{O}(|\epsilon|^2)$, with $\gamma_{j+1} = -\frac{\partial}{\partial z}\tilde{R}_{j+1,k}(\rho_k, 1)$,

- $\forall i > j+1$ (outer radicands) : $\tilde{R}_{i,k}(\rho_k - \epsilon, 1) = a_i + b_i\sqrt{\epsilon} + \mathcal{O}(|\epsilon|)$, with

$$a_{i+1} = 1 - 4(k-i)\rho_k^2 - 2\rho_k + 2\rho_k\sqrt{a_i}, \quad \text{for} \quad j+2 \le i \le k, \tag{9.12}$$
$$a_{j+2} = 1 - 4(k-j-1)\rho_k^2 - 2\rho_k, \tag{9.13}$$

and

$$b_{i+1} = \frac{b_i\rho_k}{\sqrt{a_i}} \quad \text{for} \quad j+2 \le i \le k, \tag{9.14}$$
$$b_{j+2} = 2\rho_k\sqrt{\tilde{\gamma}_{j+1}}. \tag{9.15}$$

Remark 9.11. *Note that these sequences coincide with the ones given in* (3.22) - (3.23).

Thus, we have

- $\forall i < j+1$ (inner radicands) : $\frac{1}{\sqrt{\tilde{R}_{i,k}(\rho_k-\epsilon,1)}} = \frac{1}{\sqrt{\tilde{R}_{i,k}(\rho_k,1)}} + \mathcal{O}(|\epsilon|)$,
- $\frac{1}{\sqrt{\tilde{R}_{j+1,k}(\rho_1-\epsilon,1)}} = \frac{1}{\sqrt{\gamma_{j+1}}}\epsilon^{-\frac{1}{2}} + \mathcal{O}(|\epsilon|^{\frac{1}{2}})$,
- $\forall i > j+1$ (outer radicands) : $\frac{1}{\sqrt{\tilde{R}_{i,k}(\rho_k-\epsilon,1)}} = \frac{1}{\sqrt{a_i}} - \frac{b_i}{2\sqrt{a_i^3}}\epsilon^{\frac{1}{2}} + \mathcal{O}(|\epsilon|^{\frac{3}{2}})$.

Now we have to perform a distinction of cases whether the De Bruijn level that we are focusing on is below the $(k-j)$-th level or not (*i.e.*, whether l is larger than j or not).

First case: $l > j$ First let us remember that $l > j$ implies that the u is inserted in a radicand that is located outside the $(j+1)$-th. From (9.11) we get for $\epsilon \to 0$ so that $\epsilon \in \mathbb{C} \setminus \mathbb{R}^-$

$$\left(\frac{\partial}{\partial u}\ {}_{k-l}\tilde{H}_k(\rho_k - \epsilon, u)\right)\Big|_{u=1} =$$
$$\rho_k^{k-l+1}(k-l)\left(\prod_{i=l+1}^{k+1}\frac{1}{\sqrt{a_i}} - \sum_{m=l+1}^{k+1}\left(\frac{b_m}{2\sqrt{a_m^3}}\prod_{i=l+1,i\neq m}^{k+1}\frac{1}{\sqrt{a_i}}\right)\epsilon^{\frac{1}{2}} + \mathcal{O}(|\epsilon|)\right).$$

By denoting the sum in the equation above with $\tilde{\delta}_l$ we can determine the coefficient of z^n asymptotically for $n \to \infty$ by

$$[z^n]\left(\frac{\partial}{\partial u}\ {}_{k-l}\tilde{H}_k(z,u)\right)\Big|_{u=1} = -\rho_k^{k-l+1}(k-l)\tilde{\delta}_l\left(\frac{1}{\rho_k}\right)^n\frac{n^{-\frac{3}{2}}}{\Gamma(-\frac{1}{2})}\left(1+\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right),$$

and by using the asymptotics of the n-th coefficient of ${}_{k-l}\tilde{H}_k(z,1) = H_k(z,1)$ (see Theorem 3.19) we finally get for the mean, asymptotically as $n \to \infty$,

$$\frac{[z^n]\left(\frac{\partial}{\partial u}\ {}_{k-l}\tilde{H}_k(z,u)\right)|_{u=1}}{[z^n]_{k-l}\tilde{H}_k(z,1)} = \frac{-\rho_k^{k-l+1}(k-l)\tilde{\delta}_l}{h_k}\left(1+\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right).$$

Thus, we showed that there is only a small number of leaves in the De Bruijn levels below the $(k-j)$-th level. More precisely, the asymptotic mean of the number of leaves is $O(1)$ for all these lower levels.

Remark 9.12. *The constant h_k can be expressed by means of the sequences a_i and b_i defined in Equations* (9.12) – (9.15)*, thereby enabling a representation of the constant* $C_{k,l} := \frac{-\rho_k^{k-l+1}(k-l)\tilde{\delta}_l}{h_k}$ *that reads as*

$$C_{k,l} = \frac{2(k-l)\rho_k^2}{a_{l+1}}\left(1 + \frac{\rho_k a_{l+1}}{a_{l+2}\sqrt{a_{l+1}}} + \frac{\rho_k^2 a_{l+1}}{a_{l+3}\sqrt{a_{l+2}}\sqrt{a_{l+1}}} + \dots\right). \tag{9.16}$$

This term can be used in order to investigate the asymptotic number of leaves in the lower De Bruijn levels more thoroughly.

Second case: $l \leq j$ Similar to the first case we get

$$\left(\frac{\partial}{\partial u}\ {}_{k-l}\tilde{H}_k(\rho_k - \epsilon, u)\right)\Big|_{u=1} =$$

$$\rho_k^{k-l+1}(k-l)\left(\left(\prod_{i=l+1}^{j}\frac{1}{\sqrt{\tilde{R}_{i,k}(\rho_k,1)}}\right)\left(\prod_{i=j+2}^{k+1}\frac{1}{\sqrt{a_i}}\right)\frac{1}{\sqrt{\gamma_{j+1}}}\epsilon^{-\frac{1}{2}} + \text{const. term} + \mathcal{O}(|\epsilon|^{\frac{1}{2}})\right).$$

By setting $\tilde{\phi}_{j+1,l} := \left(\prod_{i=l+1}^{j}\frac{1}{\sqrt{\tilde{R}_{i,k}(\rho_k,1)}}\right)\left(\prod_{i=j+2}^{k+1}\frac{1}{\sqrt{a_i}}\right)\frac{1}{\sqrt{\gamma_{j+1}}}$, we obtain for $n \to \infty$

$$[z^n]\left(\frac{\partial}{\partial u}\ {}_{k-l}\tilde{H}_k(z,u)\right)\Big|_{u=1} = \rho_k^{k-l+1}(k-l)\tilde{\phi}_{j+1,l}\left(\frac{1}{\rho_k}\right)^n \frac{n^{-\frac{1}{2}}}{\Gamma(\frac{1}{2})}\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right).$$

Thus, we get for the mean, asymptotically as $n \to \infty$,

$$\frac{[z^n]\left(\frac{\partial}{\partial u}\ {}_{k-l}\tilde{H}_k(z,u)\right)|_{u=1}}{[z^n]_{k-l}\tilde{H}_k(z,1)} = \frac{\rho_k^{k-l+1}(k-l)\Gamma(-\frac{1}{2})\tilde{\phi}_{j+1,l}}{\Gamma(\frac{1}{2})h_k}\cdot n\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right).$$

Hence, we proved that the asymptotic mean for the number of leaves in the De Bruijn levels above the $(k-j)$-th is $\Theta(n)$. So, altogether we can see that almost all of the leaves are located in the upper $j+1$ De Bruijn levels.

The case: $k = N_j$**.** Now we will deal with the second case, where the bound k is an element of the sequence $(N_j)_{j\in\mathbb{N}}$.

We start by determining the expansions of the radicands around the dominant singularity $\tilde{\rho}_k(u)$ of ${}_{k-l}\tilde{H}_k(z,u)$ for $u = 1$ and $\epsilon \to 0$ so that $\epsilon \in \mathbb{C} \setminus \mathbb{R}^-$ (*cf.* [11, Proposition 9]):

- $\forall i < j$ (inner radicands) : $\tilde{R}_{i,k}(\rho_k - \epsilon, 1) = \tilde{R}_{i,k}(\rho_k, 1) + \mathcal{O}(|\epsilon|)$,
- $\tilde{R}_{j,k}(\rho_k - \epsilon, 1) = \tilde{\gamma}_j\epsilon + \mathcal{O}(|\epsilon|^2)$ with $\gamma_j = -\frac{\partial}{\partial z}\tilde{R}_{j,k}(\rho_k, 1)$,
- $\tilde{R}_{j+1,k}(\rho_k - \epsilon, 1) = 2\tilde{\rho}_k\sqrt{\gamma_j}\epsilon^{\frac{1}{2}} + \mathcal{O}(|\epsilon|)$,
- $\forall i > j+1$ (outer radicands) : $\tilde{R}_{i,k}(\rho_k - \epsilon, 1) = a_i + b_i\epsilon^{\frac{1}{4}} + \mathcal{O}(|\epsilon|)$, with a_i and b_i as defined in (9.12) - (9.15).

Thus, we get

- $\forall i < j$ (inner radicands) : $\frac{1}{\sqrt{\tilde{R}_{i,k}(\rho_k-\epsilon,1)}} = \frac{1}{\sqrt{\tilde{R}_{i,k}(\rho_k,1)}} + \mathcal{O}(|\epsilon|)$,
- $\frac{1}{\sqrt{\tilde{R}_{j,k}(\rho_k-\epsilon,1)}} = \frac{1}{\sqrt{\gamma_j}}\epsilon^{-\frac{1}{2}} + \mathcal{O}(|\epsilon|^{\frac{1}{2}})$,
- $\frac{1}{\sqrt{\tilde{R}_{j+1,k}(\tilde{\rho}_k-\epsilon,1)}} = \frac{1}{\sqrt{2\rho_k}\sqrt[4]{\tilde{\gamma}_j}}\epsilon^{-\frac{1}{4}} + \mathcal{O}(|\epsilon|^{\frac{1}{4}})$,
- $\forall i > j+1$ (outer radicands) : $\frac{1}{\sqrt{\tilde{R}_{i,k}(\rho_k-\epsilon,1)}} = \frac{1}{\sqrt{a_i}} - \frac{b_i}{2\sqrt{a_i^3}}\epsilon^{\frac{1}{4}} + \mathcal{O}(|\epsilon|)$.

We proceed analogously to the case where $N_j < k < N_{j+1}$, with the only difference that we have to distinguish between three cases now and since for $u = 1$ the j-th and the $(j+1)$-th radicand vanish simultaneously, we get a closed formula for the dominant singularity $\rho_k = \frac{1}{1+\sqrt{1+4(k-j)}}$.

First case: $l > j$ Let us again remember that $l > j$ implies that the u is inserted in the p-th radicand with $p > j+1$. From (9.11) we get for $\epsilon \in \mathbb{C} \setminus \mathbb{R}^-$ with $|\epsilon| \to 0$

$$\begin{aligned}\left(\frac{\partial}{\partial u}\,{}_{k-l}\tilde{H}_k(\rho_k-\epsilon,u)\right)\Big|_{u=1} &= \rho_k^{k-l+1}(k-l)\prod_{i=l+1}^{k+1}\left(\frac{1}{\sqrt{\tilde{a}_i}} - \frac{\tilde{b}_i}{2\sqrt{\tilde{a}_i^3}}\epsilon^{\frac{1}{4}} + \mathcal{O}(|\epsilon|)\right)\\ &= \left(\frac{1}{1+\sqrt{1+4(k-j)}}\right)^{k-l+1}(k-l)\Bigg(\prod_{i=l+1}^{k+1}\frac{1}{\sqrt{\tilde{a}_i}}\\ &\quad - \sum_{m=l+1}^{k+1}\left(\frac{\tilde{b}_m}{2\sqrt{\tilde{a}_m^3}}\prod_{i=l+1,i\neq m}^{k+1}\frac{1}{\sqrt{\tilde{a}_i}}\right)\epsilon^{\frac{1}{4}} + \mathcal{O}(|\epsilon|^{\frac{1}{2}})\Bigg).\end{aligned}$$

As in the previous case, we set $\tilde{\delta}_l := \sum_{m=l+1}^{k+1}\left(\frac{\tilde{b}_m}{2\sqrt{\tilde{a}_m^3}}\prod_{\substack{i=l+1\\ i\neq m}}^{k+1}\frac{1}{\sqrt{\tilde{a}_i}}\right)$. Extracting the n-th coefficient and using the asymptotics of $[z^n]_{k-l}\tilde{H}_k(z,1) = [z^n]H_k(z,1)$ given in Theorem 3.19 we have for $n \to \infty$

$$\frac{[z^n]\left(\frac{\partial}{\partial u}\,{}_{k-l}\tilde{H}_k(z,u)\right)|_{u=1}}{[z^n]_{k-l}\tilde{H}_k(z,1)} = \frac{-4\rho_k^{j-l+3}(k-l)\tilde{\delta}_l\prod_{m=j+2}^{k+1}\sqrt{a_m}}{b_{j+2}}\left(1+\mathcal{O}\left(n^{-\frac{1}{4}}\right)\right).$$

Thus, as in the previous case ($N_j < k < N_{j+1}$) the asymptotic mean for the number of leaves in the De Bruijn levels below the $(k-j)$-th level is $O(1)$.

Furthermore the constant $D_{k,l} := \frac{-4\rho_k^{j-l+3}(1)(k-l)\tilde{\delta}_l\prod_{m=j+2}^{k+1}\sqrt{a_m}}{b_{j+2}}$ can be simplified to

$$D_{k,l} = \frac{k-l}{2\lambda_{l-j}}\left(1 + \frac{\sqrt{\lambda_{l-j}}}{2\lambda_{l-j+1}} + \frac{\sqrt{\lambda_{l-j}}}{4\lambda_{l-j+2}\sqrt{\lambda_{l-j+1}}} + \frac{\sqrt{\lambda_{l-j}}}{8\lambda_{l-j+3}\sqrt{\lambda_{l-j+2}}\sqrt{\lambda_{l-j+1}}} + \ldots\right) \tag{9.17}$$

with the sequence λ_i defined by $\lambda_0 = 0$ and $\lambda_{i+1} = i + 1 + \sqrt{\lambda_i}$ for $i \geq 0$.

Second case: $l = j$ Here the u is inserted in the $(j+1)$-th radicand. In this case we get

$$\frac{[z^n]\left(\frac{\partial}{\partial u}\ {}_{k-l}\tilde{H}_k(z,u)\right)|_{u=1}}{[z^n]_{k-l}\tilde{H}_k(z,1)} =$$

$$\frac{-4\rho_k^3(k-j)\Gamma(-1/4)\psi_j\prod_{m=j+2}^{k+1}\sqrt{a_m}}{\Gamma(\frac{1}{4})b_{j+2}}\cdot\sqrt{n}\ \left(1+\mathcal{O}\left(n^{-\frac{1}{4}}\right)\right) \tag{9.18}$$

with

$$\psi_j = \frac{1}{\sqrt{2\rho_k}\sqrt[4]{\tilde{\gamma}_j}}\prod_{i=j+2}^{k+1}\frac{1}{\sqrt{a_i}}. \tag{9.19}$$

The constant $\hat{D}_{k,l} := \frac{-4\rho_k^3(k-j)\Gamma(-1/4)\psi_j\prod_{m=j+2}^{k+1}\sqrt{a_m}}{\Gamma(\frac{1}{4})b_{j+2}}$ simplifies to

$$\hat{D}_{k,l} = \frac{-\Gamma(-1/4)(k-j)\sqrt{\rho_k}}{\Gamma(1/4)\sqrt{\tilde{\gamma}_j}}.$$

In order to get some information on the magnitude of this factor we would have to investigate $\tilde{\gamma}_j = -\frac{\partial}{\partial z}\tilde{R}_{j,k}(\rho_k,1)$, which seems to get rather involved. However, taking a look at Equation (9.18) we can see that there are already considerably more unary nodes in the $(k-j)$-th De Bruijn level, namely $\Theta(\sqrt{n})$.

Third case: $l < j$ The third case gives for $n \to \infty$

$$\frac{[z^n]\left(\frac{\partial}{\partial u}\ {}_{k-l}\tilde{H}_k(z,u)\right)|_{u=1}}{[z^n]_{k-l}\tilde{H}_k(z,1)} =$$

$$\frac{-4\rho_k^{j-l+3}(k-l)\Gamma(-1/4)\chi_j\prod_{m=j+2}^{k+1}\sqrt{a_m}}{\Gamma(3/4)b_{j+2}}\cdot n\ \left(1+\mathcal{O}\left(n^{-\frac{1}{4}}\right)\right),$$

with

$$\chi_j = \frac{1}{\sqrt{\tilde{\gamma}_j}}\psi_j\left(\prod_{i=l+1}^{j-1}\frac{1}{\sqrt{\tilde{R}_{i,k}(\rho_k,1)}}\right),$$

where ψ_j is defined as in (9.19). Thus, we proved that on average there are $\Theta(n)$ leaves in the upper j De Bruijn levels. The constant $\tilde{D}_{k,l} := \frac{-4\rho_k^{j-l+3}(k-l)\Gamma(-1/4)\chi_j\prod_{m=j+2}^{k+1}\sqrt{a_m}}{\Gamma(3/4)b_{j+2}}$ can be rewritten as

$$\tilde{D}_{k,l} = \frac{-\Gamma(-1/4)(k-l)\rho_k^{j-l}}{\Gamma(3/4)\tilde{\gamma}_j}\prod_{i=l+1}^{j-1}\frac{1}{\sqrt{\tilde{R}_{i,k}(\rho_k,1)}}.$$

The following proposition sums up all the results that we obtained within this section.

Proposition 9.13. *Let X_n denote the number of leaves in the $(k-l)$-th De Bruijn level in a random lambda term of size n with at most k De Bruijn levels.*

If $k \in (N_j, N_{j+1})$, then we get for the asymptotic mean when $n \to \infty$

- *in the case $l > j$:*

$$\mathbb{E}(X_n) = \frac{[z^n]\left(\frac{\partial}{\partial u}\ {}_{k-l}H_k(z,u)\right)|_{u=1}}{[z^n]_{k-l}H_k(z,1)} = C_{k,l}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right),$$

- *and in the case $l \leq j$:*

$$\mathbb{E}(X_n) = \frac{[z^n]\left(\frac{\partial}{\partial u}\ {}_{k-l}H_k(z,u)\right)|_{u=1}}{[z^n]_{k-l}H_k(z,1)} = \tilde{C}_{k,l}\cdot n\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right),$$

with constants $C_{k,l}$ and $\tilde{C}_{k,l}$ depending on l and k.

If $k = N_j$, then the asymptotic mean for $n \to \infty$ reads as

- *in the case $l > j$:*

$$\mathbb{E}(X_n) = \frac{[z^n]\left(\frac{\partial}{\partial u}\ {}_{k-l}H_k(z,u)\right)|_{u=1}}{[z^n]_{k-l}H_k(z,1)} = D_{k,l}\left(1+\mathcal{O}\left(n^{-\frac{1}{4}}\right)\right),$$

- *in the case $l = j$:*

$$\mathbb{E}(X_n) = \frac{[z^n]\left(\frac{\partial}{\partial u}\ {}_{k-l}H_k(z,u)\right)|_{u=1}}{[z^n]_{k-l}H_k(z,1)} = \hat{D}_{k,l}\cdot\sqrt{n}\left(1+\mathcal{O}\left(n^{-\frac{1}{4}}\right)\right),$$

- *and in the case $l < j$:*

$$\mathbb{E}(X_n) = \frac{[z^n]\left(\frac{\partial}{\partial u}\ {}_{k-l}H_k(z,u)\right)|_{u=1}}{[z^n]_{k-l}H_k(z,1)} = \tilde{D}_{k,l}\cdot n\left(1+\mathcal{O}\left(\left(n^{-\frac{1}{4}}\right)\right)\right),$$

with constants $D_{k,l}$, $\hat{D}_{k,l}$ and $\tilde{D}_{k,l}$ depending on l and k.

All the constants occurring in Proposition 9.13 have been calculated explicitly and can be obtained for every fixed k. In particular, we investigated $D_{k,l}$ in order to show that for large k the number of leaves in the De Bruijn levels that are closer to the root is smaller (*cf.* Figure 9.4). In fact, they rapidly tend to zero for k tending to infinity.

Proposition 9.14. *Let us consider a random closed lambda term of size n with at most k De Bruijn levels and let us consider the case $k = N_j$. Then the average number of leaves in De Bruijn level L, with $0 \leq L \leq k-j-1$, is asymptotically equal to $D_{k,k-L}$, defined in* (9.17)*. It behaves like*

$$D_{k,k-L} \sim \frac{L}{2(k-j-L)} \qquad \text{as } k \to \infty.$$

Proof. The proposition follows directly by investigating this constant $D_{k,l}$. The asymptotics for the sequence λ_i (defined by $\lambda_0 = 0$ and $\lambda_{i+1} = i+1+\sqrt{\lambda_i}$ for $i \geq 0$) can be obtained by bootstrapping (see [11]). We obtain $\lambda_i \sim i$, as $i \to \infty$, see (3.25). □

Remark 9.15. *Note that the expression for $D_{k,l}$ (cf. Equ.* (9.17)*) can be obtained by plugging $\tilde{a}_{j+l} = 4\rho_k^2\lambda_{l-1}$ into the equation for $C_{k,l}$ (cf. Equ.* (9.16)*). However, this relation is solely valid for the case $k = N_j$ and thus, Proposition 9.14 holds just for the constants $D_{k,l}$. Nonetheless, we expect that by means of some suitable estimates for the a_i's one can obtain a similar behavior for the constants $C_{k,l}$. Since computations get rather involved, we omitted any further investigations of these constants within this paper. Anyway, we can conclude that in both cases, whether k is an element of $(N_i)_{i>0}$ or not, a random closed lambda term with at most k De Bruijn levels has very few leaves in its lowest levels if k is large.*

Distributions

Now that we derived the mean values for the number of leaves in the different De Bruijn levels, we are interested in their distribution. Therefore we distinguish again between the cases of k being an element of the sequence $(N_i)_{i\geq 0}$ or not.

The case: $N_j < k < N_{j+1}$ We know that the generating function ${}_{k-l}\tilde{H}_k(z,u)$ consists of $k+1$ nested radicals, where a u is inserted in the $(l+1)$-th radicand counted from the innermost one. Additionally we know that for $N_j < k < N_{j+1}$ the dominant singularity $\tilde{\rho}_k(u)$ comes from the $(j+1)$-th radicand. Therefore, for $l > j$ the function $\tilde{\rho}_k(u)$ is independent of u, which is the reason why we do not get a quasi-power in that case. Thus, for the first $k-j$ levels of the lambda-PDAG (*i.e.* the case $l > j$), where there are just a few leaves, we can not say anything about the distribution of the leaves so far. It might be a degenerated distribution.

However, in case that $l \leq j$ (*i.e.*, for the upper levels where there are a lot of leaves) we will use the Quasi-Powers Theorem to show that the number of leaves in the $(k-j)$-th until the k-th level is asymptotically normally distributed.

Analogously as we did in Section 9.1.1 we can show that

$$\frac{[z^n]\ {}_{k-l}\tilde{H}_k(z,u)}{[z^n]\ {}_{k-l}\tilde{H}_k(z,1)} = \frac{\tilde{h}_k(u)}{h_k}\left(\frac{\rho_k}{\tilde{\rho}_k(u)}\right)^n\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right). \tag{9.20}$$

We can easily see that Equation (9.20) has the desired shape for the Quasi-Powers Theorem. Hence, assuming that $\tilde{B}''(1)+\tilde{B}'(1)-\tilde{B}'(1)^2 \neq 0$, where $\tilde{B}(u) = \frac{\rho_k}{\tilde{\rho}_k(u)}$, the Quasi-Powers Theorem can be applied, which proves that the number of leaves in a De Bruijn level that is above the $(k-j-1)$-th level is asymptotically normally distributed.

The case: $k = N_j$ As is the previous case we do not know the distribution of the number of leaves in the lowest $k-j$ De Bruijn levels (*i.e.*, the levels 0 to $k-j-1$), due to the fact that for these levels the function $\tilde{\rho}_k(u)$ does not depend on u. It might as well be a degenerated distribution.

In Section 9.1.2 we showed that the dominant singularity comes from the j-th radicand when u is in a neighbourhood of 1. Thus, for the case that $l = j$, where we insert a u in the $(j+1)$-th radicand, the dominant singularity $\rho_k(u)$ does still do not depend on u. Therefore we also do not know the distribution of the leaves in the $(k-j)$-th De Bruijn level. It seems very unlikely that the number of leaves in this level will be asymptotically normally distributed, but further studies on this subject might be very interesting.

Now we are going to show that the number of leaves in the upper j De Bruijn levels (*i.e.*, from the $(k-j+1)$-th to the k-th level) is asymptotically normally distributed. In order to do so we proceed analogously as in Section 9.1.2 for the total number of leaves. Therefore, for $l < j$ we set again $z = \tilde{\rho}_k(u)(1+\frac{t}{n})$ and $u = 1+\frac{s}{\sqrt{n}}$ and obtain expansions that behave just as the ones in Lemma 9.6. The only differences that occur concern the constants and therefore do not alter our results for the normal distribution.

Thus, Theorem 9.10 is proved. Figure 9.4 summarizes the results that we obtained in Section 9.2.1 and illustrates a combinatorial interpretation of the occurring phenomena. Sections 9.2.2 and 9.2.3 are concerned with the investigation of the number

of unary nodes, and binary nodes respectively, among the De Bruijn levels. Using the same techniques as in Section 9.2.1 we can show that their number behaves in fact very similar to the number of leaves.

Theorem 9.16. *If $k \in (N_j, N_{j+1})$, then both the average number of unary nodes and the average number of binary nodes in the first $k-j$ De Bruijn levels are $O(1)$, as $n \to \infty$, while they are $\Theta(n)$ in each of the last $j+1$ levels.*

If $k = N_j$, then both the average number of unary nodes and the average number of binary nodes in the first $k-j$ De Bruijn levels is $O(1)$, as $n \to \infty$, while the average number of nodes of the respective type in the $(k-j)$-th De Bruijn level is $\Theta(\sqrt{n})$. The last j De Bruijn levels contain each asymptotically $\Theta(n)$ unary nodes, as well as $\Theta(n)$ binary nodes.

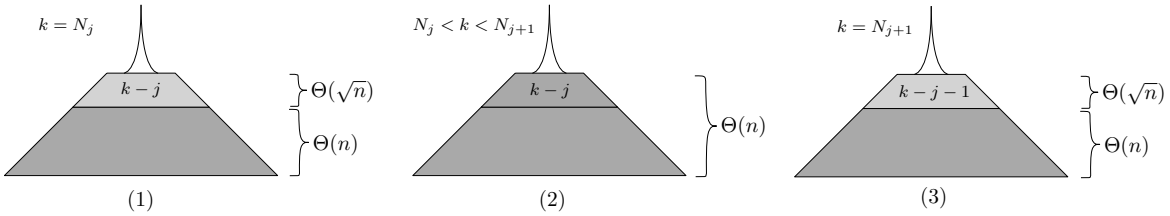


Figure 9.4: (1) In the $(k-j)$-th Be Bruijn level ($l = j$) are considerably more leaves than in the lower levels, but still less leaves then in the levels above. (2) With growing k the $(k-j)$-th Be Bruijn level gets filled with leaves, while the number of leaves in the next level below (*i.e.*, the $(k-j-1)$-th) slowly increases. (3) As soon as k reaches the next element of the sequence $(N_j)_{j\geq 0}$, namely $k = N_{j+1}$ the $(k-j-1)$-th De Bruijn level immediately contains considerably more leaves than the levels below.

9.2.2 Location of unary nodes among the De Bruijn levels

Now we want to investigate the number of unary nodes among the different De Bruijn levels. The bivariate generating function ${}_{k-l}\bar{H}_k(z,w)$ of the class of closed lambda terms with at most k De Bruijn levels, where z marks the size and w the number of unary nodes in the $(k-l)$-th De Bruijn level, can then be expressed by

$$B(z, B(z, 1 + \ldots + B(z, (k-l) + w \cdot B(z, \ldots (k-1) + B(z,k)) \ldots) \ldots)).$$

This can be rewritten to

$${}_{k-l}\bar{H}_k(z,w) = \frac{1 - \sqrt{\bar{R}_{k+1,k}(z,w)}}{2z},$$

with

$$\bar{R}_{i,k}(z,w) = \begin{cases} 1 - 4z^2 k & i = 1, \\ 1 - 4z^2(k-i+1) - 2z + 2z\sqrt{\bar{R}_{i-1,k}(z,w)} & i > 1, i \neq l+1, \\ 1 - 4z^2(k-l) - 2zw + 2zw\sqrt{\bar{R}_{l,k}(z,w)} & i = l+1. \end{cases}$$

Thus, for the derivatives we get

$$\frac{\partial \bar{R}_{i,k}(z,w)}{\partial w} = \begin{cases} 0 & i < l+1, \\ -2z + 2z\sqrt{\bar{R}_{l,k}(z,w)} & i = l+1, \\ \frac{z}{\sqrt{\bar{R}_{i-1,k}}} \frac{\partial \bar{R}_{i-1,k}(z,w)}{\partial w} & i > l+1, \end{cases}$$

which implies

$$\frac{\partial_{k-l}\bar{H}_k(z,w)}{\partial w}\Big|_{w=1} = \frac{z^{k-l}}{2} \prod_{i=l+1}^{k+1} \frac{1}{\sqrt{\bar{R}_{i,k}(z,1)}} \left(1 - \sqrt{\bar{R}_{l,k}(z,1)}\right). \tag{9.21}$$

As in the previous section we distinguish between different cases.

The case: $k = N_j$

First case: $l > j+1$ Inserting the expansion of the radicands $\bar{R}_{i,k}$ (see page 132) into (9.21) and simplifying yields for $n \to \infty$

$$[z^n]_{k-l}\bar{H}_k(z,w)\Big|_{w=1} = \rho_k^{k-l} 2\alpha_l \frac{n^{-5/4}}{\Gamma(-1/4)} \rho_k^{-n} \left(1 + \mathcal{O}\left(n^{-\frac{1}{4}}\right)\right),$$

with

$$\alpha_l = -\frac{b_l}{2\sqrt{a_l}} \prod_{i=l+1}^{k+1} \frac{1}{\sqrt{a_i}} - \sum_{m=l+1}^{k+1} \frac{b_m}{2\sqrt{a_m^3}} \prod_{\substack{i=l+1 \\ i\neq m}}^{k+1} \frac{1}{\sqrt{a_i}} \left(1 - \sqrt{a_l}\right), \tag{9.22}$$

where $a_i := \tilde{a}_i = a_i(1)$ and $b_i := \tilde{b}_i = b_i(1)$ are defined in (3.22) and (3.23). Thus, in this case the expected value of the number of unary nodes in the $(k-l)$-th De Bruijn level reads as

$$\frac{[z^n]\frac{\partial}{\partial w}{}_{k-l}\bar{H}_k(z,w)}{[z^n]_{k-l}\bar{H}_k(z,1)} = \frac{-2\alpha_l \rho_k^{j-l+2} \prod_{m=j+2}^{k+1} \sqrt{a_m}}{b_{j+2}} \left(1 + \mathcal{O}\left(n^{-\frac{1}{4}}\right)\right), \quad \text{as } n \to \infty.$$

Furthermore, the constant $\frac{-2\alpha_l \rho_k^{j-l+2} \prod_{m=j+2}^{k+1} \sqrt{a_m}}{b_{j+2}}$ can be simplified to

$$1 + \left(\frac{1}{4\rho_k \lambda_{l-j}} - \frac{\sqrt{\lambda_{l-j-1}}}{2\lambda_{l-j}}\right) \left(1 + \frac{\lambda_{l-j}}{2\lambda_{l-j+1}\sqrt{\lambda_{l-j}}} + \frac{\lambda_{l-j}}{2^2 \lambda_{l-j+2}\sqrt{\lambda_{l-j+1}}\sqrt{\lambda_{l-j}}} + \dots\right),$$

with the sequence λ_i defined by $\lambda_0 = 0$ and $\lambda_{i+1} = i + 1 + \sqrt{\lambda_i}$ for $i \geq 0$.

Since the second summand is almost zero for l being close to k and large k, this implies that the number of unary nodes in these levels (close to the root) is close to one for large k.

Second case: $l = j + 1$ For $n \to \infty$ we get

$$[z^n]_{k-l}\bar{H}_k(z,w)\Big|_{w=1} = \rho_k^{k-l} 2\zeta_l \frac{n^{-5/4}}{\Gamma(-1/4)} \rho_k^{-n} \left(1 + \mathcal{O}\left(n^{-\frac{1}{4}}\right)\right),$$

with

$$\zeta_l = -\sqrt{2\rho_k}(\rho_k \tilde{\gamma})^{1/4} \prod_{i=j+2}^{k+1} \frac{1}{\sqrt{a_i}} - \sum_{m=j+2}^{k+1} \frac{b_m}{2\sqrt{a_m^3}} \prod_{\substack{i=j+2 \\ i\neq m}}^{k+1} \frac{1}{\sqrt{a_i}}.$$

Thus,

$$\frac{[z^n]\frac{\partial}{\partial w}{}_{k-l}\bar{H}_k(z,w)}{[z^n]_{k-l}\bar{H}_k(z,1)} = \frac{-2\zeta_l \rho_k^{j-l+2} \prod_{m=j+2}^{k+1} \sqrt{a_m}}{b_{j+2}} \left(1 + \mathcal{O}\left(n^{-\frac{1}{4}}\right)\right), \quad \text{as } n \to \infty.$$

In this case the constant $\frac{-2\zeta_l \rho_k^{j-l+2} \prod_{m=j+2}^{k+1} \sqrt{a_m}}{b_{j+2}}$ simplifies to

$$1 + \frac{1}{4\rho_k}\left(1 + \frac{1}{2\lambda_2} + \frac{1}{2^2\lambda_3\sqrt{\lambda_2}} + \dots\right).$$

So, the expected number of unary nodes in the $(k-j-1)$-th De Bruijn level behaves exactly like in the lower levels. Starting from the next level a change in the behavior can be determined, as we will see in the following.

Third case: $l = j$ For $n \to \infty$ we have

$$[z^n]_{k-l}\bar{H}_k(z,w)\Big|_{w=1} = \rho_k^{k-l} 2\beta_l \frac{n^{-3/4}}{\Gamma(1/4)} \rho_k^{-n} \left(1 + \mathcal{O}\left(n^{-\frac{1}{2}}\right)\right),$$

with

$$\beta_l = \frac{1}{\sqrt{2\rho_k}\sqrt[4]{\rho_k\tilde{\gamma}_j}} \prod_{i=j+2}^{k+1} \frac{1}{\sqrt{a_i}}.$$

Thus, as $n \to \infty$,

$$\frac{[z^n]\frac{\partial}{\partial w}{}_{k-l}\bar{H}_k(z,w)}{[z^n]_{k-l}\bar{H}_k(z,1)} = \frac{-2\beta_l \rho_k^2 \Gamma(-1/4) \prod_{m=j+2}^{k+1} \sqrt{a_m}}{\Gamma(1/4) b_{j+2}} \cdot \sqrt{n}\left(1 + \mathcal{O}\left(n^{-\frac{1}{2}}\right)\right).$$

The constant $\frac{-2\beta_l \rho_k^2 \Gamma(-1/4) \prod_{m=j+2}^{k+1} \sqrt{a_m}}{\Gamma(1/4) b_{j+2}}$ can be written as

$$\frac{-\Gamma(-1/4)}{2\Gamma(1/4)\sqrt{\rho_k\tilde{\gamma}_j}}.$$

The expected number of unary nodes in this "separating level" is therefore asymptotically $\Theta(\sqrt{n})$ (as was the number of leaves).

Fourth case: $l < j$ For $n \to \infty$ we get

$$[z^n]_{k-l}\bar{H}_k(z,w)\Big|_{w=1} = \rho_k^{k-l} 2\epsilon_l \frac{n^{-1/4}}{\Gamma(3/4)} \rho_k^{-n} \left(1 + \mathcal{O}\left(n^{-\frac{1}{4}}\right)\right),$$

with

$$\epsilon_l = \frac{1}{\sqrt{2\rho_k}\sqrt[4]{\rho_k\tilde{\gamma}_j}\sqrt{\rho_k\tilde{\gamma}_j}} \prod_{i=j+2}^{k+1} \frac{1}{\sqrt{a_i}} \prod_{m=l+1}^{j-1} \frac{1}{\sqrt{\tilde{R}_{m,k}}} \left(1 - \sqrt{\bar{R}_{l,k}}\right).$$

Thus, as $n \to \infty$,

$$\frac{[z^n]\frac{\partial}{\partial w}{}_{k-l}\bar{H}_k(z,w)}{[z^n]_{k-l}\bar{H}_k(z,1)} = \frac{-2\epsilon_l \rho_k^{j-l+2}\Gamma(-1/4)\prod_{m=j+2}^{k+1}\sqrt{a_m}}{\Gamma(3/4)b_{j+2}} \cdot n\left(1 + \mathcal{O}\left(n^{-\frac{1}{4}}\right)\right).$$

The constant $\frac{-2\epsilon_l \rho_k^{j-l+2}\Gamma(-1/4)\prod_{m=j+2}^{k+1}\sqrt{a_m}}{\Gamma(3/4)b_{j+2}}$ can be simplified to

$$\frac{-\rho_k^{j-l+1}\Gamma(-1/4)}{2\Gamma(3/4)\tilde{\gamma}_j} \prod_{m=l+1}^{j-1} \frac{1}{\sqrt{\tilde{R}_{m,k}}} \left(1 - \sqrt{\bar{R}_{l,k}}\right).$$

Hence, analogously to the number of leaves, we proved that the number of unary nodes on the upper $j+1$ De Bruijn levels is $\Theta(n)$.

The case: $N_j < k < N_{j+1}$

This case works analogously to the previous one. Thus, we just give the results for the expected values.

First case: $l > j+1$ In this case, the expected value is entirely equal to the mean for the case $k = N_j$ and $l > j+1$. So, with α_l defined as in (9.22), we have for $n \to \infty$

$$\frac{[z^n]\frac{\partial}{\partial w}{}_{k-l}\bar{H}_k(z,w)}{[z^n]_{k-l}\bar{H}_k(z,1)} = \frac{-2\alpha_l \rho_k^{j-l+2}\prod_{m=j+2}^{k+1}\sqrt{a_m}}{b_{j+2}} \cdot n\left(1 + \mathcal{O}\left(n^{-\frac{1}{2}}\right)\right).$$

Second case: $l = j+1$ In the second case, the constant differs a little bit, but the result stays qualitatively unaltered. We get

$$\frac{[z^n]\frac{\partial}{\partial w}{}_{k-l}\bar{H}_k(z,w)}{[z^n]_{k-l}\bar{H}_k(z,1)} = \frac{-2\mu_l \rho_k^{j-l+2}\prod_{m=j+2}^{k+1}\sqrt{a_m}}{b_{j+2}} \cdot n\left(1 + \mathcal{O}\left(n^{-\frac{1}{2}}\right)\right), \qquad \text{as } n \to \infty,$$

with

$$\mu_l = -\sum_{m=j+2}^{k+1} \frac{b_m}{2\sqrt{a_m^3}} \prod_{\substack{i=j+2 \\ i\neq m}}^{k+1} \frac{1}{\sqrt{a_i}} + \sqrt{\rho_k\tilde{\gamma}_{j+1}} \prod_{i=j+2}^{k+1} \frac{1}{\sqrt{a_i}}.$$

Third case: $l < j+1$ For $n \to \infty$ we have

$$\frac{[z^n]\frac{\partial}{\partial w}{}_{k-l}\bar{H}_k(z,w)}{[z^n]_{k-l}\bar{H}_k(z,1)} = \frac{-2\theta_l\Gamma(-1/2)\rho_k^{j-l+2}\prod_{m=j+2}^{k+1}\sqrt{a_m}\prod_{s=l+1}^{j}\sqrt{\bar{R}_s}}{b_{j+2}\Gamma(1/2)}\cdot n\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right),$$

with

$$\theta_l = \prod_{i=j+2}^{k+1}\frac{1}{\sqrt{a_i}}\frac{1}{\sqrt{\rho_k\tilde{\gamma}_{j+1}}}\left(1-\sqrt{\bar{R}_{j,k}}\right).$$

Thus, the expected number of unary nodes in the last $j+1$ De Bruijn levels is asymptotically $\Theta(n)$.

9.2.3 Location of binary nodes among the De Bruijn levels

In this section we calculate the mean values of the number of binary nodes in the different De Bruijn levels. We denote by $B(z,v,u)$ the generating function of the class of binary trees where z marks the total number of nodes, v marks the number of binary nodes, and u marks the number of leaves. Thus, we have

$$B(z,v,u) = \frac{1-\sqrt{1-4z^2uv}}{2zv}. \tag{9.23}$$

Using this generating function, we can write the bivariate generating function of the class of closed lambda terms with z marking the size, and v marking the number of binary nodes on the $(k-l)$-th De Bruijn level as

$${}_{k-l}H_k(z,v) = B(z,1,B(z,1,1+B(z,1,2+\ldots+B(z,v,(k-l)+\ldots+B(z,1,k))\ldots)\ldots))). \tag{9.24}$$

Plugging (9.23) into (9.24) gives

$${}_{k-l}\breve{H}_k(z,v) = \frac{1-\sqrt{\breve{R}_{k+1,k}(z,v)}}{2z},$$

with

$$\breve{R}_{i,k}(z,v) = \begin{cases} 1-4z^2k & i=1 \\ 1-4z^2(k-i+1)-2z+2z\sqrt{\breve{R}_{i-1,k}(z,v)} & i>1, i\neq l+1, l+2, \\ 1-4z^2(k-l)v-2zv+2zv\sqrt{\breve{R}_{l,k}(z,v)} & i=l+1, \\ 1-4z^2(k-l-1)-\frac{2z}{v}+\frac{2z}{v}\sqrt{\breve{R}_{l+1,k}(z,v)} & i=l+2. \end{cases}$$

Thus, for the derivatives we get

$$\frac{\partial \breve{R}_{i,k}(z,v)}{\partial v} = \begin{cases} 0 & i < l+1, \\ -4z^2(k-l) - 2z + 2z\sqrt{\breve{R}_{l,k}(z,v)} & i = l+1, \\ \frac{2z}{v^2} - \frac{2z}{v^2}\sqrt{\breve{R}_{l+1,k}(z,1)} + \frac{z}{v}\frac{1}{\sqrt{\breve{R}_{l+1,k}(z,1)}}\frac{\partial \breve{R}_{l+1,k}(z,v)}{\partial v} & i = l+2, \\ z\frac{1}{\sqrt{\breve{R}_{i-1,k}(z,1)}}\frac{\partial \breve{R}_{i-1,k}(z,v)}{\partial v} & i > l+2. \end{cases}$$

Finally, we have

$$\frac{\partial_{k-l}\breve{H}_k(z,v)}{\partial v}\Big|_{v=1} = \prod_{i=l+2}^{k+1}\frac{1}{\sqrt{\breve{R}_{i,k}(z,1)}}\Big(-\frac{z^{k-l-1}}{2} + \frac{\sqrt{\breve{R}_{l+1,k}(z,1)}z^{k-l-1}}{2} + \frac{z^{k-l}}{2\sqrt{\breve{R}_{l+1,k}(z,1)}} - \frac{z^{k-l}\sqrt{\breve{R}_{l,k}(z,1)}}{2\sqrt{\breve{R}_{l+1,k}(z,1)}} + \frac{z^{k-l+1}(k-l)}{\sqrt{\breve{R}_{l+1,k}(z,1)}}\Big). \tag{9.25}$$

Analogously to the previous sections we have to distinguish between different cases. For the case $k = N_j$ and $l > j+1$ we get for $n \to \infty$

$$[z^n]_{k-l}\breve{H}_k(z,v) = \xi_l \frac{n^{-5/4}}{\Gamma(-1/4)}\rho_k^{-n}\left(1 + \mathcal{O}\left(n^{-\frac{1}{4}}\right)\right),$$

with

$$\xi_l = -\sum_{m=l+2}^{k+1}\frac{b_m}{2\sqrt{a_m^3}}\prod_{\substack{i=l+2\\ i\neq m}}\frac{1}{\sqrt{a_i}}\left(\frac{\rho_k^{k-l-1}(\sqrt{a_{l+1}}-1)}{2} + \frac{\rho_k^{k-l+1}(k-l)}{\sqrt{a_{l+1}}} + \frac{\rho_k^{k-l}(1-\sqrt{a_l})}{2\sqrt{a_{l+1}}}\right)$$
$$+ \prod_{i=l+2}^{k+1}\frac{1}{\sqrt{a_i}}\left(\frac{\rho_k^{k-l-1}b_{l+1}}{2\sqrt{a_{l+1}}} - \frac{\rho_k^{k-l+1}(k-l)b_{l+1}}{2\sqrt{a_{l+1}^3}} + \frac{b_{l+1}}{2\sqrt{a_{l+1}^3}}(\sqrt{a_l}-1) - \frac{\rho_k^{k-l}b_l}{4\sqrt{a_l}\sqrt{a_{l+1}}}\right).$$

Thus,

$$\frac{[z^n]\frac{\partial}{\partial v}{}_{k-l}\breve{H}_k(z,v)}{[z^n]_{k-l}\breve{H}_k(z,1)} = \frac{-4\xi_l\prod_{m=j+2}^{k+1}\sqrt{a_m}}{\rho_k^{k-j}b_{j+2}}\left(1 + \mathcal{O}\left(\left(n^{-\frac{1}{4}}\right)\right)\right), \quad \text{as } n \to \infty.$$

We performed a thorough investigation of the constant $\frac{-4\xi_l\prod_{m=j+2}^{k+1}\sqrt{a_m}}{\rho_k^{k-j}b_{j+2}}$ and showed that - equivalently to the constant $D_{k,l}$ given in (9.17) - it is almost zero, in case l is close to $k+1$ and k is large, *i.e.*, if we consider a very low De Bruijn level, that is close to the root.

Due to Equation (9.25) calculations get rather involved. Since the methods that are used are exactly the same as in the previous sections, we will omit further calculations. However, the results resemble the ones that we got in Section 9.2.1 for the number of leaves. The only difference appears in the constants, but qualitatively also these constants behave equally.

Chapter 10

Conclusion

We discuss Part III according to the chronological development sequence of the presented work, since the studies that led to a large part of the results presented in Chapter 8 were motivated by the results obtained in Chapter 9. However, in order to provide a clear structure of the thesis, it was beneficial to present the results in the given order.

Our investigation concerning the shape of lambda terms with bounded number of De Bruijn levels was triggered by the striking observation that the asymptotic number of these terms with n vertices is of the form $\rho^n n^{-3/2}$ except if the bound belongs to some peculiar doubly exponentially growing sequence. There was no apparent reason why bounding the number of De Bruijn levels by 8 is substantially different from setting the bound to 7 or 9.

The results in this thesis showed that the vertices corresponding to the variables in the associated lambda terms gather at the bottom of the lambda-PDAG, meaning the De Bruijn levels of highest order within the lambda-PDAG. Precisely, in each of the last ℓ_n levels, where $\ell_n = \Theta(\log \log k)$, we find $\Theta(n)$ variables. The other levels contain only a bounded number of variables. As the bound grows, the higher levels become fuller and fuller and whenever k reaches a value that makes ℓ_n jump to the next integer, a further De Bruijn level becomes populated with variables. In this stage, there are only $\Theta(\sqrt{n})$ variables, but for the next value of k this level gets densely populated with variables, just as the other levels of high order. This shows that there is a structural difference within the classes of lambda terms with at most k De Bruijn levels, depending on whether the bound belongs to $(N_i)_{i\geq 0}$ or not. The distribution of the variables, in particular the fact that a further level has to contain a larger but still fairly small number of variables apparently has some slight affects on the degrees of freedom to choose the bindings which modifies the subexponential term in the asymptotics.

Since the class of lambda terms with bounded De Bruijn indices is a proper superclass of the above mentioned class of terms with a bounded number of De Bruijn levels, we were interested in the structure of these terms to see if a similar behavior concerning the distribution of variables could be observed. Thus, we investigated their expected unary profile as well and showed that this class of lambda terms behaves actually very treelike. The expected unary profile looks like the density of a Rayleigh distribution, as it is known for trees. However, we omitted the calculation of the distribution of the number of leaves in each De Bruijn level within this thesis, since

we expect it to get rather involved, while the result might not be very surprising.

By calculating the asymptotic number of nodes by degree that occur in k-colored Motzkin trees, we get exactly the same results as in Theorem 8.3 (and therefore also as in Corollary 8.5). 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, lambda terms where all De Bruijn indices are at most k are very much alike k-colored Motzkin trees. However, their counting sequences differ significantly due to the restrictions on labelling leaves in the hats of the terms. So, there are considerably more k-colored Motzkin trees than lambda terms where all De Bruijn indices are at most k. 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.

Bibliography

[1] Alfred V. Aho and Neil J. A. Sloane. "Some doubly exponential sequences." In: *Fibonacci Quart.* 11.4 (1973), pp. 429–437. ISSN: 0015-0517.

[2] David Aldous. "Asymptotic fringe distributions for general families of random trees." In: *The Annals of Applied Probability* 1.2 (1991), pp. 228–266. ISSN: 10505164. URL: http://www.jstor.org/stable/2959767.

[3] Henk P. Barendregt. *The Lambda Calculus, its Syntax and Semantics*. Vol. 40. Studies in Logic (London). [Reprint of the 1984 revised edition, MR0774952], With addenda for the 6th imprinting, Mathematical Logic and Foundations. College Publications, London, 2012, xvi+622+E16. ISBN: 978-1-84890-066-0.

[4] Maciej Bendkowski, Olivier Bodini, and Sergey Dovgal. "Polynomial tuning of multiparametric combinatorial samplers." In: *2018 Proceedings of the Fifteenth Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*. SIAM, Philadelphia, PA, 2018, pp. 92–106. DOI: 10.1137/1.9781611975062.9. URL: https://doi.org/10.1137/1.9781611975062.9.

[5] Maciej Bendkowski, Katarzyna Grygiel, Pierre Lescanne, and Marek Zaionc. "Combinatorics of λ-terms: a natural approach." In: *J. Logic Comput.* 27.8 (2017), pp. 2611–2630. ISSN: 0955-792X. DOI: 10.1093/logcom/exx018. URL: https://doi.org/10.1093/logcom/exx018.

[6] Maciej Bendkowski, Katarzyna Grygiel, and Paul Tarau. "Boltzmann samplers for closed simply-typed lambda terms." In: *Practical aspects of declarative languages*. Vol. 10137. Lecture Notes in Comput. Sci. Springer, Cham, 2017, pp. 120–135.

[7] Maciej Bendkowski, Katarzyna Grygiel, and Marek Zaionc. "On the likelihood of normalization in combinatory logic." In: *J. Logic Comput.* 27.7 (2017), pp. 2251–2269. ISSN: 0955-792X. DOI: 10.1093/logcom/exx005. URL: https://doi.org/10.1093/logcom/exx005.

[8] Louisa Seelbach Benkner and Stephan Wagner. "On the collection of fringe subtrees in random binary trees." In: *arXiv preprint arXiv:2003.03323* (2020).

[9] François Bergeron, Philippe Flajolet, and Bruno Salvy. "Varieties of increasing trees." In: *CAAP*. 1992, pp. 24–48.

[10] Olivier Bodini, Danièle Gardy, and Bernhard Gittenberger. "Lambda terms of bounded unary height." In: *ANALCO11—Workshop on Analytic Algorithmics and Combinatorics*. SIAM, Philadelphia, PA, 2011, pp. 23–32.

[11] Olivier Bodini, Danièle Gardy, Bernhard Gittenberger, and Zbigniew Gołębiewski. "On the number of unary-binary tree-like structures with restrictions on the unary height." In: *Ann. Comb.* 22.1 (2018), pp. 45–91. ISSN: 0218-0006.

[12] Olivier Bodini, Danièle Gardy, Bernhard Gittenberger, and Alice Jacquot. "Enumeration of generalized *BCI* lambda-terms." In: *Electron. J. Combin.* 20.4 (2013), Paper 30, 23. ISSN: 1077-8926.

[13] Olivier Bodini, Danièle Gardy, and Alice Jacquot. "Asymptotics and random sampling for BCI and BCK lambda terms." In: *Theoret. Comput. Sci.* 502 (2013), pp. 227–238. ISSN: 0304-3975. URL: https://doi.org/10.1016/j.tcs.2013.01.008.

[14] Olivier Bodini, Antoine Genitrini, Bernhard Gittenberger, Isabella Larcher, and Mehdi Naima. "Average number of non-isomorphic subtree-shapes." In: (2020). Unpublished manuscript.

[15] Olivier Bodini, Antoine Genitrini, and Frédéric Peschanski. "A quantitative study of pure parallel processes." In: *Electronic Journal of Combinatorics* 23.1 (2016), P1.11, 39 pages, (electronic).

[16] Olivier Bodini and Bernhard Gittenberger. "On the asymptotic number of $BCK(2)$-terms." In: *ANALCO14—Meeting on Analytic Algorithmics and Combinatorics*. SIAM, Philadelphia, PA, 2014, pp. 25–39. URL: https://doi.org/10.1137/1.9781611973204.3.

[17] Olivier Bodini, Bernhard Gittenberger, and Zbigniew Gołębiewski. "Enumerating lambda terms by weighted length of their de Bruijn representation." In: *Discrete Appl. Math.* 239 (2018), pp. 45–61. ISSN: 0166-218X. DOI: 10.1016/j.dam.2017.12.042. URL: https://doi.org/10.1016/j.dam.2017.12.042.

[18] Olivier Bodini and Yann Ponty. "Multi-dimensional Boltzmann sampling of languages." In: *21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA'10)*. Discrete Math. Theor. Comput. Sci. Proc., AM. Assoc. Discrete Math. Theor. Comput. Sci., Nancy, 2010, pp. 49–63.

[19] Olivier Bodini and Paul Tarau. "On uniquely closable and uniquely typable skeletons of lambda terms." In: *Logic-based program synthesis and transformation*. Vol. 10855. Lecture Notes in Comput. Sci. Springer, Cham, 2018, pp. 252–268.

[20] Miklós Bóna and Boris Pittel. "On a random search tree: asymptotic enumeration of vertices by distance from leaves." In: *Advances in Applied Probability* 49.3 (2017), pp. 850–876. DOI: 10.1017/apr.2017.24.

[21] Mireille Bousquet-Mélou, Markus Lohrey, Sebastian Maneth, and Eric Noeth. "XML compression via directed acyclic graphs." In: *Theory of Computing Systems* 57.4 (2015), pp. 1322–1371.

[22] Nicolas Broutin, Luc Devroye, Erin McLeish, and Mikael de la Salle. "The height of increasing trees." In: *Random Struct. Algorithms* 32.4 (2008), pp. 494–518.

[23] Nicolaas G. de Bruijn. "Lambda calculus notation with nameless dummies, a tool for automatic formula manipulation, with application to the Church-Rosser theorem." In: *Indagationes Mathematicae (Proceedings)* 75.5 (1972), pp. 381–392.

[24] Gi-Sang Cheon and Louis W. Shapiro. "Protected points in ordered trees." In: *Appl. Math. Lett.* 21.5 (2008), pp. 516–520. DOI: 10.1016/j.aml.2007.07.001. URL: https://doi.org/10.1016/j.aml.2007.07.001.

[25] YS Chow, Sigaiti Moriguti, Herbert Robbins, and SM Samuels. "Optimal selection based on relative rank (the "secretary problem")." In: *Israel Journal of mathematics* 2.2 (1964), pp. 81–90.

[26] Keith Copenhaver. "k-protected vertices in unlabeled rooted plane trees." In: *Graph. Comb.* 33.2 (2017), pp. 347–355. ISSN: 0911-0119. DOI: 10.1007/s00373-017-1772-9. URL: https://doi.org/10.1007/s00373-017-1772-9.

[27] Robert M. Corless, Gaston H. Gonnet, David E. G. Hare, David J. Jeffrey, and Donald E. Knuth. "On the Lambert W function." In: *Advances in Computational mathematics* 5.1 (1996), pp. 329–359.

[28] Haskell B. Curry, Robert Feys, William Craig, J. Roger Hindley, and Jonathan P. Seldin. *Combinatory Logic*. Vol. 1. North-Holland Amsterdam, 1958.

[29] René David, Katarzyna Grygiel, Jakub Kozik, Christophe Raffalli, Guillaume Theyssier, and Marek Zaionc. "Asymptotically almost all λ-terms are strongly normalizing." In: *Log. Methods Comput. Sci.* 9.1 (2013), 1:02, 30. ISSN: 1860-5974. DOI: 10.2168/LMCS-9(1:2)2013. URL: https://doi.org/10.2168/LMCS-9(1:2)2013.

[30] Jean-François Delmas, Jean-Stéphane Dhersin, and Marion Sciauveau. "Cost functionals for large (uniform and simply generated) random trees." In: *Electron. J. Probab.* 23 (2018), Paper No. 87, 36. ISSN: 1083-6489. DOI: 10.1214/18-EJP213. URL: https://doi.org/10.1214/18-EJP213.

[31] Luc Devroye. "A note on the probabilistic analysis of patricia trees." In: *Random Structures & Algorithms* 3.2 (1992), pp. 203–214.

[32] Luc Devroye and Svante Janson. "Protected nodes and fringe subtrees in some random trees." In: *Electron. Commun. Probab.* 19 (2014), 10 pp. DOI: 10.1214/ECP.v19-3048. URL: https://doi.org/10.1214/ECP.v19-3048.

[33] Michael Drmota. "An analytic approach to the height of binary search trees II." In: *J. ACM* 50.3 (2003), pp. 333–374.

[34] Michael Drmota. *Random Trees: An Interplay between Combinatorics and Probability*. Springer Science & Business Media, 2009.

[35] Michael Drmota. "The saturation level in binary search tree." In: *Mathematics and Computer Science*. Springer, 2000, pp. 41–51.

[36] Michael Drmota and Bernhard Gittenberger. "The shape of unlabeled rooted trees." In: *European J. Combinat.* 31 (2010), pp. 2028–2063.

[37] Michael Drmota, Alex Iksanov, Martin Moehle, and Uwe Roesler. "A limiting distribution for the number of cuts needed to isolate the root of a random recursive tree." In: *Random Struct. Algorithms* 34.3 (2009), pp. 319–336.

[38] Michael Drmota, Anna de Mier, and Marc Noy. "Extremal statistics on non-crossing configurations." In: *Discrete Mathematics* 327 (2014), pp. 103–117. ISSN: 0012-365X. DOI: 10.1016/j.disc.2014.03.016.

[39] Rosena R. X. Du and Helmut Prodinger. "Notes on protected nodes in digital search trees." In: *Appl. Math. Lett.* 25.6 (2012), pp. 1025–1028. DOI: 10.1016/j.aml.2011.11.017. URL: https://doi.org/10.1016/j.aml.2011.11.017.

[40] Philippe Duchon, Philippe Flajolet, Guy Louchard, and Gilles Schaeffer. "Boltzmann samplers for the random generation of combinatorial structures." In: *Combin. Probab. Comput.* 13.4-5 (2004), pp. 577–625. ISSN: 0963-5483. DOI: 10.1017/S0963548304006315. URL: https://doi.org/10.1017/S0963548304006315.

[41] Eugene B. Dynkin. "The optimum choice of the instant for stopping a Markov process." In: *Soviet Mathematics* 4 (1963), pp. 627–629.

[42] Steven R. Finch. *Mathematical Constants*. Vol. 94. Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge, 2003, pp. xx+602. ISBN: 0-521-81805-2.

[43] Philippe Flajolet, Éric Fusy, and Carine Pivoteau. "Boltzmann sampling of unlabelled structures." In: *Proceedings of the Ninth Workshop on Algorithm Engineering and Experiments and the Fourth Workshop on Analytic Algorithmics and Combinatorics*. SIAM, Philadelphia, PA, 2007, pp. 201–211.

[44] Philippe Flajolet and Andrew Odlyzko. "Singularity analysis of generating functions." In: *SIAM J. Discrete Math.* 3.2 (1990), pp. 216–240. ISSN: 0895-4801. DOI: 10.1137/0403019. URL: https://doi.org/10.1137/0403019.

[45] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, Cambridge, 2009, pp. xiv+810. ISBN: 978-0-521-89806-5. DOI: 10.1017/CBO9780511801655. URL: https://doi.org/10.1017/CBO9780511801655.

[46] Philippe Flajolet, Paolo Sipala, and Jean-Marc Steyaert. "Analytic variations on the common subexpression problem." In: *Automata, languages and programming (Coventry, 1990)*. Vol. 443. Lecture Notes in Comput. Sci. Springer, New York, 1990, pp. 220–234. DOI: 10.1007/BFb0032034. URL: http://dx.doi.org/10.1007/BFb0032034.

[47] Peter R. Freeman. "The secretary problem and its extensions: A review." In: *International Statistical Review/Revue Internationale de Statistique* (1983), pp. 189–206.

[48] Walter Gautschi. "Some elementary inequalities relating to the Gamma and incomplete Gamma function." In: *Journal of Mathematics and Physics* 38 (1959), pp. 77–81. DOI: 10.1002/sapm195938177. URL: https://doi.org/10.1002/sapm195938177.

[49] Antoine Genitrini, Bernhard Gittenberger, Manuel Kauers, and Michael Wallner. "Asymptotic enumeration of compacted binary trees." In: *arXiv preprint arXiv:1703.10031* (2017).

[50] Antoine Genitrini, Bernhard Gittenberger, Manuel Kauers, and Michael Wallner. "Asymptotic enumeration of compacted binary trees of bounded right height." In: *Journal of Combinatorial Theory, Series A* 172 (2020), p. 105177.

[51] Nicholas Georgiou. "Embeddings and other mappings of rooted trees into complete trees." In: *Order* 22.3 (2005), pp. 257–288. DOI: 10.1007/s11083-005-9020-y. URL: https://doi.org/10.1007/s11083-005-9020-y.

[52] John P. Gilbert and Frederick Mosteller. "Recognizing the maximum of a sequence." In: *Selected Papers of Frederick Mosteller*. Springer, 2006, pp. 355–398.

[53] Bernhard Gittenberger. "On the contour of random trees." In: *SIAM Journal on Discrete Mathematics* 12 (1999), pp. 434–458. DOI: 10.1137/S0895480195289928.

[54] Bernhard Gittenberger, Zbigniew Gołębiewski, Isabella Larcher, and Małgorzata Sulkowska. "Counting embeddings of rooted trees into families of rooted trees." In: (2020). Unpublished manuscript.

[55] Bernhard Gittenberger, Zbigniew Gołębiewski, Isabella Larcher, and Małgorzata Sulkowska. "Protection numbers in simply generated trees and Pólya trees." In: *arXiv preprint arXiv:1904.03519* (2019).

[56] Bernhard Gittenberger, Emma Yu Jin, and Michael Wallner. "On the shape of random Pólya structures." In: *Discrete Math.* 341.4 (2018), pp. 896–911. ISSN: 0012-365X. DOI: 10.1016/j.disc.2017.12.016. URL: https://doi.org/10.1016/j.disc.2017.12.016.

[57] Bernhard Gittenberger and Isabella Larcher. "Distribution of variables in lambda-terms with restrictions on De Bruijn indices and De Bruijn levels." In: *Electronic Journal of Combinatorics* 26.P4.47 (2019). DOI: 10.37236/8579.

[58] Bernhard Gittenberger and Isabella Larcher. "On the number of variables in special classes of random lambda-terms." In: *29th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms*. Vol. 110. LIPIcs. Leibniz Int. Proc. Inform. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2018, Art. No. 25, 14.

[59] Zbigniew Gołębiewski and Mateusz Klimczak. "Protection number of recursive trees." In: *2019 Proceedings of the Sixteenth Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*. SIAM. 2019, pp. 45–53.

[60] Clemens Grabmayer. "Linear depth increase of lambda terms in leftmost-outermost rewrite sequences." In: *A tribute to Albert Visser*. Vol. 30. Tributes. Coll. Publ., [London], 2016, pp. 125–139.

[61] Katarzyna Grygiel, Paweł M. Idziak, and Marek Zaionc. "How big is BCI fragment of BCK logic." In: *J. Logic Comput.* 23.3 (2013), pp. 673–691. ISSN: 0955-792X. URL: https://doi.org/10.1093/logcom/exs017.

[62] Katarzyna Grygiel and Isabella Larcher. "Unary profile of lambda terms with restricted De Bruijn indices." working paper or preprint. Oct. 2019. URL: https://hal.archives-ouvertes.fr/hal-02313735.

[63] Katarzyna Grygiel and Pierre Lescanne. "Counting and generating terms in the binary lambda calculus." In: *J. Funct. Programming* 25 (2015), e24, 25. ISSN: 0956-7968. DOI: 10.1017/S0956796815000271. URL: https://doi.org/10.1017/S0956796815000271.

[64] Ryu Hasegawa. "The generating functions of lambda terms (extended abstract)." In: *Combinatorics, complexity, & logic (Auckland, 1996)*. Springer Ser. Discrete Math. Theor. Comput. Sci. Springer, Singapore, 1997, pp. 253–263.

[65] Clemens Heuberger and Helmut Prodinger. "Protection number in plane trees." In: *Applicable Analysis and Discrete Mathematics* 11.2 (2017), pp. 314–326.

[66] Cecilia Holmgren and Svante Janson. "Limit laws for functions of fringe trees for binary search trees and random recursive trees." In: *Electron. J. Probab.* 20 (2015), 51 pp. DOI: 10.1214/EJP.v20-3627. URL: https://doi.org/10.1214/EJP.v20-3627.

[67] Hsien-Kuei Hwang. "On convergence rates in the central limit theorems for combinatorial structures." In: *European Journal of Combinatorics* 19.3 (1998), pp. 329–343.

[68] Edward L. Ince. *Ordinary Differential Equations*. Dover Publications, New York, 1944, pp. viii+558.

[69] Svante Janson. "Simply generated trees, conditioned Galton-Watson trees, random allocations and condensation." In: *Probab. Surv.* 9 (2012), pp. 103–252. ISSN: 1549-5787. DOI: 10.1214/11-PS188. URL: https://doi.org/10.1214/11-PS188.

[70] Donald E. Knuth. *The Art of Computer Programming, Volume 3: (2nd ed.) Sorting and Searching*. Redwood City, CA, USA: Addison Wesley Longman Publishing Co., Inc., 1998. ISBN: 0-201-89685-0.

[71] Markus Kuba and Alois Panholzer. "On the degree distribution of the nodes in increasing trees." In: *J. Comb. Theory, Ser. A* 114.4 (2007), pp. 597–618.

[72] Grzegorz Kubicki, Jenö Lehel, and Michal Morayne. "A ratio inequality for binary trees and the best secretary." In: *Combinatorics, Probability & Computing* 11.2 (2002), pp. 149–161. DOI: 10.1017/S0963548301004977. URL: https://doi.org/10.1017/S0963548301004977.

[73] Grzegorz Kubicki, Jenö Lehel, and Michał Morayne. "An asymptotic ratio in the complete binary tree." In: *Order* 20.2 (2003), pp. 91–97. DOI: 10.1023/B:ORDE.0000009243.79750.85. URL: https://doi.org/10.1023/B:ORDE.0000009243.79750.85.

[74] Grzegorz Kubicki, Jeno Lehel, and Michal Morayne. "Counting chains and antichains in the complete binary tree." In: *Ars Comb.* 79 (2006).

[75] Malgorzata Kuchta, Michal Morayne, and Jaroslaw Niemiec. "Counting embeddings of a chain into a binary tree." In: *Ars Comb.* 91 (2009).

[76] Malgorzata Kuchta, Michal Morayne, and Jaroslaw Niemiec. "Counting embeddings of a chain into a tree." In: *Discrete Mathematics* 297.1-3 (2005), pp. 49–59. DOI: 10.1016/j.disc.2005.04.008. URL: https://doi.org/10.1016/j.disc.2005.04.008.

[77] Michael Lavine. "Some aspects of Pólya tree distributions for statistical modelling." In: *The annals of statistics* 20.3 (1992), pp. 1222–1235.

[78] Pierre Lescanne. "On counting untyped lambda terms." In: *Theoret. Comput. Sci.* 474 (2013), pp. 80–97. ISSN: 0304-3975. URL: https://doi.org/10.1016/j.tcs.2012.11.019.

[79] Xueliang Li, Yiyang Li, and Yongtang Shi. "The asymptotic number of non-isomorphic rooted trees obtained by rooting a tree." In: *Journal of Mathematical Analysis and Applications* 434.1 (2016), pp. 1–11.

[80] Hosam M. Mahmoud and Robert T. Smythe. "A survey of recursive trees." In: *Theo. Probability and Mathematical Statistics* 51 (1995), pp. 1–37.

[81] Hosam M. Mahmoud and Mark D. Ward. "Asymptotic distribution of two-protected nodes in random binary search trees." In: *Appl. Math. Lett.* 25.12 (2012), pp. 2218–2222. DOI: 10.1016/j.aml.2012.06.005. URL: https://doi.org/10.1016/j.aml.2012.06.005.

[82] Hosam M. Mahmoud and Mark D. Ward. "Asymptotic properties of protected nodes in random recursive trees." In: *J. Applied Probability* 52.1 (2015), pp. 290–297. DOI: 10.1017/S0021900200012365. URL: https://doi.org/10.1017/S0021900200012365.

[83] Toufik Mansour. "Protected points in k-ary trees." In: *Appl. Math. Lett.* 24.4 (2011), pp. 478–480. DOI: 10.1016/j.aml.2010.10.045. URL: https://doi.org/10.1016/j.aml.2010.10.045.

[84] Richard D. Mauldin, William D. Sudderth, and Stanley C. Williams. "Pólya trees and random distributions." In: *The Annals of Statistics* (1992), pp. 1203–1221.

[85] Amram Meir and John W. Moon. "On the altitude of nodes in random trees." In: *Canadian Journal of Mathematics* 30 (1978), pp. 997–1015.

[86] Peter D. Miller. *Applied Asymptotic Analysis*. Graduate studies in mathematics. American Mathematical Society, 2006. ISBN: 9780821840788.

[87] Małgorzata Moczurad, Jerzy Tyszkiewicz, and Marek Zaionc. "Statistical properties of simple types." In: *Math. Structures Comput. Sci.* 10.5 (2000), pp. 575–594. ISSN: 0960-1295. DOI: 10.1017/S0960129599002959. URL: https://doi.org/10.1017/S0960129599002959.

[88] Michał Morayne. "Partial-order analogue of the secretary problem - the binary tree case." In: *Discrete Mathematics* 184.1-3 (1998), pp. 165–181. DOI: 10.1016/S0012-365x(97)00091-5. URL: https://doi.org/10.1016/S0012-365x(97)00091-5.

[89] Richard Otter. "The number of trees." In: *Annals of Mathematics* 49.3 (1948), pp. 583–599. ISSN: 0003486X. URL: http://www.jstor.org/stable/1969046.

[90] Zbigniew Paiłka. "Testing an optimising compiler by generating random lambda terms." In: *Proceedings of the 6th International Workshop on Automation of Software Test*. ACM, 2011, pp. 91–97.

[91] Konstantinos Panagiotou and Benedikt Stufler. "Scaling limits of random Pólya trees." In: *Probab. Theory Related Fields* 170.3-4 (2018), pp. 801–820. ISSN: 0178-8051. DOI: 10.1007/s00440-017-0770-4. URL: https://doi.org/10.1007/s00440-017-0770-4.

[92] Alois Panholzer and Helmut Prodinger. "Level of nodes in increasing trees revisited." In: *Random Struct. Algorithms* 31.2 (2007), pp. 203–226.

[93] George Pólya. "Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen." In: *Acta Mathematica* 68.1 (1937), pp. 145–254. ISSN: 1871-2509. DOI: 10.1007/BF02546665. URL: https://doi.org/10.1007/BF02546665.

[94] Ryoma Sin'ya, Kazuyuki Asada, Naoki Kobayashi, and Takeshi Tsukada. "Almost every simply typed λ-term has a long β-reduction sequence." In: *Foundations of software science and computation structures*. Vol. 10203. Lecture Notes in Comput. Sci. Springer, Berlin, 2017, pp. 53–68.

[95] Neil J. A. Sloane. *The On-Line Encyclopedia of Integer Sequences (OEIS)*. http://oeis.org.

[96] Paul Tarau. "A hiking trip through the orders of magnitude: deriving efficient generators for closed simply-typed lambda terms and normal forms." In: *Logic-based program synthesis and transformation*. Vol. 10184. Lecture Notes in Comput. Sci. Springer, Cham, 2017, pp. 240–255.

[97] Stephan Wagner. "Central limit theorems for additive tree parameters with small toll functions." In: *Combin. Probab. Comput.* 24.1 (2015), pp. 329–353. ISSN: 0963-5483. DOI: 10.1017/S0963548314000443. URL: https://doi.org/10.1017/S0963548314000443.

[98] Jue Wang. "Generating random lambda calculus terms." In: (2005). Unpublished manuscript.

[99] Xuejun Yang, Yang Chen, Eric Eide, and John Regehr. "Finding and understanding bugs in C compilers." In: *ACM SIGPLAN Notices*. Vol. 46. 6. ACM. 2011, pp. 283–294.

[100] Noam Zeilberger. "Linear lambda terms as invariants of rooted trivalent maps." In: *J. Funct. Programming* 26 (2016), e21, 20. ISSN: 0956-7968. DOI: 10.1017/S095679681600023X. URL: https://doi.org/10.1017/S095679681600023X.

[101] Noam Zeilberger and Alain Giorgetti. "A correspondence between rooted planar maps and normal planar lambda terms." In: *Log. Methods Comput. Sci.* 11.3 (2015), 3:22, 39. ISSN: 1860-5974. DOI: 10.2168/LMCS-11(3:22)2015. URL: https://doi.org/10.2168/LMCS-11(3:22)2015.