UNIVERSITÉ PARIS NORD – PARIS 13
Institut Galilée

Laboratoire d'Informatique de Paris Nord

DOCTORAT
Spécialité informatique

Thèse pour obtenir le grade de docteur de l'Université Paris 13

Une image interdisciplinaire de la Combinatoire Analytique

An interdisciplinary image of Analytic Combinatorics

Présentée et soutenue publiquement le 18 septembre 2019 par

Sergey DOVGAL

devant le jury composé de

M.	Olivier BODINI
Mme.	Mireille BOUSQUET-MÉLOU
M.	Éric FUSY
M.	Vlady RAVELOMANANA
M.	Andrea SPORTIELLO
Mme.	Brigitte VALLÉE

Directeurs de thèse	M. Olivier BODINI
	M. Vlady RAVELOMANANA
Rapporteurs	M. Éric FUSY
	M. Valeriy LISKOVETS
	M. Konstantinos PANAGIOTOU

Dedicated to my students,
from whom I learned a lot.

Abstract

This thesis is devoted to the development of tools and the use of methods from Analytic Combinatorics, including exact and asymptotic enumeration, statistical properties of random objects, and random generation. The key ingredient is the multidisciplinarity of the domain, which is emphasised by using examples from computational logic, statistical mechanics, biology, mathematical statistics, networks and queueing theory.

Résumé

Cette thèse est consacrée au développement des outils et à l'utilisation des méthodes de la combinatoire analytique, notamment l'énumération exacte et asymptotique, les propriétés statistiques des objets aléatoires et la génération aléatoire. L'ingrédient clé est la multidisciplinarité du domaine, qui est soulignée par des exemples tirés de la programmation logique, de la mécanique statistique, de la biologie, de la statistique mathématique, des réseaux et de la théorie des files d'attente.

Preface

In my impression, a doctoral thesis in France is a mixture of personal and professional matters. The ALÉA community[1] impressed me a lot and has become my second family. Since my arrival in France, I have discovered many wonderful people with beautiful ideas and challenging projects.

While trying to stick to the scientific part of this thesis, I would like to briefly explain what motivated me to learn Analytic Combinatorics and try to do my best to motivate others. I got hooked on generating functions when I failed to solve the following combinatorial problem from the International Mathematical Olympiad Shortlist (2007) during the national selection days in Belarus in 2008 at the age of 15.

Problem C3 from IMO Shortlist-2007 (adapted). Assume that the set $S = \{1, 2, \dots, n\}$ can be colored red and blue, with the following condition being satisfied: the set $S \times S \times S$ contains exactly 2007 ordered triples (x, y, z) such that (i) x, y, z are of the same color, and (ii) $x + y + z$ is divisible by n. Prove that if the number of red and blue colored numbers is, respectively, r and b, then $r^2 - rb + b^2 = 2007$.

Solution (sketch). Let $k_1, k_2, \dots, k_r$ be the indices colored red, and $\ell_1, \dots, \ell_b$ be the indices colored blue. Consider a polynomial

$$F(x) = (x^{k_1} + \dots + x^{k_r})^3 + (x^{\ell_1} + \dots + x^{\ell_b})^3.$$

By expanding the brackets, we discover that the mth coefficient of $F(x)$ is equal to the number of same-colored triples of indices such that their sum equals m. Using a well-known formula for the sum of cubes

$$A^3 + B^3 = (A + B)(A^2 - AB + B^2),$$

we observe that $F(x)$ can be represented as a product

$$F(x) = (x^1 + \dots + x^n)\Big((x^{k_1} + \dots + x^{k_r})^2 - (x^{k_1} + \dots + x^{k_r})(x^{\ell_1} + \dots + x^{\ell_b}) + (x^{\ell_1} + \dots + x^{\ell_b})^2\Big).$$

By taking the sum of nth, $2n$th and $3n$th coefficients of $F(x)$ (i.e. the sum of coefficients whose indices are divisible by n), we observe that each coefficient of $(x^{k_1} + \dots + x^{k_r})^2 - (x^{k_1} + \dots + x^{k_r})(x^{\ell_1} + \dots + x^{\ell_b}) + (x^{\ell_1} + \dots + x^{\ell_b})^2$ contributes to the sum exactly once, and therefore, the desired sum is equal to the sum of coefficients of the second multiple term in $F(x)$. By substituting $x = 1$ into the second bracket, we obtain the desired identity. □

The fact that a solution to a combinatorial problem involved *polynomials*, impressed me a lot. The polynomials used in the proof are called *generating functions* and are now taught in almost every higher education mathematical program. Since then, I started to create my own collection of the problems using generating functions and share it with people, until in 2013 I discovered the course called *Analytic Combinatorics* led by Robert Sedgewick on the on-line educational platform *Coursera*. This course, based on the book with the same title written in co-authorship with Philippe Flajolet, explained how to lift formal power series manipulation to a combinatorial level, and showed how to use complex analysis to extract the asymptotics, which seemed a highly exotic but beautiful technique for me at that point.

[1] ALÉA, groupe de travail du GDR-IM in France and abroad http://gt-alea.math.cnrs.fr

Later, I have had a chance to teach this material at LESh (ecological summer school) and in MIPT (Moscow Institute of Physics and Technology) while I was a Master student in MIPT, at the department of Mathematical Foundations of Control. It was a great happiness to teach discrete mathematics to other students, trying to present the mathematical beauty of the subject and to reveal its applications. It was my duty to show the reasons why the subject has been present in their curriculum. This motivated me to search for more techniques and applications, until I finally discovered the French community and a rich heritage of Philippe Flajolet. It turned out that the contribution of analytic combinatorics extended far beyond what I previously knew: the applications include analysis of algorithms, number theory, statistical mechanics, quantum gravity, maps on surfaces, words, automata, graphs and phase transitions, random walks, queues and networks, concurrent processes, DNA and RNA, lambda terms, tilings, molecules... The variety of tools also very impressive: differential equations in several variables, algebraic geometry, topology and Morse theory, convex optimisation, special functions, matrix integrals, complex martingales, etc.

The division between the methods and applications can be highly subjective, as e.g. combinatorial objects such as graphs and directed graphs are, on one hand, objects of study by themselves, and on the other hand, serve as a language for describing objects from other disciplines. This thesis is about both, tools and their applications.

Acknowledgements

I would like to express my deepest gratitude:

- to Philippe Flajolet for putting so much energy into Analytic Combinatorics and teaching us through his wizard works;
- to the referees Éric Fusy, Valeriy Liskovets and Konstantinos Panagiotou for accepting to review this manuscript and for their valuable remarks; to the anonymous referees reviewing the conference papers and journal papers comprising parts of this thesis;
- to the jury members Mireille Bousquet-Mélou, Éric Fusy, Andrea Sportiello, Brigitte Vallée for accepting to be the part of the public defence;
- to my advisors, Olivier Bodini and Vlady Ravelomanana, for taking me into the wonderful community and teaching me many things;
- to my co-authors Maciej Bendkowski, Olivier Bodini, Julien Courtiel, Élie de Panafieu, Hsien-Kuei Hwang and Vlady Ravelomanana, please know that I am proud to have worked with you;
- for their insightful remarks and happy interesting discussions to Cyril Banderier, Anne Bouillard, Nicolas Curien, Ilya Galanov, Danièle Gardy, Bernhard Gittenberger, Christina Goldschmidt, Zbigniew Golębiewski, Leonid Kaganov, Mihyun Kang, Natasha Kharchenko, Isabella Larcher, Valeriy Liskovets, Marc Noy, Khaydar Nurligareev, Fedor Petrov, Yann Ponty, Juanjo Rué, Andrea Sportiello, Sergey Tarasov, Brigitte Vallée, Michael Wallner, Lutz Warnke, Noam Zeilberger;
- to Éric Fusy, Etienne Grandjean, Mireille Bousquet-Mélou, and Gérard Duchamp for honoring me with talk invitations at various occasions;
- à Nathalie Tavares et Coraline Nelson pour leur aide administrative décisive aux moments cruciaux;
- to Lionel Pournin and Olivier Bodini for supporting me in teaching for French students;
- to my teachers and students to have taught me mathematics;
- to my friends and colleagues for trusting me with their difficulties;
- to my family for letting me be myself;

The attributions are not necessarily disjoint.

Contents

Chapter 1

Introduction

Contents

According to Flajolet and Sedgewick [FS09], analytic combinatorics can be divided into two large parts: the *symbolic method* and *asymptotic analysis*. The first part is a language, or a certain algebra, which allows to turn natural combinatorial descriptions of objects into the equations on their respective generating functions. For example, the sequence a_n counting the number of chemical isomers of alcohols $C_nH_{2n+1}OH$ without asymmetric carbon atoms [Pól36] has generating function $F(z)$ satisfying a functional equation

$$F(z) = \frac{1}{1 - zF(z^2)}.$$

The structures whose generating functions satisfy equations involving substitutions, are sometimes now called *Pólya structures*. Due to universality and flexibility of the method, it has many applications outside the field of pure combinatorics. The classes of possible combinatorial objects that can be studied using the tools from analytic combinatorics include: (*i*) atonal musical combinatorics, which is study of combinations of chords and scales in the dihedral group of order 24 [Kei91]; (*ii*) objects from computational logic: propositional statements, "and/or" trees, conjunctive normal forms, lambda terms; (*iii*) random graphs and their phase

transitions, metric properties of graphs (diameter, circumference), planarity, colorability, etc; (*iv*) objects related to statistical mechanics, Ising model on various lattices, tilings and Bose–Einstein condensations; (*v*) random maps on surfaces with possible decorations and degree constraints; random walks with various types of constraints; permutations and chord diagrams; (*vi*) queues and networks; (*vii*) biological objects: DNA and RNA sequences, philogenetic trees in the context of evolution; (*viii*) in general, any objects whose multivariate generating functions can be specified using systems of equations, possibly involving substitutions, derivatives, integrals and non-linearities.

The second part, asymptotic analysis, is required to process the equations obtained on the previous step and to extract the asymptotic properties. The involved techniques come mostly from real-valued or complex-valued analysis, and include, for example, *Tauberian analysis*, *singularity analysis* and *saddle point method*.

An exemplary execution of the singularity analysis is an asymptotic expansion of $n!$ in Stirling's formula. Applying standard techniques to generating function of Cayley trees

$$T(z) = ze^{T(z)}, \quad T(z) = \sum_{n \geqslant 0} \frac{n^{n-1}}{n!} z^n$$

one obtains

$$\frac{n^{n-1}}{n!} \sim \frac{e^n}{\sqrt{2\pi n^3}} \left(1 - \frac{1}{12n} + O(n^{-2})\right).$$

The methods allow to obtain the asymptotics in the form of the main asymptotic term or often the full asymptotic expansions, for the enumeration of objects and for limiting distributions of numerous parameters inside these objects.

From "bird's eye view", the three most typical problems from enumerative combinatorics include

1. *Enumeration.* To each object a *size* is assigned, and it is asked, *how many objects of size n does there exist*?

2. *Statistical properties.* To each object a set of *parameters* is assigned (e.g. number of vertices and root face degree in a map), and it is asked, *what is the distribution (joint distribution) of these parameters*?

3. *Random generation.* Generate uniformly at random an object of given size n. Sometimes the distribution is not uniform and additional assumptions are superimposed.

Our contributions extend the techniques of analytic combinatorics and the range of their applicability in several ways.

The present thesis develops on the contributions written in cooperation with my coauthors. Here, a broader context of rapidly developing methods and applications is given. The picture of how different methods and applications match is summarised in Table 1.1.

	λ-calculus	networks	mathematical statistics	queueing theory	biology	statistical mechanics
infinite algebraic systems	de Bruijn size notion					
Boltzmann tuning	formal verification	Bianconi–Barabási model	combinatorial learning	multi-customer w/proc. sharing	RNA design simulation	tilings, BEC
partial diff. eq.	linear λ-terms					random maps
Airy function	height profile	graphs with degree constraints				2-SAT

Table 1.1: Crossroad of methods and their applications discussed in the current thesis

1.1 Elements of computational logic

Computational logic forms a basis of the modern proof systems. The *propositional logic*, also called *zeroth-order logic* operates with Boolean expressions and their transformations. Using the axioms and inference rules, it is possible to obtain an inductively defined language of statements. The first-order logic and second-order logic have increasing expressivity allowing universal and existential quantification $\forall$ and $\exists$. The difference between first-order and second-order logics is that the latter allows to quantify over sequences and subsets, while the first allows only to quantify over finite sets of variables.

1.1.1 Propositional logic

Propositional logic is sometimes called *zeroth-order logic* and does not use universal and existential quantifiers $\exists$ and $\forall$. Each formula is a Boolean expression in several variables such as $P \to Q$, $(P \vee Q) \to R$, etc. In propositional calculus, one of the major focuses belongs to the *proofs* which are tightly related to automatic proofs generated by proof assistants.

A set of *premises* is a set of boolean formulae like, for example, a set $\{P, P \to Q\}$. These formulae are not necessarily tautologies. Then, a set of *axioms* is defined. An example of such an axiom is *modus ponens* which is symbolically written as $\dfrac{A, A \to B}{B}$ meaning that if the statements A and $A \to B$ belong to the set of premises, then a formula B can be deduced. Various sets of deduction axioms give different examples of propositional calculus systems. Finally, a *proof* is a sequence of applications of the inference rules to an initial set of premises.

One of the examples is the *natural deduction system* which uses only the unary operation of negation $\{\neg\}$ and binary operators $\{\wedge, \vee, \to, \leftrightarrow\}$. The set of inference rules contain eleven different rules. We shall return to the natural deduction system later in the context of Curry–Howard correspondence.

1.1.2 Lambda calculus

Lambda terms represent anonymous functions which can, in their turn, take functions as arguments, and return functions. Recursively, each function may take functions as arguments, et cetera. For example, in a rather informal way,

$$f(x, y) := x(y)$$

defines a function f which takes as an input two arguments x and y, where x is also a *function*, and returns a result of application of x to y. In a dedicated *lambda notation* this function is written as

$$f := \lambda x.\lambda y.xy := \lambda x.\lambda y.(xy),$$

where $\lambda x.$ and $\lambda y.$ are sorts of "declarations" that x and y are going to be the arguments of the function, and xy denotes the application of x and y, just as in group theory, the composition is sometimes denoted as a product and no symbol between the two elements is used. When three terms are applied to one another, the associative rule does not apply and parenthesis may be necessary.

Result 1. In the paper [BBD18b] in co-authorship with Maciej Bendkowski and Olivier Bodini, we present a quantitative, statistical analysis of random lambda terms in the de Bruijn notation. Following an analytic approach using multivariate generating functions, we investigate the distribution of various combinatorial parameters of random open and closed lambda terms, including the number of redexes, head abstractions, free variables or the de Bruijn index value profile. Moreover, we conduct an average-case complexity analysis of finding the leftmost-outermost redex in random lambda terms showing that it is on average constant. The main technical ingredient of our analysis is a novel method of dealing with combinatorial parameters inside certain infinite, algebraic systems of multivariate generating functions. Finally, we briefly discuss the random generation of lambda terms following a given skewed parameter distribution and provide empirical results regarding a series of more involved combinatorial parameters such as the number of open subterms and binding abstractions in closed lambda terms.

The *types* are then introduced, so that a lambda expression can be correctly interpreted. In the above example, we might say that y is of type a, and therefore, x should be of type $a \to b$, because it takes y as an argument. Combining, we could write that the function f has the type $(a \to b, a) \to b$. It is not common to use tuples when writing types, in type theory of lambda terms and a different notation is used.

In fact, if a function takes several arguments, like $f(x, y, z)$, it can be represented as a function that takes one argument x and returns a function of two arguments y and z: $x \mapsto g(y, z)$. The same applies to the function g. Therefore, if x has type a, y has type b, z has type c, and f returns a value of type d, then it can be said that $f(x, y, z)$ has type $a \to b \to c \to d := a \to (b \to (c \to d))$. Continuing the previous example, the lambda term $\lambda x.\lambda y.xy$ can be assigned a type $(a \to b) \to (a \to b)$.

Some lambda terms cannot be assigned any type, as, for example, $\lambda x.xx$, as it involves substituting x into itself, which should mean that x is a function, and is therefore simultaneously of types $a \to a$ and a. The problem of typeability of a lambda term is undecidable.

1.1.3 Expressivity of lambda calculus

Lambda calculus with types by itself turned out to be a universal model of computation, along with a Turing machine. In fact, the celebrated Church–Turing thesis can be viewed as the equivalence of the definitions of the computable functions in terms of either Turing machines or lambda calculus. One step of computation in lambda calculus corresponds to a certain operation called *beta reduction* which can be viewed as a substitution step of the function composition. An example of such substitution can be performed on the term $f := \lambda y.((\lambda x.xx)y)$, despite the latter being untyped. The function represented by lambda term $\lambda x.xx$ applies its argument to itself, and therefore, application to y yields y. Consequently, beta-reduction of f is a lambda term $\lambda y.(yy)$. Basing on a choice of which application to expand, several beta-reduction strategies are possible. The language of lambda terms, equipped with beta reduction, is of the same expressivity as Turing machines with a notion of a computation step.

A final touch on computation logic is *Curry–Howard isomorphism*. Eventually, the typing rules can be written in a similar manner as the deduction rules from logic systems. Assuming the notation $f\colon \alpha$ meaning that a lambda term f is of type α, one of the examples of type deduction rules can be written as follows:

$$\frac{\Gamma \vdash F_1\colon \alpha \to \beta; \quad \Gamma \vdash F_2\colon \alpha}{\Gamma \vdash F_1F_2\colon \beta}$$

which means that an application of two lambda terms of types $\alpha \to \beta$ and α inherits the type β. Γ denotes a set of premises from which type deduction is performed. A corresponding inference rule is the *modus ponens* rule mentioned above:

$$\frac{\Gamma \vdash \alpha \to \beta; \quad \Gamma \vdash \alpha}{\Gamma \vdash \beta}.$$

This correspondence or isomorphism allows to build a bijection between *statements* and *lambda terms*, between *true statements* and *typeable lambda terms*, and finally, between *proofs* and *types*.

One of the major challenges in formal verification and enumerative combinatorics is to enumerate and randomly generate typeable lambda terms, and consequently, lambda terms of given type. Given that the typeability problem is undecidable, no simple solution is available.

1.2 Phase transitions

Phase transitions in random graphs, directed graphs, hypergraphs, random geometric complexes, percolations, real-world networks and constraint satisfiability problems (CSP) have become rapidly developing areas with a huge spread of interconnected publications.

The term "phase transition" was mostly used by physicists, but can be also applied to many combinatorial situations, when a small change of a certain parameter results in a huge asymptotic change of some other parameter. The original studies of the physical phase transitions, including Ising and Potts model, considered graphs which formed certain regular lattices: from rectangular ones to more complicated including maps on

surfaces. Of close relation is the percolation theory which is sometimes called "the simplest model displaying a phase transition". Friendly introductions into percolation theory and Potts models can be found in a PhD Thesis [Dom13], a survey paper [Bea+10] and in a lecture course [Dum17].

1.2.1 Zero-one laws

Zero-one laws in random graphs, mappings and other structures form an interesting mixture between combinatorics and logics. A wonderful introduction to the subject is given in the book [Spe13]. Zero-one laws can be roughly viewed as the precursors of phase transitions, while the latter may be considered as the refined versions. *Per contra*, not every property is monotonic in the number of edges, and the existence of a threshold function for a property is not always guaranteed.

In a random graph $G(n, p(n))$, a property A holds *almost surely* if

$$\lim_{n\to\infty} \mathbb{P}[G(n, p(n)) \models A] = 1,$$

and *almost never* if

$$\lim_{n\to\infty} \mathbb{P}[G(n, p(n)) \models A] = 0.$$

For $p = 1/2$, for the case when all the graphs with given number of vertices are drawn uniformly at random, the following statement holds.

Zero-one property for labelled graphs (Fagin's theorem). If A is a property expressible in first-order logic, then either this property A holds for $G(n, 1/2)$ *almost never* or *almost surely*.

Remarkably, connectedness and bipartiteness are not expressible in first order logic, and such properties as the presence of a triangle, are. The above property breaks for *random mappings*, so that the zero-one law does not anymore necessarily hold, but a different theorem can be stated. The details can be found in [FS09, p. 467].

First-order sentences for random mappings (Lynch's theorem). Let $\mu_n(A)$ be the probability that a random mapping $\varphi\colon \{1, 2, \dots, n\} \to \{1, 2, \dots, n\}$ satisfies A. Then, the limit $\lim_{n\to\infty} \mu_n(P)$ exists and its value is given by an expression constructed from integer constants and the operations $+, -, \times, \div$, and e^x.

For some edge probabilities like $p(n) = 1/n$, the limit of the probability that a random labelled graph satisfies property A, exists, but its value is not equal to 0 and 1. One of the major results of [SS88] and [Spe13] is that the exponent of the critical probability is forced to be a rational number.

Critical exponent characterisation. If $\alpha \in (0, 1)$ and α is irrational, then $p(n) = n^{-\alpha}$ satisfies the zero-one law relative to any first-order sentence A, that is, either the property A holds for $G(n, p(n))$ *almost never* or *almost surely*.

The aforementioned results can be regarded as a broader guideline to a narrower refine analysis of phase transitions in various combinatorial structures.

1.2.2 Giant component and the Airy function

Airy function plays a central role in the analysis of asymptotics of the phase transitions [Ban+01; FSS04] and also played its role in Kontsevitch's proof of Witten's conjecture [Kon92]. It is defined as

$$A(y) = \int_{-\infty}^{\infty} e^{i(x^3/3 - xy)} dx$$

and satisfies a second-order linear differential equation

$$A''(y) + yA(y) = 0.$$

A matrix version of the Airy function also appears in the doctoral thesis of Kontsevich [Kon92], which constitutes the proof of celebrated Witten's conjecture, enumerative proof of the equivalence of two models of

quantum gravity. The omnipresence of Airy function in combinatorics is of the same nature as omnipresence of Gaussian law in probability: while the philosophical reason for Gaussian law is the disappearance of the first derivative in the log-density of the distribution:

$$\log f(z) \approx g(z_0) + \frac{g''(z_0)}{2!}(z - z_0)^2, \quad g(z) = \log f(z),$$

Airy-related distributions appear when two first derivatives vanish:

$$\log f(z) \approx g(z_0) + \frac{g'''(z_0)}{3!}(z - z_0)^3, \quad g(z) = \log f(z).$$

The reason why the second derivative may vanish, is that systems prone to phase transitions contain a variable parameter α (in the case of simple graphs, ratio between the number of edges and vertices), and the second derivative of $\log f(z)$ depends on this parameter α. For some systems, there exists a parameter $\alpha = \alpha_0$ such that the second derivative vanishes at this value of the parameter. This explains "Airy phenomenon" in such systems.

An example of Airy phenomenon in the context of phase transitions is the probability that a graph consists only of trees and unicycles which is expressed in terms of the Airy function when the ratio of the number of edges m and the number of vertices n approaches $1/2$ at speed $n^{-1/3}$. The same phenomenon can be observed in graphs with degree constraints, with a different value of the ratio determined by a system of functional equations.

The term *giant component* is used to define the largest connected component of size $\Theta(n)$ in a graph with n vertices and $m = cn$ edges, $c > 1/2$. This component is not a tree neither a unicycle, so it becomes identical to the *complex component* of the graph, which is the set of connected components whose number of edges is greater than its number of vertices. Airy function plays the central role in the description of the period when $c \approx 1/2$ and the giant component starts to appear [Jan+93].

Result 2. In the paper [DR18] jointly with Vlady Ravelomanana, we show that by restricting the degrees of the vertices of a graph to an arbitrary set Δ, the threshold point $\alpha(\Delta)$ of the phase transition for a random graph with n vertices and $m = \alpha(\Delta)n$ edges can be either accelerated (e.g., $\alpha(\Delta) \approx 0.381$ for $\Delta = \{0, 1, 4, 5\}$) or postponed (e.g., $\alpha(\{2^0, 2^1, \cdots, 2^k, \cdots\}) \approx 0.795$) compared to a classical Erdős–Rényi random graph with $\alpha(\mathbb{Z}_{\geqslant 0}) = \frac{1}{2}$. In particular, we prove that the probability of graph being non-planar and the probability of having a complex component, goes from 0 to 1 as m passes $\alpha(\Delta)n$. We investigate these probabilities and also different graph statistics inside the critical window of transition (diameter, longest path and circumference of a complex component).

A different result using similar techniques can be obtained for the phase transition of 2-SAT.

Result 3. In the paper [Dov19] we prove that, with high probability, the contradictory components of a random 2-SAT formula in the subcritical phase of the phase transition have only 3-regular kernels. This follows from the relation between these kernels and the complex component of a random graph in the subcritical phase. This partly settles the question about the structural similarity between the phase transitions in 2-SAT and random graphs. As a byproduct, we describe the technique that allows to obtain a full asymptotic expansion of the satisfiability in the subcritical phase. We also obtain the distribution of the number of contradictory variables and the structure of the spine in the subcritical phase.

1.2.3 Related studies

Apart from the existing practical applications in hardware in software engineering (e.g. cuckoo hashing [DM03; Die+10]), phase transitions in graphs and digraphs are studied in their own right. Of the most recent references on random graphs and networks can be mentioned the books [JŁR11; Van16]. Concerning the phase transition in *directed* graphs, the width of the transitions window has been described in [ŁS09] and a description of the giant core is given in [PP17]. Recent studies reveal the properties of phase transitions in random hypergraphs [dPan15; CKP18] and simplicial complexes [Coo+18].

One of the most recent surveys on satisfiability is a part of the excellent "Art of Computer Programming, Volume 4" by Knuth [Knu15]. The k-SAT problem has also played its role on the intersection of theoretical computer science and theoretical physics. One of the surprising connections is that several NP-complete decision problems can be reformulated in terms of positivity of the ground state energy of an Ising model and that the techniques like belief propagation and cavity method used in statistical physics can be also used to predict the behaviour of random formulae, see [MZ97; BCM02; MPZ02]. At the same time, phase transitions in random CSP seem to be closely related to their algorithmic complexity, see [AC08]. While the existence of satisfiability threshold and its value for k-SAT with $k \geqslant 3$ remains an unsolved open problem, several advances have been made, including a precise asymptotic location of the transition point for $k \to \infty$ [CP16], phase transitions of k-XORSAT [HR11; PS16], and more generally, systems of linear equations over F_q [Ayr+17], improvements in k-colorability threshold [CV13].

1.3 Permutations and maps on surfaces

Maps or *graphs embedded into surfaces*, are now one of the central objects of study of modern mathematical physics. The recurrences related to maps appear in such fields as quantum field theory, quantum gravity, string theories, integrable systems. The popularity of maps has led to creation of multiple research groups, conferences and dedicated events during a couple of last decades.

Theory of partial differential equations is closely related to enumeration through construction of multivariate generating function. The derivative and multiplication operations have their own combinatorial meaning though the symbolic method. On the other hand, along with a combinatorial interpretation of the differential equations, it is somewhat curious to learn about their physical interpretation as well.

For example, let

$$F = \frac{1}{24}t_1 + \frac{1}{6}t_0^3 + \frac{1}{48}t_1^2 + \cdots$$

denote a formal power series of infinitely many variables $(t_0, t_1, \ldots)$ called *Witten's generating function* [Kon92]. We do not present its rigorous definition here, as it necessitates certain background from Riemann surfaces, moduli curves and Chern classes. However, the role of maps in the definition of this generating function consists in the correspondence between ribbon graphs and certain moduli spaces. Witten's conjecture, and later, Kontsevitch's theorem, states that the second derivative $u = \partial^2 F/\partial t_0^2$ satisfies Korteweg–de Vries equation

$$\partial_t u = uu_x + \frac{1}{12}u_{xxx}, \quad x := t_0, \quad t := t_1.$$

At the same time, if t denotes the time variable and x denotes the spacial coordinate, this equation has a physical meaning of wave model on shallow water surface. This connection seems now to be only a top of the iceberg and people continue discovering new connections between hierarchies of physical equations and multivariate generating functions of combinatorial families [GJ08; Cha].

While most of the map families appearing in mathematical physics can be embedded into a surface with bounded genus, the question of behaviour of maps without fixed genus is also relevant. Thus said, the enumeration of such map families gives another non-linear equation known as *Riccati equation* appearing in physical expansions [Kud91] and population dynamics [NR02]. A different viewpoint on maps consists in representing a map as a pair of permutations (σ, α) such that α is an involution and the subgroup generated by σ and α acts transitively on $\{1, 2, \ldots, n\}$.

Result 4. In the paper [Bod+18a], written jointly with Olivier Bodini, Julien Courtiel and Hsien-Kuei Hwang, we consider random rooted maps without regard to their genus, with fixed large number of edges, and address the problem of limiting distributions for six different parameters: vertices, leaves, loops, root edges, root isthmus, and root vertex degree. Each of these leads to a different limiting distribution, varying from (discrete) geometric and Poisson distributions to different continuous ones: Beta, normal, uniform, and an unusual distribution whose moments are characterised by a recursive triangular array.

1.4 Boltzmann sampling

The principle of recursive generation and Boltzmann sampling can be most easily illustrated using the example of *Catalan binary trees*, which are defined as rooted binary trees. It is known that the number of Catalan trees with n nodes satisfies the recurrence

$$T_n = \sum_{k=1}^{n-1} T_k T_{n-k-1}, \quad T_1 = 1. \tag{1.4.1}$$

Moreover, this recurrence contains *structural information* about these binary trees: the product operation between T_k and T_{n-k-1} corresponds to joining together two trees of sizes k and $n-k-1$ to form a new tree of size n. A natural way to generate a Catalan tree of size n uniformly at random in a recursive manner is to generate a random variable Ξ_n from discrete distribution

$$\mathbb{P}(\Xi_n = k) = \frac{T_k T_{n-k-1}}{\sum_{k=1}^{n-1} T_k T_{n-k-1}}, \tag{1.4.2}$$

and to form a binary tree by linking two recursively generated trees with $k = \Xi_n$ and $n-k-1$ vertices.

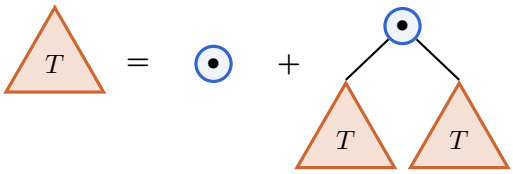


Figure 1.1: Combinatorial representation of the recursion for binary Catalan trees.

The complexity of such a method (called *recursive method*) is quadratic in target size n. Boltzmann sampler is a different algorithm allowing a faster generation, in quasi-linear time, but with an approximate size. The method consists of taking the equation which generation function $T(z) = \sum_{n \geqslant 0} T_n z^n$ satisfies

$$T(z) = z + zT^2(z), \tag{1.4.3}$$

fixing a positive value $z \in (0, 1/2)$, and making a Bernoulli choice Ξ_z such that

$$\mathbb{P}(\Xi_z = 0) = \frac{z}{z + zT^2(z)}, \quad \mathbb{P}(\Xi_z = 1) = \frac{zT^2(z)}{z + zT^2(z)}. \tag{1.4.4}$$

If $\Xi_z = 0$, the generation outputs a single node and stops, otherwise it recursively calls two Boltzmann samplers with a parameter z and outputs a tree constructed by linking two recursively generated trees.

This elegant algorithm was designed [Duc+04] and provides a fast generation tool in the cases when the exact parameter value constraint can be relaxed.

Recursive generation of combinatorial structures is discussed in more detail in Chapter 5. Several applications of multiparametric sampling are discusses in Chapter 10.

1.4.1 Generation from context-free grammars

Generation of words uniformly at random from the language defined by an unambiguous context free grammar can be settled using either the recursive approach or Boltzmann sampling approach. A *context free grammar* is defined as the set of its production rules, such as the context-free grammar for well-formed parentheses

$$S \to SS \mid (S) \mid ().$$

A context-free grammar is *unambiguous* if every word derived from the grammar has a unique leftmost derivation.

When sampling words from context-free grammars using Boltzmann sampler, the output size of the word is not fixed, and in fact, is a random variable. The dedicated *tuning* procedure consists in choosing the algorithm parameter z, which is the argument of the generating functions, giving the desired expected size of the generated word. The case of unambiguous context free grammars corresponds to systems of algebraic equations

$$\boldsymbol{F}(z) = \boldsymbol{\Phi}(\boldsymbol{F}(z), z),$$

where $\boldsymbol{F}(z) = \big(F_1(z), \dots, F_m(z)\big)$ denotes a vector of generating functions, where

$$F_k(z) = \sum_{n \geqslant 0} a_{n,k} z^n,$$

and $a_{n,k}$ denotes the number of words of length n which can be produced from kth non-terminal.

As shown in the work [PSS12] tightly related with previous works on numerical iterations for species theory [BLL98], Newton's iteration provides fast convergence towards the target value $\boldsymbol{F}(z)$, once the argument z is given. Despite the large body of work on tuning and sampling from context-free grammars, including [BGR15; BP10; Bod10; BRS12; BBJ13; BFR17], the particular choice of the schemes: singular vs. pointed Boltzmann sampler, approximate vs. exact, tuning precision, behaviour of the Newton oracle near the singularity point, etc. can be far from obvious.

1.4.2 Multiparametric sampling

While the principles of Boltzmann sampling can be quickly generalised to the multiparametric setting corresponding to weighted generation, the process of finding the arguments of generating functions corresponding to given weight, is not trivial. In the multiparametric setting, for the case of context free unambiguous grammars, the system of algebraic equations takes form

$$\boldsymbol{F}(z_1, \dots, z_n) = \boldsymbol{\Phi}(\boldsymbol{F}(z_1, \dots, z_n), z_1, \dots, z_n). \tag{1.4.5}$$

Instead of marking the total length of the word as in the case of univariate generating functions, the variables $z_1, \dots, z_n$ can mark the proportions of certain symbols inside the generated words, or some local patterns.

Result 5. In the paper [BBD18a], written in co-authorship with Maciej Bendkowski and Olivier Bodini, we propose an efficient polynomial-time, with respect to the number of tuned parameters, tuning algorithm based on convex optimisation techniques. We illustrate the efficiency of our approach using several applications of rational, algebraic and Pólya structures including polyomino tilings with prescribed tile frequencies, planar trees with a given specific node degree distribution, and weighted partitions. Our implementation is available on-line, `http://github.com/maciej-bendkowski/boltzmann-brain`.

While the family of languages derived from unambiguous context-free grammars may seem not very expressive, it turns out that it is possible to carry out the analysis by specifying only a part of a non context free language, for example, the set of words of a certain fixed length, by incorporating this length parameter into the context free grammar itself. This technique may require very long and complex descriptions, potentially, exponentially long, but tractable in some cases, [Ham+19]. We consider several such examples in Chapter 10.

Some examples of families, including graphs, tilings, lambda terms and weighted integer partitions have been implemented, see Section 10.6.

1.4.3 Miscellaneous applications

The multidimensional tuning presented in Chapter 5 can be applied in many different contexts, including software verification, queueing theory, biology and statistical mechanics. Several such applications are discussed in detail in Chapter 10. In particular, Bianconi–Barabási model, along with related Bose–Einstein condensation model is discussed in Section 10.3; *combinatorial learning* which is a new technique mixing

mathematical statistics and enumerative combinatorics, is introduced in Section 10.5; queueing networks with multiple customer types and processor sharing is introduced in Section 10.4; one more application communicated to me by Yann Ponty and Sebastian Will, is presented in Section 10.2. Apart from these various applications, several proof of concept implementations for tilings, random graphs, lambda terms and weighted integer partitions have been brought into life and are available at https://github.com/maciej-bendkowski/multiparametric-combinatorial-samplers. They are discussed in detail in Section 10.6.

1.5 A few words about the tools

1.5.1 Tree-like and graph-like combinatorial structures

The first noticeable feature of formal power series is that if the analysed generating function $f(z)$ is analytic at zero, then its coefficients $[z^n]f(z)$ grow not faster than C^n for some constant $C > 0$. For exponential generating functions, which enumerate labelled objects, i.e. generating functions of the kind

$$f(z) = \sum_{n\geqslant 0} \frac{a_n z^n}{n!},$$

where a_n denotes the number of labelled objects of size n, the growth of a_n is bounded by $n!C^n = e^{n\log n + O(n)}$. This behaviour is typically the case for *simply-generated trees* whose generating function satisfies

$$f(z) = z\phi(f(z)),$$

where $\phi(t)$ is an entire function of t. At the same time, very simple combinatorics tells us that the number of *simple graphs* on n labelled nodes with arbitrary number of edges is $2^{\binom{n}{2}} = e^{\Theta(n^2)}$. Such coefficient growth implies that the generating function of simple graphs is not analytic.

This observation shows a gap between the enumeration of tree-like and graph-like structures. The word "analytic" from the term "analytic combinatorics" does not anymore apply to such a divergent series

$$f(z) = \sum_{n\geqslant 0} 2^{\binom{n}{2}} z^n,$$

since it is not analytic at $z = 0$. There are several ways to handle this difficulty.

First approach. *Consider the series formally and carry out certain manipulations on the coefficient level.* This approach can be illustrated by looking at the asymptotics of the coefficients of $f(z)^2$ where $f(z) = \sum_{n\geqslant 0} n!z^n$. The coefficient at monomial z^n of $f^2(z)$ can be written as the convolution sum

$$[z^n]f(z)^2 = \sum_{k=0}^{n} k!(n-k)!.$$

Denoting $a_k := k!(n-k)!$, we note that $a_1 = O(a_0/n)$, $a_2 = O(a_1/n)$, and each next summand is negligible by a factor $O(n^{-1})$. This allows to write

$$[z^n]f(z)^2 \sim 2n!(1 + O(n^{-1})).$$

Certainly, this method requires further rigorous justification for the error terms.

Second approach. *Represent the function as analytic at $z \neq 0$.* The absence of analyticity at $z = 0$ does not necessarily mean that the function cannot be analytically continued to the domain where the real part of the argument is, for example, negative. The formula

$$\int_0^\infty \frac{e^{-t}}{1-zt} dt \sim \sum_{n=0}^{\infty} n!z^n$$

holds by coefficient-wise differentiation at $z = 0$ as $z < 0$, $z \to 0$.

Third approach. *Decompose the generating function into an infinite sum of analytic ones.* Each of the contributions can be analysed separately, and then summed up back again. Such an approach is employed for analysis of simple graphs near the point of phase transition: the bivariate generating function of graphs $G(z, w)$ with z marking vertices and w marking edges is decomposed into

$$G(z, w) = \sum_{r \geqslant 0} e^{U(z,w)+V(z,w)} E_r(z, w)$$

where $U(z, w)$ is the bivariate generating function for unrooted labelled trees, $V(z, w)$ is the bivariate generating function for unicyclic connected graphs, and $E_r(z, w)$ stands for graphs whose each connected component has at least one edge more than the number of vertices (and represents a so-called *complex component* of a graph). It turns out that such a decomposition allows to catch the main asymptotic contribution. Such decompositions require, as a rule, noticing certain structural patterns.

This list of three approaches is not extensive, as coefficient normalising techniques such as *Borel transform* can be applied. In one of the papers we develop a symbolic approach for directed graphs which consists of introducing a normalising factor $(1 + w)^{\binom{n}{2}}$, where n denotes the number of vertices of a directed graph, and w marks the number of the directed edges.

Result 6. In paper [dPD19] together with Élie de Panafieu, we introduce the arrow product, a new generating function technique for directed graph enumeration. It provides new short proofs for previous results of Gessel on the number of directed acyclic graphs and of Liskovets, Robinson and Wright on the number of strongly connected directed graphs. We also obtain new enumerative results on directed graphs with given numbers of strongly connected components and source-like components using this new technique.

We apply the third approach to study of the 2-SAT phase transition in [Dov19].

1.5.2 Catalytic equations

One particular kind of equations appearing in enumeration of walks and maps, is called *catalytic equations*.

Consider an example of such an equation. Let $F(z, u)$ be the generating function of *rooted planar maps* with z marking vertices and u marking the degree of the root face. Then,

$$F(z, u) = 1 + zu^2 F(z, u)^2 + zu \frac{uF(z, u) - F(z, 1)}{u - 1}.$$

Both bivariate and univariate generating functions $F(z, u)$ and $F(z, 1)$ are unavailable directly from this equation. The reason why such equations are called catalytic, is that the extra variable u is necessary to define an equation for the generating function, but putting $u = 1$ does not completely eliminate this variable, as it introduces a new term $\partial_u F(z, 1)$. Using a dedicated elimination procedure [BJ06], it is possible to prove that the univariate generating function for maps $H(z) = F(z, 1)$ satisfies a polynomial equation

$$H(z) = 1 - 16z + 18zH(z) - 27z^2 H(z)^2.$$

The general algebraic case with one catalytic variable has been successfully settled in [BJ06]. Algebraic catalytic equations with two or more variables, and also differential catalytic equations present a challenging problem.

While direct solution of the catalytic equation is unavailable, method of moments can be still applied to obtain the distribution of parameters. One example of such an application is our paper [Bod+18a] written in co-authorship with Olivier Bodini, Julien Courtiel and Hsien-Kuei Hwang. The catalytic differential equation for generating function $Y(z, v, w)$ of rooted connected maps with z marking the number of edges, v marking root degree and w marking the number of loops, is

$$Y(z, v, w) = 1 + v^2 zY(z, v, w) + vzY(z, 1, w)Y(z, v, w) + 2vz^2 \partial_z Y(z, v, w) + v^2 z(vw - 1)\partial_v Y(z, v, w).$$

1.5.3 Infinite systems

The tree-like structures which appear from a generalised version of the simply-generated tree equation $f(z) = z\phi(f(z))$, satisfy systems of polynomial equations

$$\begin{aligned} F_1(z_1, \dots, z_d) &= \Phi_1(F_1, \dots, F_n, z_1, \dots, z_d), \\ &\dots \\ F_n(z_1, \dots, z_d) &= \Phi_n(F_1, \dots, F_n, z_1, \dots, z_d). \end{aligned}$$

The analysis of such multivariate systems already became classic, typically handled through the celebrated *Drmota–Lalley–Woods theorem* which is a variation of the implicit function theorem for the case of algebraic systems of combinatorial origin.

The above theorem can be generalised for the case when the number of variables and the number of functions is infinite, [DGM12]. Among several technical conditions required in the referenced paper is the condition that the Jacobian operator of the infinite system can be represented as a sum of a scaled identity operator and an operator whose power is compact. Geometrically speaking, a *compact operator* is the one that can be approximated by a finite-dimensional one; an example of a non-compact operator in an infinite-dimensional space is the identity operator.

In the paper [BBD18b] written in co-authorship with Maciej Bendkowski and Olivier Bodini, we design a new tool for infinite systems of type

$$L_m(z, \boldsymbol{u}) = \Phi_m(L_m(z, \boldsymbol{u}), L_{m+1}(z, \boldsymbol{u}), z, \boldsymbol{u}), \quad m = 0, 1, 2, \dots$$

which arises in the enumeration of closed lambda terms. Among the listed condition is the condition of *exponential convergence* which requires convergence of Φ_m to a limiting operator Φ_∞ at an exponential speed. Arguably, from the viewpoint of analytic combinatorics, our novel result complements the existing result of Drmota, Gittenberger and Morgenbesser [DGM12] on infinite systems. In our formulation, the infinite system is not required to be strongly connected. Consequently, we conjecture that the properties of the Jacobian operator of the infinite system are not sufficient to deduce the result, in contrast with the mentioned paper. In other words, we conjecture that the condition of exponential convergence is essential.

1.5.4 Convex optimisation

The problem of convex mathematical programming is formulated as

$$\begin{aligned} f_0(\boldsymbol{z}) &\to \max_{\boldsymbol{z}}, \\ f_k(\boldsymbol{z}) &\leqslant 0, \quad k = 1, \dots, n, \end{aligned}$$

where the functions $f_0(\boldsymbol{z})$ and $f_k(\boldsymbol{z})$ are convex in $\boldsymbol{z}$ for $k = 1, \dots, n$ and $\boldsymbol{z} \in \mathbb{R}^d$ is a d-dimensional variable. It is known that if the functions are convex, the solution $\boldsymbol{z}^*$ where $f_0(z)$ attains its maximum, belongs to a convex set. Under certain regularity conditions, the solution is unique. In a more computer-interpretable form, where the solution is required up to a real precision $\varepsilon > 0$, this formulation may be interpreted in several different forms:

(*i*) Find $\boldsymbol{z}_\varepsilon$ such that $\|\boldsymbol{z}_\varepsilon - \boldsymbol{z}^*\| < \varepsilon$; the choice of the norm may vary;

(*ii*) Find $\boldsymbol{z}_\varepsilon$ such that $|f_0(\boldsymbol{z}_\varepsilon) - f_0(\boldsymbol{z}^*)| < \varepsilon$.

The first and the second formulations can be related through the matrix of second derivatives of $f_0(\boldsymbol{z})$ at $\boldsymbol{z} = \boldsymbol{z}^*$, and also that of $f_k(\boldsymbol{z})$ near that point. Every convex programming problem can be reformulated in a way that $f_0(\boldsymbol{z}) := \boldsymbol{z}$ by introducing an additional inequality constraint.

It is widely believed that *non-convex* problems are computationally hard to solve, while the *convex* ones are polynomially easy to solve. In practice, the complexity of the solution of a convex programming problem

depends on the curvature matrix of the problem specification and this curvature matrix norm can experience an exponential blow-up in the external parameters of the input.

One of the results obtained in [BBD18a] is the formulation of the multiparametric Boltzmann tuning problem as a convex optimisation problem. A peculiar feature of this reformulation is the use of a *log-sum-exp* transformation which exhibits the aforementioned blow-up phenomenon. This makes any practical implementation directly using black-box convex optimisation techniques extremely slow in some cases and does not allow for any provable bounds. Luckily, there exists a *disciplined convex programming* framework which creates a *reformulation of this reformulation*, and efficiently solves such programs with log-sum-exp convex restrictions.

1.6 Structure of this thesis

The material from the following papers has been used while writing this thesis:

M. Bendkowski, O. Bodini, *S. Dovgal*	*"Statistical properties of lambda terms"*. Accepted to Electronic Journal of Combinatorics. Available on-line at https://arxiv.org/abs/1805.09419	Chapter 4 Chapter 8
S. Dovgal	*"The birth of the contradictory component in random 2-SAT"*, submitted. Available on-line at https://arxiv.org/abs/1904.10266	Chapter 2 Chapter 7
O. Bodini, J. Courtiel, *S. Dovgal*, H.-K. Hwang	*"Asymptotic distribution of parameters in random maps"*. International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms (AofA 2018). Available on-line at https://arxiv.org/abs/1802.07112	Chapter 9
É. de Panafieu, *S. Dovgal*	*"Symbolic method and directed graph enumeration"*. Accepted to EUROCOMB 2019. Available on-line at https://arxiv.org/abs/1903.09454	Chapter 2
S. Dovgal, V. Ravelomanana	*"Shifting the Phase Transition Threshold for Random Graphs Using Degree Set Constraints"*, Latin American Symposium on Theoretical Informatics 2018. Available on-line at https://arxiv.org/abs/1704.06683	Chapter 2 Chapter 3 Chapter 6
M. Bendkowski, O. Bodini, *S. Dovgal*	*"Polynomial tuning of multiparametric combinatorial samplers"*, 2018 Proceedings of the Fifteenth Workshop on Analytic Algorithmics and Combinatorics (ANALCO). Available on-line at https://arxiv.org/abs/1708.01212	Chapter 5 Sections 10.6 and 10.7

Table 1.2: Personal bibliography

This thesis is divided into two parts. In the first part we focus mostly on pure analytic combinatorics, and on the new and existing methods developed, starting with Chapter 2 where the symbolic method is explained. This part is necessary for understanding the source of the functional equations considered in the thesis. In this chapter we also present a new branch of the symbolic method applicable for directed graphs, see Section 2.4. Chapters 3 to 5 can be read in arbitrary order, as they almost do not depend upon one another. The material related to Airy function, generalisations of saddle point lemma for graphs and semi-large powers theorem is collected in Chapter 3. The infinite-dimensional generalisations of Drmota–

Lalley–Woods theorem are discussed in Chapter 4. The multiparametric Boltzmann samplers and a novel polynomial-time multiparametric tuning algorithm are described in Chapter 5.

Part II, presents several new results with the main focus on the application of already developed tools in Part I. The results concerning graph enumeration are collected in Chapters 6 and 7. Chapter 6 contains an application of the saddle point analysis in the context of graphs with degree constraints. Chapter 7 contains a recent development on the structure of a random 2-SAT formula in the subcritical phase, presenting several new connections between random graphs and the random 2-SAT. The next two chapters concern two different combinatorial families that can be put into the tree-like category. In Chapter 8 we present the analysis of multiple parameters of closed and plain lambda terms, applying the tools for infinite systems developed in Chapter 4. In Chapter 9, asymptotic of the parameters in random maps is presented, using a dedicated analysis of the non-linear Riccati equations and a combinatorial linearising transform. The concluding section Chapter 10 contains several new applications of multiparametric Boltzmann tuning. Sections 10.6 and 10.7 are part of [BBD18a].

Part I

Methods

Chapter 2

Symbolic method and related constructions

Contents

The *symbolic method*[1] is an essential part of Analytic Combinatorics and has its roots at group theory and theory of categories. This method, covered in details in [BLL98; FS09], allows to construct combinatorial objects (like graphs, tilings, triangulations, sets, multisets, partitions, 2-SAT formulae, etc) by combining them in various different ways using a set of allowed operations, and to translate these constructions into the language of generating functions. In particular, the method allows to manipulate the generating functions without going into the coefficient level. A spectacular advantage of the symbolic method is its reusability. Once a "library" of specifications is written, each item of this library can then be reused to construct further objects on the basis of already described ones.

2.1 Preliminaries

The symbolic method, or its close relative, theory of species, can be described as an *algebra on combinatorial families*. The definitive treatment of the symbolic method with lots of references and wonderful examples is the *purple book* [FS09]. A formal treatment of a notion of combinatorial family goes into very intricate

[1] The term should not be confused with another Symbolic Method from invariant theory.

formalities which are out of the scope of the thesis [BLL98]. Nevertheless, let us provide an informal definition which can be used throughout the text.

Definition 2.1.1 (Combinatorial family)**.** A *combinatorial family*, or a *combinatorial class* is a denumerable set of objects equipped with a *size* function. The size of each object takes a finite non-negative value, and the number of objects of a given size is finite.

Example 2.1.1. A class of *simple graphs* with size defined as the number of vertices, forms a combinatorial family, while the class of *multigraphs* does not, since the number of multigraphs with size one is infinite, as such a graph may contain infinitely many loops.

In most applications, it is not sufficient to consider a uniparametric size function, as seen in the example of multigraphs above. In such cases, a multiparametric combinatorial family is introduced.

Definition 2.1.2 (Multiparametric combinatorial family)**.** A *multiparametric combinatorial family*, or a *multiparametric combinatorial class* is a denumerable set of objects equipped with a vector size function. The size of each object is a vector of non-negative integer values, and the number of objects with a given vector size is finite.

Example 2.1.2. Consider a class of *multigraphs* equipped with a biparametric size function, where the size of the multigraph is defined as a tuple (n, m), where n equals the number of vertices of the graph, and m the number of its edges. The number of multigraphs with n labelled vertices and m labeleed edges does not exceed $\binom{n+1}{2}^m$, and therefore, is finite.

For combinatorial families such as graphs and multigraphs, the problem of enumeration can be posed in many different ways, depending on the way how two given graphs are distinguished. Enumeration of labelled graphs, for example, is different from the enumeration of graphs up to automorphisms. Such aspects require going deeper into a structure of an *object*.

2.1.1 Unlabelled objects

To a family $\mathcal{A}$ we associate an *ordinary generating function* $A(z)$ which is a formal power series defined as

$$A(z) := \sum_{a \in \mathcal{A}} z^{|a|} = \sum_{n \geqslant 0} a_n z^n, \tag{2.1.1}$$

where $|a|$ denotes the size of an object $a \in \mathcal{A}$, and a_n denotes the number of object of size n from $\mathcal{A}$. The nth coefficient of a series $A(z)$ with respect to the variable z is denoted by $[z^n]A(z)$, so $A(z) = \sum_{n \geqslant 0}([x^n]A(x))z^n$.

For multiparametric combinatorial family $\mathcal{A}$ we also associate a multivariate ordinary generating function $A(z_1, \ldots, z_d)$ defined as

$$A(z_1, \ldots, z_d) := \sum_{a \in \mathcal{A}} \boldsymbol{z}^{|a|}, \tag{2.1.2}$$

where $|a|$ is a vector size of $a \in \mathcal{A}$, and $\boldsymbol{z}^{\boldsymbol{n}} := z_1^{n_1} z_2^{n_2} \cdots z_d^{n_d}$. Similarly, the coefficient with vector index $\boldsymbol{n}$ with respect to the variable $\boldsymbol{z}$ is denoted by $[\boldsymbol{z}^{\boldsymbol{n}}]A(\boldsymbol{z})$, so $A(\boldsymbol{z}) = \sum_{\boldsymbol{n} \geq 0}([\boldsymbol{x}^{\boldsymbol{n}}]A(\boldsymbol{x}))\boldsymbol{z}^{\boldsymbol{n}}$.

Given two families $\mathcal{A}$ and $\mathcal{B}$, the sum of their respective generating functions $A(\boldsymbol{z})$ and $B(\boldsymbol{z})$ represents the dijnoint union of the two families $\mathcal{A}$ and $\mathcal{B}$. The family $\mathcal{C} := \mathcal{A} \times \mathcal{B}$ is defined as the family of pairs equipped with an additive size function

$$\mathcal{C} := \{(a, b) \mid a \in \mathcal{A},\ b \in \mathcal{B}\}, \quad |(a, b)| := |a| + |b|. \tag{2.1.3}$$

This definition applies also to multiparametric families whose parameter vectors have the same dimensions. The generating function of $\mathcal{C}$ is the product of generating functions of $\mathcal{A}$ and $\mathcal{B}$.

A *sequence* operator is defined as a disjoint union of tuples formed of combinatorial object:

$$\mathsf{Seq}(\mathcal{A}) := \bigsqcup_{k \geqslant 0} \mathcal{A}^k \tag{2.1.4}$$

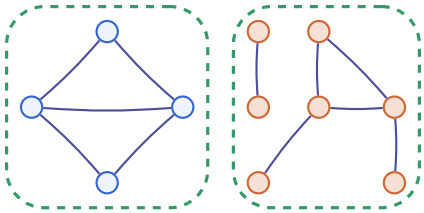


Figure 2.1: Product of combinatorial families.

where $\mathcal{A}^0$ is defined as a combinatorial class containing a single object ε of size zero. Assuming that $\mathcal{A}$ does not contain any objects of size zero, we readily obtain a formal expression for the generating function of the new class $\mathcal{B} := \mathsf{Seq}(\mathcal{A})$:

$$B(z) = \sum_{k \geqslant 0} A(z)^k = \frac{1}{1 - A(z)}. \tag{2.1.5}$$

2.1.2 Labelled objects

Intuitively, a *labelled object* of size n is an object carrying n distinguished nodes with labels $\{1, 2, \ldots, n\}$. The most typical objects of such type are labelled graphs, permutations, and their generalisations. The labels can serve several purposes: one of them is to facilitate the enumeration of objects by making the labels follow specific rules. Alternatively, labels may mark the sequence of moments in time corresponding to some steps of a parallel computation [BGG18; Bod+18b; BGN19].

The cartesian product of labelled families is defined in a different way than in the case of the unlabelled objects. Given two families $\mathcal{A}$ and $\mathcal{B}$ of labelled objects, a *pair* of labelled objects $a \in \mathcal{A}$ and $b \in \mathcal{B}$ with respective sets of labels $\{1, 2, \ldots, k\}$ and $\{1, 2, \ldots, n-k\}$ can be formed in several different ways; in order to form a pair, a subset of k labels of $\{1, 2, \ldots, n\}$ should be chosen and inserted at the positions of the labels of a; the remaining $(n - k)$ labels are then inserted at the positions of the labels of b. Since there are $\binom{n}{k}$ ways to choose k labels, the number of pairs of size n is equal to

$$c_n := \sum_{k=0}^{n} \binom{n}{k} a_k b_{n-k}, \tag{2.1.6}$$

where a_n is the number of objects of size n from $\mathcal{A}$, and b_n is the number of objects of size n from $\mathcal{B}$. This convolution rule requires considering a different type of generating function. To a combinatorial family $\mathcal{A}$ we assign an *exponential generating function*

$$A(z) := \sum_{a \in \mathcal{A}} \frac{z^{|a|}}{|a|!} = \sum_{n \geqslant 0} a_n \frac{z^n}{n!}, \tag{2.1.7}$$

where $|a|$ denotes the size of an object $a \in \mathcal{A}$, and a_n denotes the number of objects of size n from $\mathcal{A}$.

As in the unlabelled case, the generating function of the disjoint unioun of two labelled classes $\mathcal{A}$ and $\mathcal{B}$ is the sum of generating functions $A(z)$ and $B(z)$ of these classes. The generating function of the product $\mathcal{A} \times \mathcal{B}$ is the product $A(z) \cdot B(z)$ of the corresponding generating functions.

Additionally, for labelled classes three more operators are defined.

(i) the *sequence* operator of $\mathcal{A}$ is defined as the disjoint union of all k-tuples of objects from $\mathcal{A}$:

$$\mathcal{B} := \mathsf{Seq}(\mathcal{A}) := \bigsqcup_{k \geqslant 0} \mathcal{A}^k, \quad B(z) = \frac{1}{1 - A(z)},$$

where $A(z)$ is the exponential generating function of $\mathcal{A}$ and $B(z)$ is the exponential generating function of $\mathsf{Seq}(\mathcal{A})$;

(*ii*) the *set* operator of $\mathcal{A}$, $\mathsf{Set}(\mathcal{A})$ is defined as the disjoint union of all k-sets of objects from $\mathcal{A}$. If $A(z)$ is the exponential generating function of $\mathcal{A}$, then the exponential generating function of k-set is $A^k(z)/k!$, and the exponential generating function $B(z)$ of the set operator is

$$B(z) = \sum_{k\geqslant 0} \frac{A(z)^k}{k!} = e^{A(z)};$$

(*iii*) the *cycle* operator of $\mathcal{A}$, $\mathsf{Cyc}(\mathcal{A})$ is defined as a disjoint union of all k-cycles of objects from $\mathcal{A}$; each k-cycle is a sequence of k objects from $\mathcal{A}$, equivalent to itself under cyclic shifts. If $\mathcal{B} = \mathsf{Cyc}(\mathcal{A})$ and $A(z)$ and $B(z)$ are the exponential generating function of $\mathcal{A}$ and $\mathcal{B}$, then

$$B(z) = \sum_{k\geqslant 0} \frac{A(z)^k}{k} = \log \frac{1}{1-A(z)}.$$

These operators are used to construct families of labelled graphs in Section 2.2.

2.1.3 Graphic generating functions

The materials of this subsection follow [dPD19]. In this section, we introduce the arrow product, a new generating function technique for directed graph enumeration. It provides new short proofs for previous results of Gessel on the number of directed acyclic graphs and of Liskovets, Robinson and Wright on the number of strongly connected directed graphs. We also obtain new enumerative results on directed graphs with given numbers of strongly connected components and source-like components using this new technique.

Consider a sequence $(a_n(w))_{n=0}^{\infty}$. Define the *graphic generating function* (GGF) (introduced in [Ges95]) of the sequence $(a_n(w))_{n=0}^{\infty}$ as

$$\mathbf{A}(z,w) := \sum_{n\geqslant 0} \frac{a_n(w)}{(1+w)^{\binom{n}{2}}} \frac{z^n}{n!}.$$

To distinguish exponential generating functions (EGF) from GGF, the latter are written in bold characters.

The *exponential Hadamard product* of two series $A(z) = \sum_{n\geqslant 0} a_n \frac{z^n}{n!}$ and $B(z) = \sum_{n\geqslant 0} b_n \frac{z^n}{n!}$ is denoted by and defined as

$$A(z) \odot B(z) = \Big(\sum_{n\geqslant 0} a_n \frac{z^n}{n!}\Big) \odot \Big(\sum_{n\geqslant 0} b_n \frac{z^n}{n!}\Big) := \sum_{n\geqslant 0} a_n b_n \frac{z^n}{n!}.$$

All Hadamard products are taken with respect to the variable z. The Hadamard product can be used to convert between EGF and GGF (see Corollary 2.4.1). The exponential Hadamard product should not be confused with the ordinary Hadamard product $\sum_n ([z^n]A(z))([z^n]B(z))z^n$.

If $\mathcal{A}$ is a certain family of digraphs or graphs, we can associate to it a sequence of series $(a_n(w))_{n=0}^{\infty}$, such that $[w^m]a_n(w)$ is equal to the number of elements in $\mathcal{A}$ with n vertices and m directed edges. Consequently, we can associate both EGF and GGF to the same family of digraphs or graphs.

An advantage of the symbolic method is its ability to keep track of a collection of *parameters* in combinatorial objects. The two default parameters are the numbers of vertices and edges, and the arguments z and w of a generating function $F(z,w)$ correspond to these parameters. As a generalization, we consider multivariate generating functions

$$\mathrm{A}(z,w,\boldsymbol{u}) := \sum_{n,\boldsymbol{p}} a_{n,\boldsymbol{p}}(w)\boldsymbol{u}^{\boldsymbol{p}} \frac{z^n}{n!} \quad \text{and} \quad \mathbf{A}(z,w,\boldsymbol{u}) := \sum_{n,\boldsymbol{p}} \frac{a_{n,\boldsymbol{p}}(w)\boldsymbol{u}^{\boldsymbol{p}}}{(1+w)^{\binom{n}{2}}} \frac{z^n}{n!},$$

where $\boldsymbol{u} = (u_1, \cdots, u_d)$ is the vector of variables, $\boldsymbol{p} = (p_1, \cdots, p_d)$ denotes a vector of parameters, and the notation $\boldsymbol{u}^{\boldsymbol{p}} := \prod_k u_k^{p_k}$ is used. We say that the variable u_k *marks* its corresponding parameter p_k.

The next proposition recalls classic operations on EGFs (see [FS09]), which extend naturally to GGFs.

Proposition 2.1.1. Consider two digraph (or graph) families $\mathcal{A}$ and $\mathcal{B}$. The EGF and GGF of the disjoint union of $\mathcal{A}$ and $\mathcal{B}$ are $\mathrm{A}(z,w)+\mathrm{B}(z,w)$ and $\mathbf{A}(z,w)+\mathbf{B}(z,w)$. The EGF and GGF of the digraphs from $\mathcal{A}$ where one vertex is distinguished are $z\partial_z \mathrm{A}(z)$ and $z\partial_z\mathbf{A}(z,w)$. The EGF of sets of digraphs from $\mathcal{A}$ is $e^{\mathrm{A}(z,w)}$. The EGF of pairs of digraphs (a,b) with $a\in\mathcal{A}$ and $b\in\mathcal{B}$ (relabeled so that the vertex labels of a and b are disjoint, see [FS09]) is $\mathrm{A}(z,w)\,\mathrm{B}(z,w)$. If a variable u marks the number of specific items in the EGF $\mathrm{A}(z,w,u)$ or the GGF $\mathbf{A}(z,w,u)$ of the family $\mathcal{A}$, then the EGF and GGF for the objects $a\in\mathcal{A}$ which have a distinguished subset of these specific items are $\mathrm{A}(z,w,u+1)$ and $\mathbf{A}(z,w,u+1)$.

The next proposition presents our new combinatorial interpretation of the product of GGFs. It was implicitly used by Gessel in several proofs (e.g. [Ges96]) at coefficient level, but we have not found it expressed at the generating function level. However, a combinatorial interpretation of the exponential of GGFs can be found in [Ges95; GS96].

Proposition 2.1.2. We define the *arrow product* of $\mathcal{A}$ and $\mathcal{B}$ as the family $\mathcal{C}$ of pairs (a,b), with $a\in\mathcal{A}$, $b\in\mathcal{B}$ (relabeld so that a and b have disjoint labels), where an arbitrary number of edges oriented from vertices of a to vertices of b are added (see Figure 2.2). The GGF of $\mathcal{C}$ is equal to $\mathbf{A}(z,w)\mathbf{B}(z,w)$.

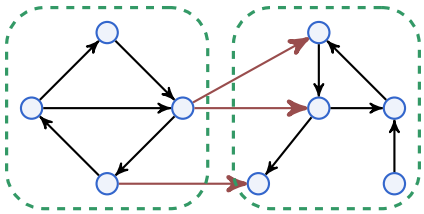


Figure 2.2: The arrow product

Proof. Consider two digraph families $\mathcal{A}$ and $\mathcal{B}$, with associated sequences $(a_n(w))$, $(b_n(w))$. Then the sequence associated to the GGF $\mathbf{A}(z,w)\mathbf{B}(z,w)$ is

$$c_n(w)=(1+w)^{\binom{n}{2}}n![z^n]\left(\sum_k \frac{a_k(w)}{(1+w)^{\binom{k}{2}}}\frac{z^k}{k!}\right)\left(\sum_\ell \frac{b_\ell(w)}{(1+w)^{\binom{\ell}{2}}}\frac{z^\ell}{\ell!}\right)=\sum_{k+\ell=n}\binom{n}{k}(1+w)^{k\ell}a_k(w)b_\ell(w).$$

This series $c_n(w)$ has the following combinatorial interpretation: it is the generating function (the variable w marks the edges) of digraphs with n vertices, obtained by

- choosing a digraph a of size k in $\mathcal{A}$, and a digraph b of size ℓ in $\mathcal{B}$, such that $k+\ell=n$,
- choosing a subset of $\{1,\dots,n\}$ for the labels of a (b receives the complementary set for its labels),
- each oriented edge (u,v) with u vertex from a, and v vertex from b, is or not added.

Hence, $(c_n(w))_{n\geqslant 0}$ is the sequence associated to the arrow product of $\mathcal{A}$ and $\mathcal{B}$. □

2.1.4 Other types of generating functions

The key ingredient of the design of the generating function is the behaviour of the convolution of the coefficients while taking the product of two generating functions. In principle, there is an unlimited supply of the number of ways in which a generating function of a combinatorial class can be designed using an appropriate weighting factor. Graphic generating function implements one of such alternative convolutions required for very particular purposes. Related well-known example is *Dirichlet generating function* of a sequence $(a_n)_{n\geqslant 1}$ which is defined as

$$A(s)=\sum_{n\geqslant 1}\frac{a_n}{n^s},$$

and the convolution rule is tightly related to number-theoretic features

$$A(s)\cdot B(s)=\sum_{n\geqslant 1}\frac{1}{n^s}\sum_{k|n}a_k b_{n/k}.$$

An interesting treatment of different type of convolution products related to acyclic algebras is given in [Han81]. An example of Knuth [Knu92] is also remarkable in this context: he considers families of polynomials $(F_k(x))_{k\geqslant 0}$ such that

$$F_n(x+y)=\sum_{k=0}^{n}F_k(x)F_{n-k}(y)$$

and then, relations between these polynomials closely resemble that of cycle index series.

For the enumeration of Pólya structures and unlabelled objects, *cycle index series* are introduced, which are formal power series in an infinite number of arguments, whose projections contain both ordinary and exponential generating functions [BLL98].

2.2 Constructing simple graphs and multigraphs

2.2.1 Graphs, digraphs and 2-CNF

The materials of this subsection follow [Dov19].

A *simple graph* $G=(V,E)$, $E\subset\{\{x,y\}\mid x,y\in V,\ x\neq y\}$ is a graph without loops and multiple edges, so that every edge connects a set of two distinct vertices and for every pair of vertices there is at most one edge connecting them. Contrary to a simple graph, a *multigraph* may contain an arbitrary number of loops and multiple edges. By $\mathcal{G}(n,m)$ we denote the set of all simple graphs with n vertices and m edges. We say that a graph has *size* n if it contains n vertices.

A *simple digraph* is a synonym for a *simple directed graph* which we define as a pair $D=(V,E)$ of the set of vertices V and the set of edges $E\subset\{(x,y)\mid x,y\in V,\ x\neq y\}$ such that every directed edge $x\to y$ is represented by a pair of vertices (x,y); a simple digraph contains no loops $x\to x$, and no multiple edges; at the same time, any two distinct vertices x and y can have simultaneously both directed edges $x\to y$ and $y\to x$ between them. The *unoriented projection* of a simple digraph is a multigraph (possibly containing double edges) which is obtained by dropping the orientations of the edges. We denote by $\mathcal{D}(2n,m)$ the set of simple digraphs with $2n$ vertices and m oriented edges.

Below we give the definitions related to the 2-SAT model which contains the formulae in *2-conjunctive normal form* (2-CNF).

Definition 2.2.1 (Variables, literals and clauses)**.** Consider n *Boolean variables* $\{x_1,x_2,\dots x_n\}$, so that $x_i\in\{0,1\}$ for all $i\in\{1,\dots,n\}$. The logical negation of x_i is denoted by $\overline{x}_i$. Each of $\{x_i,\overline{x}_i\}$ is called a *literal*, while the two literals $\{x_i,\overline{x}_i\}$ refer to the same *variable* x_i. We say that two literals ξ and η are *complementary* if $\xi=\overline{\eta}$. Two literals ξ and η are said to be *strictly distinct* if their underlying variables are distinct. The 2-*clauses* are disjunctions of two literals, corresponding to distinct Boolean variables, in other words, each 2-clause is of the form $(\xi_j\vee\eta_j)$ where ξ_j and η_j belong to the set of $2n$ possible literals $\{x_1,\dots,x_n,\overline{x}_1,\dots,\overline{x}_n\}$. We do not distinguish clauses obtained by a change of the order of variables inside a disjunction, so $(\xi_j\vee\eta_j)\equiv(\eta_j\vee\xi_j)$. A *2-CNF* is a conjunction of clauses, i.e. a formula of the form $\wedge_{j=1}^{m}(\xi_j\vee\eta_j)$ where each of the clauses $(\xi_j\vee\eta_j)$ is distinct for $j\in\{1,\dots,m\}$. A formula is *satisfiable* (SAT) if there exists variable assignment yielding the truth value of the formula. By $\mathcal{F}(n,m)$ we denote the set of 2-CNF formulae with n Boolean variables and m clauses.

In different models, the distinctness condition for clauses is not required, but we show that it is not essential for the phase transition and for our techniques. More explicitly, we require for technical reasons that in a 2-CNF formula in our model the conditions (C1)–(C3) hold:

(C1) clauses $(x_i\vee x_i)$ are not allowed;

(C2) clauses $(x_i \vee \overline{x}_i)$ are not allowed;

(C3) all the clauses are distinct.

Each of the conditions (C1)–(C3) may be violated without changing the main results, see Remark 7.3.3.

Definition 2.2.2 (Implication digraphs)**.** Any 2-CNF with n Boolean variables and m clauses can be represented in the form of an *implication digraph* where every clause $(x \vee y)$ is replaced with two directed edges $(\overline{x} \to y)$ and $(\overline{y} \to x)$.

If there is a directed path from a vertex x to a vertex y in a directed graph D, we write $x \rightsquigarrow y$. In the case when D is an implication digraph, we also say that a literal x *implies* literal y.

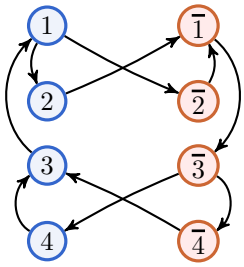


Figure 2.3: Implication digraph.

Definition 2.2.3 (Spine, contradictory variables and circuits)**.** The *spine* of a formula F (denoted $\mathcal{S}(F)$) is the set of literals that imply their complementary literals, i.e.

$$\mathcal{S}(F) := \{x \mid x \rightsquigarrow \overline{x} \text{ in } F\}.$$

A variable x is *contradictory* if $x \in \mathcal{S}(F)$ and $\overline{x} \in \mathcal{S}(F)$. The *contradictory component* $\mathcal{C}(F)$ of a formula F is then formed of all the contradictory variables, i.e.

$$\mathcal{C}(F) := \{x \mid x \rightsquigarrow \overline{x} \rightsquigarrow x \text{ in } F\}.$$

A *contradictory circuit* is a distinguished directed path passing through x, $\overline{x}$ and x. It can be easily shown that any variable belonging to a contradictory circuit is also contradictory (possibly via a different contradictory circuit). It is well known and has been proven in [APT82], that a formula is satisfiable if and only if it contains a so-called *contradictory circuit*, that is, there exists a literal x such that $x \rightsquigarrow \overline{x}$ and $\overline{x} \rightsquigarrow x$. Consequently, a formula is satisfiable if and only if its contradictory component is empty.

Example 2.2.1. Figure 2.3 contains an implication digraph of a formula $(\overline{x}_1 \vee \overline{x}_2)(\overline{x}_1 \vee x_2)(x_1 \vee \overline{x}_3)(x_3 \vee x_4)(x_3 \vee \overline{x}_4)$. This formula is unsatisfiable because its implication digraph containts a contradictory circuit $1 \to \overline{2} \to \overline{1} \to \overline{3} \to 4 \to 3 \to 1$ passing through 1 and $\overline{1}$.

Remark 2.2.1. The contradictory component forms a set of strongly connected components such that there is no directed path between any two strongly connected components. Indeed, if $\mathcal{C}_1$ and $\mathcal{C}_2$ are such strongly connected components (where each variable is contradictory), and there is a path from $x \in \mathcal{C}_1$ to $y \in \mathcal{C}_2$, then, the complementary path $\overline{y} \rightsquigarrow \overline{x}$ is also a part of the implication digraph, while $\overline{y} \in \mathcal{C}_2$, $\overline{x} \in \mathcal{C}_1$, since each of the strongly connected components is contradictory. Therefore, the presence of a directed path from $\mathcal{C}_1$ to $\mathcal{C}_2$ implies a directed path from $\mathcal{C}_2$ to $\mathcal{C}_1$ and they form a single strongly connected component.

2.2.2 Symbolic method for graphs

Recall that the *size* of a graph (or a directed graph) is defined to be equal to the number of its vertices. In this section, labelled graphs are considered: every vertex has a distinct label from the set $\{1, 2, \ldots, n\}$ where n is the total number of vertices of the graph.

Given a family of graphs (or directed graphs) $\mathcal{A}$, we construct its corresponding *exponential generating function* (EGF) which is

$$A(z) := \sum_{n \geqslant 0} a_n \frac{z^n}{n!},$$

where a_n is equal to the number of graphs or digraphs of size n from $\mathcal{A}$. The operator $[z^n]$ of taking nth coefficient is then defined as $[z^n]A(z) = a_n/n!$. All the indefinite integrals $\int F(z)dz$ should be interpreted as $\int_0^z F(\tau)d\tau$.

Given two families $\mathcal{A}$ and $\mathcal{B}$, the sum of their respective generating functions $A(z)$ and $B(z)$ represents the union of the two families $\mathcal{A}$ and $\mathcal{B}$. The family $\mathcal{C} := \mathcal{A} \times \mathcal{B}$ is defined as the family of labelled graphs (or digraphs) whose vertices are partitioned into two sets such that there are no edges between the parts, the underlying graph of the first part is from $\mathcal{A}$, and the graph from the second part is from $\mathcal{B}$. We can also say that $\mathcal{A} \times \mathcal{B}$ is the labelled family of ordered tuples of graphs from $\mathcal{A}$ and $\mathcal{B}$. If the respective EGFs of these families are A(z) and B(z) then the EFG of $\mathcal{C}$ is $A(z) \cdot B(z)$.

Consequently, if $A(z)$ is an EGF for a graph family $\mathcal{A}$ not containing an empty graph, then $A^k(z)$ is the EGF for sequences of length k of graphs from $\mathcal{A}$. Summing over all k, we obtain the EGF for all possible sequences (possibly empty) of graphs from $\mathcal{A}$, which is $\frac{1}{1-A(z)}$. If the family $\mathcal{A}$ does not contain an empty graph, then the EGF for non-ordered sequences of length k (i.e. sets) of graphs from $\mathcal{A}$ is $A(z)^k/k!$. Summing over all k, we obtain the EGF of all unordered sequences of graphs from $\mathcal{A}$, which is $e^{A(z)}$. Finally, a oriented cyclic composition of graphs from $\mathcal{A}$ has EGF $\sum_{k \geqslant 1} A^k(z)/k = \log \frac{1}{1-A(z)}$. If a cycle is not oriented, then the corresponding EGF is $\frac{1}{2} \log \frac{1}{1-A(z)}$.

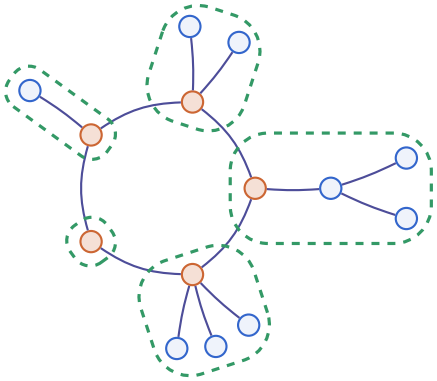


Figure 2.4: Cyclic composition.

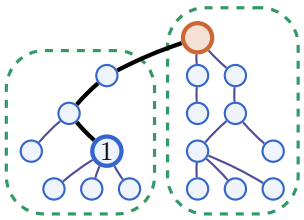


Figure 2.5: Constructing unrooted trees: the case when the label of the root is not equal to 1.

If all the graphs in a family $\mathcal{A}$ are connected, then $e^{A(z)}$ enumerates labelled graphs whose connected components are from $\mathcal{A}$. The product of exponential generating functions for labelled graphs can be given further interpretations: if two graph families $\mathcal{A}$ and $\mathcal{B}$ do not intersect and the graphs from these families are connected, then the product of their EGFs enumerates graphs with two connected components, one from $\mathcal{A}$ and the second from $\mathcal{B}$. In the same manner, one can multiply generating functions of not necessarily connected graph families provided that the underlying connected components corresponding to the two families, are always distinct. Then, the product can be interpreted as a new family of graphs whose vertices can be partitioned uniquely into two sets, and the graphs constructed on the respective sets, belong to the first and the second family, respectively.

It is well known that the EGF $T(z)$ for rooted trees, also known as *Cayley trees*, satisfies

$$T(z) = ze^{T(z)} = \sum_{n \geqslant 0} n^{n-1} \frac{z^n}{n!}.$$

Unicyclic graphs are defined as connected graphs whose number of vertices is equal to their number of edges. Every such graph has exactly one undirected cycle inside it and can be represented as a sequence of trees arranged in a cycle of length at least 3, which results in the following EGF $V(z)$:

$$V(z) = \frac{1}{2}\sum_{k\geqslant 3}\frac{T(z)^k}{2k} = \frac{1}{2}\left[\log\frac{1}{1-T(z)} - T(z) - \frac{T^2(z)}{2}\right].$$

The EGF for *unrooted trees* has the form $U(z) = T(z) - \frac{T(z)^2}{2}$. An elegant proof of this fact is given in [FKP89]. We present a sketch of the proof for completeness. The label of the root of a rooted tree can be equal to 1 or greater than 1. The generating function of rooted trees whose root has label 1 is $U(z)$ because an unrooted tree can be canonically rooted at vertex with label 1. Otherwise, a tree whose root label is greater than 1 can be represented as a pair consisting of the subtree whose parent is the root and which contains the vertex with label 1, and the remaining tree which is formed by removing the aforementioned subtree from the initial tree. The generating function of such (unordered) tuples is $T^2(z)/2$. Adding up the two cases, we obtain an identity $T(z) = U(z) + T^2(z)/2$.

Similarly to trees and unicycles, more complex structures can be defined. An *excess* of a connected graph is equal to the number of its edges minus the number of its vertices. A connected graph with minimal possible excess is a tree, and its excess is equal to -1. Next, a unicycle is a connected graph of excess 0. A connected graph of excess 1 is called a *bicycle*.

Suppose that a connected graph G has excess r. The *pruning* procedure is described as repeated removal of vertices of degree 1 until no vertices of degree 1 remain. The resulting graph is then called the *2-core* or just the *core* of G. Consequently, each vertex of degree 2 of G can be removed, while connecting its former neighbours by an edge; this procedure is called *cancelling*. The resulting graph obtained by pruning and cancelling is called the *3-core*, or the *kernel* of G. After cancelling, a simple graph may gain some multiple edges and loops, i.e. become a multigraph. Note that pruning and cancelling do not change the excess of G. It is known (see e.g. [Jan+93]) that after pruning and cancelling, the resulting multigraph belongs to a finite set of multigraphs with given excess r, see also Section 2.2.3.

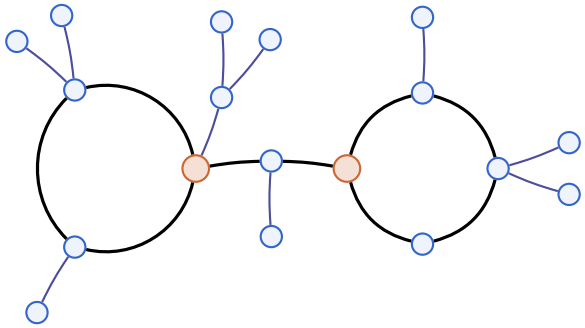


Figure 2.6: Pruning and cancelling.

Remark 2.2.2 (The philosophy of symbolic composition)**.** In order to illustrate the concept of symbolic method on the example of random graphs, let us briefly consider an average height of a dangling tree of the complex component in a critical random graph, see Figure 2.6. It is known [FO82] that the height of a random tree with n vertices is $\Theta(\sqrt{n})$ and that the number of vertices of a dangling tree in a critical random graph with n vertices and $n/2$ edges is $\Theta(n^{1/3})$. However, the distribution of such trees does not coincide with a uniform distribution on rooted trees, so their average height is not necessarily $\Theta(\sqrt{n^{1/3}}) = \Theta(n^{1/6})$, as a naive heuristic would suggest. One of the ways to settle the question of the average height of a dangling tree is to construct the generating function of subcritical graphs weighted according to the height of the dangling trees of the complex component, using admissible constructions from the symbolic method [DR18] and then apply the standard techniques for coefficient extraction. The resulting average tree height is of

order $\Theta(\log n)$, where the logarithm in the expression is closely related to the logarithmic form of height-weighted generating function of trees discussed in [FO82]. A detailed treatment of this case can be found in Section 6.6.3.

It turns out that in the subcritical and in the critical phases of the random graphs, for example, when the number of edges m equals the number of vertices n divided by 2, the contribution of connected graphs whose kernel is not cubic, is negligible: if a kernel is not cubic, it can be obtained as a "limiting case" from some cubic kernel by contracting some of its edges. This corresponds to the case when the corresponding paths in the initial graph have length zero. Heuristically, since the average length of a path on the cubic core is of order $\Theta(n^{1/3})$ in the critical phase $m = n/2$ (see also [JLR11]), the probability that a path has zero vertices, i.e. that the kernel is non-cubic, is of order $O(n^{-1/3})$.

Turning to the case of digraphs, pruning is not defined for a strongly connected component, because the degree of every vertex in a strongly connected component is at least 1 (except when a component consists of an isolated vertex). Cancellation of a strongly connected component is then defined in the following way: if a vertex has in-degree 1 and out-degree 1, it is removed and a directed edge between the in-neighbour and the out-neighbour is added. It is easy to see that a directed multigraph obtained after a cancellation procedure with given excess belongs to a finite sets of reduced directed multigraphs of given excess (see also Problem 7.4.1). If in a directed multigraph every vertex has the sum of in- and out-degrees equal 3, it is also called cubic.

2.2.3 Compensation factors

The compensation factor of a multigraph M is a certain coefficient depending on M which is required to obtain the EGF of the family of graphs reducing to M under pruning and cancelling. While for the purposes of the current section we take a relatively low-level classical approach, there also exists an elegant unifying framework [dPan16] for dealing with compensation factors by introducing bivariate generating functions which are exponential with respect to two variables.

Consider the three possible kernels of excess 1 in Figure 2.7. Each of these kernels a certain symmetry group which acts on the set of its half-edges and vertices. Since the objects are labelled, the cardinality of the automorphism group acting on its vertices equals $n!$ divided by the number of distinct labelled graphs. The remaining compensation factor can be determinted solely on the base of the incidence matrix.

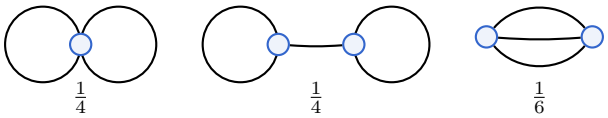


Figure 2.7: Kernels of complex components of excess 1 and their respective compensation factors.

Definition 2.2.4. Consider a multigraph M with n labelled vertices and with m_{xy} edges between vertices with labels x and y. In particular, m_{xx} denotes the number of loops of a vertex x, and $m_{xy} = m_{yx}$.

A *compensation factor* $\varkappa(M)$ is defined as follows:

$$\frac{1}{\varkappa(M)} := \prod_{x=1}^{n} 2^{m_{xx}} \prod_{y=x}^{n} m_{xy}!. \tag{2.2.1}$$

The number of ways to construct a single multigraph M with n vertices and m edges by gluing the half-edges of the labelled vertices can be viewed as $2^m m! \varkappa(M)$. The EGF of multigraphs G reducing under pruning and cancelling to a given multigraph M is then written as

$$\frac{\varkappa(M)}{n!} \cdot \frac{T(z)^n}{(1-T(z))^m},$$

which can be obtained by substitution of trees into the corresponding sequences and vertices. For a rigorous proof of this expression, see [Jan+93].

For contradictory components in implication digraphs, we define the compensation factors in a very similar way.

Definition 2.2.5 (Compensation factor of sum-representations and implication digraphs)**.** Consider a sum-representation digraph $D \in \mathcal{D}^{\circ}(2n, m)$ and a contradictory implication digraph M with $2n$ conventionally labelled vertices and $2m$ directed edges. Suppose that for each literals x and y there are d_{xy} oriented edges in D and m_{xy} oriented edges in M between vertices with labels x and y. The compensation factors of D and M are then defined as

$$\frac{1}{\varkappa(D)} := \prod_{x=1}^{2n}\prod_{y=1}^{2n} d_{xy}!, \quad \frac{1}{\varkappa(M)} := \prod_{x=1}^{2n}\prod_{y=1}^{2n} m_{xy}!!, \tag{2.2.2}$$

where for even N, the double factorial is defined as $N!! = 2 \cdot 4 \cdot \cdots \cdot N$.

For the case of simple graphs, the sum over all possible labelled kernels M with n vertices and given excess r, $\sum_M \frac{\varkappa(M)}{n!}$ expresses the coefficient at $|\mu|^{-3r}$ in the probability that a random graph has a complex component of such excess (c.f. the asymptotic probability $5/24 \cdot |\mu|^{-3}$ of having a bicyclic complex component, where $5/24$ equals the sum of the compensation factors $1/4$ and $1/6$ divided by 2!, see also Remark 3.2.2) in the subcritical phase, μ is defined by the relation $m = n/2(1 + \mu n^{-1/3})$, see Chapters 6 and 7.

As we show in Theorem 7.3.1, the compensation factor for contradictory components plays exactly the same role. The compensation factor for contradictory components has the following additional interpretation.

Lemma 2.2.1. Consider a contradictory component $\mathcal{C}$ with n Boolean variables and one of its sum-representations $\pi(\mathcal{C})$. We shall say that a sum-representation digraph π_1 is *isomorphic* to a digraph π_2 if π_1 can be obtained from π_2 by a permutation of Boolean variables. Then, the number of sum-representations, both equivalent and isomorphic to $\pi(\mathcal{C})$ is equal to $n!\varkappa(\pi)/\varkappa(\mathcal{C})$.

Proof. If $\mathcal{C}_1, \ldots, \mathcal{C}_K$ are K possible isomorphic implication digraphs obtained by label permutations, then each of the digraphs has the same number of 2^m sum-representations, and therefore, each isomorphic sum-representation is counted with multiplicity K.

Let us compute the number K of possible isomorphic digraphs obtained by label permutations. All the edges of $\mathcal{C}$ come in pairs, therefore, within each pair there is a cyclic group of order 2, and there are $(m_{xy}/2)!$ permutations between the $(m_{xy}/2)$ pairs of oriented edges between the vertices x and y. Multiplying the orders of these groups, we obtain $(m_{xy}/2)! \cdot 2^{m_{xy}/2} = (m_{xy})!!$. Each loop is directed, so there is no additional factors corresponding to the loops. The number K then equal to $n!/\varkappa(\mathcal{C})$ since this is equal to the cardinality of the automorphism group. □

Example 2.2.2. As an illustration of this concept, let us compute the compensation factor of the contradictory component of excess 1 from Figure 7.6. There are $n = 2$ Boolean variables and two multiple edges, between, respectively, x and $\overline{x}$, and y and $\overline{y}$. Therefore, the compensation factor is $1/2!!^2 = 1/4$. The number of isomorphic equivalent sum-representations is therefore $2! \cdot 4 = 8$, which explains the factor $1/8$ in the second summand of the expression for ξ_1 in Section 7.2.3. The factor $1/4$ corresponding to Figure 7.5 is computed as $1! \cdot 2!!^2$ because there is one Boolean variable and two double edges.

It is natural to expect that the concept of compensation factor of a contradictory component will prove helpful not only for the subcritical phase of the phase transition, but will allow to give a complete description of the transition curve. See also Section 7.4 for discussions and open questions.

2.3 Variations on graphs

2.3.1 Graphs with degree constraints

The materials of this subsection follow [DR18].

To a given arbitrary set $\Delta \subseteq \{0, 1, 2, \ldots\}$, we associate the *exponential generating function* (EGF) $\omega(z)$:

$$\mathrm{SET}_\Delta(z) = \omega(z) = \sum_{d \in \Delta} \frac{z^d}{d!} \ . \tag{2.3.1}$$

We only consider sets Δ such that $1 \in \Delta$.

For each $\ell \geqslant 0$, let us define *ℓ-sprouted trees*: rooted trees whose vertex degrees belong to the set Δ, except the root (see Figure 2.8), whose degree belongs to the set $\Delta - \ell = \{\delta \geqslant 0 \colon \delta + \ell \in \Delta\}$.

Their EGF $T_\ell(z)$ can be defined recursively

$$T_\ell(z) = z\omega^{(\ell)}(T_1(z)),\ T_1(z) = z\omega'(T_1(z)) \ , \quad \ell \geqslant 0 \ . \tag{2.3.2}$$

Lemma 2.3.1. Let $U(z)$ be the EGF for *unrooted* trees and $V(z)$ the EGF of *unicycles* whose vertices have degrees $\in \Delta$. Then

$$U(z) = T_0(z) - \frac{T_1(z)^2}{2} \ , \quad V(z) = \frac{1}{2}\left[\log\frac{1}{1 - T_2(z)} - T_2(z) - \frac{T_2^2(z)}{2}\right] \ , \tag{2.3.3}$$

where $T_0(z)$, $T_1(z)$, and $T_2(z)$ are by (2.3.2).

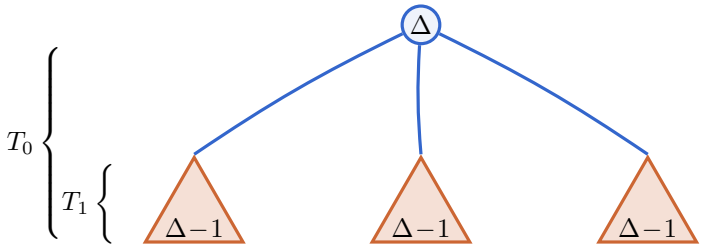


Figure 2.8: Recursive construction of $T_0(z)$: the degree of the root of each subtree should belong to the set $\Delta - 1$

Remark 2.3.1. The above statement for $U(z)$ can be proven using the *dissymetry theorem for trees*, adapted for the case with degree constraints (see [BLL98, Section 4.1], [FPK89]). In short, we can consider rooted trees T_0 and mark the vertex with label 1. Then we consider two cases, when this vertex is the root one, or not. The first case corresponds to $U(z)$ and in the second situation we can consider two subcases, whether the label 1 belongs to the subtree induced by the first child or not (see Figure 2.9). Summarizing the argument, we obtain

$$T_0(z) = U(z) + \frac{1}{2}T_1(z)^2 \ .$$

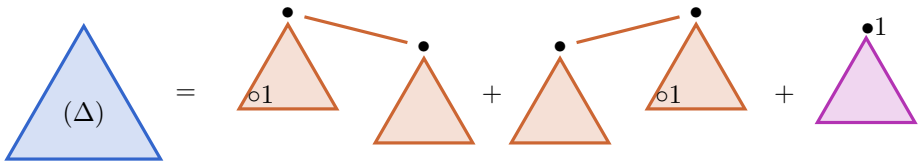


Figure 2.9: Variant of dissymetry theorem for unrooted trees with degree constraints

The expression for $V(z)$ is an application of the symbolic method of EGFs in the case of undirected cycles ($\mathrm{CYC}_{\geqslant 3}$) of 2-sprouted trees, see Figure 2.4.

Any multigraph M on n labeled vertices can be defined by a symmetric $n \times n$ matrix of nonnegative integers m_{xy}, where $m_{xy} = m_{yx}$ is the number of edges $x-y$ in M. The *compensation factor* $\kappa(M)$ is defined by

$$\kappa(M) = 1 \Big/ \prod_{x=1}^{n} \left(2^{m_{xx}} \prod_{y=x}^{n} m_{xy}! \right) \quad . \tag{2.3.4}$$

A *multigraph process* is a sequence of $2m$ independent random vertices

$$(v_1, v_2, \ldots, v_{2m}) \ , \quad v_k \in \{1, 2, \ldots, n\} \ ,$$

and output multigraph with the set of vertices $\{1, 2, \ldots, n\}$ and the set of edges $\{\{v_{2i-1}, v_{2i}\} \colon 1 \leqslant i \leqslant m\}$. The number of sequences that lead to the same multigraph M is exactly $2^m m! \kappa(M)$.

Lemma 2.3.2. Let $\overline{\overline{M}}$ be some 3-core multigraph with a vertex set V, $|V| = n$, having μ edges, and compensation factor $\kappa(\overline{\overline{M}})$. Let μ_{xy} be the number of edges between vertices x and y for $1 \leqslant x \leqslant y \leqslant n$. The generating function for all *simple* graphs G that lead to $\overline{\overline{M}}$ under reduction is

$$\frac{\kappa(\overline{\overline{M}}) \prod\limits_{v \in V} T_{\deg(v)}(z)}{n!} \cdot \frac{P(\overline{\overline{M}}, T_2(z))}{(1 - T_2(z))^{\mu}}; \tag{2.3.5}$$

$$P(\overline{\overline{M}}, z) = \prod_{x=1}^{n} \Big(z^{2\mu_{xx}} \prod_{y=x+1}^{n} z^{\mu_{xy}-1} (\mu_{xy} - (\mu_{xy} - 1)z) \Big). \tag{2.3.6}$$

Corollary 2.3.1. Assume that $\phi_1(\widehat{z}) = 1$. Near the singularity $z \sim \widehat{z}$, i.e. $T_2(z) \approx 1$, some of the summands from Lemma 2.3.2 are negligible. Dominant summands correspond to graphs $\overline{\overline{M}}$ with maximal number of edges, i.e. graphs with $3r$ edges and $2r$ vertices. The vertices of degree greater than 3 can be splitted into more vertices with additional edges. Due to [Jan+93, Section 7, Eq. (7.2)], the sum of the compensation factors is expressed as

$$e_{r0} = \frac{(6r)!}{2^{5r} 3^{2r} (3r)! (2r)!} \quad . \tag{2.3.7}$$

and the sum of major summands is asymptotically

$$e_{r0} \frac{T_3(z)^{2r}}{(1 - T_2(z))^{3r}} \quad . \tag{2.3.8}$$

The proofs of the two previous statements are postponed until the end of Remark 2.3.2.

Remark 2.3.2. Let's give an example of application of Lemma 2.3.2, first in less technical multigraph form, then for simple graphs.

Each multi-edge in the 3-core $\overline{\overline{M}}$ corresponds to a sequence of trees in the initial graph M. Therefore, the generating function for multigraphs M which reduce to one of the three depicted (see Figure 2.10) 3-core multigraphs consists of 3 summands:

$$W_{\Delta}(z) = \frac{1}{4} \frac{T_4(z)}{(1 - T_2(z))^2} + \frac{1}{4} \frac{T_3(z)^2}{(1 - T_2(z))^3} + \frac{1}{6} \frac{T_3(z)^2}{(1 - T_2(z))^3} \quad . \tag{2.3.9}$$

We write $T_2(z)$ because if we attach a tree on any path, the degree of the root decreases by 2. For the same reason there appear $T_3(z)$ and $T_4(z)$. If we evaluate $W_{\Delta}(z)$ near the pole $z = \widehat{z}$, or equivalently at $T_2(z) = 1$, the first summand goes to ∞ slower than the second and the third. This yields asymptotic approximation

$$W_{\Delta}(z) = \frac{5}{24} \frac{T_3(z)^2}{(1 - T_2(z))^3} + O\left(\frac{1}{(1 - T_2(z))^2} \right) \quad . \tag{2.3.10}$$

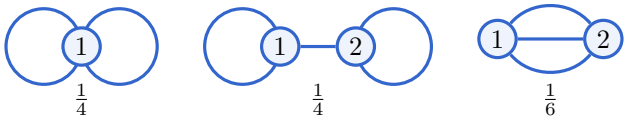


Figure 2.10: All possible 3-core multigraphs of excess 1 and their compensation factors. The first one has negligible contribution because it is non-cubic

The big-O notation with the generating functions means:

$$F(z) = O(B(z)) \text{ if } [z^n]F(z) \leqslant c[z^n]B(z) \tag{2.3.11}$$

for sufficiently large n, so from Eq. (2.3.10) we know that

$$[z^n]W_\Delta(z) \sim [z^n]\frac{5}{24}\frac{T_3(z)^2}{(1-T_2(z))^3} \ . \tag{2.3.12}$$

With simple graphs (not multigraphs) the situation is similar. For the first core we want the path to be non-empty (because simple graphs don't contain loops), so the generating function is $\frac{1}{4}\frac{T_4(z)T_2(z)^4}{(1-T_2(z))^2}$. For the second graph we also require that both paths obtained from loops, contain at least one node inside: $\frac{1}{4}\frac{T_3(z)^2T_2(z)^4}{(1-T_2(z))^3}$. Then, for the third core, we need at least two of the paths contain at least one node. Collecting all the summands we obtain

$$\widetilde{W}_\Delta(z) = \frac{1}{4}\frac{T_4(z)T_2(z)^4}{(1-T_2(z))^2} + \frac{1}{4}\frac{T_3(z)^2T_2(z)^4}{(1-T_2(z))^3} + \frac{1}{6}\frac{T_3(z)^2[3T_2(z)^2-2T_2^3(z)]}{(1-T_2(z))^3} \ . \tag{2.3.13}$$

At z near $\widehat{z}$, $T_2(z) = 1$, so the asymptotics of this term is again

$$\widetilde{W}_\Delta(z) = \frac{5}{24}\frac{T_3(\widehat{z})^2}{(1-T_2(z))^3} + O\left(\frac{1}{(1-T_2(z))^2}\right) \ . \tag{2.3.14}$$

In the similar manner as was done in [Jan+93, Lemma 2, Eq. (9.21)], we can prove that the dominant summand in the case of simple graphs and multigraphs is the same and equals the total compensation factor of cubic kernels e_{r0} times the generating function $\frac{T_3(z)^{\#\text{nodes}}}{(1-T_2(z))^{\#\text{edges}}}$. We omit the factor $T_2(z)^{\#\text{edges}}$ because it is equal to 1 (as $z = \widehat{z}$).

Proof of Lemma 2.3.2. The proof is similar to [Jan+93, Lemma 2]. We need to count the junctions of different degrees. All the paths contain vertices of degree at least 2, so we plug $T_2(z)$ into $P(\overline{\overline{M}}, z)$. □

Proof of Corollary 2.3.1. (From [Bol85, Chapter 2]) In a cubic multigraph each vertex has 3 half-edges that need to be paired, there are $6r$ half-edges in total. The number of such pairings is $(6r)!/\left((3r)!2^{3r}\right)$. In each vertex the three half-edges can be permuted in $3! = 6$ ways, so we divide by 6^{2r} to obtain finally that

$$\text{the number of cubic multigraphs} = \frac{(6r)!}{(3r)!2^{3r}6^{2r}} \ . \tag{2.3.15}$$

The multiple $(2r)!$ appears because the graph has $2r$ vertices and we deal with exponential generating functions. □

2.3.2 Weakly connected directed graphs

A *weakly connected* digraph is a directed graph whose underlying non-oriented projection is connected. We construct the EGFs of the analogs of the trees and unicycles in the world of directed graphs.

A rooted tree of size n has $n-1$ edges, and for each of the edges there are two orientation choices. For unrooted trees, we can take the vertex with label 1 as a canonical root, so that all of the 2^{n-1} edge orientations can be also distinguished and result in distinct oriented unrooted trees. Therefore, the EGFs for, respectively, oriented rooted and unrooted trees (i.e. directed weakly connected graphs whose non-oriented projections are, respectively, rooted and unrooted trees) are, respectively,

$$T_{\to}(z) = \frac{1}{2}T(2z), \quad \text{and} \quad U_{\to}(z) = \frac{1}{2}U(2z) = \frac{1}{2}T(2z) - \frac{1}{4}T^2(2z). \tag{2.3.16}$$

The EGF for simple digraphs whose underlying non-oriented projections are cycles of length at least 3, is obtained by substitution $z \mapsto 2z$ into that of cyclic non-oriented graphs, and equals $\frac{1}{2}\left[\log\frac{1}{1-2z} - (2z) - \frac{(2z)^2}{2}\right]$. In simple digraphs, it is allowed to have a circuit of length 2 on condition that it has the form $x \to \overline{x} \to x$, i.e. connects two vertices in both directions. Adding this case corresponds to the summand $z^2/2$.

By substitution, it follows that the EGF for unicyclic directed graphs (directed graphs whose non-oriented projections are unicyclic graphs) is

$$V_{\to}(z) = \frac{1}{2}\left[\log\frac{1}{1-2T_{\to}(z)} - (2T_{\to}(z)) - \frac{(2T_{\to}(z))^2}{2}\right] + \frac{T_{\to}(z)^2}{2}. \tag{2.3.17}$$

The last summand is taken out intentionally. Essentially, $V_{\to}(z) = V(2z) + T^2(2z)/8$, where $V(z)$ is the previously obtained EGF for unicyclic simple graphs. EGFs for directed graphs with higher excess can be constructed in the same way.

2.4 Acyclic and strongly connected digraphs

The materials of this section follow [dPD19]. The enumeration of two important digraph families, the *Directed Acyclic Graphs* (DAGs) and the strongly connected digraphs, has been successfully approached at least since 1969. Apparently, it was Liskovets [Lis69a; Lis70] who first deduced a recurrence for the number of strongly connected digraphs and also introduced and studied the concept of initially connected digraph, a helpful tool for their enumeration. Subsequently, Wright [Wri71] derived a simpler recurrence for strongly connected digraphs and Liskovets [Lis73] extended his techniques to the unlabeled case. Stanley counted labeled DAGs in [Sta73], and Robinson, in his paper [Rob77b], counted unlabeled DAGs with a given number of sources, which was the culmination of a series of publications he started in 1970 independently of Stanley. In the unlabeled case, his approach is very much related to the Species Theory [BLL98] which systematises the usage of cycle index series. Robinson also announced [Rob77a] a simple combinatorial explanation for the generating function of strongly connected digraphs in terms of the cycle index function. Publications on the exact enumeration of digraphs slowed down, until Gessel [Ges95], in 1995, returned to the problem with a new approach, based on *graphic generating functions*. It allowed him to enumerate initially connected components of a digraph and to immediately extend the analysis of DAGs by marking sources and sinks [Ges96]. The first English version paper we found containing the elegant expression for the generating function of strongly connected digraphs recalled in Theorem 2.4.1 is [Lis00]. It points to an earlier publication [Lis73] in Russian, which contains the proof.

In this section, we consider directed graphs (digraphs) with labeled vertices without loops and multiple edges. Two vertices u, v can be simultaneously linked by both edges $u \to v$ and $v \to u$. We also consider simple graphs which are undirected graphs with neither multiple edges nor loops.

We start by defining the building bricks for the symbolic method of the directed graphs.

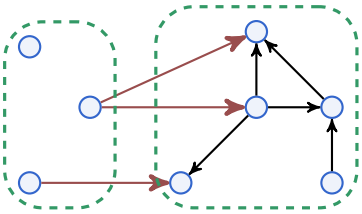


Figure 2.11: Symbolic method for DAG

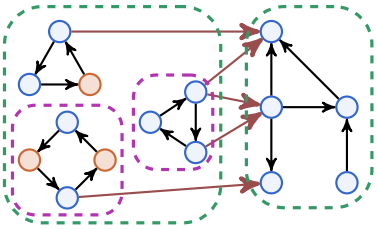


Figure 2.12: Digraphs and SCC

Proposition 2.4.1. The EGF of graphs $\mathrm{G}(z,w)$, GGF of digraphs $\mathbf{D}(z,w)$, and GGF of sets $\mathbf{Set}(z,w)$ (graphs that contain no edge) are

$$\mathrm{G}(z,w)=\mathbf{D}(z,w)=\sum_{n\geqslant 0}(1+w)^{\binom{n}{2}}\frac{z^n}{n!}\quad\text{and}\quad\mathbf{Set}(z,w)=\sum_{n\geqslant 0}\frac{1}{(1+w)^{\binom{n}{2}}}\frac{z^n}{n!}.$$

Proof. The sequences associated to the families of graphs, digraphs and sets are

$$g_n(w)=(1+w)^{\binom{n}{2}},\quad d_n(w)=(1+w)^{n(n-1)},\quad\text{and}\quad set_n(w)=1.$$

The expressions of the EGFs and GGFs follow. □

Corollary 2.4.1. The EGF and GGF of a family $\mathcal{A}$ are linked by the relations

$$\mathrm{A}(z,w)=\mathrm{G}(z)\odot\mathbf{A}(z,w)\quad\text{and}\quad\mathbf{A}(z)=\mathbf{Set}(z,w)\odot\mathrm{A}(z,w).$$

2.4.1 Directed acyclic graphs.

The next proposition illustrates the power of the symbolic method. The first result and its proof are classic. The second comes from [Sta73; Rob77b; Ges96].

Proposition 2.4.2. The EGF $\mathrm{C}(z,w)$ of connected graphs and the GGF $\mathbf{DAG}(z,w,u)$ of directed acyclic graphs (DAGs) with an additional variable u marking the sources (i.e. there are no oriented edge pointing to those vertices) are

$$\mathrm{C}(z,w)=\log(\mathrm{G}(z,w))\quad\text{and}\quad\mathbf{DAG}(z,w,u)=\frac{\mathbf{Set}((u-1)z,w)}{\mathbf{Set}(-z,w)}.$$

Proof. Since a graph is a set of connected graphs, we have, applying the symbolic method (Proposition 2.1.1),

$$\mathrm{G}(z,w)=e^{\mathrm{C}(z,w)},\quad\text{so}\quad\mathrm{C}(z,w)=\log(\mathrm{G}(z,w)).$$

The GGF of DAGs where each source is either marked, or left unmarked by the variable u, is $\mathbf{DAG}(z,w,u+1)$ (see Proposition 2.1.1). Such a DAG is decomposed as the arrow product of a set (the marked sources) with a digraph (Figure 2.11), so

$$\mathbf{DAG}(z,w,u+1)=\mathbf{Set}(zu,w)\mathbf{DAG}(z,w).$$

Taking $u=-1$ gives $1=\mathbf{Set}(-z,w)\mathbf{DAG}(z,w)$, so $\mathbf{DAG}(z,w)=1/\mathbf{Set}(-z,w)$. Replacing u with $u-1$ gives $\mathbf{DAG}(z,w,u)=\mathbf{Set}((u-1)z,w)/\mathbf{Set}(-z,w)$. This second proof also illustrates the translation into the generating function world of the inclusion-exclusion principle. □

2.4.2 Strongly connected graphs.

Let us recall that the *condensation* of a digraph is the directed acyclic graph (DAG) obtained from it by contracting each strongly connected component (SCC) to a vertex. The SCCs of the digraph corresponding to sources of the condensation are called *source-like SCCs*.

Lemma 2.4.1. The GGF $\mathbf{D}(z, w, u, v)$ of digraphs where u marks the number of source-like SCCs, and v the sum of their sizes, is equal to

$$\mathbf{D}(z, w, u, v) = \left(\mathbf{Set}(z, w) \odot e^{u\,\mathrm{SCC}(zv,w)-\mathrm{SCC}(z,w)}\right) \mathrm{G}(z, w).$$

Proof. Let $\mathcal{W}$ denote the family of sets of strongly connected components, each of them either

- marked by u, and where each vertex is either marked by v or left unmarked,
- not marked by u, but containing at least one vertex marked by v.

This construction translates into the relation

$$\mathrm{W}(z, w, u, v) = e^{u\,\mathrm{SCC}(z(v+1),w)} \cdot e^{\mathrm{SCC}(z(v+1),w)-\mathrm{SCC}(z,w)}.$$

By substituting $u \mapsto u+1$ and $v \mapsto v+1$ in $\mathbf{D}(z, w, u, v)$, we obtain the generating function of digraphs with variables u marking distinguished source-like components, and v marking distinguished vertices in source-like components (see Proposition 2.1.1). As illustrated by Figure 2.12, such digraphs can be represented as an arrow product of the family $\mathcal{W}$ and the family of all digraphs $\mathcal{D}$. On the level of generating functions, it writes as

$$\mathbf{D}(z, w, u+1, v+1) = \mathbf{W}(z, w, u, v)\mathbf{D}(z, w) = (\mathbf{Set}(z, w) \odot \mathrm{W}(z, w, u, v))\, \mathbf{D}(z, w),$$

using Corollary 2.4.1. According to Proposition 2.4.1, the digraphs GGF $\mathbf{D}(z, w)$ is equal to the graphs EGF $\mathrm{G}(z, w)$. Finally, u is replaced by $u-1$ and v by $v-1$ to obtain the claimed result. □

The *initially connected digraphs* are defined as digraphs where any vertex is reachable from the vertex with label 1 via an oriented path. Such digraphs have exactly one source-like SCC.

Lemma 2.4.2. The GGFs $\mathbf{IC}(z, w)$ and $[u^1]\mathbf{D}(z, w, u, v)$ of initially connected digraphs and, respectively, digraphs containing only one source-like SCC, whose size is marked by v, are linked by the relation

$$\partial_{v=1}[u^1]\mathbf{D}(z, w, u, v) = z\partial_z\mathbf{IC}(z, w).$$

Proof. As noted in [BLL98], the derivative $\partial_z\mathbf{IC}(z, w)$ corresponds to the class of initially connected digraphs with omitted vertex of label 1. Multiplying by z implies inserting a node with arbitrary label in place of it. We thus obtain digraphs reachable from a distinguished node. On the other hand, the operation $v\partial_v$ distinguishes a vertex in a source-like SCC. Taking $[u^1]v\partial_v\mathbf{D}(z, w, u, v)|_{v=1}$ brings us to the case where there is exactly one source-like SCC. It implies that every vertex is reachable from the distinguished one. Therefore, the two generating functions enumerate the same family. □

Our motivation for introducing initially connected digraphs is that there is a relation between their enumeration and the number of connected graphs ([Lis69b], proof also available in the conclusion of [Jan+93]).

Lemma 2.4.3. The GGF of initially connected digraphs is equal to the EGF of connected graphs

$$\mathbf{IC}(z, w) = \mathrm{C}(z, w) = \log(\mathrm{G}(z, w)).$$

Combining the previous results, we finally provide our new proof for the EGF of strongly connected digraphs (original result from [Lis00; Lis73]).

Theorem 2.4.1. The exponential generating function of strongly connected digraphs is equal to

$$\mathrm{SCC}(z,w) = -\log\left(G(z,w) \odot \frac{1}{G(z,w)}\right), \quad \text{where} \quad \mathrm{G}(z,w) = \sum_{n\geqslant 0} (1+w)^{\binom{n}{2}} \frac{z^n}{n!}.$$

Proof. Combining Lemma 2.4.2 and Lemma 2.4.3, we obtain

$$\partial_{v=1}[u^1]\mathbf{D}(z,w,u,v) = z\partial_z \mathbf{IC}(z,w) = z\partial_z \mathrm{C}(z,w).$$

Since $\mathrm{C}(z,w) = \log(\mathrm{G}(z,w))$ (see Proposition 2.4.2), the right hand-side is equal to $z(\partial_z \mathrm{G}(z,w))/\mathrm{G}(z,w)$. Injecting the expression of $\mathbf{D}(z,w,u,v)$ from Lemma 2.4.1 gives

$$\partial_{v=1}[u^1]\left(\mathbf{Set}(z,w) \odot e^{u\,\mathrm{SCC}(zv,w)-\mathrm{SCC}(z,w)}\right)\mathrm{G}(z,w) = \frac{z\partial_z \mathrm{G}(z,w)}{\mathrm{G}(z,w)}.$$

Since

$$\partial_{v=1}[u^1]e^{u\,\mathrm{SCC}(zv,w)-\mathrm{SCC}(z,w)} = (z\partial_z \mathrm{SCC}(z,w))\, e^{-\mathrm{SCC}(z,w)} = -z\partial_z e^{-\mathrm{SCC}(z,w)},$$

this relation becomes, after dividing both sides by $\mathrm{G}(z,w)$,

$$\mathbf{Set}(z,w) \odot z\partial_z e^{-\mathrm{SCC}(z,w)} = -\frac{z\partial_z \mathrm{G}(z)}{\mathrm{G}(z,w)^2} = z\partial_z \frac{1}{\mathrm{G}(z,w)}$$

By construction, the product $z\partial_z A(z) \odot B(z)$ is always equal to $z\partial_z(A(z)\odot B(z))$. Thus, the series $\mathbf{Set}(z,w)\odot e^{-\mathrm{SCC}(z,w)}$ and $1/\mathrm{G}(z,w)$ have equal derivatives and constant term 1, so they are equal. The neutral element for the exponential Hadamard product is the exponential function (i.e. for any series $A(z)$, we have $A(z)\odot e^z = A(z)$), so the inverse of $\mathbf{Set}(z,w)$, with respect to the Hadamard product, is $\mathrm{G}(z,w)$, implying

$$e^{-\mathrm{SCC}(z,w)} = \mathrm{G}(z,w) \odot \frac{1}{\mathrm{G}(z,w)}.$$

Applying the logarithm to both sides concludes the proof. □

This formula enables fast computation of the numbers of strongly connected digraphs: $\mathcal{O}(nm\log(n+m))$ arithmetic operations to compute the array of SCCs with at most n vertices and at most m edges, $\mathcal{O}(n\log(n))$ for the SCCs with at most n vertices without edge constraint. The following theorem is an original result. It could be a key element to the investigation of the structure of critical digraphs.

Theorem 2.4.2. The GGF of digraphs where the numbers of source-like SCCs and all SCCs are marked by the variables u and s is equal to

$$\frac{\mathbf{Set}(z,w) \odot e^{(u-1)s\,\mathrm{SCC}(z,w)}}{\mathbf{Set}(z,w) \odot e^{-s\,\mathrm{SCC}(z,w)}},$$

where $\mathrm{SCC}(z,w)$ has been expressed in Theorem 2.4.1.

Proof. In this proof, let $\mathbf{D}(z,w,u,s)$ denote the GGF of the digraph family expressed in the theorem. Then $\mathbf{D}(z,w,u+1,s)$ is the GGF of digraphs where each source-like SCC is either marked by u or left unmarked. Such a digraph is decomposed as the arrow product of a set of marked source-like SCCs with a digraph, so

$$\mathbf{D}(z,w,u+1,s) = \left(\mathbf{Set}(z,w) \odot e^{us\,\mathrm{SCC}(z,w)}\right)\mathbf{D}(z,w,1,s).$$

Replacing u with -1 gives

$$1 = \left(\mathbf{Set}(z,w) \odot e^{-s\,\mathrm{SCC}(z,w)}\right)\mathbf{D}(z,w,1,s) \quad \text{so} \quad \mathbf{D}(z,w,1,s) = \frac{1}{\mathbf{Set}(z,w) \odot e^{-s\,\mathrm{SCC}(z,w)}}.$$

Injecting this in the previous expression, with u replaced by $u-1$, finishes the proof. □

Conclusion. Many other digraph families could be enumerated using the same technique: symbolic method enriched with the arrow product, translation from EGF to GGF using exponential Hadamard product. The next challenge is the extraction of the coefficient asymptotics of such series. This would give access to a precise analysis of the phase transition of digraphs, following [Jan+93].

Chapter 3

Airy function and saddle point analysis

Contents

3.1 Introducing Airy function

In this chapter, we focus on the extraction of coefficients of type

$$[z^n]A(z)B(z)^m \tag{3.1.1}$$

when m is a function of n and $m \to \infty$ with n. There are so many things to be said about this expression! Starting from the simplest case, when $A(z) = 1$, $B(z) = 1 + z$, we fall into the case of Binomial theorem: $[z^n](1+z)^m = \binom{m}{n}$. Such a simple case allows an elegant and short expression, and also a similar case $B(z) = (1-z)^{-1}$: $[z^n](1-z)^{-m} = (-1)^n\binom{-m}{n}$. When m is constant, the result asymptotically appears to be a polynomial of n of degree $(m-1)$.

Another significant case is when m is fixed and $B(z)$ can be represented by a fractional power $B(z) \sim (1-z)^\alpha$ in a Delta-domain.

Proposition 3.1.1 (Transfer theorem [FS09, Section VI.3])**.** Suppose that $f(z/\rho)$ is a function analytic in the so-called *delta-domain* $\Delta(R, \phi)$ for some $R > 1$ and $0 < \phi < \frac{\pi}{2}$, where

$$\Delta(R,\phi) = \{\zeta \colon |\zeta| < R,\ \zeta \neq 1,\ \arg(\zeta - 1) > \phi\}. \tag{3.1.2}$$

Suppose that as $z \to \rho$, for $z/\rho \in \Delta(R,\phi)$, it holds

$$f(z) = h(z) - g(z)\left(1 - \frac{z}{\rho}\right)^{-\alpha} + O\left(\left|1 - \frac{z}{\rho}\right|^{-\beta}\right) \tag{3.1.3}$$

where $\alpha, \beta \in \mathbb{C} \setminus \mathbb{Z}_{\leqslant 0}$, and $h(z)$ and $g(z)$ are functions analytic in $|z| < R$.

Then, as $n \to \infty$, the coefficients $[z^n]f(z)$ admit an asymptotic approximation in form of

$$[z^n]f(z) \sim g(\rho) \cdot \rho^{-n} \cdot \frac{n^{\alpha-1}}{\Gamma(\alpha)} + O\left(\rho^{-n} n^{\beta-1}\right) \tag{3.1.4}$$

where $\Gamma \colon \mathbb{C} \setminus \mathbb{Z}_{\leqslant 0} \to \mathbb{C}$ is the Gamma function defined as $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$.

The proof of this case already requires a sophisticated contour in Cauchy's integral theorem. Moving further, considering the case when m is linear in n and when $A(z)$ and $B(z)$ are entire functions, the coefficient extraction can be heuristically described by the following scheme:

(i) Represent the coefficient $[z^n]A(z)B(z)^m$ as

$$[z^n]A(z)B(z)^m = \frac{1}{2\pi i}\oint_{|z|=R} A(z)B(z)^m \frac{dz}{z^{n+1}};$$

(ii) The contour integral is represented as

$$\frac{1}{2\pi i}\oint_{|z|=R} A(z)B(z)^m \frac{dz}{z^{n+1}} = \frac{1}{2\pi i}\oint_{|z|=R} e^{h_{n,m}(z)}\frac{dz}{z},$$

where $h_{n,m}(z) := \log A(z)B(z)^m z^{-n}$;

(iii) The circular contour of the integration $|z| = R$ is chosen in such a way that

$$h'_{n,m}(R) = 0;$$

(iv) The integral is approximated by Gaussian integral

$$\frac{1}{2\pi i}\oint e^{nh(z)}\frac{dz}{z} \sim \left(\frac{1}{2\pi n}\right)^{1/2} \frac{1}{R\sqrt{h''(R)}} e^{nh(R)}.$$

The case when $h''(R)$ turns to zero at the point where $h'(R) = 0$ is of special interest, as the above approximation is not available, yielding an expression tending to infinity at an a priori unknown speed.

Therefore,

$$\frac{1}{2\pi i}\oint g(z)e^{nh(z)}\frac{dz}{z} \sim \frac{1}{2\pi i}\oint g(z)\exp\left(nh(R) + nh'''(R)\frac{(z-R)^3}{6}\right)\frac{dz}{z}\ , \tag{3.1.5}$$

and this is the point where Airy function

$$A(y) = \int_{-\infty}^{\infty} e^{i(x^3/3 - xy)}dx$$

and related functions appear.

Furthermore, depending on the ratio m/n and depending on the nature of the functions $A(z)$ and $B(z)$ as complex-valued functions, the circular contour $|z| = R$ is subject to change. One possible solution is to use a *Lagrangian substitution* $t = z(t)e^t$, and a different solution is to adapt the contour. The situation becomes even more involved in case of two or more dimensions. On this account, the book [PW13] is entirely devoted to understanding of the topology of such multi-dimensional contour integrals.

The case when Hankel integration contour should be applied is covered by *semi-large powers theorem*.

Proposition 3.1.2 (Semi-large powers theorem, [FS09, Theorem IX.16], [Ban+01])**.** Suppose that $f(z/\rho)$ is a function analytic in delta-domain $\Delta(R, \phi)$, see Proposition 3.1.1, for some $R > 1$, and $f(z)$ admits asymptotic expansion as $z \to \rho$ for z/ρ staying in $\Delta(R, \phi)$:

$$f(z) \sim 1 - a\sqrt{1 - \frac{z}{\rho}}. \tag{3.1.6}$$

Then, for x in any compact subinterval of $(0, +\infty)$ the coefficient standing by z^n in $f(z)^k$ admits an asymptotic estimate

$$[z^n]f(z)^k \sim \frac{\rho^{-n}}{n}S(ax) \tag{3.1.7}$$

where $S(x)$ is the Rayleigh function satisfying

$$S(x) = \frac{xe^{-x^2/4}}{2\sqrt{\pi}} \quad \text{and} \quad x = \frac{k}{\sqrt{n}}. \tag{3.1.8}$$

3.2 Saddle point lemma for graphs

This section follows [DR18] and [Dov19].

3.2.1 Connected graphs and weakly connected digraphs

Among many tools introduced in [Jan+93], the central one was the saddle point lemma allowing to study the fine structure of a random graph near the point of its phase transition using the generating functions. We give it in a slightly reformulated form without a proof.

Lemma 3.2.1 ([Jan+93, Lemma 3])**.** Let $U(z)$ be the EGF of unrooted trees, $T(z)$ be the EGF of rooted Cayley trees, and $H(z)$ be an entire function of z. Suppose that $m = \frac{n}{2}(1+\mu n^{-1/3})$, and $H(t)$ is a function analytic at $t = 1$. Then, as $\mu \to -\infty$, and $|\mu| \leqslant n^{1/12}$,

$$\frac{n!}{|\mathcal{G}(n,m)|}[z^n]\frac{U(z)^{n-m}}{(n-m)!}e^{V(z)}\frac{H(T(z))}{(1-T(z))^y} = H(1)\left(\frac{n^{1/3}}{|\mu|}\right)^y(1+O(|\mu|^{-3})). \tag{3.2.1}$$

Remark 3.2.1. For the subcritical phase $\mu \to -\infty$, the condition $|\mu| \leqslant n^{1/12}$ can be replaced by $|\mu| \ll n^{1/3}$, as shown in [HR11] by a refined technical analysis of the asymptotics. We do not apply this refinement in the current paper as it would require a certain amount of additional technical details.

In a similar context, the enumeration of connected graphs has been pushed even further in [FSS04] using purely analytic methods. The resulting expansions are very tightly related with the Airy function and complex contour integration in the context of the previous result [Jan+93]. The properties of the Airy function and the phenomenon of coalescing saddles in combinatorics are out of the scope of the current paper, and the interested reader can find additional details e.g. in [Ald97; Ban+01].

Remark 3.2.2. Let us present an exemplary application of Lemma 3.2.1 analogous to presented in [Jan+93]. If a simple graph with n vertices and m edges contains only trees and unicycles, then the number of trees is equal to $m-n$, because each tree has one edge less than its number of vertices. The EGF for such graphs is then $e^{V(z)}U(z)^{n-m}/(n-m)!$. The factor $e^{V(z)}$ can be then rewritten as

$$e^{V(z)} = \frac{1}{\sqrt{1-T(z)}}e^{-T(z)/2-T^2(z)/4}.$$

When $m = \frac{n}{2}(1+\mu n^{-1/3})$, $\mu \to -\infty$ with n, and $|\mu| \leqslant n^{1/12}$, the probability that a graph consists only of trees and unicycles can be further refined using either the result from [Jan+93], or even further, using Lemma 3.3.1 (see below):

$$\frac{n!}{|\mathcal{G}(n,m)|}[z^n]\frac{U(z)^{n-m}}{(n-m)!}e^{V(z)} = 1 - \frac{5+o(1)}{24|\mu|^3}.$$

The interpretation of the factor $5/24$ is at least twofold. Formally, it appears as the second expansion term in the saddle point lemma with $5/24 = \frac{3y^2+3y-1}{6}\Big|_{y=1/2}$. At the same time, it denotes the sum of the compensation factors of the two possible cubic bicyclic multigraphs (see Figure 2.7), also expressed as the sum of the inverse cardinalities of their automorphism groups. Furthermore, the probability of having only one complex component which is bicyclic, is $\frac{5}{24}|\mu|^{-3} + O(|\mu|^{-6})$, and the probability of having a complex component of excess r is $e_r|\mu|^{-3r}$, where $e_r = \frac{(6r)!}{2^{5r}3^{2r}(3r)!(2r)!}$ corresponds to the sum of compensation factors of cubic multigraphs of excess r, see e.g. [Bol85, Chapter 2]. For complete asymptotic expansions in powers of $|\mu|^{-3}$ we refer to Section 3.3.

We give a variant of Lemma 3.2.1 for the case of directed graphs.

Lemma 3.2.2. As $\mu \to -\infty$, and $|\mu| \leqslant n^{1/12}$,

$$\frac{(2n)!}{|\mathcal{D}(2n,m)|}[z^{2n}]\frac{U_{\to}(z)^{2n-m}}{(2n-m)!}e^{V_{\to}(z)}\frac{H(T_{\to}(z))}{(1-2T_{\to}(z))^y} = H(1/2)\left(\frac{n^{1/3}}{|\mu|}\right)^y(1+O(|\mu|^{-3})) \tag{3.2.2}$$

$H(t)$ is a function analytic at $t = 1/2$.

Proof. As $\frac{m}{n} \to 1$ with n going to infinity, the number of simple digraphs with $2n$ vertices and m edges can be asymptotically related to the number of simple graphs with the same number of vertices and edges using the Stirling's formula

$$\frac{|\mathcal{D}(2n,m)|}{2^m|\mathcal{G}(2n,m)|} = \frac{\binom{2\binom{n}{2}}{m}}{2^m\binom{\binom{2n}{2}}{m}} \to e^{1/8}, \quad \text{as } n \to \infty.$$

Replacing $|\mathcal{D}(2n,m)|$ by its asymptotic equivalent, $U_{\to}(z)$ and $T_{\to}(z)$ by $\frac{1}{2}U(2z)$ and $\frac{1}{2}T(2z)$, and also $V_{\to}(z)$ by $V(2z) + T(2z)^2/8$, we obtain the expression

$$e^{-1/8}\frac{(2n)!2^{-m}}{|\mathcal{G}(2n,m)|}[z^{2n}]\frac{2^{-2n+m}U(2z)^{2n-m}}{(2n-m)!}e^{V(2z)+T(2z)/8}\frac{H(T(2z)/2)}{(1-T(2z))^y}.$$

After substituting $2z = x$, we get an additional 2^{2n} from the operator of coefficient extraction $[z^{2n}]$, and therefore, all the powers of two cancel out. We proceed by applying Lemma 3.2.2 and replacing each $T(x)$ by 1. Since the lemma is designed for a different regime, namely $m = \frac{n}{2}(1 + \mu n^{-1/3})$, we can adapt by replacing $n \mapsto 2n$, $\mu \mapsto 2^{1/3}\mu$. Finally, the factor $e^{-1/8}$ also cancels out, because of the presence of the multiple $e^{T(x)/8}$, and the ratio of $(2n)^{1/3}$ to $2^{1/3}|\mu|$ becomes again $n^{1/3}/|\mu|$. □

3.2.2 Graphs with degree constraints

In [DR18] we extend this lemma for the case of unrooted trees with degree constraints and ℓ-sprouted trees (see Section 2.3.1).

Lemma 3.2.3. Let $m = rn = \alpha n(1 + \mu\nu)$, where $\nu = n^{-1/3}$, $|\mu| = O(n^{1/12})$, $n \to \infty$, and $\widehat{z}$ be a unique real positive solution of $\phi_1(\widehat{z}) = 1$. Let

$$C_2 = \frac{t_3\alpha\widehat{z}}{2(1-\alpha)}, \quad C_3 = \frac{2t_3\alpha\widehat{z}}{3}, \quad t_3 = \frac{\widehat{z}\omega'''(\widehat{z})}{\omega'(\widehat{z})}.$$

Then for any function $\tau(z)$ analytic in $|z| \leqslant \widehat{z}$ the contour integral encircling complex zero, admits asymptotic representation

$$\frac{1}{2\pi i}\oint(1-\phi_1(z))^{1-y}e^{nh(z;r)}\tau(z)\frac{dz}{z} \sim \nu^{2-y}(zt_3)^{1-y}\tau(z)e^{nh(z;\alpha)} \times B_\Delta(y,\mu)\Big|_{z=\widehat{z}}\ , \tag{3.2.3}$$

$$\begin{aligned} B_\Delta(y,\mu) &= \frac{1}{3}C_3^{(y-2)/3}\sum_{k\geqslant 0}\frac{\left(C_2C_3^{-2/3}\mu\right)^k}{k!\Gamma\left(\frac{y+1-2k}{3}\right)} \\ h(z;r) &= \log\omega' - r\log z + (1-r)\log\left(2\frac{\omega}{\omega'} - z\right) \end{aligned}$$

The proof of the current lemma will be given below.

Corollary 3.2.1. If $m = \alpha n(1 + \mu n^{-1/3})$ and $y \in \mathbb{R}$, $y \geqslant \frac{1}{2}$, then for any $\Psi(t)$ analytic at $t = 1$ we have

$$\frac{n!}{(n-m)!|\mathcal{G}_{n,m,\Delta}|}[z^n]\frac{U(z)^{n-m}\Psi(T_2(z))}{(1-T_2(z))^y} = \sqrt{2\pi}\Psi(1)A_\Delta(y,\mu)n^{y/3-1/6} + O(R)\ , \tag{3.2.4}$$

$B_\Delta(y,\mu)$ is from Lemma 3.2.3 and the error term R is given by $R = (1 + |\mu|^4)n^{y/3-1/2}$. This function $A_\Delta(y,\mu)$ can be expressed in terms of $A(y,\mu) = A_{\mathbb{Z}_{\geqslant 0}}(y,\mu)$ introduced in [Jan+93]:

1. $A_\Delta(y,\mu) = (t_3\widehat{z})^{1-y}(3C_3)^{\frac{y-2}{3}}A\left(y, \frac{2C_2}{\sqrt[3]{(3C_3)^2}}\mu\right) = e^{-\mu^3/6}(\widehat{z}t_3)^{1-y}B_\Delta(y,\mu)$;

2. As $\mu \to -\infty$, we have $A(y,\mu) = \dfrac{1}{\sqrt{2\pi}|\mu|^{y-1/2}}\left(1 - \dfrac{3y^2+3y-1}{6|\mu|^3} + O(\mu^{-6})\right)$;

3. As $\mu \to +\infty$, we have $A(y,\mu) = \dfrac{e^{-\mu^3/6}}{2^{y/2}\mu^{1-y/2}}\left(\dfrac{1}{\Gamma(y/2)} + \dfrac{4\mu^{-3/2}}{3\sqrt{2}\Gamma(y/2-3/2)} + O(\mu^{-2})\right)$.

Proof of Lemma 3.2.3 and Corollary 3.2.1. Let us prove the corollary first. We start with "Stirling" approximation part. In case of classical random graphs it would be enough to apply the Stirling approximation, but in the case of degree constraints we apply the asymptotic result of [dPR16]:

$$\frac{n!}{(n-m)!|\mathcal{F}_{n,m,\Delta}|} = \sqrt{2\pi n}\frac{z_0^{2m}\sqrt{2\alpha}}{p\cdot\omega(z_0)^n}\times \exp(n\log n + (n-m)\log(n-m) - m\log 2m)\times \exp(\tfrac{1}{2}\phi_1(z_0) + \tfrac{1}{4}\phi_1^2(z_0))(1+O(n^{-1})) \ .$$

It happens that the exponential part of Stirling and some terms that will appear in Cauchy approximation, cancel out:

$$\exp\left(n\log n + (n-m)\log(n-m) - m\log 2m\right) = e^{-\mu^3/6}\times\frac{\omega(z_0)^n}{z_0^{2m}}2^{m-n}e^{nh(\widehat{z};\alpha)} \ . \tag{3.2.5}$$

Let us move to the Cauchy part for obtaining formal series coefficients. After Lagrangian variable change $t = T_1(z)$ we obtain:

$$[z^n]\frac{U(z)^{n-m}\Psi(T_2(z))}{(1-T_2(z))^y} = \frac{1}{2\pi i}\oint\frac{\Psi(T_2)U(z)^{n-m}dz}{(1-T_2(z))^y z^{n+1}} = \frac{2^{m-n}}{2\pi i}\oint\Psi(\phi_1(t))(1-\phi_1(t))^{1-y}e^{nh(t;r)}\frac{dt}{t} \ .$$

The statement readily follows from Lemma 3.2.3.

Let us prove the lemma then. We start with specifying an integration contour, namely the circle $z = \widehat{z}e^{-s\nu}$ where $s = \beta + it$, $\beta > 0$, $t \in [-\pi n^{1/3}, \pi n^{1/3}]$. We need $\beta \to 0$ with $n \to \infty$. Technically, for correct error estimate, β can be chosen from

$$\mu = \beta^{-1} - \beta \ , \tag{3.2.6}$$

as suggested by [Jan+93]. We need to switch to contour $t \in (-\infty, +\infty)$ with the price of exponentially small error $O(e^{-\max(2,|\mu|)n^{1/6}/3})$, we omit the details of this approximation since they are already considered in the mentioned article.

Next, there will be two variable changes. The first change of variables is $z = \widehat{z}e^{-s\nu}$. We use an approximation for $nh(z;r)$ near the double saddle $\widehat{z}$ and critical ratio α. From Lemma 3.2.5 it follows that maximum value of $|e^{nh(z,r)}|$ for $t \in [-\pi n^{1/3}, \pi n^{1/3}]$ is attained for $t = 0$ (and also at the points $\frac{2\pi i k}{d}\nu$ where d is a period of Δ, but we can assume without loss of generality that $d = 1$, because otherwise, extra terms cancel out when we count the probability, since the denominator is given by expression from [dPR16]). Thus, we can choose a small $t_0 > 0$, such that $nh''(\widehat{z})(\nu t_0)^2 \to \infty$, $nh'''(\widehat{z})(\nu t_0)^3 \to 0$, and the absolute value of integral for $|t| > t_0$ is negligible.

Since there is a relation $r = \alpha(1+\mu\nu)$, we can use a Taylor expansion for $h(z,r)$ for z around $\widehat{z}$, which is uniform with respect to $(\alpha - r)$:

$$h(z;r) = \sum_{k=0}^{3}\frac{h_z^{(k)}(\widehat{z};r)(z-\widehat{z})^k}{k!} + O\left((s\nu)^4\right) \ . \tag{3.2.7}$$

The first derivative turns to zero, the second and the third can be written as

$$h''(\widehat{z},r) = \frac{(\phi_0(\widehat{z})-2r)\phi_1'(\widehat{z})}{\widehat{z}(\phi_0(\widehat{z})-2)} = \frac{t_3(\alpha-r)}{\widehat{z}(\alpha-1)} \ , \tag{3.2.8}$$

$$h'''(\widehat{z},r) = \frac{\phi_0'(\widehat{z})\phi_1'(\widehat{z})}{\widehat{z}(\alpha-1)} + O(\mu\nu) \sim -\frac{4t_3\alpha}{\widehat{z}^2} \ , \tag{3.2.9}$$

hence the final approximation takes the form

$$nh(z;r) = nh(\widehat{z};\alpha) + C_2\mu s^2 + C_3 s^3 + O\left((\mu^2 s^2 + s^4)\nu\right), \tag{3.2.10}$$

where $C_2 = h''(\widehat{z};\alpha)\widehat{z}^2/2$ and $C_3 = -h'''(\widehat{z};\alpha)\widehat{z}^3/6$ are given in the formulation. We also have

$$(1-\phi_1(z))^{1-y} = s^{1-y}\nu^{1-y}(1+O(s\nu)) \ ,$$

so when $s = O(n^{1/12})$, the integrand can be approximated

$$(1 \ - \ \phi_1(z))^{1-y} e^{nh(z,r)} \quad = \quad \nu^{1-y} s^{1-y} e^{nh(\widehat{z};\alpha)} \ \times \ e^{C_2\mu s^2 + C_3 s^3}(1 \ + \ O(s\nu) \ + \ O(\mu^2 s^2 \nu) \ + \ O(s^4\nu)) \ ,$$

therefore

$$\frac{1}{2\pi i}\oint (1 \ - \ \phi_1(z))^{1-y} e^{nh(z;r)}\tau(z)\frac{dz}{z} \quad = \quad \frac{1}{2\pi i}(\nu\widehat{z}\phi_1'(\widehat{z}))^{1-y}\tau(\widehat{z})e^{nh(\widehat{z})} \ \times \ \oint s^{1-y} e^{C_2\mu s^2 + C_3 s^3} \ \cdot \ (-\nu)ds \ .$$

Then we perform a second change of variable $s = u^{1/3}C_3^{-1/3}$. We need to be careful with the contour of integration: note that the integral doesn't change if we take instead of $t \in [-\infty, +\infty]$ any path $\Pi(\beta), \beta > 0$ given by

$$s(t) = \begin{cases} -e^{-\pi i/3}t, & -\infty < t \leqslant 2\beta; \\ \beta + it\sin\pi/3, & -2\beta \leqslant t \leqslant 2\beta; \\ e^{\pi i/3}t, & 2\beta \leqslant t < +\infty. \end{cases} \tag{3.2.11}$$

After variable transform we obtain Hankel contour Γ extending from $-\infty$, circling the origin counterclockwise, and returning to $-\infty$, and $ds = \frac{1}{3}(C_3u^2)^{-1/3}du$.

$$\frac{1}{2\pi i}\int_{\Pi(\beta)}(\cdots) \quad = \quad \nu^{2-y}(\widehat{z}\phi_1'(\widehat{z}))^{1-y}\tau(\widehat{z})e^{nh(\widehat{z};\alpha)}C_3^{-y/3} \ \times \ \frac{1}{2\pi i}\int_\Gamma u^{y/3}e^u e^{C_2\mu u^{2/3}C_3^{-2/3}}du \ \cdot \ \tfrac{1}{3}C_3^{-1/3} \ .$$

Expanding the exponent

$$e^{C_2C_3^{-2/3}\mu u^{2/3}} = \sum_{k\geqslant 0}(C_2C_3^{-2/3}\mu u^{2/3})^k/k! \tag{3.2.12}$$

and applying the formula for inverse Gamma function on approximate Hankel contour we obtain the final statement. □

Lemma 3.2.4. Assume that $m = rn$. The coefficient at z^n of an EGF for graphs from $\mathcal{F}_{n,m,\Delta}$ given by an equation

$$[z^n]\frac{U(z)^{n-m+q}}{(n-m+q)!}e^{V(z)}E_{\mathbf{q}}(z) \ , \tag{3.2.13}$$

can be expressed as

$$\frac{2^{m-n}}{2\pi i(n-m+q)!}\oint e^{nh(z;r)}g(z)\tau(z)\frac{dz}{z} \ , \tag{3.2.14}$$

where the contour contains 0, the functions $h(z;r)$, $g(z)$ are given by

$$\begin{aligned} h(z;r) &= r\log\omega'(z) - r\log z + (1-r)\log(2\omega - z\omega'), \\ g(z) &= (1-\phi_1(z))^{1-y} \ , \ \ y = 3q + \tfrac{1}{2} \end{aligned}$$

and $\tau(z)$ doesn't have singularities in $|z| \leqslant \widehat{z}$.

Proof. We can do a variable change $T_1(z) = t \mapsto z$.
From the equation $T_1(z) = z\omega'(T_1(z))$ we obtain:

$$\begin{aligned} z &= t\omega'(t)^{-1} \mapsto z\omega'(z)^{-1}, \\ dz &= (\omega')^{-1}(1-\phi_1)dt, \\ T_\ell &= z\omega^{(\ell)}(t) = t\omega^{(\ell)}(t)\omega'(t)^{-1}, \\ T_2(z) &\mapsto \phi_1(z)\ , \\ U(z)^{n-m} &\mapsto 2^{m-n} z^{n-m}(2\omega(\omega')^{-1} - z)^{n-m} \end{aligned}$$

Then, we separate out the singular part:

$$\begin{aligned} U(z)^q e^{V(z)} E_{\mathbf{q}}(z)\frac{dz}{z} = &\underbrace{(1-T_2(z))(\omega')^{-1}}_{dz/dt}\underbrace{\omega'(t)t^{-1}}_{z^{-1}} \\ &\times \underbrace{U(z)^q\sqrt{1-T_2(z)}e^{V(z)}}_{\tau_1(z)}\underbrace{(1-T_2(z))^{3q}E_{\mathbf{q}}(z)}_{\tau_2(z)} \\ &\times \frac{1}{(1-T_2(z))^{3q}}\cdot\frac{1}{\sqrt{1-T_2(z)}}dt \mapsto \tau(z)g(z)\ , \end{aligned}$$

and the exponential one:

$$U(z)^{n-m}z^{-n} \mapsto 2^{m-n}\left(2\frac{\omega}{\omega'} - z\right)^{n-m} z^{n-m}\left(\frac{\omega'}{z}\right)^n = 2^{m-n}e^{nh(z;r)}\ . \tag{3.2.15}$$

□

In this section we mainly establish some asymptotic properties of $h(z;r)$ around $z = \widehat{z}$ and $r = r_0 = \alpha$. Its behaviour is important for saddle-point techniques. At arbitrary point $r = \alpha(1 + \mu n^{-1/3})$ its derivative factors as

$$h'_z(z;r) = \frac{(\phi_0(z) - 2r)(\phi_1(z) - 1)}{z(\phi_0(z) - 2)}\ , \tag{3.2.16}$$

and the dominant complex root of $h'_z(z;r)$ (closest to zero) is a positive real number which is either the solution of $\phi_0(z) = 2r$ or the solution of $\phi_1(z) = 1$. Each of the equations has unique real positive solution which we denote by $\mathrm{Root}_1(r)$ and $\mathrm{Root}_2 = \widehat{z}$, see Figure 3.1.

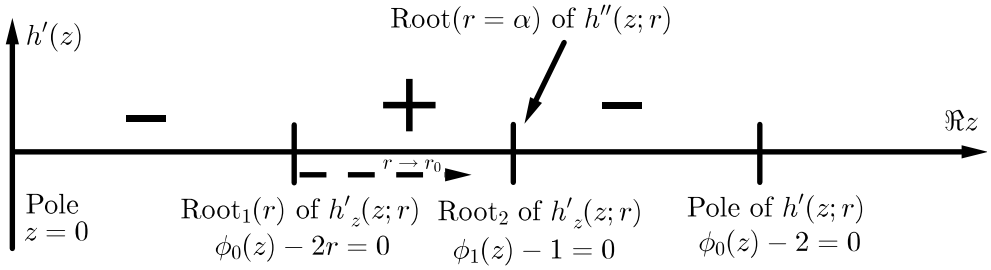


Figure 3.1: Configuration of roots of $h_z(z;r)$

Lemma 3.2.5. Let $z_0 > 0$, $z_0 \leqslant \min(\mathrm{Root}_1(r), \mathrm{Root}_2)$, the periodicity of Δ is p. Then the function

$$\Phi(\theta; r) = \mathrm{Re}\ h(z_0 e^{i\theta}; r) \tag{3.2.17}$$

attains its global maximums for $\theta \in [0, 2\pi)$ at p points $\theta_k = \frac{2\pi k}{p}$, $k = 0, 1, \ldots, p-1$.

Proof. Denote $z_0 e^{i\theta}$ by z. Without loss of generality we will treat the case of aperiodic $\omega(z)$, since any p-periodic function $\pi(z)$ can be reduced to an aperiodic one $\varphi(z)$ by a variable change $\pi(z) = z^\ell \varphi(z^p)$. $\Phi(\theta; r)$ can be rewritten as

$$\Phi(\theta; r) = r \log |\omega'(z)| + (1-r) \log |2\omega - z\omega'| + C \ . \tag{3.2.18}$$

We apply a version of Gibbs inequality for Kullback-Leibler divergence, also known as cross-entropy inequality: if p_1, p_2, q_1, q_2 are positive real numbers and $q_1 + q_2 \leqslant p_1 + p_2$ then

$$p_1 \log \frac{p_1}{q_1} + p_2 \log \frac{p_2}{q_2} \geqslant 0 \ . \tag{3.2.19}$$

It is now sufficient to prove that

$$r \cdot \left| \frac{\omega'(z)}{\omega'(z_0)} \right| + (1-r) \cdot \left| \frac{2\omega(z) - z\omega'(z)}{2\omega(z_0) - z_0\omega'(z_0)} \right| \leqslant 1 \ . \tag{3.2.20}$$

Note that $\phi_0(z_0) \leqslant 2r$. We first prove the inequality for $\widetilde{r} = \frac{1}{2}\phi_0(z_0)$. Since the function $\omega'(z)$ has non-negative coefficients, we always have $|\omega'(z)/\omega'(z_0)| \leqslant 1$, therefore if r increases, the inequality still remains true, thus for all $r \geqslant \widetilde{r}$ it is also true.

Substituting $\phi_0(z_0) = z_0\omega'(z_0)\omega(z_0)^{-1} = 2r$ we arrive to more simple inequality

$$|z\omega'(z)| + |2\omega(z) - z\omega'(z)| \leqslant 2\omega(z_0) \ , \quad z_0 \leqslant \alpha \ . \tag{3.2.21}$$

This inequality was proven by Fedor Petrov at mathoverflow [Pet16] using a beautiful geometric statement.

Let $\gamma > \beta > 0$ and $1/\beta - 1/\gamma \geqslant 2$, then for any vector z with $|z| = 1$

$$|1 + \gamma z| + |1 - \beta z| \leqslant 2 + \gamma - \beta \ . \tag{3.2.22}$$

Let us denote $z = e^{it}$. Differentiating the expression by θ and finding the zeros, we obtain

$$\frac{-2\gamma \sin\theta}{|1+\gamma z|} + \frac{2\beta \sin\theta}{|1+\beta z|} = 0 \ , \tag{3.2.23}$$

which is equivalent to

$$|z + 1/\gamma| = |z - 1/\beta| \ , \tag{3.2.24}$$

but the middle point of the segment $[-1/\gamma, 1/\beta]$ has value greater than or equal to 1 provided that $1/\beta - 1/\gamma \geqslant 2$, so the perpendicular bisector to this segment doesn't contain non-real points. The geometric statement in now proven.

Let $\omega(z) = \sum_{k\geqslant 0} c_k z^k$. Since $\phi_1(\widehat{z}) = 1$ and $0 < |z| \leqslant \widehat{z}$, the inequality $\phi_1(z) \leqslant 1$ can be expanded as

$$c_1 \geqslant \sum_{k\geqslant 2} (k^2-1) c_{k+1} |z|^k \ , \tag{3.2.25}$$

and we need to prove (3.2.21), which is equivalent to

$$\left| \sum_{k\geqslant 1} k c_k z^k \right| + |2c_0 + c_1 z - c_3 z^3 - 2c_4 z^4 - \dots| \leqslant 2c_0 + 2c_1|z| + 2c_2|z|^2 + \dots \tag{3.2.26}$$

This is reduced by applying triangle inequality for removing terms with c_0 and c_2 and dividing by $|z|$:

$$|c_1 + 3c_3 z^3 + \dots| + |c_1 - c_3 z^2 - 2c_4 z^3 - \dots| \leqslant 2c_1 + 2c_3|z|^2 + \dots \tag{3.2.27}$$

Repeatedly using triangle inequalities, the above can be reduced to a family of inequalities

$$|(k^2-1)|z|^k + (k+1)z^k| + |(k^2-1)|z|^k - (k-1)z^k| \leqslant (2(k^2-1)+2)|z|^k \ , \tag{3.2.28}$$

which is a partial case of the geometric statement with $\gamma = \frac{1}{k-1}$, $\beta = \frac{1}{k+1}$. □

3.3 Complete asymptotic expansion for the saddle point lemma

Let us show how a complete asymptotic expansion in powers of μ^{-3} can be obtained in Lemma 3.2.1. In [Jan+93], the following formulation is given.

Lemma 3.3.1 ([Jan+93, Lemma 3]). If $m = \frac{n}{2}(1+\mu n^{-1/3})$ and y is any real constant, $\mathcal{MG}(n,m)$ is the set of all labelled multigraphs with n vertices and m edges, then, as $\mu \to -\infty$ with n while remaining $|\mu| \leqslant n^{1/12}$,

$$\frac{n!}{|\mathcal{MG}(n,m)|}[z^n]\frac{U(z)^{n-m}}{(n-m)!(1-T(z))^y} = \sqrt{2\pi}A(y,\mu)n^{y/3-1/6}\Big(1+O(\mu^4 n^{-1/3})\Big),$$

where

$$A(y,\mu) = \frac{1}{\sqrt{2\pi}|\mu|^{y-1/2}}\left(1-\frac{3y^2+3y-1}{6|\mu|^3}+O(\mu^{-6})\right).$$

Complete asymptotic expansion. In the proof, it is mentioned that in principle, it is possible to obtain a complete asymptotic series of $|\mu|^{-3}$. Let us describe the procedure that can be used to compute these coefficients.

Let $\alpha = -\mu$. As shown in the proof of [Jan+93, Lemma 3], the function $A(y,\mu)$ can be represented in the form

$$A(y,\mu) = \frac{1}{2\pi\alpha^{y-1/2}}\int_{-\infty}^{\infty}\left(1+\frac{it}{\alpha^{3/2}}\right)^{1-y} e^{-t^2/2-it^3/(3\alpha^{3/2})}dt.$$

In order to express $A(y,\mu)$ in the form of a complete asymptotic expansion, we introduce $\beta := i\alpha^{-3/2}$ and obtain:

$$A(y,\mu) = \frac{1}{\sqrt{2\pi}|\mu|^{y-1/2}}\int_{-\infty}^{+\infty}(1+\beta t)^{1-y}\,e^{-\beta t^3/3}dt \sim \frac{1}{2\pi|\mu|^{y-1/2}}\sum_{r\geqslant 0}c_r(y)\beta^r,$$

where $(c_r(y))_{r=0}^{\infty}$ are polynomials in y. The coefficient $[\beta^k]\,(1+\beta t)^{1-y}\,e^{-\beta t^3/3}$ can be expressed as the convolution of two generating functions

$$[\beta^r]\,(1+\beta t)^{1-y}\,e^{-\beta t^3/3} = \sum_{k=0}^{r} t^k\binom{1-y}{k}\frac{(-t^3/3)^{r-k}}{(r-k)!} = \sum_{k=0}^{r}\frac{(-1/3)^{r-k}}{(r-k)!}t^{3r-2k}\binom{1-y}{k}$$

A formal term by term integration (the series is most likely not convergent, but the expansion could be extended up to rth term for any finite r) yields for even r

$$c_{2r}(y) = \sqrt{2\pi}\sum_{k=0}^{2r}\frac{(-1/3)^{2r-k}}{(2r-k)!}\binom{1-y}{k}\int_{-\infty}^{\infty}e^{-t^2/2}t^{6r-2k}dt = \sum_{k=0}^{2r}\frac{(-1/3)^{2r-k}}{(2r-k)!}\binom{1-y}{k}(6r-2k-1)!!,$$

where the double factorial notation is used $(2n-1)!! := 1\cdot 3\cdot\dots\cdot(2n-1)$; for odd r the principal value of the integral equals zero, and so, $c_{2r+1}(y) = 0$.

As an example, the first nontrivial term $c_2(y)$ can be computed as

$$c_2(y) = \frac{1/9}{2!}\cdot 5!! - \frac{1/3}{1!}\cdot(1-y)\cdot 3!! + \frac{(1-y)(-y)}{2}\cdot 1!! = \frac{3y^2+3y-1}{6}$$

and a factor -1 in the terms $c_{4r+2}(y)$ in the expansion of $A(y,\mu)$ appears because a multiple $i^{4r+2} = -1$ should be extracted from $\beta = i\alpha^{-3/2}$.

Chapter 4

Infinite systems of algebraic equations

Contents

This chapter follows [BBD18b]. In this chapter we present our technical contribution meant for dealing with certain recursive, infinite systems of generating functions, i.e. Theorem 4.4.1. Although our results admits broader applications than counting of closed lambda terms, for consistency, we focus only on this application. Our proof is motivated by the papers [BGG17; GG16] where the authors consider the enumeration problem of closed λ-terms without additional marking parameters. For that purpose, they construct a series of sequences $(L_{m,N})_{m\geqslant 0}$ that approximate m-open λ-terms with convergence rate of order $O\left(\frac{1}{\sqrt{N}}\right)$. In what follows we improve this rate to an exponential one. Moreover, we abstract the considered system away from λ-terms in the de Bruijn notation, allowing for an analysis of more general varieties of combinatorial systems.

Remarkably, in [GG16, Section 5] the authors obtain the asymptotic estimate $\Theta(n^{-3/2}\rho^{-n})$ for the number of closed lambda terms of size n with the difference between upper and lower bounds on the constant multiple within 10^{-7}. Let us notice that this technique is quite different from the technique used in [BGG17]. Our approach is more similar to the former method, however admits certain improvements. Specifically, we simplify the procedure of improved constant estimation, give a rigorous proof about the exponential convergence rate of the constant multiple, and provide more general and simpler tools based on the properties of the Jacobian of the limiting system (instead of exploiting the particular form of this equation in the case of λ-terms).

4.1 Systems of algebraic equations

4.1.1 Drmota–Lalley–Woods Theorem

The following theorem, commonly known as the Drmota–Lalley–Woods theorem, is a fundamental result obtained independently by several authors [Drm97; Woo97; Lal93] in order to establish limit laws in various families of tree structures specified by context-free grammars. In our exposition, we reference Drmota's book [Drm09, Section 2.2.5], and the papers [Drm97; DGM12; PSS12; BBY10].

Definition 4.1.1. Consider a polynomial system of equations

$$\boldsymbol{y} = \boldsymbol{\Phi}(z, \boldsymbol{y}, \boldsymbol{u}) \tag{4.1.1}$$

which is a vector notation for $(y_j = \Phi_j(z, y_1, \ldots, y_m, u_1, \ldots, u_d))$ with j ranging over $1, \ldots, m$. Assume that $\boldsymbol{\Phi}(0, \mathbf{0}, \mathbf{0}) = \mathbf{0}$. Then,

- $\boldsymbol{\Phi}(z, \boldsymbol{y}, \boldsymbol{u})$ is said to be *non-linear* if at least one of its component polynomials Φ_j is non-linear in one of the formal variables $y_1, \ldots, y_m$;
- $\boldsymbol{\Phi}(z, \boldsymbol{y}, \boldsymbol{u})$ is said to be *algebraic positive* if all of its component polynomials Φ_j have non-negative coefficients;
- $\boldsymbol{\Phi}(z, \boldsymbol{y}, \boldsymbol{u})$ is said to be *algebraic proper* if it admits a unique formal power series solution to which the iteration
$$\boldsymbol{y}_0(z, \boldsymbol{u}) = \mathbf{0} \quad \text{and} \quad \boldsymbol{y}_{k+1}(z, \boldsymbol{u}) = \boldsymbol{\Phi}(z, \boldsymbol{y}_k(z, \boldsymbol{u}), \boldsymbol{u}) \tag{4.1.2}$$
considered in the metric space of formal power series with valuation, converges as $k \to \infty$, and the Jacobian matrix $\dfrac{\partial \boldsymbol{\Phi}}{\partial \boldsymbol{y}}$ is nilpotent at $(z, \boldsymbol{y}) = (0, \mathbf{0})$.
- $\boldsymbol{\Phi}(z, \boldsymbol{y}, \boldsymbol{u})$ is said to be *algebraic irreducible* if its dependency graph (i.e. a graph whose vertices are the integers $1, \ldots, m$ and there exists a directed edge $k \to j$ if y_j figures in a monomial of Φ_k) is strongly connected;
- $\boldsymbol{\Phi}(z, \boldsymbol{y}, \boldsymbol{u})$ is said to be *algebraic aperiodic* if each of its component solutions $y_j(z, \mathbf{1})$ for $j = 1, \ldots, m$ is aperiodic in the sense that the greatest common divisor of the pairwise differences of the set of exponent indices of z within $y_j(z, \mathbf{1})$ is equal to 1.

Remark 4.1.1. The notion of algebraic properness of systems, also referred to as *well-foundedness*, is extensively studied in [PSS12, Section 5]. As discussed in [PSS12; Joy81], the system has combinatorial meaning only if the Jacobian is nilpotent, i.e. if the recursive definition is well-defined and allows to inductively construct all the instances of combinatorial species.

Let us note that the condition $\boldsymbol{\Phi}(0, \mathbf{0}, \mathbf{0}) = \mathbf{0}$ is a technical assumption of the Drmota–Lalley–Woods theorem. Pivoteau, Salvy and Soria consider, *inter alia*, well-founded systems for which $\boldsymbol{y}(0, \mathbf{0}) \neq \mathbf{0}$. One possible characterisation of such systems is the condition that the limit of a suitable iterative approximation procedure yields the solution of the initial functional system.

The assumption that a system is *polynomial* can be replaced by a more general assumption that the functions are *analytic* at zero, see [Drm97]. For a detailed and non-trivial study of the conditions regarding analytic functions and the configuration of the critical points in this more general case, see [BBY10].

Proposition 4.1.1 (Irreducible positive polynomial systems)**.** Let

$$\boldsymbol{y} = \boldsymbol{\Phi}(z, \boldsymbol{y}, \boldsymbol{u}) = (y_j = \Phi_j(z, y_1, \ldots, y_m, \boldsymbol{u})), \quad j = 1, \ldots, m \tag{4.1.3}$$

be a non-linear polynomial system of equations which is algebraic positive, proper, and irreducible. Then there exists $\varepsilon > 0$ such that all component solutions $y_j(z, \boldsymbol{u})$ admit representation of the form

$$y_j(z, \boldsymbol{u}) = h_j\left(\sqrt{1 - \frac{z}{\rho(\boldsymbol{u})}}, \boldsymbol{u}\right) = \sum_{k \geqslant 0} c_{k,j}(\boldsymbol{u}) \left(1 - \frac{z}{\rho(\boldsymbol{u})}\right)^{k/2} \tag{4.1.4}$$

for $\boldsymbol{u}$ in a neighbourhood of $\mathbf{1}$, $|z - \rho(\boldsymbol{u})| < \varepsilon$ and $\arg(z - \rho(\boldsymbol{u})) \neq 0$, where $c_{k,j}(\boldsymbol{u})$ and $\rho(\boldsymbol{u})$ are analytic functions of $\boldsymbol{u}$, and the functions $h_j(t, \boldsymbol{u})$ are analytic at $(t, \boldsymbol{u}) = (0, \mathbf{1})$. In addition, if the system is algebraic aperiodic, then all y_j have $\rho(\boldsymbol{u})$ as their unique dominant singularity, and there exist constants $0 < \delta < \pi/2$ and $\eta > 0$ such that $\boldsymbol{y}(z, \boldsymbol{u})$ is analytic in a region of the form

$$\Delta := \{z \colon |z| < \rho(\mathbf{1}) + \eta, |\arg(z/\rho(\boldsymbol{u}) - 1)| > \delta\}. \tag{4.1.5}$$

Remark 4.1.2. The above Drmota–Lalley–Woods theorem has been further generalised by Drmota, Gittenberger and Morgenbesser in the case of strongly connected systems with infinitely many equations [DGM12]. In their generalisation, the authors require that the Jacobian of the system (or some of its power) is a compact operator. Alas, as the system corresponding to closed λ-terms is not strongly connected, it does not fit into their framework. In the current section we introduce a new condition of *exponential convergence* which is independent of the Jacobian and conjecture that it is crucial for obtaining the respective Puiseux expansions of generating functions.

Proposition 4.1.2 (Differential condition for the systems of equations [DGM12, see proof of Theorem 1])**.** Let $\boldsymbol{y} = \boldsymbol{\Phi}(z, \boldsymbol{y})$ be a non-linear system of polynomial equations $y_j = \Phi_j(z, y_1, \ldots, y_m)$ with j ranging over $1, \ldots, m$. Assume that $\boldsymbol{y} = \boldsymbol{\Phi}(z, \boldsymbol{y})$ is algebraic positive, proper and irreducible. Let ρ be the common singularity of its solution vector y_j. Then, the spectral radius (largest absolute value of its eigenvalues) of the Jacobian matrix $\dfrac{\partial \boldsymbol{\Phi}}{\partial \boldsymbol{y}}$ is a strictly increasing function of z on the interval $[0, \rho]$ and is bounded from above by 1, with the equality holding if and only if $z = \rho$.

4.1.2 Limit laws

Consider a bivariate generating function $L(z, u)$ with non-negative coefficients and a sequence of random variables $(X_n)_{n \geqslant 0}$ such that

$$L(z, u) = \sum_{n,k \geqslant 0} a_{n,k} z^n u^k \quad \text{and} \quad \mathbb{P}(X_n = k) = \frac{a_{n,k}}{\sum_{j \geqslant 0} a_{n,j}}. \tag{4.1.6}$$

We say that X_n *is associated with variable* u. In order to understand the limiting behaviour of X_n we investigate the *probability generating function* $p_n(u)$ of X_n defined as

$$p_n(u) := \sum_{k \geqslant 0} \mathbb{P}(X_n = k) u^k = \frac{[z^n] L(z, u)}{[z^n] L(z, 1)}. \tag{4.1.7}$$

Once accessed, $p_n(u)$ proves extremely useful in establishing the traits of X_n as n tends to infinity. In what follows, we focus on two types of limiting distributions. The first type is related to the case of a so-called *fixed* singularity, which results in discrete limit law; the second type is related to so-called *moving* singularity, and typically results in a Gaussian limit law.

Discrete limit laws

Proposition 4.1.3 ([FS09, Section IX.2])**.** Suppose that bivariate power series $L(z, u)$ admits in a complex neighbourhood of $u = 1$ a Puiseux series expansion in form of

$$L(z, u) = \alpha(u) - \beta(u) \sqrt{1 - \frac{z}{\rho}} + O\left(\left|1 - \frac{z}{\rho}\right|\right) \tag{4.1.8}$$

as $z \to \rho$ uniformly in delta-domain $\Delta(R)$ for some $R > \rho$ (see Proposition 3.1.1). Then, the random variable X_n associated with the marking variable u converges in distribution to a discrete limiting distribution with probability generating function

$$p(u) = \lim_{n \to \infty} p_n(u) = \frac{\beta(u)}{\beta(1)}. \tag{4.1.9}$$

The corresponding mean values satisfy

$$\lim_{n \to \infty} \mathbb{E} X_n = \frac{\beta'(1)}{\beta(1)}. \tag{4.1.10}$$

Central limit theorem

Remark 4.1.3. In 1983, Bender and Richmond [BR83] proved a multi-dimensional variant of the central limit theorem for probability generating functions taking the quasi-power form $p_n(\boldsymbol{u}) \sim A(\boldsymbol{u})B(\boldsymbol{u})^n$. This line of research was later continued by Hwang [Hwa98] who established precise rates of convergence in the one-dimensional case. The two-dimensional case by was next investigated by Heuberger [Heu07]. More recently, in 2016, the full multi-dimensional version has been resolved by Heuberger and Kropf [HK16] using a multi-dimensional version of the Berry–Esseen inequality. Although we do not touch on the rates of convergence in the current section, let us mention that they can be obtained using the above results.

In order to formulate the multivariate central limit theorem, it is convenient to introduce the notion of logarithmic derivative which enters the mean value and the covariance matrix of the resulting random variable.

Definition 4.1.2. The logarithmic derivative of $A(z, u)$ is given by the expression

$$\frac{\partial}{\partial \log u} A(z,u) := \frac{\partial}{\partial \eta} A(z, e^{\eta})\Big|_{\eta=\log u} = u \frac{\partial}{\partial u} A(z,u). \tag{4.1.11}$$

Proposition 4.1.4 (Multivariate central limit theorem, [BR83, Theorem 1])**.** Let $(\boldsymbol{X}_n)_{n=1}^{\infty}$ be a sequence of coordinate-wise non-negative d-dimensional discrete random vectors with probability generating functions $p_n(\boldsymbol{u}) := \mathbb{E}(\boldsymbol{u}^{X_n})$, $\boldsymbol{u} = (u_1, u_2, \dots, u_d)$. Suppose that uniformly in a fixed complex neighbourhood of $\boldsymbol{u} = \mathbf{1}$ one has

$$p_n(\boldsymbol{u}) \sim A(\boldsymbol{u}) \cdot B(\boldsymbol{u})^n \tag{4.1.12}$$

where $A(\boldsymbol{u})$ is uniformly continuous and $B(\boldsymbol{u})$ has a quadratic Taylor series expansion with error term $O\big(\sum |u_k - 1|^3\big)$. Assume that $B(\boldsymbol{u})$ satisfies the following *variability condition*:

$$\det\left[\frac{\partial^2 \log B(\boldsymbol{u})}{\partial \log u_i \, \partial \log u_j}\right]_{i,j} > 0. \tag{4.1.13}$$

Then, the sequence of random variables $\boldsymbol{X}_n$, after standardization, converges in law to Gaussian random variable satisfying

$$\frac{\boldsymbol{X}_n - \mathbb{E}\boldsymbol{X}_n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma). \tag{4.1.14}$$

The mean vector and the covariance matrix satisfy

$$\mathbb{E}\boldsymbol{X}_n \sim n \cdot \frac{\partial B}{\partial \boldsymbol{u}}\bigg|_{\boldsymbol{u}=\mathbf{1}} \quad \text{and} \quad \operatorname{Cov} \boldsymbol{X}_n \sim n \cdot \left[\frac{\partial^2 \log B(\boldsymbol{u})}{\partial \log u_i \partial \log u_j}\right]_{i,j}\bigg|_{\boldsymbol{u}=1}. \tag{4.1.15}$$

In the one-dimensional case (4.1.15) simplifies to

$$\mathbb{E}(X_n) \sim nB'(1) \quad \text{and} \quad \mathbb{V}(X_n) \sim n\left(B''(1) + B'(1) - B'(1)^2\right). \tag{4.1.16}$$

Remark 4.1.4. Typically, when the singularity is moving (see Proposition 4.1.1) the bivariate generating function takes the form

$$A(z, \boldsymbol{u}) = \alpha(\boldsymbol{u}) - \beta(\boldsymbol{u})\sqrt{1 - \frac{z}{\rho(\boldsymbol{u})}} + O\left(\left|1 - \frac{z}{\rho(\boldsymbol{u})}\right|\right) \tag{4.1.17}$$

uniformly as $z \to \rho(\boldsymbol{u})$ for $\boldsymbol{u}$ in a vicinity of $\mathbf{1}$.

Consequently, the probability generating function takes form

$$p_n(\boldsymbol{u}) \sim \frac{\beta(\boldsymbol{u})}{\beta(\mathbf{1})} \left(\frac{\rho(\mathbf{1})}{\rho(\boldsymbol{u})}\right)^n. \tag{4.1.18}$$

In this form, the probability generating function satisfies the premises of the multivariate quasi-powers theorem (see Remark 4.1.3) and so one can also obtain the speed of convergence. In our situations this speed is typically of order $O\left(\frac{1}{\sqrt{n}}\right)$.

Remark 4.1.5. For convenience, we say that a random vector $\boldsymbol{X}_n$ converges in law to multivariate Gaussian distribution with mean $n\boldsymbol{\mu}$ and variance $n\Sigma$ writing

$$\boldsymbol{X}_n \xrightarrow{d} \mathcal{N}(n\boldsymbol{\mu}, n\Sigma) \tag{4.1.19}$$

to denote that $\dfrac{\boldsymbol{X}_n - n\boldsymbol{\mu}}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma)$.

4.2 Calculus techniques for formal power series

We start with some basic notation and properties of coefficient-wise inequalities on formal power series and some geometric results on matrices of formal power series. In what follows, we denote the *spectral radius* of matrix $\boldsymbol{\mathcal{A}}$, i.e. the largest absolute value of its eigenvalues, by $r(\boldsymbol{\mathcal{A}})$.

Definition 4.2.1 (Formal power series domination)**.** We say that $f(z)$ *is dominated by* $g(z)$, denoted as $f(z) \preceq g(z)$, if for each $n \geqslant 0$ we have $[z^n]f(z) \leqslant [z^n]g(z)$. For multivariate formal power series

$$f(\boldsymbol{z}) = \sum_{\boldsymbol{n} \geqslant \mathbf{0}} a_{\boldsymbol{n}} \boldsymbol{z}^{\boldsymbol{n}} \quad \text{and} \quad g(\boldsymbol{z}) = \sum_{\boldsymbol{n} \geqslant \mathbf{0}} b_{\boldsymbol{n}} \boldsymbol{z}^{\boldsymbol{n}} \tag{4.2.1}$$

where $\boldsymbol{z} = (z_1, \ldots, z_d)$, $\boldsymbol{n} = (n_1, \ldots, n_d)$ and $\boldsymbol{z}^{\boldsymbol{n}} = z_1^{n_1} z_2^{n_2} \ldots z_d^{n_d}$ the domination $f(\boldsymbol{z}) \preceq g(\boldsymbol{z})$ means that for each vector of indices $\boldsymbol{n}$ it holds $a_{\boldsymbol{n}} \leqslant b_{\boldsymbol{n}}$. Certainly, if a combinatorial class $\mathcal{F}$ is included in a combinatorial class $\mathcal{G}$, then the generating functions $f(z)$ and $g(z)$ corresponding to respective classes satisfy $f(z) \preceq g(z)$. The same holds for marked classes and associated multivariate generating functions. Finally, for vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ of identical (however not necessarily finite) dimension, we write $\boldsymbol{A} \preceq \boldsymbol{B}$ to denote a coordinate-wise domination of respective components.

In real analysis, the *squeeze lemma* is a theorem regarding the limit of the sequence which is upper- and lower-bounded by two sequences with the same limit value. The following statement is a variant of this lemma, stated in the context of formal power series admitting coefficient asymptotics suitable for analysis of corresponding limit laws, see Section 4.1.2.

Lemma 4.2.1 (Squeeze lemma for formal power series)**.** Let $z \in \mathbb{C}$ and $\boldsymbol{u} = (u_1, \ldots, u_r) \in \mathbb{C}^r$. Assume that $f(z, \boldsymbol{u})$, $(h_m(z, \boldsymbol{u}))_{m \geqslant 0}$, and $(g_m(z, \boldsymbol{u}))_{m \geqslant 0}$ are multivariate formal power series in $(z, \boldsymbol{u})$ with non-negative coefficients, such that for every n and m, the functions

$$[z^n]f(z, \boldsymbol{u}), \ [z^n]h_m(z, \boldsymbol{u}) \text{ and } [z^n]g_m(z, \boldsymbol{u}) \tag{4.2.2}$$

are polynomials in $\boldsymbol{u}$.

Moreover, assume that the following conditions hold:

- for each $m \geqslant 0$, we have
$$h_m(z, \boldsymbol{u}) \preceq f(z, \boldsymbol{u}) \preceq g_m(z, \boldsymbol{u}) \tag{4.2.3}$$
in the sense of multivariate formal power series domination;
- there exists a sequence of real positive numbers C_n and functions $\left(\underline{A_m}(\boldsymbol{u})\right)_{m \geqslant 0}$, $\left(\overline{A_m}(\boldsymbol{u})\right)_{m \geqslant 0}$, $B(\boldsymbol{u})$ analytic in a common neighbourhood of $\boldsymbol{u} = \mathbf{1}$, such that uniformly for $m \geqslant 0$ and uniformly in a fixed complex vicinity of $\boldsymbol{u} = \mathbf{1}$, it holds
$$\lim_{n \to \infty} \frac{[z^n]h_m(z, \boldsymbol{u})}{C_n \underline{A_m}(\boldsymbol{u}) B(\boldsymbol{u})^n} = 1 \quad \text{and} \quad \lim_{n \to \infty} \frac{[z^n]g_m(z, \boldsymbol{u})}{C_n \overline{A_m}(\boldsymbol{u}) B(\boldsymbol{u})^n} = 1; \tag{4.2.4}$$

- there exists a function $A(\boldsymbol{u})$ analytic near $\boldsymbol{u}=\mathbf{1}$ such that $A(\mathbf{1})\neq 0$ satisfying uniformly in a complex vicinity of $\boldsymbol{u}=\mathbf{1}$

$$\lim_{m\to\infty}\underline{A_m}(\boldsymbol{u})=\lim_{m\to\infty}\overline{A_m}(\boldsymbol{u})=A(\boldsymbol{u}). \tag{4.2.5}$$

Then, uniformly in a complex vicinity of $\boldsymbol{u}=\mathbf{1}$, as $n\to\infty$:

$$[z^n]f(z,\boldsymbol{u})\sim C_nA(\boldsymbol{u})B(\boldsymbol{u})^n. \tag{4.2.6}$$

Proof. We divide the proof into two parts. We start with showing that the statement holds for vectors $\boldsymbol{u}$ whose components are real positive numbers. Next, we extend this property onto all complex components of $\boldsymbol{u}$.

First, take $\boldsymbol{u}\in\mathbb{R}^r$ in the vicinity where the functions $\underline{A_m}(\boldsymbol{u})$, $\overline{A_m}(\boldsymbol{u})$, $B(\boldsymbol{u})$ are analytic. Then, following (4.2.3) and (4.2.4), for every positive ε there exists $N:=N(\varepsilon)$, independent of m and $\boldsymbol{u}$, such that

$$\forall n>N\quad \underline{A_m}(\boldsymbol{u})(1-\varepsilon)\leqslant\frac{[z^n]f(z,\boldsymbol{u})}{C_nB(\boldsymbol{u})^n}\leqslant\overline{A_m}(\boldsymbol{u})(1+\varepsilon). \tag{4.2.7}$$

Taking the limit with respect to m we note that condition (4.2.5) guarantees that for arbitrary small $\varepsilon>0$ and sufficiently large N (again, independent of $\boldsymbol{u}$) we have

$$\forall n>N\quad A(\boldsymbol{u})(1-\varepsilon)\leqslant\frac{[z^n]f(z,\boldsymbol{u})}{C_nB(\boldsymbol{u})^n}\leqslant A(\boldsymbol{u})(1+\varepsilon). \tag{4.2.8}$$

In other words, for values of $\boldsymbol{u}\in\mathbb{R}^r$ within a fixed vicinity of $\boldsymbol{u}=\mathbf{1}$ it holds

$$[z^n]f(z,\boldsymbol{u})\sim C_nA(\boldsymbol{u})B(\boldsymbol{u})^n. \tag{4.2.9}$$

Now, let us consider $\boldsymbol{u}\in\mathbb{C}^r$. Note that since for each $n,m\geqslant 0$ the formal power series $[z^n]f(z,\boldsymbol{u})$, $[z^n]h_m(z,\boldsymbol{u})$ and $[z^n]g_m(z,\boldsymbol{u})$ are polynomials in $\boldsymbol{u}$, they are analytic in $\mathbb{C}^r$. Moreover, as $\psi_{n,m}(\boldsymbol{u}):=[z^n]f(z,\boldsymbol{u})-[z^n]h_m(z,\boldsymbol{u})$ is a polynomial with non-negative coefficients, for every $\boldsymbol{u}\in\mathbb{C}^r$ we have $|\psi_{n,m}(\boldsymbol{u})|\leqslant\psi_{n,m}(|\boldsymbol{u}|)$ and, consequently,

$$\left|[z^n]f(z,\boldsymbol{u})-[z^n]h_m(z,\boldsymbol{u})\right|\leqslant[z^n]f(z,|\boldsymbol{u}|)-[z^n]h_m(z,|\boldsymbol{u}|). \tag{4.2.10}$$

After dividing both parts by $C_n|A(\boldsymbol{u})||B(\boldsymbol{u})^n|$ we obtain

$$\left|\frac{[z^n]f(z,\boldsymbol{u})}{C_nA(\boldsymbol{u})B(\boldsymbol{u})^n}-\frac{[z^n]h_m(z,\boldsymbol{u})}{C_nA(\boldsymbol{u})B(\boldsymbol{u})^n}\right|\leqslant\frac{[z^n]f(z,|\boldsymbol{u}|)}{C_n|A(\boldsymbol{u})||B(\boldsymbol{u})^n|}-\frac{[z^n]h_m(z,|\boldsymbol{u}|)}{C_n|A(\boldsymbol{u})||B(\boldsymbol{u})^n|}. \tag{4.2.11}$$

Following condition (4.2.4) and the estimate (4.2.9) for $\boldsymbol{u}\in\mathbb{R}^r$, we note that for every $\varepsilon>0$ there exists $N:=N(\varepsilon)$ independent of m and $\boldsymbol{u}$, such that for all $n>N$ we further have

$$\left|\frac{[z^n]f(z,\boldsymbol{u})}{C_nA(\boldsymbol{u})B(\boldsymbol{u})^n}-\frac{\underline{A_m}(\boldsymbol{u})}{A(\boldsymbol{u})}\right|\leqslant\frac{A(|\boldsymbol{u}|)-\underline{A_m}(|\boldsymbol{u}|)}{|A(\boldsymbol{u})|}\cdot\frac{B(|\boldsymbol{u}|)^n}{|B(\boldsymbol{u})^n|}+\varepsilon. \tag{4.2.12}$$

Since ε does not depend on m, we can take the limit with respect to m. Given condition (4.2.5) we note that for sufficiently large n we have

$$\left|\frac{[z^n]f(z,\boldsymbol{u})}{C_nA(\boldsymbol{u})B(\boldsymbol{u})^n}-1\right|\leqslant\varepsilon. \tag{4.2.13}$$

Hence, uniformly in a fixed complex vicinity of $\boldsymbol{u}=\mathbf{1}$

$$\lim_{n\to\infty}\frac{[z^n]f(z,\boldsymbol{u})}{C_nA(\boldsymbol{u})B(\boldsymbol{u})^n}=1, \tag{4.2.14}$$

which finishes the proof. □

Remark 4.2.1. Using the same technique, higher-order error terms can also be transferred, provided that the function is squeezed between two sequences of formal power series with known Puiseux expansions. In such a situation, higher-order terms correspond to coefficients obtained from the summands of Puiseux expansion

$$f(z, \boldsymbol{u}) \sim \sum_{k \geqslant 0} c_k(\boldsymbol{u}) \left(1 - \frac{z}{\rho(\boldsymbol{u})}\right)^{k/2}. \tag{4.2.15}$$

The next lemma is a formal power series analogue of Lagrange's mean value theorem.

Lemma 4.2.2 (Mean value lemma for formal power series)**.** Let $f(z)$ and $g(z)$ be two formal power series such that $f(z) \preceq g(z)$. Assume that $\Psi(t) = \sum_{n \geqslant 0} \psi_n t^n$ is a formal power series with non-negative coefficients. Then,

$$\Psi(g(z)) - \Psi(f(z)) \preceq (g(z) - f(z))\, \Psi'(g(z)). \tag{4.2.16}$$

Likewise, the statement holds for vectors of formal power series in a coordinate-wise manner.

Proof. Coefficient-wise subtraction of the left-hand side of (4.2.16) yields

$$\begin{aligned} \Psi(g(z)) - \Psi(f(z)) &= \sum_{n \geqslant 0} \psi_n \left(g(z)^n - f(z)^n\right) \\ &= (g(z) - f(z)) \sum_{n \geqslant 0} \psi_n \left(\sum_{i=0}^{n-1} g(z)^i f(z)^{n-i-1}\right) \\ &\preceq (g(z) - f(z)) \sum_{n \geqslant 0} \psi_n n g(z)^{n-1} \\ &= (g(z) - f(z))\, \Psi'(g(z)). \end{aligned} \tag{4.2.17}$$

The first two equalities hold as a consequence of formal power series composition and the identity $a^n - b^n = (a-b)\sum_{i=0}^{n-1} a^i b^{n-i-1}$. The subsequent domination follows from the assumption that $f(z) \preceq g(z)$. □

4.3 Forward recursive systems

The following definition of *forward recursive systems* encapsulates the general, abstract features of the infinite systems that we consider in the current section. Core characteristics of the infinite system corresponding to closed λ-terms are abstracted and divided into three general conditions which are sufficient to access the asymptotic form of respective coefficients.

Definition 4.3.1 (Forward recursive systems)**.** Let z be a formal variable and $\boldsymbol{u} = (u_1, \ldots, u_r)$ be a vector of r formal variables. Consider infinite sequences $\left(\boldsymbol{L}^{\langle m \rangle}\right)_{m \geqslant 0}$ and $\left(\boldsymbol{\mathcal{K}}^{\langle m \rangle}\right)_{m \geqslant 0}$ of d-dimensional vectors

$$\boldsymbol{L}^{\langle m \rangle} = \left(L_1^{\langle m \rangle}, \ldots, L_d^{\langle m \rangle}\right) \quad \text{and} \quad \boldsymbol{\mathcal{K}}^{\langle m \rangle} = \left(K_1^{\langle m \rangle}, \ldots, K_d^{\langle m \rangle}\right) \tag{4.3.1}$$

consisting of formal power series $L_i^{\langle m \rangle}(z, \boldsymbol{u})$ and $K_i^{\langle m \rangle}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, z, \boldsymbol{u})$ where $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$ are vectors of d variables and $i = 1, \ldots, d$.

Assume that $\left(\boldsymbol{L}^{\langle m \rangle}\right)_{m \geqslant 0}$ and $\left(\boldsymbol{\mathcal{K}}^{\langle m \rangle}\right)_{m \geqslant 0}$ satisfy

$$\boldsymbol{L}^{\langle m \rangle} = \boldsymbol{\mathcal{K}}^{\langle m \rangle}\left(\boldsymbol{L}^{\langle m \rangle}, \boldsymbol{L}^{\langle m+1 \rangle}, z, \boldsymbol{u}\right). \tag{4.3.2}$$

Then, we say that the system (4.3.2) is *forward recursive*.

Furthermore, consider a *limiting system* in form of

$$\boldsymbol{L}^{\langle \infty \rangle} = \boldsymbol{\mathcal{K}}^{\langle \infty \rangle}\left(\boldsymbol{L}^{\langle \infty \rangle}, \boldsymbol{L}^{\langle \infty \rangle}, z, \boldsymbol{u}\right) \tag{4.3.3}$$

where $\boldsymbol{L}^{\langle\infty\rangle}$ and $\boldsymbol{\mathcal{K}}^{\langle\infty\rangle}$ are d-dimensional vectors of formal power series $L_i^{\langle\infty\rangle}(z, \boldsymbol{u})$ and $K_i^{\langle\infty\rangle}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, z, \boldsymbol{u})$, respectively, and moreover all series $K_i^{\langle\infty\rangle}$ are analytic at $(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, z, \boldsymbol{u}) = (\mathbf{0}, \mathbf{0}, 0, \mathbf{1})$. In this setting, we say that the system (4.3.2):

1. is *infinitely nested* if $\boldsymbol{\mathcal{K}}^{\langle m\rangle}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, z, \boldsymbol{u}) \preceq \boldsymbol{\mathcal{K}}^{\langle\infty\rangle}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, z, \boldsymbol{u})$ for each $m \geqslant 0$;

2. *tends to an irreducible context-free schema* if it is infinitely nested and its corresponding limiting system (4.3.3) satisfies the premises of the Drmota–Lalley–Woods theorem (see Proposition 4.1.1) i.e. is a polynomial, non-linear system of functional equations which is algebraic positive, proper, irreducible and aperiodic;

3. is *exponentially converging* if there exists a vector $\boldsymbol{A}(z, \boldsymbol{u}) = (A_1(z, \boldsymbol{u}), \dots, A_d(z, \boldsymbol{u}))$ and a function $B(z, \boldsymbol{u})$ such that:

   - for each $m \geqslant 0$ we have

   $$\boldsymbol{\mathcal{K}}^{\langle\infty\rangle}(\boldsymbol{L}^{\langle\infty\rangle}, \boldsymbol{L}^{\langle\infty\rangle}, z, \boldsymbol{u}) - \boldsymbol{\mathcal{K}}^{\langle m\rangle}(\boldsymbol{L}^{\langle\infty\rangle}, \boldsymbol{L}^{\langle\infty\rangle}, z, \boldsymbol{u}) \preceq \boldsymbol{A}(z, \boldsymbol{u}) \cdot B(z, \boldsymbol{u})^m; \tag{4.3.4}$$

   - $A_1(z, \boldsymbol{u}), \dots, A_d(z, \boldsymbol{u})$ and $B(z, \boldsymbol{u})$ are analytic functions in the disk $|z| < \rho + \varepsilon$ for some $\varepsilon > 0$ and $\boldsymbol{u} = \mathbf{1}$ where ρ is the dominant singularity of the limit system (4.3.3) at point $\boldsymbol{u} = \mathbf{1}$. Moreover, at $\boldsymbol{u} = \mathbf{1}$ we have $|B(\rho(\boldsymbol{u}), \boldsymbol{u})| < 1$ where $\rho(\boldsymbol{u})$ is the singularity of the limiting system (4.3.3).

Example 4.3.1. Consider the infinite system corresponding to m-open λ-terms, see (8.1.3). Recall that the sequence $(L_m(z))_{m\geqslant 0}$ of respective generating functions satisfies

$$\begin{aligned} L_0(z) &= zL_1(z) + zL_0(z)^2 \\ L_1(z) &= zL_2(z) + zL_1(z)^2 + z \\ &\dots \\ L_m(z) &= zL_{m+1}(z) + zL_m(z)^2 + z\frac{1-z^m}{1-z} \\ &\dots \end{aligned} \tag{4.3.5}$$

Let us show that (4.3.5) is an infinitely nested, forward recursive system which tends to an irreducible context-free schema of $L_\infty(z)$ at an exponential convergence rate. Here, each intermediate system $\boldsymbol{L}^{\langle m\rangle}$ consists of a single equation defining $L_m(z)$. Note that there are no additional marking variables $\boldsymbol{u}$. The vectors $\boldsymbol{\mathcal{K}}^{\langle m\rangle}$ are one-dimensional and the corresponding functions K_m are given by

$$K_m(\ell_1, \ell_2, z) := z\ell_2 + z\ell_1^2 + z\frac{1-z^m}{1-z}. \tag{4.3.6}$$

The limiting system $L_\infty(z)$ satisfies

$$L_\infty(z) = zL_\infty(z) + zL_\infty(z)^2 + \frac{z}{1-z}. \tag{4.3.7}$$

One can easily check that it also satisfies the premises of Proposition 4.1.1; hence, the considered system (4.3.5) tends to an irreducible context-free schema. Since the trivariate formal power series $K_\infty(\ell_1, \ell_2, z) - K_m(\ell_1, \ell_2, z)$ has non-negative coefficients, the system (4.3.5) is also infinitely nested. Moreover, the difference between the limiting equation and the mth equation computed at $\ell_1 = \ell_2 = L_\infty(z)$ is equal to $\frac{z^{m+1}}{1-z}$ and corresponds to a subset of de Bruijn indices. Certainly, as m tends to infinity, this difference converges to zero exponentially fast.

Given the combinatorial relation between m-open λ-terms and plain terms, we readily obtain the requested condition $L_m(z) \preceq L_\infty(z)$. However, for arbitrary forward recursive systems (8.3.2) establishing that $\boldsymbol{L}^{\langle m\rangle} \preceq \boldsymbol{L}^{\langle\infty\rangle}$ is no longer so straightforward. In what follows, we prove that for this inequality to hold it is sufficient that the limiting system is well-founded.

Lemma 4.3.1. Let $\mathcal{S}$ be an infinitely nested, forward recursive system (4.3.2). Assume that the coefficients of the formal power series $\boldsymbol{\mathcal{K}}^{\langle\infty\rangle}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, z, \boldsymbol{u})$ corresponding to the limiting system are non-negative and the limiting system (4.3.3) is well-founded (i.e. algebraic proper in the sense of Definition 4.1.1) and has a non-zero solution $\boldsymbol{L}^{\langle\infty\rangle}(z, \boldsymbol{u})$. Finally, assume that $\boldsymbol{\mathcal{K}}^{\langle\infty\rangle}(\mathbf{0}, \mathbf{0}, 0, \boldsymbol{u}) = \mathbf{0}$. Then,

$$\boldsymbol{L}^{\langle m\rangle}(z, \boldsymbol{u}) \preceq \boldsymbol{L}^{\langle\infty\rangle}(z, \boldsymbol{u}). \tag{4.3.8}$$

Proof. Consider the vectors $\boldsymbol{\mathcal{L}} = (\boldsymbol{L}^{\langle 0\rangle}, \boldsymbol{L}^{\langle 1\rangle}, \ldots)$ and $\boldsymbol{\mathcal{L}}^{+} = (\boldsymbol{L}^{\langle\infty\rangle}, \boldsymbol{L}^{\langle\infty\rangle}, \ldots)$ consisting of aptly concatenated and flattened systems $\left(\boldsymbol{L}^{\langle m\rangle}\right)_{m\geqslant 0}$ and $\boldsymbol{L}^{\langle\infty\rangle}$, respectively. Intuitively, $\boldsymbol{\mathcal{L}}$ and $\boldsymbol{\mathcal{L}}^{+}$ are in a sense *vectors of vectors*, but for convenience we call them just *vectors*. Note that both $\boldsymbol{\mathcal{L}}(z, \boldsymbol{u})$ and $\boldsymbol{\mathcal{L}}^{+}(z, \boldsymbol{u})$ satisfy

$$\begin{aligned} \boldsymbol{\mathcal{L}}(z, \boldsymbol{u}) &= \boldsymbol{\Phi}(\boldsymbol{\mathcal{L}}(z, \boldsymbol{u}), z, \boldsymbol{u}) \\ \boldsymbol{\mathcal{L}}^{+}(z, \boldsymbol{u}) &= \boldsymbol{\Psi}(\boldsymbol{\mathcal{L}}^{+}(z, \boldsymbol{u}), z, \boldsymbol{u}) \end{aligned} \tag{4.3.9}$$

where

$$\begin{aligned} \boldsymbol{\Phi}(\boldsymbol{\lambda}, z, \boldsymbol{u}) &= (\boldsymbol{\mathcal{K}}^{\langle 0\rangle}(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1, z, \boldsymbol{u}), \boldsymbol{\mathcal{K}}^{\langle 1\rangle}(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, z, \boldsymbol{u}), \ldots) \\ \boldsymbol{\Psi}(\boldsymbol{\lambda}, z, \boldsymbol{u}) &= (\boldsymbol{\mathcal{K}}^{\langle\infty\rangle}(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1, z, \boldsymbol{u}), \boldsymbol{\mathcal{K}}^{\langle\infty\rangle}(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, z, \boldsymbol{u}), \ldots) \end{aligned} \tag{4.3.10}$$

with $\boldsymbol{\lambda}$ taken as a flattening concatenation of d-dimensional vectors of free variables $(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1, \ldots)$.

Since for each m we have $\boldsymbol{\mathcal{K}}^{\langle m\rangle}(\boldsymbol{\lambda}_m, \boldsymbol{\lambda}_{m+1}, z, \boldsymbol{u}) \preceq \boldsymbol{\mathcal{K}}^{\langle\infty\rangle}(\boldsymbol{\lambda}_m, \boldsymbol{\lambda}_{m+1}, z, \boldsymbol{u})$ it also holds

$$\boldsymbol{\Phi}(\boldsymbol{\lambda}, z, \boldsymbol{u}) \preceq \boldsymbol{\Psi}(\boldsymbol{\lambda}, z, \boldsymbol{u}). \tag{4.3.11}$$

The idea of the current proof is to consider the difference $\boldsymbol{\mathcal{L}}^{+}(z, \boldsymbol{u}) - \boldsymbol{\mathcal{L}}(z, \boldsymbol{u})$ and show that it is non-negative. According to (4.3.9) this difference can be represented as

$$\boldsymbol{\mathcal{L}}^{+}(z, \boldsymbol{u}) - \boldsymbol{\mathcal{L}}(z, \boldsymbol{u}) = \boldsymbol{\Psi}(\boldsymbol{\mathcal{L}}^{+}(z, \boldsymbol{u}), z, \boldsymbol{u}) - \boldsymbol{\Phi}(\boldsymbol{\mathcal{L}}(z, \boldsymbol{u}), z, \boldsymbol{u}). \tag{4.3.12}$$

Since $\boldsymbol{\Psi}(\boldsymbol{\lambda}, z, \boldsymbol{u}) \succeq \boldsymbol{\Phi}(\boldsymbol{\lambda}, z, \boldsymbol{u})$, the formal power series $\boldsymbol{\Psi}$ can be represented as a sum $\boldsymbol{\Psi}(\boldsymbol{\lambda}, z, \boldsymbol{u}) = \boldsymbol{\Phi}(\boldsymbol{\lambda}, z, \boldsymbol{u}) + \boldsymbol{\Theta}(\boldsymbol{\lambda}, z, \boldsymbol{u})$ with $\boldsymbol{\Theta}(\boldsymbol{\lambda}, z, \boldsymbol{u}) \succeq \mathbf{0}$. Hence, the difference (4.3.12) becomes

$$\boldsymbol{\mathcal{L}}^{+}(z, \boldsymbol{u}) - \boldsymbol{\mathcal{L}}(z, \boldsymbol{u}) = \boldsymbol{\Theta}(\boldsymbol{\mathcal{L}}^{+}, z, \boldsymbol{u}) + \left(\boldsymbol{\Phi}(\boldsymbol{\mathcal{L}}^{+}, z, \boldsymbol{u}) - \boldsymbol{\Phi}(\boldsymbol{\mathcal{L}}, z, \boldsymbol{u})\right). \tag{4.3.13}$$

At this point, our tactic is to apply an analog of the mean value theorem to the right-hand side difference and obtain an equation of the form

$$\boldsymbol{\mathcal{L}}^{+}(z, \boldsymbol{u}) - \boldsymbol{\mathcal{L}}(z, \boldsymbol{u}) = \boldsymbol{\Theta}(\boldsymbol{\mathcal{L}}^{+}, z, \boldsymbol{u}) + \boldsymbol{\mathcal{J}} \cdot \left(\boldsymbol{\mathcal{L}}^{+} - \boldsymbol{\mathcal{L}}\right) \tag{4.3.14}$$

and consequently

$$\boldsymbol{\mathcal{L}}^{+}(z, \boldsymbol{u}) - \boldsymbol{\mathcal{L}}(z, \boldsymbol{u}) = (\mathbf{I} - \boldsymbol{\mathcal{J}})^{-1}\boldsymbol{\Theta} = \sum_{k\geqslant 0} \boldsymbol{\mathcal{J}}^{k}\boldsymbol{\Theta} \succeq \mathbf{0} \tag{4.3.15}$$

where $\boldsymbol{\mathcal{J}}$ is some non-negative operator whereas $\mathbf{I}$ is the corresponding identity. The rest of the proof is dedicated to formalising the above approach, in particular showing that the Neumann series $\sum_{k\geqslant 0} \boldsymbol{\mathcal{J}}^{k}$ is well-defined.

We start by noticing that due to the well-foundedness of the limiting system (4.3.3) we have $\boldsymbol{\mathcal{L}}^{+}(0, \boldsymbol{u}) = \mathbf{0}$. Since $\boldsymbol{\Phi}(\boldsymbol{\lambda}, z, \boldsymbol{u}) \preceq \boldsymbol{\Psi}(\boldsymbol{\lambda}, z, \boldsymbol{u})$ there also holds $\boldsymbol{\Phi}(\mathbf{0}, 0, \boldsymbol{u}) = \mathbf{0}$. Furthermore, since there exists a unique formal power series solution of the equation $\boldsymbol{\mathcal{L}}(0, \boldsymbol{u}) = \boldsymbol{\Phi}(\boldsymbol{\mathcal{L}}(0, \boldsymbol{u}), 0, \boldsymbol{u})$ and $\boldsymbol{\mathcal{L}}(0, \boldsymbol{u}) = \mathbf{0}$ satisfies this equation, we note that indeed $\boldsymbol{\mathcal{L}}(0, \boldsymbol{u}) = \mathbf{0}$.

Consider two formal infinite-dimensional variables $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^{+}$ which are both flattened concatenations of d-dimensional vectors of free variables. Let us show that the difference $\boldsymbol{\Phi}(\boldsymbol{\lambda}^{+}, z, \boldsymbol{u}) - \boldsymbol{\Phi}(\boldsymbol{\lambda}, z, \boldsymbol{u})$ can be represented as

$$\boldsymbol{\Phi}(\boldsymbol{\lambda}^{+}, z, \boldsymbol{u}) - \boldsymbol{\Phi}(\boldsymbol{\lambda}, z, \boldsymbol{u}) = \boldsymbol{\mathcal{J}}(z, \boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\lambda}^{+})(\boldsymbol{\lambda}^{+} - \boldsymbol{\lambda}) \tag{4.3.16}$$

where $\boldsymbol{\mathcal{J}} = \boldsymbol{\mathcal{J}}(z, \boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\lambda}^+)$ is some non-negative operator (i.e. infinite-dimensional matrix). Moreover, after substituting $\boldsymbol{\lambda} = \boldsymbol{\mathcal{L}}(z, \boldsymbol{u})$ and $\boldsymbol{\lambda}^+ = \boldsymbol{\mathcal{L}}^+(z, \boldsymbol{u})$ into $\boldsymbol{\mathcal{J}}$, there exists a non-negative integer $K > 0$ such that $\boldsymbol{\mathcal{J}}^K$ is element-wise divisible by z. Since we have established that both $\boldsymbol{\mathcal{L}}^+(0, \boldsymbol{u}) = \mathbf{0}$ and $\boldsymbol{\mathcal{L}}(0, \boldsymbol{u}) = \mathbf{0}$, the latter condition is equivalent to the nilpotency of the operator $\boldsymbol{\mathcal{J}}$ evaluated at $\boldsymbol{\lambda} = \boldsymbol{\mathcal{L}}(z, \boldsymbol{u})$, $\boldsymbol{\lambda}^+ = \boldsymbol{\mathcal{L}}^+(z, \boldsymbol{u})$ and $z = 0$.

Note that the function $\boldsymbol{\Phi}(\boldsymbol{\lambda}, z, \boldsymbol{u})$ is a sum of (finite) monomials in formal variables $(\boldsymbol{\lambda}, z, \boldsymbol{u})$; although $\boldsymbol{\lambda}$ is infinitely-dimensional, each of the monomials involves only finitely many factors of $\boldsymbol{\lambda}$. Let us consider the difference of arbitrary monomials in form of

$$x_1^{n_1} \cdots x_k^{n_k} - y_1^{n_1} \cdots y_k^{n_k}. \tag{4.3.17}$$

Note that we can rewrite (4.3.17) as

$$\begin{aligned} x_1^{n_1} \cdots x_k^{n_k} - y_1^{n_1} \cdots y_k^{n_k} &= (x_1^{n_1} \cdots x_k^{n_k} - y_1^{n_1} x_2^{n_2} \cdots x_k^{n_k}) \\ &\quad + (y_1^{n_1} x_2^{n_2} \cdots x_k^{n_k} - y_1^{n_1} y_2^{n_2} x_3^{n_3} \cdots x_k^{n_k}) + \cdots \\ &\quad + \left(y_1^{n_1} \cdots y_{k-1}^{n_{k-1}} x_k^{n_k} - y_1^{n_1} \cdots y_k^{n_k}\right) \\ &= (x_1 - y_1)\, x_2^{n_2} \cdots x_k^{n_k} \frac{(x_1^{n_1} - y_1^{n_1})}{x_1 - y_1} \\ &\quad + (x_2 - y_2)\, y_1^{n_1} x_3^{n_3} \cdots x_k^{n_k} \frac{(x_2^{n_2} - y_2^{n_2})}{x_2 - y_2} + \cdots \\ &\quad + (x_k - y_k)\, y_1^{n_1} \cdots y_{k-1}^{n_{k-1}} \frac{(x_k^{n_k} - y_k^{n_k})}{x_k - y_k}. \end{aligned} \tag{4.3.18}$$

Note that each factor $\dfrac{(x_i^{n_i} - y_i^{n_i})}{x_i - y_i}$ in the final sum is in fact a polynomial $\displaystyle\sum_{j=0}^{n_i - 1} x_i^j y_i^{n_i - j - 1}$. Therefore, the difference $x_1^{n_1} \cdots x_k^{n_k} - y_1^{n_1} \cdots y_k^{n_k}$ can be represented as a scalar product of $\boldsymbol{x} - \boldsymbol{y} := (x_i - y_i)_{i=1}^k$ and a vector of formal power series in $(\boldsymbol{x}, \boldsymbol{y})$. Furthermore, the difference $\boldsymbol{\Phi}(\boldsymbol{\lambda}^+, z, \boldsymbol{u}) - \boldsymbol{\Phi}(\boldsymbol{\lambda}, z, \boldsymbol{u})$ consists of the sums of such differences of monomials multiplied by appropriate non-negative coefficients. Grouping these differences together, we obtain the desired form (4.3.16).

Next, as an intermediate step, let us now show that the Jacobian operator $\dfrac{\partial \boldsymbol{\Psi}}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}, z, \boldsymbol{u})$ is nilpotent at $(z, \boldsymbol{\lambda}) = (0, \mathbf{0})$. For convenience, set

$$J_1(\boldsymbol{u}) := \left.\frac{\partial \boldsymbol{\mathcal{K}}^{\langle \infty \rangle}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, z, \boldsymbol{u})}{\partial \boldsymbol{\ell}_1}\right|_{\substack{z=0 \\ \boldsymbol{\ell}_1 = \mathbf{0} \\ \boldsymbol{\ell}_2 = \mathbf{0}}} \quad \text{and} \quad J_2(\boldsymbol{u}) := \left.\frac{\partial \boldsymbol{\mathcal{K}}^{\langle \infty \rangle}(\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, z, \boldsymbol{u})}{\partial \boldsymbol{\ell}_2}\right|_{\substack{z=0 \\ \boldsymbol{\ell}_1 = \mathbf{0} \\ \boldsymbol{\ell}_2 = \mathbf{0}}}. \tag{4.3.19}$$

Then,

$$J_1(\boldsymbol{u}) + J_2(\boldsymbol{u}) = \left.\frac{\partial \boldsymbol{\mathcal{K}}^{\langle \infty \rangle}(\boldsymbol{\ell}, \boldsymbol{\ell}, z, \boldsymbol{u})}{\partial \boldsymbol{\ell}}\right|_{\substack{z=0 \\ \boldsymbol{\ell} = \mathbf{0}}}. \tag{4.3.20}$$

Since the limiting system is well-founded (see Definition 4.1.1) the sum $J_1(\boldsymbol{u}) + J_2(\boldsymbol{u})$ is nilpotent. Moreover, since each of the matrices $J_1(\boldsymbol{u})$ and $J_2(\boldsymbol{u})$ is non-negative, there exists K such that all the summands of the expanded binomial $(J_1(\boldsymbol{u}) + J_2(\boldsymbol{u}))^K$ are zero.

On the other hand, note that following the definition of $\boldsymbol{\Psi}(\boldsymbol{\lambda}, z, \boldsymbol{u})$ its Jacobian operator $\dfrac{\partial \boldsymbol{\Psi}}{\partial \boldsymbol{\lambda}}$ evaluated at $(z, \boldsymbol{\lambda}) = (0, \mathbf{0})$ admits the following block structure:

$$\left.\frac{\partial \boldsymbol{\Psi}}{\partial \boldsymbol{\lambda}}\right|_{\substack{z=0 \\ \boldsymbol{\lambda} = \mathbf{0}}} = \left[\begin{array}{c|c|c|c} J_1 & J_2 & & \\ \hline & J_1 & J_2 & \\ \hline & & J_1 & J_2 \\ \hline & & & \ddots \end{array}\right]. \tag{4.3.21}$$

If we take the Kth power of this operator, it will have a block structure in which each block element consists of a sum of certain summands from the binomial expansion of $(J_1+J_2)^K$. Since the latter is a zero matrix and the summands corresponding to the blocks are non-negative and dominated by $(J_1+J_2)^K$, all such summands are also zero. This implies that the Jacobian operator $\dfrac{\partial \boldsymbol{\Psi}}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}, z, \boldsymbol{u})$ evaluated at $(z, \boldsymbol{\lambda}) = (0, \mathbf{0})$ is nilpotent with a nilpotence index at most K, i.e. the nilpotence index of the Jacobian operator $\dfrac{\partial \boldsymbol{\mathcal{K}}^{\langle\infty\rangle}(\boldsymbol{\ell}, \boldsymbol{\ell}, z, \boldsymbol{u})}{\partial \boldsymbol{\ell}}$ evaluated at $(z, \boldsymbol{\ell}) = (0, \mathbf{0})$.

Now, let us show that the infinitely-dimensional matrix $\boldsymbol{\mathcal{J}}(z, \boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\lambda}^+)$ evaluated at $(z, \boldsymbol{\lambda}, \boldsymbol{\lambda}^+) = (0, \mathbf{0}, \mathbf{0})$ is equal to the Jacobian operator $\dfrac{\partial \boldsymbol{\Phi}}{\partial \boldsymbol{\lambda}}$ evaluated at $(z, \boldsymbol{\lambda}) = (0, \mathbf{0})$. Recall that the operator $\boldsymbol{\mathcal{J}}$ is determined by the differences of monomials in $\boldsymbol{\Phi}(\boldsymbol{\lambda}^+, z, \boldsymbol{u}) - \boldsymbol{\Phi}(\boldsymbol{\lambda}, z, \boldsymbol{u})$. Monomials that have degree zero in $\boldsymbol{\lambda}$ or $\boldsymbol{\lambda}^+$ cancel out because they depend only on the arguments z and $\boldsymbol{u}$. Likewise, monomials with degree two or more in $\boldsymbol{\lambda}$ or $\boldsymbol{\lambda}^+$ vanish after the substitution $\boldsymbol{\lambda} = \boldsymbol{\lambda}^+ = \mathbf{0}$. The only type of the terms that do not turn to zero upon substitution $\boldsymbol{\lambda} = \boldsymbol{\lambda}^+ = \mathbf{0}$ are terms coming from differences of monomials linear in $\boldsymbol{\lambda}$ or $\boldsymbol{\lambda}^+$. Note that such terms have the same contribution to the infinitely-dimensional matrix $\boldsymbol{\mathcal{J}}$ as the corresponding terms of the infinitely-dimensional matrix $\dfrac{\partial \boldsymbol{\Phi}}{\partial \boldsymbol{\lambda}}$. Hence, $\boldsymbol{\mathcal{J}}(z, \boldsymbol{u}, \boldsymbol{\lambda}, \boldsymbol{\lambda}^+)$ evaluated at $(z, \boldsymbol{\lambda}, \boldsymbol{\lambda}^+) = (0, \mathbf{0}, \mathbf{0})$ is indeed equal to the Jacobian operator $\dfrac{\partial \boldsymbol{\Phi}}{\partial \boldsymbol{\lambda}}$ evaluated at $(z, \boldsymbol{\lambda}) = (0, \mathbf{0})$.

The nilpotence of $\boldsymbol{\mathcal{J}}$ evaluated at $(z, \boldsymbol{\lambda}, \boldsymbol{\lambda}^+) = (0, \mathbf{0}, \mathbf{0})$ follows from the fact that $\boldsymbol{\Psi}(\boldsymbol{\lambda}, z, \boldsymbol{u})$ is dominating $\boldsymbol{\Phi}(\boldsymbol{\lambda}, z, \boldsymbol{u})$, and therefore, the corresponding Jacobian operator $\dfrac{\partial \boldsymbol{\Psi}}{\partial \boldsymbol{\lambda}}$ (respectively its Nth power) is dominating the operator $\dfrac{\partial \boldsymbol{\Phi}}{\partial \boldsymbol{\lambda}}$ (respectively its Nth power). For this reason, the latter Jacobian operator $\dfrac{\partial \boldsymbol{\Phi}}{\partial \boldsymbol{\lambda}}$, evaluated at $(z, \boldsymbol{\lambda}, \boldsymbol{\lambda}^+) = (0, \mathbf{0}, \mathbf{0})$ is nilpotent with the nilpotence index at most the corresponding nilpotence operator of the former Jacobian operator $\dfrac{\partial \boldsymbol{\Psi}}{\partial \boldsymbol{\lambda}}$.

And so, we have established that $\boldsymbol{\mathcal{J}}$ evaluated at $\boldsymbol{\lambda} = \boldsymbol{\mathcal{L}}(z, \boldsymbol{u})$, $\boldsymbol{\lambda}^+ = \boldsymbol{\mathcal{L}}^+(z, \boldsymbol{u})$ and $z = 0$ is nilpotent. Equivalently, it meas that after substituting $\boldsymbol{\lambda} = \boldsymbol{\mathcal{L}}(z, \boldsymbol{u})$ and $\boldsymbol{\lambda}^+ = \boldsymbol{\mathcal{L}}^+(z, \boldsymbol{u})$ into $\boldsymbol{\mathcal{J}}$ there exists a non-negative integer $K > 0$ such that $\boldsymbol{\mathcal{J}}^K$ is element-wise divisible by z. Consequently, each coefficient in z of the formal sum $\sum_{j \geqslant 0} \boldsymbol{\mathcal{J}}^j$ is finite. Indeed, for each integer $N \geqslant 0$, the coefficient at z^N in this formal series is a sum of coefficients at z^N in the finite sum $\sum_{j=0}^{K \cdot N} \boldsymbol{\mathcal{J}}^j$. Moreover, since $\boldsymbol{\mathcal{J}}$ is non-negative, this sum is also non-negative. Finally, this infinite formal series is equal to $(\boldsymbol{I} - \boldsymbol{\mathcal{J}})^{-1}$ where $\boldsymbol{I}$ is the identity operator of appropriate dimension. □

Remark 4.3.1. The condition that $\boldsymbol{\mathcal{K}}^{\langle\infty\rangle}(\mathbf{0}, \mathbf{0}, 0, \boldsymbol{u}) = \mathbf{0}$ can be omitted but we keep it for the simplicity of the proof. For the above proof, it is enough to guarantee that each coefficient in z of the infinite formal sum $\sum_{j \geqslant 0} \boldsymbol{\mathcal{J}}^j$ is finite, which is equivalent to saying that some power of $\boldsymbol{\mathcal{J}}$ is divisible by z. More details on well-founded systems can be found in [PSS12].

4.4 Coefficient transfer for infinite systems

Finally, we give our main theorem on the transfer of coefficients for infinitely nested forward-recursive systems.

Theorem 4.4.1. Let $\mathcal{S}$ be an infinitely nested, forward recursive system (4.3.2) which tends to an irreducible context-free schema at an exponential convergence rate. Then, the respective solutions $L_j^{\langle m \rangle}(z, \boldsymbol{u})$ of $\mathcal{S}$ admit for each $m \geqslant 0$ an asymptotic expansion of their coefficients as $n \to \infty$ in form of

$$[z^n] L_j^{\langle m \rangle}(z, \boldsymbol{u}) \sim [z^n] \sum_{k \geqslant 0} c_{j,k}^{\langle m \rangle}(\boldsymbol{u}) \left(1 - \frac{z}{\rho(\boldsymbol{u})}\right)^{k/2} \tag{4.4.1}$$

where $\rho(\boldsymbol{u})$ is the dominant singularity of the corresponding limiting system (4.3.3) and the coefficients $c_{j,k}^{\langle m\rangle}(\boldsymbol{u})$ are analytic at $\boldsymbol{u}=\mathbf{1}$. Furthermore, $\rho(\boldsymbol{u})$ is analytic near $\boldsymbol{u}=\mathbf{1}$.

The coefficients $c_{j,k}^{\langle m\rangle}(\boldsymbol{u})$ can be approximated by taking first $(M-1)$ equations from (4.3.2) and replacing the Mth equation by its following limit variant:

$$\boldsymbol{L}^{\langle M\rangle} = \boldsymbol{\mathcal{K}}^{\langle\infty\rangle}\left(\boldsymbol{L}^{\langle M\rangle}, \boldsymbol{L}^{\langle M\rangle}, z, \boldsymbol{u}\right). \tag{4.4.2}$$

Such a truncated system can be solved recursively. Consequently, the coefficients of respective Puiseux expansions (4.4.1) are estimated with an error which is exponentially small in M.

Remark 4.4.1.

- The condition that the system $\mathcal{S}$ tends to an irreducible context-free schema can be replaced by a weaker condition asserting that the limiting system (4.3.3) admits a suitable Puiseux expansion.
- Instead of limiting systems with square-root type singularities, it is also possible to consider rational systems or other types of systems. The same set of conditions is sufficient to establish the transfer of behaviours around the dominant singular point.

Proof. (Theorem 4.4.1) We divide the proof into three conceptual parts.

1. First, we show that each component of the difference vector $\boldsymbol{L}^{\langle\infty\rangle}(z,\boldsymbol{u}) - \boldsymbol{L}^{\langle m\rangle}(z,\boldsymbol{u})$ can be upper bounded by a Puiseux series expansion whose coefficients decay exponentially fast as m tends to infinity;
2. Next, we show that if for $m \geqslant 1$ there exist coordinate-wise upper and lower bounds
$$\underline{\boldsymbol{L}^{\langle m\rangle}}(z,\boldsymbol{u}) \preceq \boldsymbol{L}^{\langle m\rangle}(z,\boldsymbol{u}) \preceq \overline{\boldsymbol{L}^{\langle m\rangle}}(z,\boldsymbol{u}), \tag{4.4.3}$$
then the vector of functions $\boldsymbol{L}^{\langle m-1\rangle}(z,\boldsymbol{u})$ obtained from the infinite system $\mathcal{S}$ admits upper and lower bounds $\overline{\boldsymbol{L}^{\langle m-1\rangle}}(z,\boldsymbol{u})$ and $\underline{\boldsymbol{L}^{\langle m-1\rangle}}(z,\boldsymbol{u})$ satisfying
$$\overline{\boldsymbol{L}^{\langle m-1\rangle}}(z,\boldsymbol{u}) - \underline{\boldsymbol{L}^{\langle m-1\rangle}}(z,\boldsymbol{u}) \preceq \boldsymbol{\mathcal{R}}(z,\boldsymbol{u})\left(\overline{\boldsymbol{L}^{\langle m\rangle}}(z,\boldsymbol{u}) - \underline{\boldsymbol{L}^{\langle m\rangle}}(z,\boldsymbol{u})\right) \tag{4.4.4}$$
for some matrix $\boldsymbol{\mathcal{R}}(z,\boldsymbol{u})$ with spectral radius satisfying $r(\boldsymbol{\mathcal{R}}(z,\boldsymbol{u})) \leqslant 1$ for $z \in [0,\rho(\boldsymbol{u})]$ where $\rho(\boldsymbol{u})$ is the dominant singularity of $\boldsymbol{L}^{\langle\infty\rangle}(z,\boldsymbol{u})$;
3. Finally, we combine two previous results and prove that the difference between the Puiseux coefficients of upper and lower bounds of $\boldsymbol{L}^{\langle m\rangle}(z,\boldsymbol{u})$ can be reduced to zero.

First part. According to Lemma 4.3.1 we have $\boldsymbol{L}^{\langle m\rangle} \preceq \boldsymbol{L}^{\langle\infty\rangle}$. Following the functional definitions of $\boldsymbol{L}^{\langle\infty\rangle}$ and $\boldsymbol{L}^{\langle m\rangle}$ from the infinite system of equations, their difference can be represented as

$$\boldsymbol{L}^{\langle\infty\rangle} - \boldsymbol{L}^{\langle m\rangle} = \boldsymbol{\mathcal{K}}^{\langle\infty\rangle}(\boldsymbol{L}^{\langle\infty\rangle}, \boldsymbol{L}^{\langle\infty\rangle}, z, \boldsymbol{u}) - \boldsymbol{\mathcal{K}}^{\langle m\rangle}(\boldsymbol{L}^{\langle m\rangle}, \boldsymbol{L}^{\langle m+1\rangle}, z, \boldsymbol{u}). \tag{4.4.5}$$

For convenience, henceforth we omit the arguments z and $\boldsymbol{u}$. Moreover, $\boldsymbol{\mathcal{K}}^{\langle\infty\rangle}$ and $\boldsymbol{\mathcal{K}}^{\langle m\rangle}$ become functions of two vector arguments

$$\boldsymbol{\mathcal{K}}^{\langle\infty\rangle}\colon \mathbb{C}^d \times \mathbb{C}^d \to \mathbb{C}^d \quad \text{and} \quad \boldsymbol{\mathcal{K}}^{\langle m\rangle}\colon \mathbb{C}^d \times \mathbb{C}^d \to \mathbb{C}^d. \tag{4.4.6}$$

In addition, we use the nabla notation to denote the Jacobian operator

$$\nabla_{\boldsymbol{x}}\boldsymbol{\mathcal{K}}(\boldsymbol{x},\boldsymbol{y}) = \left(\frac{\partial}{\partial x_1}\boldsymbol{\mathcal{K}}, \ldots, \frac{\partial}{\partial x_d}\boldsymbol{\mathcal{K}}\right) \quad \text{and} \quad \nabla_{\boldsymbol{y}}\boldsymbol{\mathcal{K}}(\boldsymbol{x},\boldsymbol{y}) = \left(\frac{\partial}{\partial y_1}\boldsymbol{\mathcal{K}}, \ldots, \frac{\partial}{\partial y_d}\boldsymbol{\mathcal{K}}\right). \tag{4.4.7}$$

We start with the following subtraction-addition trick. Each component of the vector difference is evaluated through Lemma 4.2.2 (mean value lemma for formal power series) and then upper-bounded by the value of the functional $\boldsymbol{\mathcal{K}}^{\langle\infty\rangle}$ at $\boldsymbol{L}^{\langle\infty\rangle}$. Specifically,

$$\begin{aligned}\boldsymbol{L}^{\langle\infty\rangle} - \boldsymbol{L}^{\langle m\rangle} &= \boldsymbol{\mathcal{K}}^{\langle\infty\rangle}(\boldsymbol{L}^{\langle\infty\rangle}, \boldsymbol{L}^{\langle\infty\rangle}) - \boldsymbol{\mathcal{K}}^{\langle\infty\rangle}(\boldsymbol{L}^{\langle m\rangle}, \boldsymbol{L}^{\langle\infty\rangle}) + \boldsymbol{\mathcal{K}}^{\langle\infty\rangle}(\boldsymbol{L}^{\langle m\rangle}, \boldsymbol{L}^{\langle\infty\rangle}) \\ &\quad - \boldsymbol{\mathcal{K}}^{\langle\infty\rangle}(\boldsymbol{L}^{\langle m\rangle}, \boldsymbol{L}^{\langle m+1\rangle}) + \boldsymbol{\mathcal{K}}^{\langle\infty\rangle}(\boldsymbol{L}^{\langle m\rangle}, \boldsymbol{L}^{\langle m+1\rangle}) - \boldsymbol{\mathcal{K}}^{\langle m\rangle}(\boldsymbol{L}^{\langle m\rangle}, \boldsymbol{L}^{\langle m+1\rangle}) \\ &\preceq \nabla_{\boldsymbol{x}}\boldsymbol{\mathcal{K}}^{\langle\infty\rangle}(\boldsymbol{x},\boldsymbol{y})\Big|_{\substack{\boldsymbol{x}=\boldsymbol{L}^{\langle\infty\rangle}\\ \boldsymbol{y}=\boldsymbol{L}^{\langle\infty\rangle}}}(\boldsymbol{L}^{\langle\infty\rangle} - \boldsymbol{L}^{\langle m\rangle}) + \nabla_{\boldsymbol{y}}\boldsymbol{\mathcal{K}}^{\langle\infty\rangle}(\boldsymbol{x},\boldsymbol{y})\Big|_{\substack{\boldsymbol{x}=\boldsymbol{L}^{\langle\infty\rangle}\\ \boldsymbol{y}=\boldsymbol{L}^{\langle\infty\rangle}}}(\boldsymbol{L}^{\langle\infty\rangle} - \boldsymbol{L}^{\langle m+1\rangle}) \\ &\quad + \Big(\boldsymbol{\mathcal{K}}^{\langle\infty\rangle}(\boldsymbol{L}^{\langle m\rangle}, \boldsymbol{L}^{\langle m+1\rangle}) - \boldsymbol{\mathcal{K}}^{\langle m\rangle}(\boldsymbol{L}^{\langle m\rangle}, \boldsymbol{L}^{\langle m+1\rangle})\Big).\end{aligned}$$

Since $\boldsymbol{\mathcal{K}}^{\langle m\rangle} \preceq \boldsymbol{\mathcal{K}}^{\langle\infty\rangle}$, the final difference in the above sum is a vector of formal power series with non-negative coefficients. Consequently, the last summand can be bounded by $\boldsymbol{\mathcal{K}}^{\langle\infty\rangle}(\boldsymbol{L}^{\langle\infty\rangle}, \boldsymbol{L}^{\langle\infty\rangle}) - \boldsymbol{\mathcal{K}}^{\langle m\rangle}(\boldsymbol{L}^{\langle\infty\rangle}, \boldsymbol{L}^{\langle\infty\rangle})$. By the condition of exponential convergence, this difference can be even further bounded by a vector of exponentially decaying functions $\boldsymbol{A}(z,\boldsymbol{u})(B(z,\boldsymbol{u}))^m$. For brevity, set

$$\partial_1\boldsymbol{\mathcal{K}} := \nabla_{\boldsymbol{x}}\boldsymbol{\mathcal{K}}^{\langle\infty\rangle}(\boldsymbol{x},\boldsymbol{y})\Big|_{(\boldsymbol{x},\boldsymbol{y})=(\boldsymbol{L}^{\langle\infty\rangle},\boldsymbol{L}^{\langle\infty\rangle})} \quad \text{and} \quad \partial_2\boldsymbol{\mathcal{K}} := \nabla_{\boldsymbol{y}}\boldsymbol{\mathcal{K}}^{\langle\infty\rangle}(\boldsymbol{x},\boldsymbol{y})\Big|_{(\boldsymbol{x},\boldsymbol{y})=(\boldsymbol{L}^{\langle\infty\rangle},\boldsymbol{L}^{\langle\infty\rangle})}.$$

Note that $\partial_1\boldsymbol{\mathcal{K}}, \partial_2\boldsymbol{\mathcal{K}} \in \mathbb{C}^{d\times d}$. Now, the upper bound on $\boldsymbol{L}^{\langle\infty\rangle} - \boldsymbol{L}^{\langle m\rangle}$ can be stated as

$$(\boldsymbol{L}^{\langle\infty\rangle} - \boldsymbol{L}^{\langle m\rangle}) \preceq \boldsymbol{A}(z,\boldsymbol{u})B(z,\boldsymbol{u})^m + \partial_1\boldsymbol{\mathcal{K}}\cdot(\boldsymbol{L}^{\langle\infty\rangle} - \boldsymbol{L}^{\langle m\rangle}) + \partial_2\boldsymbol{\mathcal{K}}\cdot(\boldsymbol{L}^{\langle\infty\rangle} - \boldsymbol{L}^{\langle m+1\rangle}) \tag{4.4.8}$$

or, equivalently,

$$(\boldsymbol{I}_d - \partial_1\boldsymbol{\mathcal{K}})(\boldsymbol{L}^{\langle\infty\rangle} - \boldsymbol{L}^{\langle m\rangle}) \preceq \boldsymbol{A}(z,\boldsymbol{u})B(z,\boldsymbol{u})^m + \partial_2\boldsymbol{\mathcal{K}}\cdot(\boldsymbol{L}^{\langle\infty\rangle} - \boldsymbol{L}^{\langle m+1\rangle}) \tag{4.4.9}$$

where $\boldsymbol{I}_d \in \mathbb{C}^{d\times d}$ is the $d\times d$ identity matrix.

Let us show that the inverse matrix $(\boldsymbol{I}_d - \partial_1\boldsymbol{\mathcal{K}})^{-1}$ exists and has non-negative coefficients in the sense of formal power series. As discussed in the proof of Lemma 4.3.1, the matrix $\partial_1\boldsymbol{\mathcal{K}}$ is nilpotent at $z=0$. Equivalently, there exists a non-negative integer K such that $(\partial_1\boldsymbol{\mathcal{K}})^K$ is divisible by z. It means that the formal series $\sum_{j\geqslant 0}(\partial_1\boldsymbol{\mathcal{K}})^j$ is well-defined and hence so is the formal inverse $(\boldsymbol{I}_d - \partial_1\boldsymbol{\mathcal{K}})^{-1}$. Moreover, since $\partial_1\boldsymbol{\mathcal{K}}$ has non-negative coefficients, the same holds for the investigated inverse matrix $(\boldsymbol{I}_d - \partial_1\boldsymbol{\mathcal{K}})^{-1}$.

Now, let us focus on the behaviour of $(\boldsymbol{I}_d - \partial_1\boldsymbol{\mathcal{K}})^{-1}$ as a function near $z=\rho(\boldsymbol{u})$. As follows from Proposition 4.1.2 applied to the system of equations $\boldsymbol{L}^{\langle\infty\rangle} = \boldsymbol{\mathcal{K}}^{\langle\infty\rangle}(\boldsymbol{L}^{\langle\infty\rangle}, \boldsymbol{L}^{\langle\infty\rangle}, z, \boldsymbol{u})$, for each real $0 < z < \rho(\boldsymbol{u})$ we have the following inequality:

$$r(\partial_1\boldsymbol{\mathcal{K}} + \partial_2\boldsymbol{\mathcal{K}}) < 1. \tag{4.4.10}$$

By Perron–Frobenius theorem (see e.g. [Drm09, section 2.2.5] and references therein) the spectral radius of a matrix with positive entries is monotonic in its coefficients. Hence, for all real $0 < z < \rho(\boldsymbol{u})$

$$r(\partial_1\boldsymbol{\mathcal{K}}) < 1 \tag{4.4.11}$$

and so, in the same interval

$$(\boldsymbol{I}_d - \partial_1\boldsymbol{\mathcal{K}})^{-1} = \sum_{j\geqslant 0}(\partial_1\boldsymbol{\mathcal{K}})^j. \tag{4.4.12}$$

Moreover, due to the continuity of the spectral radius, the same identity can be extended to some complex neighbourhood of $\boldsymbol{u}=\boldsymbol{1}$.

Consequently, we can multiply both sides of (4.4.9) by $(\boldsymbol{I}_d - \partial_1\boldsymbol{\mathcal{K}})^{-1}$ and obtain

$$\boldsymbol{L}^{\langle\infty\rangle} - \boldsymbol{L}^{\langle m\rangle} \preceq (\boldsymbol{I}_d - \partial_1\boldsymbol{\mathcal{K}})^{-1}\boldsymbol{A}(z,u)B(z,\boldsymbol{u})^m + (\boldsymbol{I}_d - \partial_1\boldsymbol{\mathcal{K}})^{-1}\partial_2\boldsymbol{\mathcal{K}}(\boldsymbol{L}^{\langle\infty\rangle} - \boldsymbol{L}^{\langle m+1\rangle}) \tag{4.4.13}$$

Let us denote $\boldsymbol{\delta}_m := (\boldsymbol{I}_d - \partial_1\boldsymbol{\mathcal{K}})^{-1}\boldsymbol{A}(z,\boldsymbol{u})B(z,\boldsymbol{u})^m$ and $\boldsymbol{\mathcal{R}} := (\boldsymbol{I}_d - \partial_1\boldsymbol{\mathcal{K}})^{-1}\partial_2\boldsymbol{\mathcal{K}}$. Note that the inequality $\boldsymbol{L}^{\langle\infty\rangle} - \boldsymbol{L}^{\langle m\rangle} \preceq \boldsymbol{\delta}_m + \boldsymbol{\mathcal{R}}(\boldsymbol{L}^{\langle\infty\rangle} - \boldsymbol{L}^{\langle m+1\rangle})$ can be further iterated for increasing values of m. In doing so, we find that

$$\boldsymbol{L}^{\langle\infty\rangle} - \boldsymbol{L}^{\langle m\rangle} \preceq \boldsymbol{\delta}_m + \boldsymbol{\mathcal{R}}\boldsymbol{\delta}_{m+1} + \boldsymbol{\mathcal{R}}^2\boldsymbol{\delta}_{m+2} + \cdots \tag{4.4.14}$$

Hence, the difference $\boldsymbol{L}^{\langle\infty\rangle} - \boldsymbol{L}^{\langle m\rangle}$ can be bounded by the tail of a geometric progression which appears as a summation of the formal Neumann series

$$\boldsymbol{L}^{\langle\infty\rangle} - \boldsymbol{L}^{\langle m\rangle} \preceq B(z,u)^m \sum_{k\geqslant 0} (B(z,\boldsymbol{u})\boldsymbol{\mathcal{R}})^k (\boldsymbol{I}_d - \partial_1\boldsymbol{\mathcal{K}})^{-1}\boldsymbol{A}(z,\boldsymbol{u}). \tag{4.4.15}$$

Let us now focus on the above formal Neumann series. Note that applying Proposition 4.1.1 (Drmota–Lalley–Woods theorem) to the limiting system (4.3.3), we obtain that the vector of functions $\boldsymbol{L}^{\langle\infty\rangle}(z,\boldsymbol{u})$ admits a coordinate-wise Puiseux expansion in form of

$$\boldsymbol{L}^{\langle\infty\rangle}(z,\boldsymbol{u}) \sim \boldsymbol{\ell}_0(\boldsymbol{u}) - \boldsymbol{\ell}_1(\boldsymbol{u})\sqrt{1-\frac{z}{\rho(\boldsymbol{u})}} \tag{4.4.16}$$

where functions $\boldsymbol{\ell}_0(\boldsymbol{u}), \boldsymbol{\ell}_1(\boldsymbol{u}), \rho(\boldsymbol{u})$ are analytic near $\boldsymbol{u} = 1$. Likewise, both matrices $\partial_1\boldsymbol{\mathcal{K}}$ and $\partial_2\boldsymbol{\mathcal{K}}$ admit Puiseux expansions of the same kind.

Let us prove that coordinate-wise Puiseux expansions of the matrix $\boldsymbol{\mathcal{R}}$ near the singular point $z = \rho(\boldsymbol{u})$ have the form

$$\boldsymbol{\mathcal{R}} = (\boldsymbol{I}_d - \partial_1\boldsymbol{\mathcal{K}})^{-1}\partial_2\boldsymbol{\mathcal{K}} \sim \boldsymbol{\mathcal{R}}_0 - \boldsymbol{\mathcal{R}}_1\sqrt{1-\frac{z}{\rho(\boldsymbol{u})}}, \quad z \to \rho(\boldsymbol{u}) \tag{4.4.17}$$

where the spectral radius of $\boldsymbol{\mathcal{R}}_0$ satisfies $r(\boldsymbol{\mathcal{R}}_0) = 1$. According to Perron–Frobenius theorem, since the coefficients of $\boldsymbol{\mathcal{R}}$ are non-negative, the eigenvalue of the matrix $\boldsymbol{\mathcal{R}}$ with the largest absolute value, i.e. the eigenvalue corresponding to the spectral radius of $\boldsymbol{\mathcal{R}}$, is the largest real positive solution λ of the characteristic equation

$$\det\left((\boldsymbol{I}_d - \partial_1\boldsymbol{\mathcal{K}})^{-1}\partial_2\boldsymbol{\mathcal{K}} - \lambda\boldsymbol{I}_d\right) = 0. \tag{4.4.18}$$

Since the determinant of a matrix product is equal to the product of respective determinants and $\det(\boldsymbol{I}_d - \partial_1\boldsymbol{\mathcal{K}}) \neq 0$, as $\boldsymbol{I}_d - \partial_1\boldsymbol{\mathcal{K}}$ is invertible, the above condition is equivalent to

$$\det(\partial_2\boldsymbol{\mathcal{K}} + \lambda\partial_1\boldsymbol{\mathcal{K}} - \lambda\boldsymbol{I}_d) = 0 \quad \text{and also} \quad \det(\partial_1\boldsymbol{\mathcal{K}} + \lambda^{-1}\partial_2\boldsymbol{\mathcal{K}} - \boldsymbol{I}_d) = 0. \tag{4.4.19}$$

Let us show that the largest positive real solution (as a function of $z < \rho(\boldsymbol{u})$) of this equation, does not exceed 1, with equality when $z = \rho(\boldsymbol{u})$. Assume, by contrary, that $\lambda > 1$. The matrix $(\partial_1\boldsymbol{\mathcal{K}} + \lambda^{-1}\partial_2\boldsymbol{\mathcal{K}})$ is a matrix with non-negative coefficients, whose coefficients are strictly smaller than the coefficients of the matrix $(\partial_1\boldsymbol{\mathcal{K}} + \partial_2\boldsymbol{\mathcal{K}})$. By Perron–Frobenius theorem (see e.g. [Drm09, section 2.2.5] and references therein), the spectral radius of a matrix with positive coefficients is monotonic in its coefficients, so for $\lambda > 1$

$$r(\partial_1\boldsymbol{\mathcal{K}} + \lambda^{-1}\partial_2\boldsymbol{\mathcal{K}}) < r(\partial_1\boldsymbol{\mathcal{K}} + \partial_2\boldsymbol{\mathcal{K}}) = 1. \tag{4.4.20}$$

Therefore, the characteristic equation cannot have a solution $\lambda > 1$. Moreover, the spectral radius of $\boldsymbol{\mathcal{R}}_0$ is equal to the spectral radius of $\boldsymbol{\mathcal{R}}$ when $z = \rho(\boldsymbol{u})$ because in this case, the two matrices coincide. Therefore, $r(\boldsymbol{\mathcal{R}}_0) = 1$.

Moving back to the upper bound (4.4.15), according to the exponential convergence condition in Definition 4.3.1, in a complex vicinity of $\boldsymbol{u} = \mathbf{1}$, the absolute value of the function $B(z,\boldsymbol{u})$ is strictly smaller than 1, hence the inverse matrix $(\boldsymbol{I}_d - B(z,u)\boldsymbol{\mathcal{R}})^{-1}$ in (4.4.15) exists. Moreover, since

$$A^{-1} = \frac{1}{\det(A)} \cdot \mathbf{adj}(A) \tag{4.4.21}$$

where $\mathbf{adj}(A)$ is the adjugate matrix of A, each element of the inverse matrix $(\boldsymbol{I}_d - B(z,u)\boldsymbol{\mathcal{R}})^{-1}$ can be represented as a ratio of a sum of products of functions admiting Puiseux series expansions in form of $\boldsymbol{a}(\boldsymbol{u}) - \boldsymbol{b}(\boldsymbol{u})\sqrt{1-z/\rho(\boldsymbol{u})} + O(|1-z/\rho(\boldsymbol{u})|)$ and a non-zero determinant of $\boldsymbol{I}_d - B(z,u)\boldsymbol{\mathcal{R}}$. It means that each coordinate in the inverse matrix $(\boldsymbol{I}_d - B(z,u)\boldsymbol{\mathcal{R}})^{-1}$ also admits a Puiseix series expansion of similar form. Thus, the Neumann series in (4.4.15) converges and we obtain

$$\boldsymbol{L}^{\langle\infty\rangle} - \boldsymbol{L}^{\langle m\rangle} \preceq B(z,\boldsymbol{u})^m \times (\boldsymbol{I}_d - B(z,\boldsymbol{u})\boldsymbol{\mathcal{R}})^{-1}(\boldsymbol{I}_d - \partial_1\boldsymbol{\mathcal{K}})^{-1}\boldsymbol{A}(z,\boldsymbol{u}). \tag{4.4.22}$$

From (4.4.22) we now note that the difference $\boldsymbol{L}^{\langle\infty\rangle} - \boldsymbol{L}^{\langle m\rangle}$ can be bounded by a vector of functions having the same singularity $\rho(\boldsymbol{u})$ as the components of the vector $\boldsymbol{L}^{\langle\infty\rangle}$. The Puiseux coefficients of this upper bound decay exponentially fast as $m \to \infty$. Moreover, these coefficients are analytic functions near $\boldsymbol{u} = \mathbf{1}$.

Second part. Assume that function $\boldsymbol{L}^{\langle m\rangle}(z, \boldsymbol{u})$ admits some upper and lower bounds and denote by $\Delta_m(z, \boldsymbol{u})$ the difference between these bounds:

$$\underline{\boldsymbol{L}^{\langle m\rangle}}(z, \boldsymbol{u}) \preceq \boldsymbol{L}^{\langle m\rangle}(z, \boldsymbol{u}) \preceq \overline{\boldsymbol{L}^{\langle m\rangle}}(z, \boldsymbol{u}), \quad \Delta^{\langle m\rangle}(z, \boldsymbol{u}) := \overline{\boldsymbol{L}^{\langle m\rangle}}(z, \boldsymbol{u}) - \underline{\boldsymbol{L}^{\langle m\rangle}}(z, \boldsymbol{u}). \tag{4.4.23}$$

Then, another pair of upper and lower bound can be established for $\boldsymbol{L}^{\langle m-1\rangle}$ from (4.3.2) (infinite system of equations) with the difference $\Delta^{\langle m-1\rangle}$ satisfying

$$\Delta^{\langle m-1\rangle} = \boldsymbol{\mathcal{K}}^{\langle m-1\rangle}(\overline{\boldsymbol{L}^{\langle m-1\rangle}}, \overline{\boldsymbol{L}^{\langle m\rangle}}) - \boldsymbol{\mathcal{K}}^{\langle m-1\rangle}(\underline{\boldsymbol{L}^{\langle m-1\rangle}}, \underline{\boldsymbol{L}^{\langle m\rangle}}) \preceq \partial_1\boldsymbol{\mathcal{K}} \cdot \Delta^{\langle m-1\rangle} + \partial_2\boldsymbol{\mathcal{K}} \cdot \Delta^{\langle m\rangle}. \tag{4.4.24}$$

That is, repeating the argument that allows to multiply both sides of the inequality by the inverse matrix, we obtain

$$\Delta^{\langle m-1\rangle} \preceq (\boldsymbol{I}_d - \partial_1\boldsymbol{\mathcal{K}})^{-1}\partial_2\boldsymbol{\mathcal{K}} \cdot \Delta^{\langle m\rangle}. \tag{4.4.25}$$

As we discovered in the first part, the matrix $\boldsymbol{\mathcal{R}} := (\boldsymbol{I}_d - \partial_1\boldsymbol{\mathcal{K}})^{-1}\partial_2\boldsymbol{\mathcal{K}}$ has spectral radius 1 at the singular point $z = \rho(\boldsymbol{u})$.

Third part. As a result of the first part, we know that $\boldsymbol{L}^{\langle\infty\rangle}(z, \boldsymbol{u}) - \boldsymbol{L}^{\langle m\rangle}(z, \boldsymbol{u})$ can be bounded in the following manner:

$$\mathbf{0} \preceq \boldsymbol{L}^{\langle\infty\rangle}(z, \boldsymbol{u}) - \boldsymbol{L}^{\langle m\rangle}(z, \boldsymbol{u}) \preceq B(z, \boldsymbol{u})^m(\boldsymbol{I}_d - B(z, u)\boldsymbol{\mathcal{R}})^{-1}\boldsymbol{A}(z, \boldsymbol{u}). \tag{4.4.26}$$

Let us assign

$$\Delta_0^{\langle m\rangle} := B(z, \boldsymbol{u})^m(\boldsymbol{I}_d - B(z, u)\boldsymbol{\mathcal{R}})^{-1}\boldsymbol{A}(z, \boldsymbol{u}) \tag{4.4.27}$$

for the difference between upper and lower bounds for the vector of functions $\boldsymbol{L}^{\langle m\rangle}(z, \boldsymbol{u})$. Next, using the result of the second part, we construct a family of differences $\Delta_k^{\langle m\rangle}$ between upper and lower bounds for $\boldsymbol{L}^{\langle m\rangle}(z, \boldsymbol{u})$, so that for every $k, m \geqslant 0$ it holds

$$\underline{\boldsymbol{L}_k^{\langle m\rangle}}(z, \boldsymbol{u}) \preceq \boldsymbol{L}^{\langle m\rangle}(z, \boldsymbol{u}) \preceq \overline{\boldsymbol{L}_k^{\langle m\rangle}}(z, \boldsymbol{u}) \quad \text{and} \quad \Delta_k^{\langle m\rangle} := \overline{\boldsymbol{L}_k^{\langle m\rangle}}(z, \boldsymbol{u}) - \underline{\boldsymbol{L}_k^{\langle m\rangle}}(z, \boldsymbol{u}). \tag{4.4.28}$$

The family of upper and lower bounds is defined using the procedure described in the second part. More specifically, for every $m \geqslant 1$ and $k \geqslant 0$ these bounds satisfy the equations

$$\begin{aligned} \overline{\boldsymbol{L}_{k+1}^{\langle m-1\rangle}}(z, \boldsymbol{u}) &:= \boldsymbol{\mathcal{K}}^{\langle m-1\rangle}(\overline{\boldsymbol{L}_{k+1}^{\langle m-1\rangle}}(z, \boldsymbol{u}), \overline{\boldsymbol{L}_k^{\langle m\rangle}}(z, \boldsymbol{u}), z, \boldsymbol{u}); \\ \underline{\boldsymbol{L}_{k+1}^{\langle m-1\rangle}}(z, \boldsymbol{u}) &:= \boldsymbol{\mathcal{K}}^{\langle m-1\rangle}(\underline{\boldsymbol{L}_{k+1}^{\langle m-1\rangle}}(z, \boldsymbol{u}), \underline{\boldsymbol{L}_k^{\langle m\rangle}}(z, \boldsymbol{u}), z, \boldsymbol{u}). \end{aligned} \tag{4.4.29}$$

According to the second part, the differences $\Delta_k^{\langle m\rangle}$ satisfy formal power series inequalities $\Delta_{k+1}^{\langle m-1\rangle} \preceq \boldsymbol{\mathcal{R}}\Delta_k^{\langle m\rangle}$. By iteration, we thus obtain

$$\Delta_m^{\langle 0\rangle} \preceq \boldsymbol{\mathcal{R}}^m\Delta_0^{\langle m\rangle}. \tag{4.4.30}$$

Since the spectral radius of $\boldsymbol{\mathcal{R}}$ is bounded by 1, and $\Delta_0^{\langle m\rangle}$ is exponentially small in m, the values of Puiseux coefficients of $\boldsymbol{L}^{\langle 0\rangle}(z, \boldsymbol{u})$ can be approximated within an exponentially small in m gap, for arbitrarily large value of m.

Finally, we note that the functions $\underline{\boldsymbol{L}_m^{\langle 0\rangle}}(z, \boldsymbol{u})$ and $\overline{\boldsymbol{L}_m^{\langle 0\rangle}}(z, \boldsymbol{u})$ have Puiseux expansions of type

$$f_m(z, \boldsymbol{u}) \sim c_m(\boldsymbol{u}) - a_m(\boldsymbol{u})\sqrt{1 - \frac{z}{\rho(\boldsymbol{u})}} \tag{4.4.31}$$

in a certain delta-domain. According to Proposition 3.1.1 (transfer theorem), their coefficients admit the following asymptotic estimate:

$$f_m(z, \boldsymbol{u}) \sim C_n A_m(\boldsymbol{u})B(\boldsymbol{u})^n. \tag{4.4.32}$$

A final application of Lemma 4.2.1 (squeeze lemma for formal power series) combined with Remark 4.2.1 finishes the proof. □

Remark 4.4.2. In Theorem 4.4.1 we prove a so-called *weak* transfer theorem, i.e. prove that the asymptotics of the coefficients of each $\boldsymbol{L}^{\langle m \rangle}$ can be obtained by taking the asymptotic expansion of the correseponding Puiseux expansion. A stronger version would suggest that the generating functions $\boldsymbol{L}^{\langle m \rangle}(z, \boldsymbol{u})$ can be analytically continued beyond the circle of convergence of corresponding formal power series, in a certain delta-domain. However, for our analysis, the presented weak variant is enough. The techniques presented above, can be further extended to obtain a stronger transfer theorem, by computing the Taylor series expansions at points z_0 inside the circle of convergence.

Chapter 5

Multiparametric Boltzmann samplers

Contents

The materials of this chapter follow [BBD18a]. Uniform random generation of combinatorial structures forms a prominent research area of computer science with multiple important applications ranging from automated software testing techniques, see [CH00], to complex simulations of large physical statistical models, see [Bha+17]. Given a formal specification defining a set of combinatorial structures (for instance graphs, proteins or tree-like data structures) we are interested in their efficient random sampling ensuring the uniform distribution among all structures sharing the same size.

One of the earliest examples of a generic sampling template is Nijenhuis and Wilf's recursive method [NW78] later systematised by Flajolet, Zimmermann and Van Cutsem [FZC94]. In this approach, the generation scheme is split into two stages – an initial preprocessing phase where recursive branching probabilities dictating subsequent sampler decisions are computed, and the proper sampling phase itself. Alas, in both phases the algorithm manipulates integers of size exponential in the target size n, turning its effective bit complexity to $O(n^{3+\varepsilon})$, compared to $\Theta(n^2)$ arithmetic operations required. Denise and Zimmermann reduced later the average-case bit complexity of the recursive method to $O(n \log n)$ in time and $O(n)$ in space using a certified floating-point arithmetic optimisation [DZ99]. Regardless, worst-case space bit complexity remained $O(n^2)$ as well as bit complexity for non-algebraic languages. Remarkably, for rational languages Bernardi and Giménez [BG12] recently linked the floating-point optimisation of Denise and Zimmermann with a specialised divide-and-conquer scheme reducing further the worst-case space bit complexity and the average-case time bit complexity to $O(n)$.

A somewhat relaxed, approximate-size setting of the initial generation problem was investigated by Duchon, Flajolet, Louchard and Schaeffer who proposed a universal sampler construction framework of so-called Boltzmann samplers [Duc+04]. The key idea in their approach is to embed the generation scheme into the symbolic method of analytic combinatorics [FS09] and, in consequence, obtain an effective recursive sampling template for a wide range of existing combinatorial classes. In recent years, a series of important improvements was proposed for both unlabelled and Pólya structures. Let us mention for instance linear

approximate-size (and quadratic exact-size) Boltzmann samplers for planar graphs [Fus05], general-purpose samplers for unlabelled structures [FFP07], efficient samplers for plane partitions [BFP10] or the cycle pointing operator for Pólya structures [Bod+11]. Moreover, the framework was generalised onto differential specifications [BRS12; Bod+16]; linear exact-size samplers for Catalan and Motzkin trees were obtained, exploiting the shape of their holonomic specifications [BBJ13].

What was left open since the initial work of Duchon et al., was the development of (i) efficient Boltzmann oracles providing effective means of evaluating combinatorial systems within their disks of convergence and (ii) an automated tuning procedure controlling the expected sizes of parameter values of generated structures. The former problem was finally addressed by Pivoteau, Salvy and Soria [PSS12] who defined a rapidly converging combinatorial variant of the Newton oracle by lifting the combinatorial version of Newton's iteration of Bergeron, Labelle and Leroux [BLL98] to a new numerical level. In principle, using their Newton iteration and an appropriate use of binary search, it became possible to approximate the singularity of a given algebraic combinatorial system with arbitrarily high precision. However, even if the singularity ρ is estimated with precision 10^{-10} its approximation quality does not correspond to an equally accurate approximation of the generating function values at ρ, often not better than 10^{-2}. Precise evaluation at z close to ρ requires an extremely accurate precision of z. Fortunately, it is possible to trade-off the evaluation precision for an additional rejection phase using the idea of analytic samplers [BLR15] retaining the uniformity even with rough evaluation estimates.

Nonetheless, frequently in practical applications including for instance semi-automated software testing techniques, additional control over the internal structure of generated objects is required, see [Pał12]. In [BP10] Bodini and Ponty proposed a multidimensional Boltzmann sampler model, developing a tuning algorithm meant for the random generation of words from context-free languages with a given target letter frequency vector. However, their algorithm converges only in an *a priori* unknown vicinity of the target tuning variable vector. In practice, it is therefore possible to control no more than a few tuning parameters at the same time.

5.1 Preliminaries

Consider an unambiguous context-free grammar

$$S_i \to \sum_j T_{ij}(S_1, \ldots, S_n, \bullet), \quad i = 1, \ldots, N, \tag{5.1.1}$$

where

(*i*) $\bullet$ is the terminal symbol

(*ii*) T_{ij} are possible transitions.

Example 5.1.1. The context-free grammar for well-formed parentheses has one non-terminal S, two terminals (and), and three transitions from S:

$$S \to SS \mid (S) \mid ().$$

It is well-known that the number $a_{n,i}$ of words of length n produced by S_i has a generating function

$$S_i(z) = \sum_{n \geqslant 0} a_{n,i} z^n$$

which satisfies a system of algebraic equations

$$S_i(z) = \sum_j T_{ij}(S_1(z), \ldots, S_n(z), z), \tag{5.1.2}$$

where the functions $T_{ij}(s_1, \ldots, s_n, z)$ (despite possible abuse of notation) are monomials, where the power of each s_j and z equals to the number of occurences of the corresponding nonterminal or terminal symbol in the transition T_{ij}.

In the case when a context-free grammar has several terminals $\bullet_1, \bullet_2, \ldots, \bullet_d$, the context-free grammar can be schematically represented by a system of equations

$$S_i \to \sum_j T_{ij}(S_1, \ldots, S_n, \bullet_1, \bullet_2, \ldots, \bullet_d). \tag{5.1.3}$$

The number $a_{n_1,n_2,\ldots,n_d,i}$ of words containing n_k terminals of the color k produced by S_i has a generating function

$$S_i(z_1, z_2, \ldots, z_d) = \sum_{n \geqslant 0} a_{n_1,n_2,\ldots,n_d,i} z_1^{n_1} z_2^{n_2} \cdots z_4^{n_d}$$

which satisfies a system of polynomial equations

$$S_i(\boldsymbol{z}) = \sum_j T_{ij}(S_1(\boldsymbol{z}), \ldots, S_n(\boldsymbol{z}), z_1, z_2, \ldots, z_d), \tag{5.1.4}$$

where $\boldsymbol{z} := (z_1, z_2, \ldots, z_d)$.

Corresponding to the presented definitions, two problems can be formulated.

Problem 5.1.1. Given a positive integer n, sample a word w of length n from a context-free grammar uniformly at random;

Problem 5.1.2. Given positive integers $(n_1, n_2, \ldots, n_d)$, sample a word w with n_k literals of color k from a context-free grammar uniformly at random;

It turns out that in its full generality, the second problem is known to be $\#P$-complete, i.e. higher in the complexity hierarchy than NP-complete. The author was not able to point the exact reference concerning the context-free grammars and the statement appears to be folklore, so we present a proof for completeness.

Theorem 5.1.1. Exact sampling from context free grammars is #P-complete.

Proof by example. In the paper of Jerrum, Valiant, Vazirani [JVV86] it is proven that uniform random sampling of all cycles in a directed graph is #P-complete. This is the celebrated link between enumeration complexity and generation complexity (and it seems that we can also hope that fast generation algorithms can also give us fast enumeration algorithms). And perhaps encoding directed cycles in oriented graphs is not so easy in the language of algebraic specifications. So I decided to take another route.

Note that, however, not every hard enumeration problem is hard to sample. For example the number of satisfying assignments of a DNF (disjunctive normal form, as opposed to conjunctive normal form) is easy to sample uniformly at random, but the problem is #P-complete. In fact, it is also known [JVV86] that uniform fast sampling is possible iff an fully-polynomial randomized approximation scheme (FPRAS) exists.

Theorem ([WG01]). There is no fully polynomial randomized approximation scheme for #2SAT unless NP = RP (the latter denoting randomized polynomial time).

Theorem ([JVV86]). Almost uniform sampling (with exponentially small error) is possible iff a problem admits fully polynomial randomized approximation scheme.

Corollary. Uniform sampling of satisfying assignments of 2-SAT is #P-complete.

Consider then for example the following 2-CNF formula

$$F = \underbrace{(x_1 \vee \overline{x}_2)}_{c_1} \underbrace{(x_1 \vee \overline{x}_4)}_{c_2} \underbrace{(\overline{x}_2 \vee \overline{x}_3)}_{c_3} \underbrace{(\overline{x}_2 \vee \overline{x}_4)}_{c_4} \underbrace{(\overline{x}_3 \vee x_4)}_{c_5}$$

Construct a system of algebraic equations, where $c_1, \ldots, c_5$ correspond to clauses, and $x_1, \overline{x}_1, \ldots, x_4, \overline{x}_4$ correspond to the literals:

$$\begin{aligned} x_1 &= c_1 c_2, \\ \overline{x}_1 &= 1, \\ \overline{x}_2 &= c_1 c_3 c_4, \\ \overline{x}_3 &= c_3 c_5, \\ &\cdots, \end{aligned}$$

where each literal x is equal to the product of c_i corresponding to clauses where x enters. Then, we define a multivariate function A in the following way:

$$A(c_1, \ldots, c_5) = (x_1 + \overline{x}_1) \ldots (x_4 + \overline{x}_4)(1 + c_1) \ldots (1 + c_5).$$

Using the notation $[\boldsymbol{z^n}]F(\boldsymbol{z}) = \boldsymbol{n}$-th coefficient of $F(\boldsymbol{z})$, we can express the number of satisfying assignments:

$$\#2SAT(F) = [c_1^2 c_2^2 \ldots c_5^2] A(c_1, \ldots, c_5).$$

Since the computation of satisfying assignments of the 2-SAT is #P-complete, the problem of exact sampling from multiparametric context-free grammars is #P-complete as well. □

This high complexity of multiparametric sampling is one of the reasons to introduce a relaxation of the exact formulation, and to impose Boltzmann distribution instead of uniform distribution. Let $S(z)$ be the generating function of the language $\mathcal{S}$:

$$S(z) = \sum_{n \geqslant 0} a_n z^n$$

Consider a distribution $\mathbb{P}_z$ on words from $\mathcal{S}$:

- conditioned on word length n, the distribution is uniform;
- the distribution of the length follows

$$\mathbb{P}_z(|w| = n) = \frac{a_n z^n}{S(z)};$$

Problem 5.1.3. Given an unambiguous context-free grammar $\mathcal{S}$

$$S_i \to \sum_j T_{ij}(S_1, \ldots, S_n, \bullet)$$

and a real value $z > 0$, sample a random word from the Boltzmann distribution.

In [Duc+04], the term *Boltzmann sampler* was invented and they presented a very elegant and efficient algorithm solving this problem.

Algorithm 1: Boltzmann sampler for context-free grammars

Data: real value $z > 0$
Result: Random word from Boltzmann distribution
Function $\Gamma S_i(z)$**:**
  if S_i *is terminal* **then**
    return $\bullet$;
  for *all* j **do**
    $p_j := \dfrac{T_{ij}(S_1(z), \ldots, S_n(z), z)}{S_i(z)}$;
  Choose the transition T_{ij} with probability p_j ;
  $A_1 A_2 \ldots A_k := T_{ij}$;
  return $\Gamma A_1(z) \Gamma A_2(z) \cdots \Gamma A_k(z)$;

In this context, one of the key advantages of Boltzmann sampling is its ability to sample from multiparametric Boltzmann distribution assuming that the values $z_1, \dots, z_d$ are chosen.

Algorithm 2: Multiparametric Boltzmann sampler for context-free grammars

Data: real values $z_1, z_2, \cdots, z_\ell > 0$
Unambiguous multiparametric context free grammar

$$S_i \to \sum_j T_{ij}(S_1, \dots, S_n, \bullet_1, \bullet_2, \cdots, \bullet_\ell)$$

Result: Random word from Boltzmann distribution
Function $\Gamma S_i(z)$**:**
  if S_i *is terminal* $\bullet_k$ **then**
    return $\bullet_k$;
  for *all* j **do**
    $p_j := \dfrac{T_{ij}(S_1(\boldsymbol{z}), \dots, S_n(\boldsymbol{z}), z_1, z_2, \cdots, z_\ell)}{S_i(\boldsymbol{z})}$;
  Choose the transition T_{ij} with probability p_j ;
  $A_1 A_2 \dots A_k := T_{ij}$;
  return $\Gamma A_1(\boldsymbol{z}) \Gamma A_2(\boldsymbol{z}) \cdots \Gamma A_k(\boldsymbol{z})$;

Theorem 5.1.2 (Properties of Boltzmann sampler, [Duc+04])**.**

(*i*) Boltzmann sampler returns a word from the Boltzmann distribution

(*ii*) The expected number of terminals $\bullet_k$ is given by

$$\mathbb{E}_{\boldsymbol{z}}[\#_{of}\ \bullet_k \text{ in a random word } w] = z_k \frac{\frac{\partial}{\partial z_k} S(\boldsymbol{z})}{S(\boldsymbol{z})}$$

(*iii*) In strongly connected grammars, if

$$\mathbb{E}_{\boldsymbol{z}}[\#_{of}\ \bullet_k] = n_k = \alpha_k n, \quad n \to \infty,$$

then, under Boltzmann distribution with parameter $\boldsymbol{z}$,

$$\left[\#_{of}\ \bullet_k \text{ in } w \;\middle|\; |w| = n\right] \xrightarrow[n\to\infty]{d} \mathcal{N}(\alpha_k n, C_k n)$$

One of the challenges of multiparametric Boltzmann sampling is to design a dedicated *tuning procedure* which, given the expectation vector $(\mathbb{E}n_d)_{d=1}^{\ell}$, would return a vector of arguments of the generating functions $\boldsymbol{z} = (z_1, \dots, z_\ell)$. The main challenge of multiparametric tuning is that every argument depends on every expectation vector, and for this reason, it is not possible to tune the arguments independently, see Figure 5.1.

Previously, several results on multiparametric tuning were obtained, most notably, Newton iteration locally convergent procedure [BP10], an automated heuristic approach [DPT10] implemented in practice, and an automatic self-correction procedure with a name *analytic sampler* [BLR15] which is a variation of the Boltzmann sampler. We show how to design an efficient and provable polynomial tuning algorithm.

5.2 Samplers for regular grammars

Recall that a strongly connected regular grammar

$$\boldsymbol{F} = \boldsymbol{\Phi}(\boldsymbol{F}, \boldsymbol{z}) \tag{5.2.1}$$

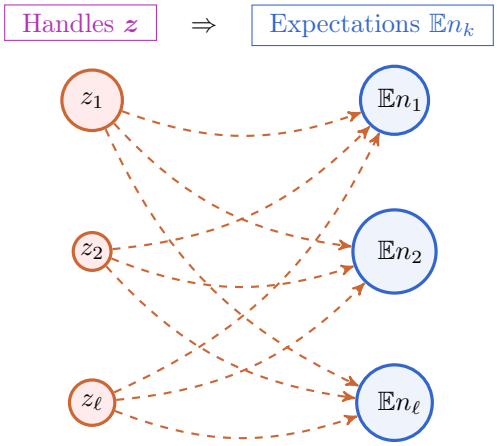


Figure 5.1: Concept of handles and expectations for multiparametric tuning

is a specification corresponding to a regular language whose dependency graph is strongly connected. State and transitions of the associated automaton correspond to classes $\boldsymbol{F} = (F_1, \dots, F_m)$ and to appropriate monomials in the system (5.2.1), respectively. We assume that a grammar is unambiguous.

For rational samplers, we decide to implement the strategy of *interruptible sampling*, introduced in [BBJ13] as the so-called *Hand of God* principle. The idea of anticipated rejection is also discussed in [BGR15]. We start with fixing two distinguished states of the automaton, a starting one and a final (terminal) one. The starting and the terminal states may coincide. Next, we construct a tuned variant of the corresponding singular sampler. Specifically, tune it with arbitrarily high, yet still feasible precision. In essence, it is enough to tune to expected quantities exceeding the target ones. While tuning, we add a constraint $\|\boldsymbol{v}\| \leqslant M$ where $\boldsymbol{v}$ contains all the variables $\boldsymbol{F}$ and $\boldsymbol{z}$, and M is a logarithm of a large number, say $M = 40$. This constraint is required because otherwise the value of associated generating functions tends to infinity as $\boldsymbol{z}$ approaches the singular point. Moreover, by doing so we will compute branching probabilities with an error no more than $O(e^{-M})$. Under these conditions, the resulting sampler is unlikely to stop with output size less than the target one. Finally, we run the sampler from its initial state and continue sampling until the target structure size n_0 is attained. We pick the maximal value n such that the object of size n visited the final state, and return this object. In the following proposition we show that such a sampling procedure is actually an efficient generation scheme. (As noted by Éric Fusy, the option to choose the "ceiling" value of n, instead of choosing the "floor" value, i.e. minimal $n \geqslant n_0$ visiting the final state, is guaranteed to be uniform, while the latter can cause unexpected side effects because of conditioning).

Proposition 5.2.1. Let n be the target size of an interruptible sampler $\Gamma\mathcal{S}$ associated with a strongly connected rational system $\mathcal{S}$. Then, the following assertions hold:

1. structures are sampled from a uniform, conditioned on the (composition) size, distribution;
2. the size of the generated structures is $n - O(1)$ in probability where the constant error term depends solely on $\mathcal{S}$.

Proof. Let ω be a structure generated by the interruptible sampler $\Gamma\mathcal{S}$. Assume w.l.o.g. that $\mathcal{S} = (S_1, \dots, S_m)$ and moreover S_1 and S_m correspond to the associated automaton's starting and final state, respectively. We split the proof into two parts.

Firstly, let us focus on the uniformity (1). We show that conditioned on the vector of quantities $\boldsymbol{n}$, the probability of a structure ω with given number of atomic classes is proportional to $\boldsymbol{z}^{\boldsymbol{n}}$. According to the underlying Boltzmann model, each transition $S_i \to S_j$ taken by $\Gamma\mathcal{S}$ happens with probability

$$\mathbb{P}_{S_i \to S_j} = \boldsymbol{z}^{\Delta \boldsymbol{n}} \frac{S_j(\boldsymbol{z})}{S_i(\boldsymbol{z})} \tag{5.2.2}$$

where $\Delta \boldsymbol{n}$ denotes the change in the size of ω following transition $S_i \to S_j$.

Note however that while we trace the interruptible sampler generating ω, the ratios of generating functions in (5.2.2) cancel out (with the exception of the final $S_m(\boldsymbol{z})$). In consequence, the probability $\mathbb{P}_\omega$ that $\Gamma\mathcal{S}$ generated the structure ω becomes

$$\mathbb{P}_\omega = \boldsymbol{z}^{\sum \Delta \boldsymbol{n}} S_m(\boldsymbol{z}) = \boldsymbol{z}^{\boldsymbol{n}} S_m(\boldsymbol{z}) \tag{5.2.3}$$

where the latter equality follows from the fact that the sum $\sum \Delta \boldsymbol{n}$ of the increments in size is equal to the final size $\boldsymbol{n}$. And so, if we condition on the composite size, i.e. the vector of numbers of atoms, the distribution is indeed uniform.

Let us turn to assertion (2). Once the sampler passes the target size, it becomes a Markov chain with a single absorbing state S_m. The chain is irreducible, as the associated system is strongly connected, whereas all of the states $S_1, \ldots, S_{m-1}$ are not absorbing. Moreover, we can assume that once the target size is reached, the sampler starts a random walk in state S_i where $i \neq m$ as otherwise our claim holds trivially.

In consequence, the expected excess outcome size is proportional to the expected absorption time starting in the transient state S_i. This time, however, is known to be finite, see [KS60, Chapter III]. In conclusion, the expected outcome excess size is necessarily finite. An application of Markov's inequality finishes the proof. □

5.3 Multiparametric sampling

In the present paper we propose a novel polynomial-time tuning algorithm based on convex optimisation techniques, overcoming the previous convergence difficulties. We demonstrate the effectiveness of our approach with several examples of rational, algebraic and Pólya structures. Remarkably, with our new method, we are easily able to handle large combinatorial systems with thousands of combinatorial classes and tuning parameters.

In order to illustrate the effectiveness of our approach, we have implemented a prototype sampler generator Boltzmann Brain (bb in short). The source code is available at Github[1]. Supplementary scripts used to generate and visualise the presented applications of this paper are available as a separate repository[2].

In § 5.3 we briefly recall the principles of Boltzmann sampling. Next, in § 5.4 we describe the tuning procedure. In § 10.6 we propose four exemplary applications and explain the interface of bb. Finally, in Section 5.5 we give the proofs of the theorems, discuss implementation details and describe a novel exact-size sampling algorithm for strongly connected rational grammars.

5.3.1 Specifiable k-parametric combinatorial classes.

Let us consider the neutral class $\mathcal{E}$ and its atomic counterpart $\mathcal{Z}$, both equipped with a finite set of admissible operators (i.e. disjoint union $+$, Cartesian product $\times$, sequence Seq, multiset MSet and cycle Cyc), see [FS09, pp. 24–30]. Combinatorial specifications are finite systems of equations (possibly recursive) built from elementary classes $\mathcal{E}$, $\mathcal{Z}$ and the admissible operators.

Example 5.3.1. Consider the following joint specification for $\mathcal{T}$ and $\mathcal{Q}$. In the combinatorial class $\mathcal{T}$ of trees, nodes of even level (the root starts at level one) have either no or two children and each node at odd level has an arbitrary number of non-planarily ordered children:

$$\begin{cases} \mathcal{T} = \mathcal{Z}\,\mathsf{MSet}(\mathcal{Q})\,, \\ \mathcal{Q} = \mathcal{Z} + \mathcal{Z}\mathcal{T}^2\,. \end{cases} \tag{5.3.1}$$

In order to distinguish (in other words *mark*) some additional combinatorial parameters we consider the following natural multivariate extension of specifiable classes.

[1] https://github.com/maciej-bendkowski/boltzmann-brain

[2] https://github.com/maciej-bendkowski/multiparametric-combinatorial-samplers

Class	Description	$C(\boldsymbol{z})$	$\Gamma\mathcal{C}(\boldsymbol{z})$
Neutral	$\mathcal{C} = \{\varepsilon\}$	$C(\boldsymbol{z}) = 1$	ε
Atom	$\mathcal{C} = \{t_i\}$	$C(\boldsymbol{z}) = z_i$	$\square_i$
Union	$\mathcal{C} = \mathcal{A} + \mathcal{B}$	$A(\boldsymbol{z}) + B(\boldsymbol{z})$	$\mathrm{Bern}\left(\frac{A(\boldsymbol{z})}{C(\boldsymbol{z})}, \frac{B(\boldsymbol{z})}{C(\boldsymbol{z})}\right) \longrightarrow \Gamma\mathcal{A}(\boldsymbol{z}) \mid \Gamma\mathcal{B}(\boldsymbol{z})$
Product	$\mathcal{C} = \mathcal{A} \times \mathcal{B}$	$A(\boldsymbol{z}) \times B(\boldsymbol{z})$	$(\Gamma\mathcal{A}(\boldsymbol{z}), \Gamma\mathcal{B}(\boldsymbol{z}))$
Sequence	$\mathcal{C} = \mathsf{Seq}(\mathcal{A})$	$(1 - A(\boldsymbol{z}))^{-1}$	$\ell := \mathrm{Geom}(1 - A(\boldsymbol{z})) \longrightarrow (\Gamma\mathcal{A}(\boldsymbol{z}))_{\times \ell \text{ times}}$
MultiSet	$\mathsf{MSet}(\mathcal{A})$	$\exp\left(\sum_{m=1}^{\infty} \frac{1}{m} A(\boldsymbol{z}^m)\right)$	see [FFP07]
Cycle	$\mathsf{Cyc}(\mathcal{A})$	$\sum_{m=1}^{\infty} \frac{\varphi(m)}{m} \ln \frac{1}{1 - A(\boldsymbol{z}^m)}$	see [FFP07]

Table 5.1: Multivariate generating functions and their Boltzmann samplers $\Gamma\mathcal{C}(\boldsymbol{z})$.

Definition 5.3.1. (Specifiable k-parametric combinatorial classes) A specifiable k-parametric combinatorial class is a combinatorial specification built, in a possibly recursive manner, from k distinct atomic classes $\mathcal{Z}_i$ ($i \in \{1, \ldots, k\}$), the neutral class $\mathcal{E}$ and admissible operators $+, \times, \mathsf{Seq}, \mathsf{MSet}$ and Cyc. In particular, a vector $\boldsymbol{\mathcal{C}} = (\mathcal{C}_1, \ldots, \mathcal{C}_m)$ forms a specifiable k-parametric combinatorial class if its specification can be written down as

$$\begin{cases} \mathcal{C}_1 = \Phi_1(\boldsymbol{\mathcal{C}}, \mathcal{Z}_1, \ldots, \mathcal{Z}_k)\,, \\ \vdots \\ \mathcal{C}_m = \Phi_m(\boldsymbol{\mathcal{C}}, \mathcal{Z}_1, \ldots, \mathcal{Z}_k) \end{cases} \tag{5.3.2}$$

where the right-hand side expressions are composed from $\boldsymbol{\mathcal{C}}, \mathcal{Z}_1, \ldots, \mathcal{Z}_k$, admissible operators and the neutral class $\mathcal{E}$. Moreover, we assume that specifiable k-parametric combinatorial specifications form well-founded aperiodic systems, see [BLL98; PSS12; Drm97].

Example 5.3.2. Let us continue our running example, see (5.3.1). Note that we can introduce two additional *marking* classes $\mathcal{U}$ and $\mathcal{V}$ into the system, of weight zero each, turning it in effect to a k-specifiable combinatorial class as follows:

$$\begin{cases} \mathcal{T} = \mathcal{U}\mathcal{Z}\,\mathsf{MSet}(\mathcal{Q}), \\ \mathcal{Q} = \mathcal{V}\mathcal{Z} + \mathcal{Z}\mathcal{T}^2\,. \end{cases} \tag{5.3.3}$$

In this example, $\mathcal{U}$ is meant to mark the occurrences of nodes at odd levels, whereas $\mathcal{V}$ is meant to mark leaves at even levels. In effect, we *decorate* the univariate specification with explicit information regarding the internal structural patterns of our interest.

Much like in their univariate variants, k-parametric combinatorial specifications are naturally linked to ordinary multivariate generating functions, see e.g [FS09].

Definition 5.3.2. (Multivariate generating functions) The multivariate ordinary generating function in variables $z_1, \ldots, z_k$ associated to a specifiable k-parametric combinatorial class $\mathcal{C}$ is defined as

$$C(z_1, \ldots, z_k) = \sum_{p_1 \geqslant 0, \ldots, p_k \geqslant 0} c_{\boldsymbol{p}} \boldsymbol{z}^{\boldsymbol{p}} \tag{5.3.4}$$

where $c_{\boldsymbol{p}} = c_{p_1, \ldots, p_k}$ denotes the number of structures with p_i atoms of type $\mathcal{Z}_i$ and $\boldsymbol{z}^{\boldsymbol{p}}$ denotes the product $z_1^{p_1} \cdots z_k^{p_k}$. In the sequel, we call $\boldsymbol{p}$ the (composition) size of the structure.

In this setting, we can easily lift the usual univariate generating function building rules to the realm of multivariate generating functions associated to specifiable k-parametric combinatorial classes. Table 5.1 summarises these rules.

5.3.2 Multiparametric Boltzmann samplers.

Consider a typical multiparametric Boltzmann sampler workflow [BP10] on our running example, see (5.3.3). We start with choosing target expectation quantities (n, k, m) of nodes from atomic classes $(\mathcal{Z}, \mathcal{U}, \mathcal{V})$. Next, using a dedicated tuning procedure we obtain a vector of three real positive numbers $\boldsymbol{z} = (z, u, v)$ depending on (n, k, m). Then, we construct a set of recursive Boltzmann samplers $\Gamma\mathcal{U}(\boldsymbol{z}), \Gamma\,\mathsf{MSet}(\mathcal{Q}(\boldsymbol{z}))$, etc. according to the building rules in Table 5.1. Finally, we use the so constructed samplers to generate structures with tuned parameters.

In order to sample from either $\mathcal{E}$ or atomic classes, we simply construct the neutral element ε or an appropriate atomic structure $\square_i$, respectively. For union classes we make a Bernoulli choice depending on the quotients of respective generating functions values and continue with sampling from the resulting class. In the case of product classes, we spawn two independent samplers, one for each class, and return a pair of built structures. Finally, for $\mathsf{Seq}(\mathcal{A})$ we draw a random value from a geometric distribution with parameter $1 - A(\boldsymbol{z})$ and spawn that many samplers corresponding to the class $\mathcal{A}$. In other words, $\mathbb{P}(\ell \text{ instances}) = A(\boldsymbol{z})^\ell(1 - A(\boldsymbol{z}))$. In the end, we collect the sampler outcomes and return their list. The more involved MSet and Cyc constructions are detailed in [FFP07].

The probability space associated to so constructed Boltzmann samplers takes then the following form. Let $\boldsymbol{z} \in (\mathbb{R}^+)^k$ be a vector inside the ball of convergence of $C(\boldsymbol{z})$ and ω be a structure of composition size $\boldsymbol{p}$ in a k-parametric class $\mathcal{C}$. Then, the probability that ω becomes the output of a multiparametric Boltzmann sampler $\Gamma\mathcal{C}(\boldsymbol{z})$ is given as

$$\mathbb{P}_{\boldsymbol{z}}(\omega) = \frac{\boldsymbol{z}^{\boldsymbol{p}}}{C(\boldsymbol{z})} \tag{5.3.5}$$

Proposition 5.3.1. Let $\boldsymbol{N} = (N_1, \ldots, N_k)$ be the random vector where N_i equals the number of atoms of type $\mathcal{Z}_i$ in a random combinatorial structure returned by the k-parametric Boltzmann sampler $\Gamma\mathcal{C}(\boldsymbol{z})$. Then, the expectation vector $\mathbb{E}_{\boldsymbol{z}}(\boldsymbol{N})$ and the covariance matrix $\mathrm{Cov}_{\boldsymbol{z}}(\boldsymbol{N})$ are given by

$$\mathbb{E}_{\boldsymbol{z}}(N_i) = \frac{\partial}{\partial \xi_i} \log C(e^{\boldsymbol{\xi}})\bigg|_{\boldsymbol{\xi}=\log \boldsymbol{z}} \quad \text{and} \quad \mathrm{Cov}_{\boldsymbol{z}}(\boldsymbol{N}) = \left[\frac{\partial^2}{\partial \xi_i \partial \xi_j} \log C(e^{\boldsymbol{\xi}})\right]_{i,j=1}^{k}\Bigg|_{\boldsymbol{\xi}=\log \boldsymbol{z}} \tag{5.3.6}$$

Hereafter, we use $e^{\boldsymbol{z}}$ to denote coordinatewise exponentiation.

Corollary 5.3.1. The function $\gamma(\boldsymbol{z}) := \log C(e^{\boldsymbol{z}})$ is convex because its matrix of second derivatives, as a covariance matrix, is positive semi-definite inside the set of convergence. This crucial assertion will later prove central to the design of our tuning algorithm.

Remark 5.3.1. Uniparametric recursive samplers of Nijenhuis and Wilf take, as well as Boltzmann samplers, a system of generating functions as their input. This system can be modified by putting fixed values of tuning variables, in effect altering the corresponding branching probabilities. The resulting distribution of the random variable corresponding to a weighted recursive sampler coincides with the distribution of the Boltzmann-generated variable conditioned on the structure size. As a corollary, the tuning procedure that we discuss in the following section is also valid for the exact-size approximate-frequency recursive sampling. In Section 5.2 we describe an algorithm for rational specifications which samples objects of size $n + O(1)$. As a by-product, we show how to convert approximate-size samplers corresponding to rational systems into exact-size samplers.

5.4 Tuning as a convex optimisation problem

We start with a general result about converting the problem of tuning arbitrary specifiable k-parametric combinatorial specifications into a convex optimisation problem, provided that one has access to an oracle yielding values and derivatives of corresponding generating functions. We note that this general technique can be applied to differential specifications as well. We write $f(\cdot) \to \min_{\boldsymbol{z}}$, $f(\cdot) \to \max_{\boldsymbol{z}}$ to denote the minimisation (maximisation, respectively) problem of the target function $f(\cdot)$ with respect to the vector

variable $\boldsymbol{z}$. All proofs are postponed until Section 5.5. Throughout this section, we assume that given tuning expectations are *admissible* in the sense that there always exists a target vector $\boldsymbol{z}^*$ corresponding to (5.3.6). Furthermore, we assume that the combinatorial system is *well-founded* and *strongly connected*. Some non-strongly connected cases fall into the scope of our framework as well, but for the core proof ideas we concentrate only on strongly connected systems.

Theorem 5.4.1. Consider a multiparametric combinatorial class $\mathcal{C}$. Fix the expectations $\mathbb{E}_{\boldsymbol{z}}\boldsymbol{N} = \boldsymbol{\nu}$, see Proposition 5.3.1. Let $C(\boldsymbol{z})$ be the generating function corresponding to $\mathcal{C}$. Then, the tuning vector $\boldsymbol{z}$, see (5.3.6), is equal to $e^{\boldsymbol{\xi}}$ where $\boldsymbol{\xi}$ comes from the following minimisation problem:

$$\log C(e^{\boldsymbol{\xi}}) - \boldsymbol{\nu}^{\top}\boldsymbol{\xi} \to \min_{\boldsymbol{\xi}} \quad . \tag{5.4.1}$$

Let us turn to the specific classes of algebraic and rational specification. In those cases, no differential-equation type systems are allowed; however, it is possible to reformulate the problem so that no extra oracles are required.

Theorem 5.4.2. Let $\boldsymbol{\mathcal{C}} = \boldsymbol{\Phi}(\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{Z}})$ be a multiparametric algebraic system with $\boldsymbol{\mathcal{C}} = (\mathcal{C}_1, \dots, \mathcal{C}_m)$. Fix the expectations N_i of the parameters of objects sampled from $\mathcal{C}_1$ to $\mathbb{E}_{\boldsymbol{z}}\boldsymbol{N} = \boldsymbol{\nu}$. Then, the tuning vector $\boldsymbol{z}$ is equal to $e^{\boldsymbol{\xi}}$ where $\boldsymbol{\xi}$ comes from the convex problem:

$$\begin{cases} c_1 - \boldsymbol{\nu}^{\top}\boldsymbol{\xi} \to \min_{\boldsymbol{\xi},\boldsymbol{c}} \ , \\ \log \boldsymbol{\Phi}(e^{\boldsymbol{c}}, e^{\boldsymbol{\xi}}) - \boldsymbol{c} \leqslant 0. \end{cases} \tag{5.4.2}$$

Hereafter, "$\leqslant$" and $\log \boldsymbol{\Phi}$ denote a set of inequalities and the coordinatewise logarithm, respectively.

Let us note that the above theorem naturally extends to the case of labelled structures with Set and Cyc operators. For unlabelled Pólya operators like MSet or Cyc, we have to truncate the specification to bound the number of substitutions. In consequence, it becomes possible to sample corresponding unlabelled structures, including partitions, functional graphs, series-parallel circuits, etc.

Singular Boltzmann samplers (also defined in [Duc+04]) are the limiting variant of ordinary Boltzmann samplers with an infinite expected size of generated structures. In their multivariate version, samplers are considered *singular* if their corresponding variable vectors belong to the boundary of the respective convergence sets.

Theorem 5.4.3. Let $\boldsymbol{\mathcal{C}} = \boldsymbol{\Phi}(\boldsymbol{\mathcal{C}}, \mathcal{Z}, \boldsymbol{\mathcal{U}})$ be a multiparametric algebraic system with $\boldsymbol{\mathcal{C}} = (\mathcal{C}_1, \dots, \mathcal{C}_m)$, the atomic class $\mathcal{Z}$ marking the corresponding structure size and $\boldsymbol{\mathcal{U}} = (\mathcal{U}_1, \dots, \mathcal{U}_k)$ being a vector (possibly empty) of distinguished atoms. Assume that the target expected frequencies of the atoms $\mathcal{U}_i$ are given by the vector $\boldsymbol{\alpha}$. Then, the variables $(z, \boldsymbol{u})$ that deliver the tuning of the corresponding singular Boltzmann sampler are the result of the following convex optimisation problem, where $z = e^{\xi}$, $\boldsymbol{u} = e^{\boldsymbol{\eta}}$:

$$\begin{cases} \xi + \boldsymbol{\alpha}^{\top}\boldsymbol{\eta} \to \max_{\xi,\boldsymbol{\eta},\boldsymbol{c}} \ , \\ \log \boldsymbol{\Phi}(e^{\boldsymbol{c}}, e^{\xi}, e^{\boldsymbol{\eta}}) - \boldsymbol{c} \leqslant 0 \, . \end{cases} \tag{5.4.3}$$

Finally, let us note that all of the above outlined convex programs can be effectively optimised using the polynomial-time interior-point method optimisation procedure of Nesterov and Nemirovskii [NN94]. The required precision ε is typically $Poly(n)$, see Section 5.5.

Theorem 5.4.4. For multiparametric combinatorial systems with description length L, the tuning problem can be solved with precision ε in time $O\big(L^{3.5} \log \frac{1}{\varepsilon}\big)$.

Let us complete this section by constructing an optimisation system for (5.3.2). Let (n, k, m) be the target expectation quantities of $(\mathcal{Z}, \mathcal{U}, \mathcal{V})$. By the rules in Table 5.1, the system of functional equations and its log-exp transformed optimisation counterpart take the form

$$\begin{cases} T(z,u,v) = uz \exp\left(\displaystyle\sum_{i=1}^{\infty} \frac{Q(z^i, u^i, v^i)}{i}\right), \\ Q(z,u,v) = vz + zT(z,u,v)^2 \, . \end{cases} \tag{5.4.4}$$

Setting $T(z^i, u^i, v^i) = e^{\tau_i}$, $Q(z^i, u^i, v^i) = e^{\kappa_i}$, $z = e^{\zeta}$, $u = e^{\eta}$, $v = e^{\phi}$, we obtain

$$\begin{cases} \tau_1 - n\zeta - k\eta - m\phi \to \min, \\ \tau_j \geqslant \eta j + \zeta j + \sum\limits_{i=1}^{\infty} \dfrac{e^{\kappa_{ij}}}{i}, \quad j \in \{1, 2, \ldots\} \\ \kappa_j \geqslant \log(e^{\phi j + \zeta j} + e^{\zeta j + 2\tau_j}), \quad j \in \{1, 2, \ldots\} \ . \end{cases} \tag{5.4.5}$$

For practical purposes, the sum can be truncated with little effect on distribution.

5.5 Convex optimisation: proofs and algorithms

Until now, we have left several important questions unanswered. Firstly, what is the required precision ε for multiparametric tuning? Secondly, what is its precise computational complexity? In order to determine the time and space complexity of our tuning procedure we need to explain some technical decisions regarding the choice of particular optimisation methods. In this section we prove that the optimisation procedures described in § 5.4 give the correct solution to the tuning problem.

5.5.1 Proofs of the theorems.

Proof of Theorem 5.4.1. Let the following *nabla-notation* denote the vector of derivatives (so-called gradient vector) with respect to the variable vector $\boldsymbol{z} = (z_1, \ldots, z_k)$:

$$\nabla_{\boldsymbol{z}} f(\boldsymbol{z}) = \left(\frac{\partial}{\partial z_1} f(\boldsymbol{z}), \ldots, \frac{\partial}{\partial z_k} f(\boldsymbol{z}) \right)^{\top} . \tag{5.5.1}$$

We start with noticing that tuning the expected number of atom occurrences is equivalent to solving the equation $\nabla_{\boldsymbol{\xi}} \log C(e^{\boldsymbol{\xi}}) = \boldsymbol{\nu}$, see Proposition 5.3.1. Here, the right-hand side is equal to $\nabla_{\boldsymbol{\xi}}(\boldsymbol{\nu}^{\top}\boldsymbol{\xi})$ so tuning is further equivalent to $\nabla_{\boldsymbol{\xi}} \left(\log C(e^{\boldsymbol{\xi}}) - \boldsymbol{\nu}^{\top}\boldsymbol{\xi} \right) = 0$. The function under the gradient is convex as it is a sum of a convex and linear function. In consequence, the problem of minimising the function is equivalent to finding the root of the derivative

$$\log C(e^{\boldsymbol{\xi}}) - \boldsymbol{\nu}^{\top}\boldsymbol{\xi} \to \min_{\boldsymbol{\xi}} \ . \tag{5.5.2}$$

Definition 5.5.1. (Feasible points) In the optimisation problem

$$\begin{cases} f(\boldsymbol{z}) \to \min, \\ \boldsymbol{z} \in \Omega \end{cases} \tag{5.5.3}$$

a point $\boldsymbol{z}$ is called feasible if it belongs to the set Ω.

Proof of Theorem 5.4.2. Let $\boldsymbol{N} = (N_1, \ldots, N_k)$ be the vector of atom occurrences of each type. Consider the vector $\boldsymbol{z}^*$ such that $\mathbb{E}_{\boldsymbol{z}^*}(\boldsymbol{N}) = \boldsymbol{\nu}$. Let $\boldsymbol{c}$ denote the logarithms of the values of generating functions at point $\boldsymbol{z}^* = e^{\boldsymbol{\xi}^*}$. Clearly, in such a case all inequalities in (5.4.2) become equalities and the point $(\boldsymbol{c}, \boldsymbol{\xi}^*)$ is feasible.

Let us show that if the point $(\boldsymbol{c}, \boldsymbol{\xi})$ is optimal, then all the inequalities in (5.4.2) become equalities. Firstly, suppose that the inequality

$$c_1 \geqslant \log \Phi_1(e^{\boldsymbol{c}}, e^{\boldsymbol{\xi}}) \tag{5.5.4}$$

does not turn to an equality. Certainly, there is a *gap* and the value c_1 can be decreased. In doing so, the target function value is decreased as well. Hence, the point $(\boldsymbol{c}, \boldsymbol{\xi})$ cannot be optimal.

Now, suppose that the initial inequality does turn to equality, however $c_k > \log \Phi_k(e^{\boldsymbol{c}}, e^{\boldsymbol{\xi}})$ for some $k \neq 1$. Since the system is strongly connected, there exists a path $P = c_1 \to c_2 \to \cdots \to c_k$ (indices are chosen

without loss of generality) in the corresponding dependency graph. Note that for pairs of consecutive variables (c_i, c_{i+1}) in P, the function $\log \Phi_i(e^{\boldsymbol{c}}, e^{\boldsymbol{\xi}})$ is strictly monotonic in c_{i+1} (as its monotonic and references c_{i+1}). In such a case we can decrease c_{i+1} so to assure that $c_i > \log \Phi_i(e^{\boldsymbol{c}}, e^{\boldsymbol{\xi}})$ while the point $(\boldsymbol{c}, \boldsymbol{\xi})$ remains feasible. Decreasing $c_{i+1}, c_i, \ldots, c_1$ in order, we finally arrive at a feasible point with a decreased target function value. In consequence, $(\boldsymbol{c}, \boldsymbol{\xi})$ could not have been optimal to begin with.

So, eventually, the optimisation problem reduces to minimising the expression subject to the system of equations $\boldsymbol{c} = \log \boldsymbol{\Phi}(e^{\boldsymbol{c}}, e^{\boldsymbol{\xi}})$ or, equivalently, $\boldsymbol{C}(\boldsymbol{z}) = \boldsymbol{\Phi}(\boldsymbol{C}(\boldsymbol{z}), \boldsymbol{z})$ and can be therefore further reduced to Theorem 5.4.1.

Proof of Theorem 5.4.3. By similar reasoning as in the previous proof, we can show that the maximum is attained when all the inequalities turn to equalities. Indeed, suppose that at least one inequality is strict, say $c_j > \log \Phi_j(e^{\boldsymbol{c}}, e^{\xi}, e^{\boldsymbol{\eta}})$. The value c_j can be slightly decreased by ε by choosing a sufficiently small distortion ε to turn all the equalities containing c_j in the right-hand side $\log \Phi_i(e^{\boldsymbol{c}}, e^{\xi}, e^{\boldsymbol{\eta}})$ to strict inequalities, because the right-hand sides of each of the inequalities are monotonic functions with respect to c_j. This procedure can be repeated until all the equalities turn into inequalities. Finally, we slightly decrease the value ξ to increase the target function while still staying inside the feasible set, because of the monotonicity of the right-hand side with respect to ξ.

Let us fix $\boldsymbol{u} = e^{\boldsymbol{\eta}}$. For rational and algebraic grammars, within the Drmota–Lalley–Woods framework, see for instance [Drm97], the corresponding generating function singular approximation takes the form

$$C(z, \boldsymbol{u}) \sim a_0(\boldsymbol{u}) - b_0(\boldsymbol{u}) \left(1 - \frac{z}{\rho(\boldsymbol{u})}\right)^t \ . \tag{5.5.5}$$

If $t < 0$, then the asymptotically dominant term becomes $-b_0 \left(1 - \frac{z}{\rho(\boldsymbol{u})}\right)^t$. In this case, tuning the target expected frequencies corresponds to solving the following equation as $z \to \rho(u)$:

$$\operatorname{diag}(\boldsymbol{u}) \frac{[z^n] \nabla_{\boldsymbol{u}} C(z, \boldsymbol{u})}{[z^n] C(z, \boldsymbol{u})} = n\boldsymbol{\alpha} \ . \tag{5.5.6}$$

Let us substitute the asymptotic expansion (5.5.5) into (5.5.6) to track how $\boldsymbol{u}$ depends on $\boldsymbol{\alpha}$:

$$\operatorname{diag}(\boldsymbol{u}) \frac{[z^n] t b_0(\boldsymbol{u}) \left(1 - \frac{z}{\rho(\boldsymbol{u})}\right)^{t-1} z \frac{\nabla_{\boldsymbol{u}} \rho(\boldsymbol{u})}{\rho^2(\boldsymbol{u})}}{[z^n] b_0(\boldsymbol{u}) \left(1 - \frac{z}{\rho(\boldsymbol{u})}\right)^t} = -n\boldsymbol{\alpha} \ . \tag{5.5.7}$$

Only dominant terms are accounted for. Then, by the binomial theorem

$$\operatorname{diag}(\boldsymbol{u}) b_0(\boldsymbol{u}) \frac{t}{n} \binom{t-1}{n} \frac{z \nabla_{\boldsymbol{u}} \rho(\boldsymbol{u})}{\rho^2(\boldsymbol{u})} b_0(\boldsymbol{u})^{-1} \binom{t}{n}^{-1} = -\boldsymbol{\alpha} \ , \tag{5.5.8}$$

With $z = \rho(\boldsymbol{u})$, as $n \to \infty$, we obtain after cancellations

$$\operatorname{diag}(\boldsymbol{u}) \frac{\nabla_{\boldsymbol{u}} \rho(\boldsymbol{u})}{\rho(\boldsymbol{u})} = -\boldsymbol{\alpha} \tag{5.5.9}$$

which can be rewritten as

$$\nabla_{\boldsymbol{\eta}} \log \rho(e^{\boldsymbol{\eta}}) = -\boldsymbol{\alpha} \ . \tag{5.5.10}$$

Passing to exponential variables (5.5.10) becomes

$$\nabla_{\boldsymbol{\eta}} (\xi(\boldsymbol{\eta}) + \boldsymbol{\alpha}^\top \boldsymbol{\eta}) = 0 \ . \tag{5.5.11}$$

As we already discovered, the dependence $\xi(\boldsymbol{\eta})$ is given by the system of equations because the maximum is achieved only when all inequalities turn to equations. That is, tuning the singular sampler is equivalent to maximising $\xi + \boldsymbol{\alpha}^\top \boldsymbol{\eta}$ over the set of feasible points.

Remark 5.5.1. For ordinary and singular samplers, the corresponding feasible set remains the same; what differs is the optimised target function. Singular samplers correspond to imposing an infinite target size. In practice, however, the required singularity is almost never known *exactly* but rather calculated up to some feasible finite precision. The tuned structure size is therefore enormously large, but still, nevertheless, finite. In this context, singular samplers provide a natural *limiting* understanding of the tuning phenomenon and as such, there are several possible ways of proving Theorem 5.4.3.

Figure 5.2 illustrates the feasible set for the class of binary trees and its transition after applying the log-exp transform, turning the set into a convex collection of feasible points. In both figures, the singular point is the rightmost point on the plot. Ordinary sampler tuning corresponds to finding the tangent line which touches the set, given the angle between the line and the abscissa axis.

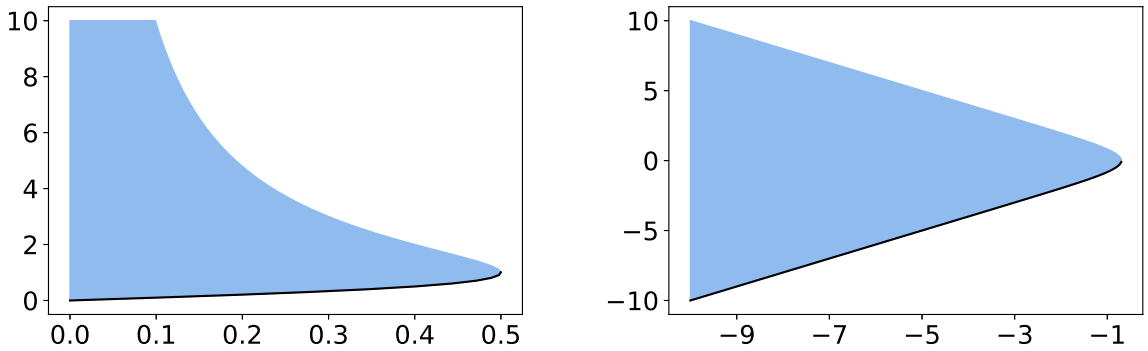


Figure 5.2: Binary trees $B \geqslant z + zB^2$ and log-exp transform of the feasible set. The black curve denotes the principal branch of the generating function $B(z)$ corresponding to the class of binary trees.

5.5.2 Disciplined convex programming and optimisation algorithms.

In the subsequent proofs, we present the framework of Disciplined Convex Programming (DCP in short) and show how to incorporate elementary combinatorial constructions into this framework. Nesterov and Nemirovskii [NN94] developed a seminal polynomial-time optimisation algorithm for convex programming which involves the construction of certain *self-concordant barriers* related to the feasible set of points. The arithmetic complexity of their method is

$$O\left(\log \frac{1}{\varepsilon}\sqrt{\vartheta}\mathcal{N}\right) \tag{5.5.12}$$

where $\mathcal{N}$ is the arithmetic complexity of a single Newton iteration step, ϑ is the so-called constant of self-concordeness of the barriers and ε is the target precision. Before we go into each of the terms, we mention that for sparse matrix representations, it is possible to accelerate the speed of the Newton iteration, i.e. the step of solving the system of linear equations

$$A\boldsymbol{x} = \boldsymbol{b} \quad \text{where} \quad A \in \mathbb{R}^{m\times m} \text{ and } \boldsymbol{x}, \boldsymbol{b} \in \mathbb{R}^m \tag{5.5.13}$$

from $O(m^3)$ to $O(m^2)$.

Unfortunately, for general convex programming problems there is no constructive general-purpose barrier construction method, merely existence proofs. Fortunately, Grant, Boyd, and Ye [GBY06] developed the DCP framework which automatically constructs suitable barriers for the user. Moreover, DCP also automatically provides the starting feasible point which is itself a nontrivial problem in general. As its price, the user is obliged to provide a certificate that the constructed problem is convex, i.e. express all convex functions in terms of a predefined set of elementary convex functions.

In our implementation, we rely on two particular solvers, a second-order (i.e. using second-order derivatives) Embedded Conic Solver (`ECOS`) [DCB13] and recently developed first-order (i.e. using only first-order derivatives) Splitting Conic Solver (`SCS`) algorithm [ODo+16]. The conversion of the DCP problem into its standard form is done using `cvxpy`, a Python-embedded modelling language for disciplined convex programming [DB16].

Proof of Theorem 5.4.4. We start with showing that the tuning procedure can be effectively represented in the framework of DCP.

In our case, every inequality takes the form

$$c_i \geqslant \log\left(\sum_{i=1}^{m} e^{\ell_i(\boldsymbol{c},\boldsymbol{z})}\right) \tag{5.5.14}$$

where $\ell_i(\boldsymbol{c},\boldsymbol{z})$ are some linear functions. Appreciably, the log-sum-exp function belongs to the set of admissible constructions of the DCP framework.

Converting the tuning problem into DCP involves creating some slack variables. For each product of two terms $X \times Y$ we create slack variables for X and Y which are represented by the variables ξ and η in the log-exp realm as

$$e^{\xi} = X \qquad \text{and} \qquad e^{\eta} = Y \ . \tag{5.5.15}$$

Next, we replace $X \times Y$ by $e^{\xi+\eta}$ as composition of addition and exponentiation is a valid DCP program. Since every expression in systems corresponding to considered combinatorial classes is a sum of products, the corresponding restriction (5.5.14) is converted to a valid DCP constraint using the elementary log-sum-exp function.

The sequence operator $\mathsf{Seq}(\mathcal{A})$ which converts a generating function $A(\boldsymbol{z})$ into $(1-A(\boldsymbol{z}))^{-1}$ is *unfolded* by adding an extra equation into the system in form of

$$D := \mathsf{Seq}\, A(\boldsymbol{z}) \quad \text{whereas} \quad D = 1 + AD\,. \tag{5.5.16}$$

Two additional constructions, MSet and Cyc are treated in a similar way. Infinite sums are replaced by finite ones because the difference in the distribution of truncated variables is a negative exponent in the truncation length, and hence negligible.

Using the DCP method, the constant of self-concordness of the barriers is equal to $\vartheta = O(L)$, where L is the length of the problem description. This includes the number of combinatorial classes, number of atoms for which we control the frequency and the sum of lengths of descriptions of each specification, i.e. their overall length. In total, the complexity of optimisation can be therefore crudely estimated as

$$O\left(L^{3.5} \log \frac{1}{\varepsilon}\right) . \tag{5.5.17}$$

Certainly, the complexity of tuning is polynomial, as stated. We emphasise that in practice, using sparse matrices this can be further reduced to $O(L^{2.5}\log(1/\varepsilon))$.

Remark 5.5.2. Weighted partitions, one of our previous applications, involves a multiset operator $\mathsf{MSet}_{\geqslant 1}(\mathcal{Z}_1 + \cdots + \mathcal{Z}_d)$ which generalises to $\mathsf{Seq}(\mathcal{C}_1)\,\mathsf{Seq}(\mathcal{C}_2)\cdots\mathsf{Seq}(\mathcal{C}_d) - 1$ and does not immediately fall into the category of admissible operators as it involves subtraction. This is a general *weak point* of Boltzmann sampling involving usually a huge amount of rejections, in consequence substantially slowing down the generation process. Moreover, it also disables our convex optimisation tuning procedure because the constructions involving the minus sign cease to be convex and therefore do not fit the DCP framework.

We present the following change of variables for this operator, involving a quadratic number of slack variables. The $\mathsf{Seq}(\mathcal{C}_i)$ operator yielding the generating function $(1-C_i(\boldsymbol{z}))^{-1}$ is replaced by $(1+\mathcal{S}_i)$ where $\mathcal{S}_i$ satisfies

$$\mathcal{S}_i = \mathcal{C}_i + \mathcal{S}_i\mathcal{C}_i \ . \tag{5.5.18}$$

Next, we expand all of the brackets in the product $\prod_{i=1}^{d}(1+\mathcal{S}_i)-1$. Consequently, we define the following arrays $\mathcal{P}_{i,j}$ and $\mathcal{Q}_{i,j}$:

$$\begin{cases} \mathcal{P}_{1,j} = \mathcal{C}_j, \quad j \in \{1,\dots,d\} \\ \mathcal{Q}_{k,d} = \mathcal{P}_{k,d}, \quad k \in \{1,\dots,d\} \\ \mathcal{Q}_{k,j} = \mathcal{P}_{k,j} + \mathcal{Q}_{k,j+1}, \quad j \in \{k,\dots,d-1\},\ k \in \{1,\dots,d\} \\ \mathcal{P}_{k,j} = \mathcal{C}_{j-k+1} \cdot \mathcal{Q}_{k-1,j}, \quad j \in \{k,\dots,d-1\},\ k \in \{2,\dots,d\}\ . \end{cases} \tag{5.5.19}$$

Semantically, as in § 10.6.4, $(\mathcal{P}_{i,j})_{j=k}^{d}$ and $(\mathcal{Q}_{i,j})_{j=k}^{d}$ denote the summands inside symmetric polynomials

$$\begin{cases} \mathcal{Q}_{1,1} = \mathcal{S}_1 + \mathcal{S}_2 + \ldots + \mathcal{S}_d\ , \\ \mathcal{Q}_{2,2} = \mathcal{S}_1(\mathcal{S}_2 + \ldots + \mathcal{S}_d) + \mathcal{S}_2(\mathcal{S}_3 + \ldots + \mathcal{S}_d) + \ldots + \mathcal{S}_{d-1}\mathcal{S}_d\ , \\ \mathcal{Q}_{3,3} = \mathcal{S}_1(\mathcal{S}_2\mathcal{S}_3 + \ldots + \mathcal{S}_{d-1}\mathcal{S}_d) + \ldots + \mathcal{S}_{d-2}\mathcal{S}_{d-1}\mathcal{S}_d \end{cases} \tag{5.5.20}$$

and the auxiliary partial sums used to recompute the consequent expressions, respectively. So for instance when $d=5$ we obtain

$$\mathcal{P} = \begin{pmatrix} \mathcal{S}_1 & \mathcal{S}_2 & \mathcal{S}_3 & \mathcal{S}_4 & \mathcal{S}_5 \\ 0 & \mathcal{S}_1(\mathcal{S}_2+\ldots+\mathcal{S}_5) & \ldots & \mathcal{S}_3(\mathcal{S}_4+\mathcal{S}_5) & \mathcal{S}_4\mathcal{S}_5 \\ 0 & 0 & \mathcal{S}_1(\ldots) & \mathcal{S}_2(\mathcal{S}_3(\mathcal{S}_4+\mathcal{S}_5)+\mathcal{S}_4\mathcal{S}_5) & \mathcal{S}_3\mathcal{S}_4\mathcal{S}_5 \\ 0 & 0 & 0 & \mathcal{S}_1(\ldots) & \mathcal{S}_2\ldots\mathcal{S}_5 \\ 0 & 0 & 0 & 0 & \mathcal{S}_1\ldots\mathcal{S}_5 \end{pmatrix}. \tag{5.5.21}$$

The union of classes in each row gives corresponding symmetric polynomial $\mathcal{Q}_{k,k}$, and the partial sum of elements in the row gives the elements of $\mathcal{Q}$.

Finally, the expression $\prod_{i=1}^{d}(1+\mathcal{S}_i)-1$ is replaced by the sum of elementary symmetric polynomials $\mathcal{Q}_{1,1}+\mathcal{Q}_{2,2}+\cdots+\mathcal{Q}_{d,d}$ where we have (combinatorially)

$$\mathcal{Q}_{j,j} = \sum_{1 \leqslant i_1 < \ldots < i_j \leqslant d} \mathcal{S}_{i_1} \cdots \mathcal{S}_{i_j}\ . \tag{5.5.22}$$

We emphasise that the last sum is not meant to be implemented in practice in a naïve way as it would take an exponential amount of time to be computed.

5.5.3 Tuning precision.

In this section, we only work with algebraic systems that meet the certain regularity conditions from Drmota–Lalley–Woods Theorem [Drm97].

Proposition 5.5.1. Consider a multiparametric combinatorial specification

$$\boldsymbol{\mathcal{Y}} = \boldsymbol{\Phi}(\boldsymbol{\mathcal{Y}}, \mathcal{Z}, \boldsymbol{\mathcal{U}}) \tag{5.5.23}$$

whose corresponding system of equations is either rational or algebraic. Suppose that we sample objects from the class $\mathcal{F} = \mathcal{Y}_1$ with target expected sizes $(n, \nu_1 n, \ldots, \nu_d n)$, where ν_i are constants, $n \to \infty$. Let $F(z, \boldsymbol{u})$ be the multivariate generating function corresponding to the class $\mathcal{F}$, and let $(z^*, \boldsymbol{u}^*)$ be the target tuning vector. Then, there exists $\varepsilon = \Theta(1/Poly(n))$ such that the points $(z, \boldsymbol{u})$ from the ε-ball centered at $(z^*, \boldsymbol{u}^*)$ intersected with the set of feasible points

$$\Big\{(z, \boldsymbol{u}) \in \mathbb{R}^{1+d} \mid \boldsymbol{Y}(z, \boldsymbol{u}) \geqslant \boldsymbol{\Phi}(\boldsymbol{Y}(z, \boldsymbol{u}), z, \boldsymbol{u}),\ \|(z^* - z, \boldsymbol{u}^* - \boldsymbol{u})\| \leqslant \varepsilon\Big\}$$

yield expectations within $O(1)$ of target expectations:

$$\begin{aligned} z F'_z(z, \boldsymbol{u})/F(z, \boldsymbol{u}) &= n + O(1)\ , \\ u_i F'_{u_i}(z, \boldsymbol{u})/F(z, \boldsymbol{u}) &= \nu_i n + O(1),\ i \in \{1,\dots,d\}\ . \end{aligned}$$

Proof. Let us show that z^*, as a function of n, satisfies

$$\begin{cases} z^*(n) \sim \rho(1 - \alpha/n), & \mathcal{F} \text{ is rational;} \\ z^*(n) \sim \rho(1 - C/n^2), & \mathcal{F} \text{ is algebraic.} \end{cases} \tag{5.5.24}$$

Here, α is a positive integer depending on the rational system, C is a generic constant. We also note that the same asymptotics is valid for each coordinate of the vector $\boldsymbol{u}^*(n)$, up to multiplicative constants depending on ν_i and the values of α and C.

For rational systems, there exist analytic functions $\beta(z, \boldsymbol{u})$, $\rho(\boldsymbol{u})$ and a positive integer α such that

$$F(z, \boldsymbol{u}) \sim \beta(z, \boldsymbol{u})(1 - z/\rho(\boldsymbol{u}))^{-\alpha} \ , \quad z \to \rho(\boldsymbol{u}) \ . \tag{5.5.25}$$

After substituting the asymptotic expansion (5.5.25) into (5.3.6), we obtain the first part of (5.5.24).

For algebraic systems, according to Drmota–Lalley–Woods Theorem [Drm97], there exist analytic functions $\alpha(z, \boldsymbol{u}), \beta(z, \boldsymbol{u}), \rho(\boldsymbol{u})$ such that as $z \to \rho(\boldsymbol{u})$,

$$F(z, \boldsymbol{u}) \sim \alpha(z, \boldsymbol{u}) - \beta(z, \boldsymbol{u})(1 - z/\rho(\boldsymbol{u}))^{1/2} \ . \tag{5.5.26}$$

Again, substituting this asymptotic expansion into (5.3.6), we obtain

$$z^*(n) \frac{\beta}{2\rho\alpha} \left(1 - \frac{z^*(n)}{\rho}\right)^{-1/2} \sim n \ . \tag{5.5.27}$$

Taking into account that $z^*(n) = \rho + o(1)$, this implies the second part of (5.5.24). Similarly, this can be applied to each coordinate of $\boldsymbol{u}$, not only to z.

Let us handle the tuning precision. We use the mean value theorem to bound ε. Let $m = n + O(1)$. Then,

$$\varepsilon^2 \geqslant \|(z^*(n) - z^*(m), \boldsymbol{u}^*(n) - \boldsymbol{u}^*(m))\|^2 = (z^*(n) - z^*(m)^2 + \sum_{i=1}^{d} (u_i^*(n) - u_i^*(m))^2 \ .$$

By the mean value theorem, there exist numbers $(n_i')_{i=0}^d$ from the interval $[n, m]$ such that

$$\begin{aligned} z^*(n) - z^*(m) &= (n - m) \frac{dz^*}{dn}(n_0') \ , \\ u_i^*(n) - u_i^*(m) &= (n - m) \frac{du_i^*}{dn}(n_i'), \ i \in \{1, \dots, d\} \ . \end{aligned}$$

Thus, as $n - m = O(1)$, we obtain

$$\varepsilon^2 \geqslant O(1) \left[\left(\frac{dz^*}{dn}(n_0') \right)^2 + \sum_{i=1}^{d} \left(\frac{du_i^*}{dn}(n_i') \right)^2 \right] \ .$$

Since $n_i' = n + O(1)$, after substituting (5.5.24) and expressing the derivatives, we obtain the bound $\varepsilon = O(n^{-2})$ for rational grammars and $\varepsilon = O(n^{-3})$ for algebraic specifications. □

Remark 5.5.3. If one uses the *anticipated rejection* principle for sampling the objects of approximate size $n + O(1)$, in effect rejecting objects smaller than $n - O(1)$ and "killing" the generation of objects whose size exceeds $n + O(1)$, it is possible to have a more relaxed bound $\varepsilon = O(n^{-2})$ for the case of algebraic specifications. Even though the expected size of generated objects will be smaller than n, so that we will need a large number of restarts, the total amount of generated atoms will be nevertheless linear in n. We refer to [BGR15, Theorem 4.1] for further discussion.

Remark 5.5.4. Under an extra frequency rejection (independently of the structure size) *it is not possible* to get rid of the assumption of strong connectivity and get a general estimate on the complexity of rejection-based sampling for arbitrary combinatorial specifications. Let us recall that Banderier, Bodini, Ponty and Bouzid give combinatorial classes with non-continuous parameter distributions [Ban+12]. For instance, consider the combinatorial class

$$\mathcal{F} = \mathsf{Seq}(\mathcal{Z}^3)\,\mathsf{Seq}(\mathcal{U}\mathcal{Z}^3) + \mathsf{Seq}(\mathcal{U}^2\mathcal{Z}^3)\,\mathsf{Seq}(\mathcal{U}^3\mathcal{Z}^3) \tag{5.5.28}$$

in which all the structures have parameter frequencies in the intervals $(0, \frac{1}{3})$ and $(\frac{2}{3}, 1)$. Certainly, tuning the sampler for a target frequency inside the interval $(\frac{1}{3}, \frac{2}{3})$ yields a rejection sampler which never stops as there is no structures of demanded frequency.

For this reason we restrict our attention on two important subclasses of combinatorial specifications, i.e. strongly connected rational and algebraic languages. Due to Bender and Richmond [BR83] both classes follow a multivariate Gaussian law with linear expectation and standard deviation. In consequence, corresponding multiparametric Boltzmann samplers work in linear time if we accept a linear tolerance for the size $[(1-\epsilon)n, (1+\epsilon)n]$ and a square root tolerance for the parameters $[f - \kappa/\sqrt{n}, f + \kappa/\sqrt{n}]$.

Part II

Applications

Chapter 6

Phase transitions in graphs with degree constraints

Contents

This chapter follows [DR18]. We show that by restricting the degrees of the vertices of a graph to an arbitrary set Δ, the threshold point $\alpha(\Delta)$ of the phase transition for a random graph with n vertices and $m = \alpha(\Delta)n$ edges can be either accelerated (e.g., $\alpha(\Delta) \approx 0.381$ for $\Delta = \{0, 1, 4, 5\}$) or postponed (e.g., $\alpha(\{2^0, 2^1, \cdots, 2^k, \cdots\}) \approx 0.795$) compared to a classical Erdős–Rényi random graph with $\alpha(\mathbb{Z}_{\geqslant 0}) = \frac{1}{2}$. In particular, we prove that the probability of graph being nonplanar and the probability of having a complex component, goes from 0 to 1 as m passes $\alpha(\Delta)n$. We investigate these probabilities and also different graph statistics inside the critical window of transition (diameter, longest path and circumference of a complex component).

6.1 Overview of the phase transition in graphs with degree constraints

6.1.1 Shifting the Phase Transition

Consider a random Erdős–Rényi graph $G(n, m)$ [ER60], that is a graph chosen uniformly at random among all simple graphs built with n vertices labeled with distinct numbers from $\{1, 2, \ldots, n\}$, and m edges. The range $m = \frac{1}{2}n(1 + \mu n^{-1/3})$ where $n \to \infty$, and μ depends on n, is of particular interest since there are three distinct regimes, according to how the crucial parameter μ grows as n is large:

- as $\mu \to -\infty$, the size of the largest component is of order $\Theta(\log n)$, and the connected components are almost surely trees and unicyclic components;

- next, inside what is known as the critical window $|\mu| = O(1)$, the largest component size is of order $\Theta(n^{2/3})$ and complex structures (nonempty set of connected components having strictly more edges than vertices) start to appear with significant probabilities;

- finally, as $\mu \to +\infty$ with n, there is typically a unique component of size $\Theta(n)$ called the giant component.

Since the article of Erdős and Rényi [ER60], various researchers have studied in depth the phase transition of the Erdős–Rényi random graph model culminating with the masterful work of Janson, Knuth, Łuczak, and Pittel [Jan+93] who used enumerative approach to analyze the fine structure of the components inside the critical window of $G(n, m)$.

The last decades have seen a growth of interest in delaying or advancing the phase transitions of random graphs. Mainly, two kinds of processes have been introduced and studied:

a) the *Achlioptas process* where models of random graph are obtained by adding edge one by one but according to a given rule which allows to choose the next edge from a set of candidate edges [BF01; RW17],

b) the *given degree sequence models* where a sequence $(d_1, \cdots, d_n)$ of degrees is given and a simple graph built on n vertices is uniformly chosen from the set of all graphs whose degrees match with the sequence d_i (see [Rio12; MR95; HM12; Joo+16]).

In [BF01; RW12; RW17], the authors studied the Achlioptas process. Bohman and Frieze [BF01] were able to show that there is a random graph process such that after adding $m = 0.535\, n > 0.5\, n$ edges the size of the largest component is (still) polylogarithmic in n which contrasts with the classical Erdős–Rényi random graphs. Initially, this process was conjectured to have a different local geometry of transition compared to classical Erdős–Rényi model, but Riordan and Warnke [RW12; RW17] were able to show that this is not the case. Next, in the model of random graphs with a fixed degree sequence $D = (d_1, \cdots, d_n)$, Joos, Perarnau, Rautenbach, and Reed [Joo+16] proved that a simple condition that a graph with degree sequence D has a connected component of linear size, is that the sum of the degrees in D which are not 2 is at least $\lambda(n)$ for some function $\lambda(n)$ that goes to infinity with n.

In the current work, our approach is rather different. We study *random graphs with degree constraints* that are graphs drawn uniformly at random from the set of all graphs with given number of vertices and edges with all vertices having degrees from a given set $\Delta \subseteq \mathbb{Z}_{\geqslant 0}$, with the only restriction $1 \in \Delta$, which we discuss below. De Panafieu and Ramos [dPR16] calculated asymptotic number of such graphs using methods from analytic combinatorics. Using their asymptotic results, we prove that random graphs with degrees from the set Δ have their phase transition shifted from the density of edges $\frac{m}{n} = \frac{1}{2}$ to $\frac{m}{n} = \alpha$ for an *explicit* and *computable* constant $\alpha = \alpha(\Delta)$ and the new critical window of transition becomes $m = \alpha n(1 \pm \mu n^{-1/3})$.

In addition, we also prove that the structure of such graphs inside this crucial window behaves as in the Erdős–Rényi case. For instance, we prove that extremal parameters such as the diameter, the circumference or the longest path are of order $\Theta(n^{1/3})$ around $m = \alpha n$. The size of complex components of our graphs are of order $\Theta(n^{2/3})$ as μ is bounded. A very similar result but about the diameter of the largest component of $G(n, p = \frac{1}{n} + \frac{\mu}{n^{4/3}})$ has been obtained by Nachmias and Peres [NP08] (using very different methods).

In the seminal paper of Erdős and Rényi, amongst other non-trivial properties, they discussed the planarity of random graphs with various edge densities [ER60]. The probabilities of planarity of Erdős-Rényi random graphs inside their window of transition have been since then computed by Noy, Ravelomanana, and Rué [RRN13]. In the current work, we extend this study by showing that the planarity threshold shifts from $\frac{n}{2}$ for classical random graphs to αn for graphs with degrees from Δ. More precisely, first we show that such objects are almost surely planar as μ goes to $-\infty$ and non-planar as μ tends to $+\infty$. Next, as function of μ, we compute the limiting probability that random graphs of degrees in Δ are planar as $\mu = O(1)$.

Our work is motivated by the following research questions: (i) what can be the contributions of analytic combinatorics to study constrained random graphs? (ii) the birth of the giant component often corresponds to drastic changes in the complexities of several algorithmic optimization / decision problems on random graphs, so by tuning the thresholds one can shift the location of hard random instances.

6.1.2 Preliminaries

The *excess* of a connected graph is the number of its edges minus the number of its vertices. For example, connected graphs with excess -1 are trees, with excess 0 — graphs with one cycle (also known as unicycles or unicyclic graphs), connected bicycles have excess 2, and so on (see Figure 6.1). Connected graph always has excess at least -1. A connected component with excess at least 1, is called a *complex component*. The *complex part* of a random graph is the union of its complex components.

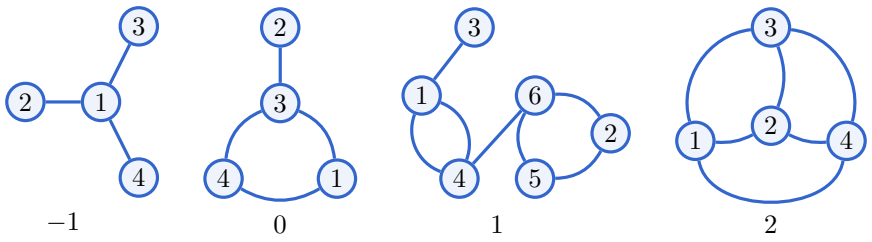


Figure 6.1: Examples of connected labeled graphs with different excess. As a whole, can be considered as a graph with total excess $-1+0+1+2=2$

Next, we introduce the notion of a 2-*core (the core)* and a 3-*core (the kernel)* of a graph. The 2-core is obtained by repeatedly removing all vertices of degree 1 (smoothing). The 3-core is obtained from 2-core by repeatedly replacing vertices of degree two with their adjacent edges by a single edge connecting the neighbors of deleted vertices (we call this a *reduction procedure*). A 3-core can be a multigraph, i.e. there can be loops and multiple edges. There is only a finite number of connected 3-cores with a given excess [Jan+93]. The inverse images of vertices of 3-core under the reduction procedure, are called *corner vertices* (cf. Figure 6.3). A *2-path* is an inverse image of an edge in a 3-core, i.e. a path connecting two corner vertices.

The *circumference* of a graph is the length of its longest cycle. A *diameter* of a graph is the maximal length of the shortest path taken over all distinct pairs of vertices. It is known that the problems of finding the longest path and the circumference are NP-hard.

Random graph with degree constraints is a graph sampled uniformly at random from the set of all possible graphs $\mathcal{G}_{n,m,\Delta}$ having m edges and n vertices all of degrees from the set $\Delta = \{\delta_1, \delta_2, \ldots\} \subseteq \{0, 1, 2, \ldots\}$, see Figure 6.2. The set Δ can be finite or infinite. **In this work, we require that** $1 \in \Delta$. This technical condition allows the existence of trees and tree-like structures in the random objects under consideration. We don't know what happens when $1 \notin \Delta$.

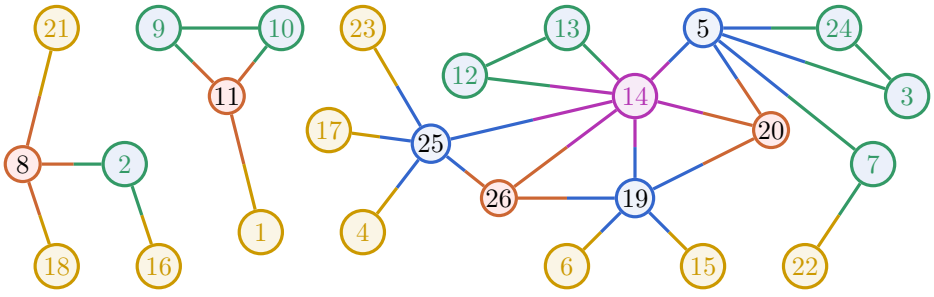


Figure 6.2: Random labeled graph from $\mathcal{G}_{26,30,\Delta}$ with the set of degree constraints $\Delta = \{1, 2, 3, 5, 7\}$

The set $\mathcal{G}_{n,m,\Delta}$ is (asymptotically) nonempty if and only if the following condition is satisfied [dPR16]:

($\mathcal{C}$) Denote $\gcd(d_1 - d_2 \colon d_1, d_2 \in \Delta)$ by *periodicity* p. Assume that the number m of edges grows linearly with the number n of vertices, with $2m/n$ staying in a fixed compact interval of $]\min(\Delta), \max(\Delta)[$, and p divides $2m - n \cdot \min(\Delta)$.

To a given arbitrary set $\Delta \subseteq \{0, 1, 2, \ldots\}$, we associate the *exponential generating function* (EGF) $\omega(z)$:

$$\mathrm{SET}_\Delta(z) = \omega(z) = \sum_{d \in \Delta} \frac{z^d}{d!} \ . \tag{6.1.1}$$

The domain of the argument z of this function can be either considered a subset $[0, R)$ of the real axis or some subset of the complex plane, depending on the context. The function $\phi_0(z) = \frac{z\omega'(z)}{\omega(z)}$, which is called the *characteristic function* of $\omega(z)$, is non-decreasing along real axis [FS09, Proposition IV.5], as well as the characteristic function $\phi_1(z) = \frac{z\omega''(z)}{\omega'(z)}$ of the derivative $\omega'(z)$.

The value of the threshold α, which is used in all our theorems, is a unique solution of the system of equations

$$\begin{cases} \phi_1(\widehat{z}) = 1, \\ \phi_0(\widehat{z}) = 2\alpha. \end{cases} \tag{6.1.2}$$

A unique solution $\widehat{z}$ of $\phi_1(z) = 1, z > 0$ always exists provided that $1 \in \Delta$. This solution is computable.

6.2 Structure of Connected Components

Recall that given a set Δ, its EGF is defined as $\omega(z) = \sum_{d \in \Delta} z^d/d!$, and characteristic function of $\omega(z)$ and its derivative $\omega'(z)$ are given by $\phi_0(z) = z\omega'(z)/\omega(z)$, $\phi_1(z) = z\omega''(z)/\omega'(z)$.

Theorem 6.2.1. Given a set Δ with $1 \in \Delta$, let α be a unique positive solution of (6.1.2). Assume that $m = \alpha n(1 + \mu n^{-1/3})$. Suppose that Condition Section 6.1.2 is satisfied and $G_{n,m,\Delta}$ is a random graph from $\mathcal{G}_{n,m,\Delta}$.

Then, as $n \to \infty$, we have

1. if $\mu \to -\infty$, $|\mu| = O(n^{1/12})$, then

$$\mathbb{P}(G_{n,m,\Delta} \textit{ has only trees and unicycles}) = 1 - \Theta(|\mu|^{-3}) \ ; \tag{6.2.1}$$

2. if $|\mu| = O(1)$, i.e. μ is fixed, then

$$\mathbb{P}(G_{n,m,\Delta} \textit{ has only trees and unicycles}) \to \text{constant} \in (0, 1) \ , \tag{6.2.2}$$
$$\mathbb{P}(G_{n,m,\Delta} \textit{ has a complex part with total excess } q) \to \text{constant} \in (0, 1), \tag{6.2.3}$$

and the constants are computable functions of μ;

3. if $\mu \to +\infty$, $|\mu| = O(n^{1/12})$, then

$$\mathbb{P}(G_{n,m,\Delta} \textit{ has only trees and unicycles}) = \Theta(e^{-\mu^3/6}\mu^{-3/4}) \ , \tag{6.2.4}$$
$$\mathbb{P}(G_{n,m,\Delta} \textit{ has a complex part with excess } q) = \Theta(e^{-\mu^3/6}\mu^{3q/2-3/4}) \ . \tag{6.2.5}$$

Proof (Sketched). Consider a graph composed of trees, unicycles and a collection of complex connected components. Fix the total excess of complex components q. Then, there are exactly $(n - m + q)$ trees, because each tree has an excess -1.

Generating functions for each of these components are given by Lemma 2.3.1 and Lemma 2.3.2: we enumerate all possible kernels and then enumerate graphs that reduce to them under pruning and smoothing.

Let $U(z)$ be the generating function for unrooted trees, $V(z)$ be the generating function for unicycles, $E_j(z)$ be the generating functions for connected graphs with excess j. We calculate the probability for each collection $(q_1, \dots, q_k)$, while the total excess is $\sum_{j=1}^{k} jq_j = q$. Accordingly, the probability that the process generates a graph with the described property can be expressed as the ratio

$$\frac{n! \cdot |\mathcal{G}_{n,m,\Delta}|^{-1}}{(n-m+q)!}[z^n]U(z)^{n-m+q}e^{V(z)}\frac{E_1^{q_1}(z)}{q_1!}\dots\frac{E_k^{q_k}(z)}{q_k!} \ . \tag{6.2.6}$$

Then we use an approximation of $E_j(z)$ from Corollary 2.3.1 and Lemma 2.3.2 and apply Corollary 3.2.1 with $y = \frac{1}{2} + 3q$ in order to extract the coefficients. Note that our approach is derived from the methods from [Jan+93], and so some of our proofs are sketched. □

6.3 Shifting the Planarity Threshold

Theorem 6.3.1. Under the same conditions as in Theorem 6.2.1 with a number of edges $m = \alpha n(1+\mu n^{-1/3})$, let $p(\mu)$ be the probability that $G_{n,m,\Delta}$ is planar.

Then, as $n \to \infty$, we have uniformly for $|\mu| = O(n^{1/12})$:

1. $p(\mu) = 1 - \Theta(|\mu|^{-3})$, as $\mu \to -\infty$;
2. $p(\mu) \to$ constant $\in (0,1)$, as $|\mu| = O(1)$, and $p(\mu)$ is computable;
3. $p(\mu) \to 0$, as $\mu \to +\infty$.

Proof. The graph is planar if and only if all the 3-cores (multigraphs) of connected complex components are planar. As $|\mu| = O(n^{1/12})$, Corollary 2.3.1 implies that for asymptotic purposes it is enough to consider only cubic regular kernels among all possible planar 3-cores. Let $G_1(z)$ be an EGF of connected planar cubic kernels. The function $G_1(z)$ is determined by the system of equations given in [RRN13], and is computable. An EGF for sets of such components is given by $G(z) = e^{G_1(z)}$. We give several first terms of $G(z)$ according to [RRN13]:

$$G(z) = \sum_{q \geqslant 0} g_q \frac{z^{2q}}{(2q)!^2} = 1 + \frac{5}{24}z^2 + \frac{385}{1152}z^4 + \frac{83933}{82944}z^6 + \frac{35002561}{7962624}z^8 + \dots \tag{6.3.1}$$

Thus, the number of planar cubic kernels with total excess q is given by

$$(2q)![z^{2q}]e^{G_1(z)} = (2q)![z^{2q}]G(z) = \frac{g_q}{(2q)!} \ .$$

In order to calculate $p(\mu)$, we sum over all possible $q \geqslant 0$ and multiply the probabilities that the 3-core is a planar cubic graph with excess q by the conditional probability that a random graph has planar cubic kernel of excess q.

The probability that $G_{n,m,\Delta}$ is planar on condition that the excess of the complex component is q, is equal to

$$\frac{n!|\mathcal{G}_{n,m,\Delta}|^{-1}}{(n-m+q)!}[z^n]U(z)^{n-m+q}e^{V(z)}\frac{g_q}{(2q)!}\frac{(T_3(z))^{2r}}{(1-T_2(z))^{3r}} \ . \tag{6.3.2}$$

We can apply Corollary 3.2.1 and sum over all $q \geqslant 0$ in order to obtain the result:

$$p(\mu) \sim \sqrt{2\pi}\sum_{q \geqslant 0} g_q t_3^{2q} A_\Delta(3q + \tfrac{1}{2}, \mu) \ , \tag{6.3.3}$$

where $A_\Delta(3q + \frac{1}{2}, \mu)$ and the constant t_3 are from Corollary 3.2.1. The probabilities on the borders of the transition window, i.e. $|\mu| \to \infty$ can be obtained from the properties of the function $A_\Delta(y, \mu)$. □

6.4 Statistics of the Complex Component Inside the Critical Window

Theorem 6.4.1. Under the same conditions as in Theorem 6.2.1, suppose that $|\mu| = O(1)$, $m = \alpha n(1 + \mu n^{-1/3})$. Then, the longest path, diameter and circumference of the complex part are of order $\Theta(n^{1/3})$ in probability, i.e. for each mentioned random parameter there exist computable (see Lemma 6.6.1) constants $A, B > 0$ depending on Δ such that the corresponding random variable X_n satisfies $\forall \lambda > 0$

$$\mathbb{P}\left(X_n \notin n^{1/3}(A \pm B\lambda)\right) = O(\lambda^{-2}) \ . \tag{6.4.1}$$

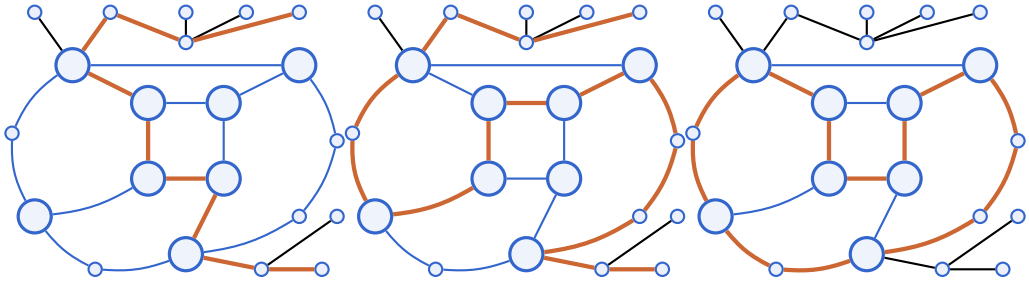


Figure 6.3: Diameter, longest path and circumference of a complex component. The large vertices like ◯ are the *corner* vertices

Proof. Recall that a *2-path* is a path connecting two corner vertices inside a complex component, see Figure 6.3. In Lemma 6.6.1 we prove that the length of a randomly uniformly chosen 2-path is $\Theta(n^{1/3})$ in probability. This lemma also gives the explicit expressions for A and B.

From Lemma 6.6.4 we obtain that the maximum height of sprouting tree over the complex part is also $\Theta(n^{1/3})$ in probability. Since the total excess of the complex component is bounded in probability as μ stays bounded, and the sizes of the kernels are finite, we can combine these two results to obtain the statement of the theorem, because all the three parameters come from adding/stitching several 2-paths and tree heights. □

6.5 Simulations

We considered random graphs with $n = 1000$ vertices, and various degree constraints. The random generation procedure of such graphs has been explained by de Panafieu and Ramos in [dPR16] and for our experiments, we implemented the *recursive method*. We note that this kind of sampling is not *exact* in the sense that the probability of obtaining a simple graph is uniform only in asymptotics.

The generator first draws a sequence of degrees and then performs a random pairing on half-edges, as in configuration model [Bol80]. We reject the pairing until the multigraph is simple, i.e. until there are no loops and multiple edges. As $|\mu| = O(1)$, expected number of rejections is asymptotically $\exp\left(-\frac{1}{2}\phi_1(\widehat{z}) - \frac{1}{4}\phi_1^2(\widehat{z})\right)$, which is $\exp(-3/4)$ in the critical window, and in the subcritical phase it is less.

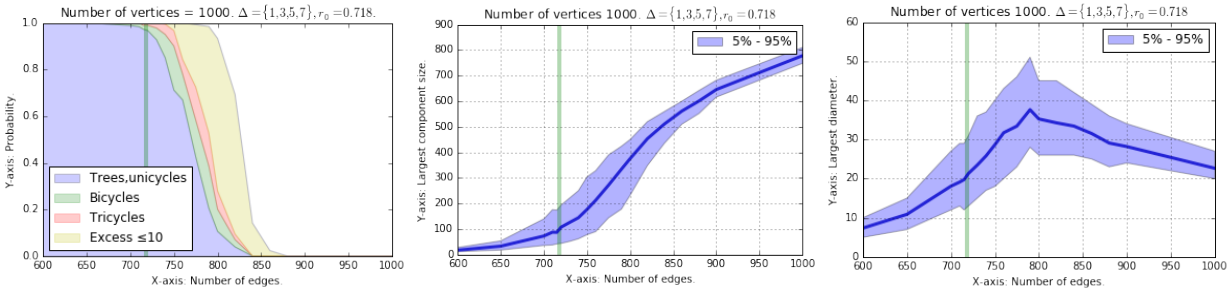
Each sequence $(d_1, \ldots, d_n)$ is drawn with weight $\prod_{v=1}^{n} 1/(d_v)!$. First, we use dynamic programming to precompute the sums of the weights $(S_{i,j})$: $i \in [0, n]$, $j \in [0, 2m]$ using initial conditions and the recursive expression:

$$S_{i,j} = \sum_{\substack{d_1+\ldots+d_i=j \\ d_1,\ldots,d_i \in \Delta}} \prod_{v=1}^{i} \frac{1}{d_v!} \ , \quad S_{i,j} = \begin{cases} 1, & (i,j) = (0,0) \ , \\ 0, & i = 0 \text{ or } j < 0 \ , \\ \sum\limits_{d \in \Delta} \dfrac{S_{i-1,j-d}}{d!}, & \text{otherwise} \ . \end{cases} \tag{6.5.1}$$

Then the sequence of degrees is generated according to the distribution

$$\mathbb{P}(d_n = d) = \frac{S_{n-1,2m-d}}{d! S_{n,2m}} \ . \tag{6.5.2}$$

We made plots for distributions of different parameters for $\Delta = \{1, 3, 5, 7\}$, see Figure 6.4.



(a) Largest excess (b) Largest component size (c) Graph diameter

Figure 6.4: Results of experiments

6.6 Analytic Tools

6.6.1 Method of Moments

In order to study the parameters of random structures, we apply the marking procedure introduced in [FS09]. We say that the variable u marks the parameter of random structure in bivariate EGF $F(z, u)$ if $n![z^n u^k]F(z, u)$ is equal to number of structures of size n and parameter equal to k. In this section we consider such parameters of a random graph as the length of 2-path, which corresponds to some edge of the 3-core, and the height of random "sprouting" tree. If we treat the parameter as a random variable X_n then the factorial moments can be calculated through an expression

$$\mathbb{E}X_n(X_n - 1)\dots(X_n - k + 1) = \frac{\dfrac{d^k}{du^k}[z^n]F(z, u)\Big|_{u=1}}{[z^n]F(z, 1)} \ . \tag{6.6.1}$$

Recall that the number of graphs having n vertices, m edges, and fixed excess vector $\mathbf{q} = (q_1, q_2, \dots)$, can be expressed as n-th coefficient of the generating function

$$\frac{U(z)^{n-m+q}}{(n-m+q)!} e^{V(z)} E_{\mathbf{q}}(z) \ , \tag{6.6.2}$$

where $E_{\mathbf{q}}(z) = \prod_{j=1}^{k} \frac{(E_j(z))^{q_j}}{q_j!}$, $q = \sum_{j=1}^{k} j q_j$. This EGF can be modified to count the moments of random variable X_n.

6.6.2 Length of a Random 2-path

Let us fix the excess vector $\mathbf{q} = (q_1, q_2, \dots, q_k)$. There are in total $q = q_1 + 2q_2 + \dots + kq_k$ connected complex components and each component has one of the finite possible number of 3-cores (see [Jan+93]). We can choose any 2-path, which is a sequence of trees, and replace it with of sequence of marked trees,

see Figure 6.5. Let random variable P_n be the length of this 2-path. Since an EGF for sequence of trees is $\frac{1}{1-T_2(z)}$, the corresponding moment-generating function $\mathbb{E}[u^{P_n}]$ becomes

$$\mathbb{E}[u^{P_n}] = \frac{n![z^n]\dfrac{U(z)^{n-m+q}}{(n-m+q)!}e^{V(z)}E_{\mathbf{q}}(z)\dfrac{1-T_2(z)}{1-uT_2(z)}}{n![z^n]\dfrac{U(z)^{n-m+q}}{(n-m+q)!}e^{V(z)}E_{\mathbf{q}}(z)} \ . \tag{6.6.3}$$

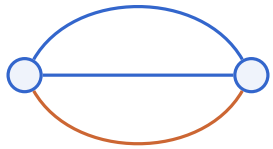


Figure 6.5: Marked 2-path inside complex component of some graph

Lemma 6.6.1. Suppose that conditions of Theorem 6.2.1 are satisfied. Suppose that there are q_j connected components of excess j for each j from 1 to k. Denote by *excess vector* a vector $\mathbf{q} = (q_1, q_2, \ldots, q_k)$. Inside the critical window $m = \alpha n(1 + \mu n^{-1/3})$, $|\mu| = O(1)$, the length P_n of a random (uniformly chosen) 2-path is $\Theta(n^{1/3})$ in probability, i.e.

$$\mathbb{P}\left(P_n \notin n^{1/3}t_3(B_1 \pm \lambda B_2)\right) \leqslant \frac{1}{(\lambda + o(1))^2} \ , \tag{6.6.4}$$

$$t_3 = \widehat{z}\frac{\omega'''(\widehat{z})}{\omega'(\widehat{z})} \ , \quad B_1 = \frac{B_\Delta(3q+\frac{3}{2},\mu)}{B_\Delta(3q+\frac{1}{2},\mu)} \ ,$$

$$B_2^2 = \frac{B_\Delta(3q+\frac{5}{2},\mu)B_\Delta(3q+\frac{1}{2},\mu) - B_\Delta^2(3q+\frac{3}{2},\mu)}{B_\Delta^2(3q+\frac{1}{2},\mu)},$$

with function $B_\Delta(y,\mu)$ from Lemma 3.2.3, $q = q_1 + 2q_2 + \ldots + kq_k$.

Proof. The statement of the lemma is just an application of Chebyshev's inequality to the first and the second moment. Essentially, we need to prove that

$$\mathbb{E}P_n \sim n^{1/3}t_3\frac{B_\Delta(3q+\frac{3}{2},\mu)}{B_\Delta(3q+\frac{1}{2},\mu)} \ , \ \mathbb{E}P_n(P_n-1) \sim n^{2/3}\ 2\,t_3^2\frac{B_\Delta(3q+\frac{5}{2},\mu)}{B_\Delta(3q+\frac{1}{2},\mu)} \ , \tag{6.6.5}$$

which is just a consequence of Lemma 3.2.3 and Eq. (6.6.1). □

6.6.3 Height of a Random Sprouting Tree

Let $\varkappa(z) = \omega'(z)$. Consider recursive definition for the generating function of simple trees whose height doesn't exceed h:

$$T^{[h+1]}(z) = z\varkappa(T^{[h]}(z)) \ , \ T^{[0]}(z) = 0 \ . \tag{6.6.6}$$

The framework of multivariate generating functions allows to mark height with a separate variable u so that the function

$$F(z,u) = \sum_{n=0}^{\infty}\frac{z^n}{n!}\sum_{h=0}^{n}A_n^{[h]}u^h \tag{6.6.7}$$

is the BGF for trees, where $A_n^{[h]}$ stands for the number of simple labelled rooted trees with n vertices, whose height equals h.

Flajolet and Odlyzko [FO82] consider the following expressions:

$$H(z) = \frac{d}{du}F(z,u)\Big|_{u=1}\ ,\ \ D_s(z) = \frac{d^s}{du^s}F(z,u)\Big|_{u=1}\ . \tag{6.6.8}$$

Generally speaking, $H(z) = D_1(z)$ is a particular case of $D_s(z)$, but their analytic behaviour is different for $s = 1$ and $s \geqslant 2$.

Lemma 6.6.2 ([FO82, pp. 42–50])**.** The functions $H(z)$ and $D_s(z)$, $s \geqslant 2$ satisfy

$$H(z) \sim \alpha \log \varepsilon(z)\ ,\ \ D_s(z) \sim (\widehat{z})^{-1} s\Gamma(s)\zeta(s)\varepsilon^{-s+1}(z)\ , \tag{6.6.9}$$

$\alpha = 2\dfrac{\varkappa'(\widehat{z})}{\varkappa''(\widehat{z})}$, $\varepsilon(z) = \widehat{z}\left(1-\dfrac{z}{\rho}\right)^{1/2}\left(\dfrac{2\varkappa''(\widehat{z})}{\varkappa(\widehat{z})}\right)^{1/2}$, $\rho = \widehat{z}\varkappa^{-1}(\widehat{z}) = (\varkappa'(\widehat{z}))^{-1}$.

Here, $\Gamma(s)$ is a gamma-function, and $\zeta(s)$ is Riemann zeta function.

We don't represent their proof here, but would like to remark that it has great methodological impact. For our purposes we need the asymptotic equivalence $\sim$ only in the circle of analyticity $|z| < \rho$.

Recall that

$$T_1(z) = z\omega'(T_1(z))\ ,\quad T_\ell(z) = z\omega^{(\ell)}(T_1(z))\ ,\quad \ell \geqslant 0 \tag{6.6.10}$$

From local expansion at $z = \rho$ of $z = z(T_1)$ it is easy to show that

$$z \sim \rho - (T_1(z) - \widehat{z})^2\left(\frac{\varkappa''(\widehat{z})\widehat{z}}{2\varkappa^2(\widehat{z})}\right) \tag{6.6.11}$$

and consequently, since $T_k(z) = z\varkappa^{(k)}(T_1(z))$,

$$T_1(z) \ = \ \widehat{z} - \sqrt{\frac{2\varkappa}{\varkappa''}}\sqrt{1-\frac{z}{\rho}} + O(1-z\rho^{-1})\ , \tag{6.6.12}$$

$$T_2(z) \ = \ 1 - \sqrt{\frac{2\widehat{z}\varkappa''}{\varkappa}}\sqrt{1-\frac{z}{\rho}} + O(1-z\rho^{-1})\ . \tag{6.6.13}$$

So we have $\varepsilon(z) \sim \widehat{z}^{1/2}(1 - T_2(z))$.

Actually, there are two kinds of sprouting trees that we have to distinguish: the first ones are attached to the vertices with degree from $\Delta - 2$, and the second — to the vertices with degree from $\Delta - 3$, we will treat these cases separately.

Now we can introduce random variables $H_{n(2)}, H_{n(3)}$ equal to the height of a randomly uniformly chosen sprouting tree (of the first and second type respectively), conditioned on excess number $\mathbf{q} = (q_1, q_2, \dots, q_k)$, and their moment generating functions:

$$\mathbb{E}[H_{n(1)}] \ = \ \frac{[z^n]U(z)^{n-m+q}e^{V(z)}E_{\mathbf{q}}(z)\dfrac{F_{(2)}(z,u)}{T_2(z)}}{[z^n]U(z)^{n-m+q}e^{V(z)}E_{\mathbf{q}}(z)}\ , \tag{6.6.14}$$

$$\mathbb{E}[H_{n(2)}] \ = \ \frac{[z^n]U(z)^{n-m+q}e^{V(z)}E_{\mathbf{q}}(z)\dfrac{F_{(3)}(z,u)}{T_2(z)}}{[z^n]U(z)^{n-m+q}e^{V(z)}E_{\mathbf{q}}(z)}\ , \tag{6.6.15}$$

where $F_{(2)}(z,u)$ and $F_{(3)}(z,u)$ are the corresponding BGF for 2- and 3-sprouted trees.

Lemma 6.6.3. Around $z = \rho$ the derivatives of $F_{(2)}$ and $F_{(3)}$ with respect to u at $u = 1$ can be expressed as

$$\frac{d^s}{du^s}F_{(2)}(z,u)\Big|_{u=1} \ \underset{z=\rho}{\sim}\ \frac{\varkappa'(\widehat{z})}{\varkappa''(\widehat{z})}\ \frac{d^s}{du^s}F(z,u)\Big|_{u=1}\ ,$$

$$\frac{d^s}{du^s}F_{(3)}(z,u)\Big|_{u=1} \ \underset{z=\rho}{\sim}\ \frac{\varkappa'(\widehat{z})}{\varkappa'''(\widehat{z})}\ \frac{d^s}{du^s}F(z,u)\Big|_{u=1}\ .$$

Proof. We only present the main idea of the proof, omitting the technical details of how the error term is treated — we refer to [FO82] for the details of transfer theorems and sum approximations.

Consider more general specification, where root degree can belong to the set Φ whose EGF is given by $\varphi(z) = \sum_{d\in\Phi}(d!)^{-1}$. As said before, let $T^{[h]}(z)$ be an EGF for trees of height $\leqslant h$ given by Eq. (6.6.6). Then the EGF $T_\Phi^{[h]}(z)$ for rooted trees, whose root belongs to Φ with height bounded by h, can be written as

$$T_\Phi^{[h+1]} = z\varphi(T_1^{[h]}(z)) \ , \ \ T_\Phi^{[0]}(z) = 0 \ . \tag{6.6.16}$$

Then, there is a second-order Taylor expansion

$$T_\Phi(z) - T_\Phi^{[h+1]}(z) = z(T_1(z) - T_1^{[h]}(z))\varphi'(T_1(z)) \times \left[1 - (T_1 - T_1^{[h]})\frac{\varphi''(T_1)}{2\varphi'(T_1)} + O\left((T_1 - T_1^{[h]})^2\right)\right] \ .$$

Denoting $T_1 - T_1^{[h]} = e_h(z)$, $T_\Phi - T_\Phi^{[h]} = \widetilde{e}_h(z)$, we get asymptotic expansions

$$F(z,u) \ \sim \ uT_1(z) + (u-1)z\sum_{h\geqslant 1} u^h e_h(z)\varkappa'(T_1) \ , \tag{6.6.17}$$

$$F_\Phi(z,u) \ \sim \ u\varphi(T_1(z)) + (u-1)z\sum_{h\geqslant 1} u^h e_h(z)\varphi'(T_1) \ , \tag{6.6.18}$$

so in order to calculate the ratio of derivatives with respect to u at the vicinity of $z = \rho$ we note that the terms $\varkappa'(\widehat{z})$ and $\varphi'(\widehat{z})$ provide the ratio of the coefficients of main asymptotics. □

Lemma 6.6.4. Inside the critical window $m = \alpha n(1 + \mu n^{-1/3})$, $|\mu| = O(1)$, the maximal height H_n of a sprouting tree, is of $O(n^{1/3})$ in probability, i.e.

$$\mathbb{P}\left(\max H_n > \lambda n^{1/3}\right) = O(\lambda^{-2}) \ . \tag{6.6.19}$$

Actually, the *average* height of a sprouting tree (if the tree is taken uniformly at random) appears to be $\Theta(\log n)$ (which seems to be a new result), but when we take the maximum over all possible $\Theta(n^{1/3})$ trees, and apply Chebyshev's inequality, this factor disappears.

Proof of Lemma 6.6.4. We prove the statement for 2-sprouting trees (with root degree from $\Delta - 2$), and for 3-sprouting trees the proof is the same up to a constant term.

The ratio of the expressions in the numerator and denominator can be treated in terms of Lemma 3.2.3. After "Lagrangian" variable change $T_1(z) = t \mapsto z$ the ratio in $\mathbb{E}H_{n(1)}$ becomes proportional to

$$\frac{C_1 \oint (1-\phi_1(z))^{1-y} e^{nh(z;r)} \log(1-\phi_1(z)) dz/z}{C_2 \oint (1-\phi_1(z))^{1-y} e^{nh(z;r)} dz/z} \tag{6.6.20}$$

with $y = 3q + \frac{1}{2}$, and after the second variable change $z = \widehat{z}e^{-s\nu}$, $s = a + it$ the main asymptotics term will become

$$\left(C_1 \oint (\cdots)\right) \Big/ \left(C_2 \oint (\cdots)\right) \sim \widetilde{C}_1(\mu)\log n \ , \tag{6.6.21}$$

For the second factorial moment we obtain

$$\widetilde{C}_2(\mu)n^{1/3} + O(1 + |\mu|^4) \ , \tag{6.6.22}$$

so from Chebyshev's inequality:

$$\mathbb{P}\left(|H_{n(1)} - \widetilde{C}_1 \log n| \geqslant \lambda C_2 n^{1/6}\right) \leqslant \frac{1}{(\lambda + o(1))^2} \ . \tag{6.6.23}$$

Since 2-path length is $\Theta(n^{1/3})$ in probability, we can control the maximal tree height:

$$\mathbb{P}(H_n \geqslant \lambda C_2 n^{1/3}) = O(\lambda^{-2}n^{-1/3}) \ , \ \ \mathbb{P}(\max H_n \geqslant \lambda C_2 n^{1/3}) = O(\lambda^{-2}) \ . \tag{6.6.24}$$

□

Conclusion. We studied how to shift the phase transition of random graphs when the degrees of the nodes are constrained by means of analytic combinatorics [dPR16; FS09].

We have shown that the planarity threshold of those constrained graphs can be shifted generalizing the results in [RRN13]. We have also shown that when our random constrained graphs are inside their critical window of transition, the size of complex components are typically of order $n^{2/3}$ and all distances inside the complex components are of order $n^{1/3}$, thus our results about these parameters complement those of Nachmias and Peres [NP08].

A few open questions are left open: for given threshold value α can we find a set $\{1\} \subset \Delta \subset \mathbb{Z}_{\geqslant 0}$ delivering the desired α? What happens if $1 \notin \Delta$, for example what is the structure of random Eulerian graphs? What happens when the generating function ω itself depends on the number of vertices? Given a sequence of degrees $d_1, \ldots, d_n$ that allows the construction of a forest of an unbounded size, a first approach to study possible relationship between the models can be the computation of the generating function $\omega(z) = \sum_{i \geqslant 0} \text{weight}(i) \frac{z^i}{i!}$ for a suitable weight function corresponding to $d_1, \ldots, d_n$.

Chapter 7

The birth of the contradictory component in random 2-SAT

Contents

This chapter follows [Dov19].

7.1 Phase transition of the 2-SAT

The 2-SAT problem (see Definition 2.2.1) is the easiest case of the more general k-SAT problem which is NP-complete for $k \geqslant 3$, and admits a linear-time algorithm for $k = 2$ [APT82]. However, already the problem of maximising the number of satisfying assignments of the 2-SAT, known as MAX-2-SAT is known to be NP-complete as well. The study of the phase transition in 2-SAT and MAX-2-SAT culminated in the papers [Bol+01] and [Cop+04], where the critical width has been determined. Kim [Kim08] improved the bounds and provided an exact constant for the subcritical case. Since then, it has been questioned whether the structural similarities between the transitions in graphs, digraphs and 2-SAT exist. As an example, it is known that for random graphs $G_{n,m}$ with n vertices and $m = \frac{n}{2}(1 + \mu n^{-1/3})$ edges,

$$\mathbb{P}(G_{n,m} \text{ contains only trees and unicycles}) = 1 - \frac{5 + o(1)}{24|\mu|^3} \quad \text{as } \mu \to -\infty \text{ with } n,$$

for a random 2-SAT formula $F_{n,m}$ with n variables and $m = n(1 + \mu n^{-1/3})$ clauses,

$$\mathbb{P}(F_{n,m} \text{ is satisfiable}) = 1 - \frac{1 + o(1)}{16|\mu|^3} \quad \text{as } \mu \to -\infty \text{ with } n,$$

and for random directed graphs $D_{n,m}$ with n vertices and $m = n(1 + \mu n^{-1/3})$ oriented edges,

$$\mathbb{P}\left(\begin{smallmatrix}\text{every strong component in } D_{n,m} \text{ is either}\\ \text{a vertex or a cycle of length } O(n^{1/3}\mu^{-1})\end{smallmatrix}\right) \to 1 \quad \text{as } \mu \to -\infty \text{ with } n,$$

where $o(1)$ goes to zero as $|\mu| \ll n^{1/3}$ (the formula for random graphs requires $|\mu| \ll n^{1/12}$ and can be further improved to $|\mu| \ll n^{1/3}$, see [HR11]). The factor 5/24 is obtained by adding the inverse numbers of automorphisms 1/12 and 1/8 of the two possible cubic (i.e. 3-regular) multigraphs with 3 edges and 2 vertices, which appear as the only two possible cores at the point when a graph doesn't anymore consist only of trees and unicycles, see Section 2.2.2 and Remark 3.2.2 for an explanation of this phenomenon. A different viewpoint on the giant component and satifsfiability has been proposed in [Mol08]. Clearly, the structure of a critical implication digraph might not be similar to that of a critical simple graph or a critical digraph, due to the presence of certain symmetries, which have been emphasised in [Kra07]. Nevertheless, it still seems to be possible to draw certain structural similarities between the models.

We follow the approach of *analytic combinatorics* which was used to obtain the structure of random graph at the point of the phase transition [FKP89; Jan+93; FSS04]. The approach makes use of the generating function technique which gives a very clear structural vision of the corresponding combinatorial objects. In addition to the aesthetic benefits (see e.g. a unifying approach for the upper bounds on satisfiability threshold [Puy04]), it allows to give very precise asymptotic descriptions, often accompanied by complete asymptotic expansions. In particular, inside the transition window, the probability that a random graph doesn't contain a complex component (i.e. consists only of trees and unicycles) has been expressed in terms of Airy function which has appeared in many other contexts in analytic combinatorics [Ban+01; FSS04]. Another interpretation, using the fact that Airy function is linked to the area under a Brownian motion, was discovered by Aldous [Ald97], see also [ABG12]. The enumeration of unsatisfiable 2-SAT formulae is equivalent to forbidding a certain contradictory pattern (namely, the contradictory circuit, see Section 2.2.1). Of the closest to our approach here is [Col+18] where the authors use the analytic approach to study the containment problem for small subgraphs. Applying the analytic methods to the phase transitions of SAT-formulae and directed graphs remains one of the important challenging problems.

The *spine* of a 2-CNF (see Definition 2.2.3) has been introduced in [Bol+01] as a useful tool for combinatorial analysis of the probability of satisfiability inside the critical window. Originally, the spine is defined as the set of variables that are forced to take the FALSE value in any satisfying assignment. Later, the spine has been shown to impact the complexity of the underlying decision problems in a more general setting for various constraint satisfaction problems including k-XORSAT, graph bipartition problem, random 3-coloring [Sol; IBP05]. In our analysis of satisfiability we do not use the properties of the spine, but we find this parameter interesting by itself. We obtain a new structural result about the spine of a random formula.

The chapter is structures as follows. In Section 7.2 we introduce the concept of *sum-representation* which is used throughout the chapter. In this section, the main comparisons between simple graphs and implication digraphs of 2-SAT are drawn, and the *excess* of a contradictory component is introduced. In Section 2.2.2 the classical symbolic method for generating functions of simple graphs is explained, along with its variations for weakly connected directed graphs. In this section, the concept of compensation factor of a contradictory component is introduced.

In Section 7.3, the actual asymptotic analysis is done. The main results of this chapter are formed into Theorems 7.3.1 to 7.3.3. In Theorem 7.3.1 we prove that in the subcritical phase of the 2-SAT, if the contradictory components are present, their kernels are typically cubic and that the asymptotic expansion of the probability of satisfiability is linked to the compensation factors of such cubic components. In Theorem 7.3.2 we prove that the number of contradictory variables, scaled by $n^{1/3}|\mu|^{-1}$ follows a mixture of Gamma laws, with the first nontrivial case being Gamma(3) with a scale parameter $\mu^{-1}n^{1/3}$ corresponding to the contradictory component of minimal possible excess 1. Finally, in Theorem 7.3.3 we classify the spine literals according to the multiplicities of their paths to the complementary variables, and prove that a negligible part of the spine can be removed in such a way that the remaining literals form a set of tree-like structures.

Several naturally arising open problems and conjectures are described in Section 7.4, along with some remarks on how the results presented in this chapter could be potentially extended. In Section 3.3 we

present a "proof of concept" full asymptotic expansion of the saddle point lemma which allows in principle to construct the full asymptotic expansion of the satisfiability.

7.2 Sum-representation of implication digraphs

For the study of the 2-SAT phase transition, it is convenient to consider digraphs with an even number of vertices $2n$ with a special vertex labelling convention. Instead of the labels $\{n+1, n+2, \dots, 2n\}$ we conventionally use the synonyms $\{\overline{1}, \overline{2}, \dots, \overline{n}\}$. Under this re-assignment, the nodes with labels $\{1, 2, \dots, n\}$ correspond to Boolean literals $\{x_1, x_2, \dots, x_n\}$, and the nodes with labels $\{\overline{1}, \dots, \overline{n}\}$ correspond to the negations of these literals $\{\overline{x}_1, \dots, \overline{x}_n\}$. Since the complementary of $\overline{x}_i$ is x_i, the same applies to the labels, $\overline{\overline{i}} = i$.

Definition 7.2.1 (Conflicts in digraphs)**.** We define a *complementary* of the edge $x \to y$ as $\overline{y} \to \overline{x}$. We shall say that a pair of complementary edges in a directed graph form a *conflict*. We say that a digraph is *conflict-free* if there are no conflicting pairs of edges inside it, i.e. if no two edges are complementary.

Note that in the model that we consider, a random 2-CNF formula contains neither clauses of type $(x \vee x)$ nor of $(x \vee \overline{x})$. We also require that every clause is present at most once. This implies that the underlying implication digraph contains neither loops nor multiple edges.

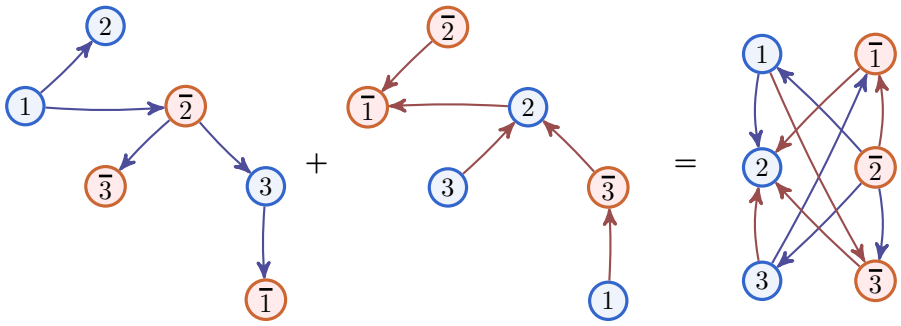


Figure 7.1: Example of a sum representation digraph G and its complementary $\overline{G}$ whose edge-union $G + \overline{G}$ gives an implication digraph.

Definition 7.2.2 (Sum-representation)**.** A digraph $G = (V, E)$ with $2n$ conventionally labelled vertices $V = \{1, \cdots, n, \overline{1}, \cdots, \overline{n}\}$ and $|E| = m$ edges is called a *sum-representation digraph* if it does not contain loops, multiple edges, edges of type $x \to \overline{x}$ and is conflict-free. The *complementary* digraph $\overline{G}$ of a sum-representation digraph is obtained by replacing all the edges by their respective complementaries. We say that G is a *sum-representation of* an implication graph $G + \overline{G}$ which is a digraph obtained by joining the sets of edges of G and $\overline{G}$ (see Figure 7.1). Every implication digraph with $2m$ edges has 2^m sum-representations, since for each of the m clauses of the corresponding formula, 2 choices of complementary edges are available. We denote the set of all sum-representation digraphs with $2n$ vertices and m oriented edges as $\mathcal{D}^\circ(2n, m)$.

Instead of implication digraphs with $2n$ vertices and $2m$ oriented edges we enumerate conflict-free digraphs with $2n$ vertices and m edges. It is worth noticing that $|\mathcal{F}(n, m)| = 2^{-m}|\mathcal{D}^\circ(2n, m)|$. Note also that $\mathcal{D}^\circ(2n, m) \subset \mathcal{D}(2n, m)$.

An *edge rotation* is a process of transformation of an edge $x \to y$ into its complementary edge $\overline{y} \to \overline{x}$ (see Figure 7.2). We say that π_1 is *equivalent* to π_2 if it can be obtained by a sequence of edge rotations. This means that π_1 and π_2 are both sum-representation of the same implication digraph.

7.2.1 The spine

Let y be a literal belonging to the spine of some formula $F \in \mathcal{F}(n, m)$, which means that $y \rightsquigarrow \overline{y}$ in the implication digraph corresponding to F. Counting such literals y with the multiplicities of the corresponding

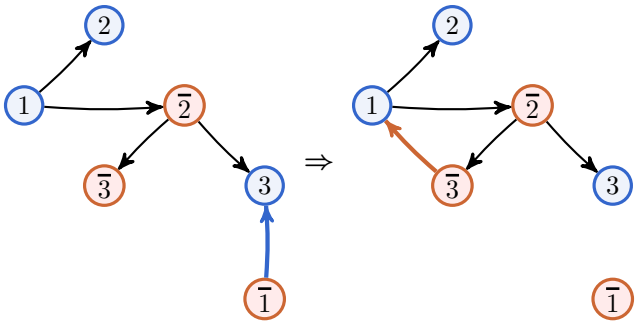


Figure 7.2: Example of two equivalent sum-representations obtain by one edge rotation. An edge $\overline{1} \to 3$ is replaced by its complementary $\overline{3} \to 1$.

paths $y \rightsquigarrow \overline{y}$ gives a larger number than just the cardinality of the spine, however, we are going to show that in the subcritical phase, i.e. when $m = n(1 + \mu n^{-1/3})$, $\mu \to -\infty$, it gives asymptotically the same result, see Theorem 7.3.3 and Corollary 7.3.4. The inherent reason behind this is that the majority of the spine components form certain tree-like structures (see Figure 7.4).

Let us say that a path $y \rightsquigarrow \overline{y}$ is *strictly distinct* if all the vertices of the path, except y and $\overline{y}$, are pairwise strictly distinct. If $y \rightsquigarrow \overline{y}$, it is not always possible to find a strictly distinct directed path from y to $\overline{y}$, but it is possible to split it into a few sections, each strictly distinct.

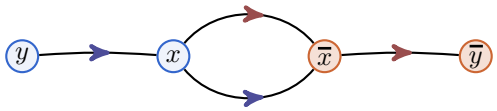


Figure 7.3: A directed path from y to $\overline{y}$ is split into three strictly distinct sections.

Lemma 7.2.1 (Minimal spinal paths)**.** Let $y \rightsquigarrow \overline{y}$ inside an implication digraph. Then there exists a literal x (not necessarily distinct from y) such that $y \rightsquigarrow x \rightsquigarrow \overline{x} \rightsquigarrow \overline{y}$, the paths $y \rightsquigarrow x$ and $x \rightsquigarrow \overline{x}$ are strictly distinct and do not intersect with each other.

As depicted in Figure 7.3, each arrow represents a directed path with strictly distinct literals. Since $x \rightsquigarrow \overline{x}$ is strictly distinct, it has no intersection with its complementary path. It is convenient to denote by $x \rightsquigarrow_1 \overline{x}$ and $x \rightsquigarrow_2 \overline{x}$ the two complementary versions of the path $x \rightsquigarrow \overline{x}$. We shall call every such quadruple $(y \rightsquigarrow x, x \rightsquigarrow_1 \overline{x}, x \rightsquigarrow_2 \overline{x}, \overline{x} \rightsquigarrow \overline{y})$ a *minimal spinal path*.

Proof. Without loss of generality assume that the path $y \rightsquigarrow \overline{y}$ does not pass through the same vertex twice. Assume that the path $y \rightsquigarrow \overline{y}$ is formed as a sequence of edges $y = a_0 \to a_1$, $a_1 \to a_2$, $\ldots$, $a_{\ell-1} \to a_\ell = \overline{y}$. If the path is strictly distinct, we can set $y = x$ and the lemma is proven. Otherwise we choose $x = a_k$ where k is the minimal index such that both a_k and $\overline{a}_k$ belong to the path $y \rightsquigarrow \overline{y}$, and the path $x \rightsquigarrow \overline{x}$ is strictly distinct. It is always possible to choose such an index: if the original path is not strictly distinct, we keep choosing a smaller directed path connecting some literal x and its complementary, until we obtain a path of length at most 2 for which the statement is obviously true since the implication digraph does not contain loops and multiple edges. Note that under such a choice, the path $y \rightsquigarrow \overline{x}$ is also strictly distinct (except for the vertices x and $\overline{x}$), because otherwise it is possible to choose a smaller k. The corresponding part $\overline{x} \rightsquigarrow \overline{y}$ can be constructed as a complementary of the path $y \rightsquigarrow x$. □

Let σ be the random variable denoting the cardinality of the spine in a random formula $F \in \mathcal{F}(n, m)$. Then, since the number of sum-representations of a single implication digraph with $2m$ edges is 2^m, the expected value of σ can be expressed as

$$\mathbb{E}\sigma = \frac{\left|\left\{(G, x) \mid G \in \mathcal{D}^\circ(2n, m),\ x \rightsquigarrow \overline{x} \text{ in } G + \overline{G}\right\}\right|}{|\mathcal{D}^\circ(2n, m)|}.$$

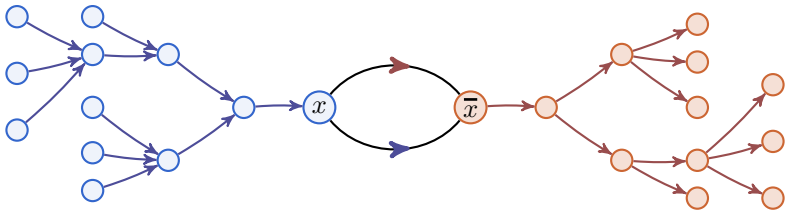


Figure 7.4: For every literal y in a tree-like spine structure, there is a unique path $y \rightsquigarrow \overline{y}$.

Since the spine literals can be characterised as the literals for which there *exists* a path connecting a literal and its complementary, the counting of such literals can be handled though inclusion-exclusion. Accordingly,

$$\mathbb{E}\sigma = \frac{\left|\left\{(G,y,p) \mid G \in \mathcal{D}^\circ(2n,m);\ p \text{ is a minimal spinal path } y \rightsquigarrow \overline{y} \text{ in } G + \overline{G}\right\}\right|}{|\mathcal{D}^\circ(2n,m)|}$$
$$- \frac{\left|\left\{(G,y,p_1,p_2) \mid G \in \mathcal{D}^\circ(2n,m);\ p_1,\ p_2 \text{ are distinct minimal spinal paths } y \rightsquigarrow \overline{y} \text{ in } G + \overline{G}\ \right\}\right|}{|\mathcal{D}^\circ(2n,m)|} + \cdots .$$

Consider a pattern $p \in G$, $G \in \mathcal{D}^\circ(2n,m)$ consisting of two paths $y \rightsquigarrow x$, $x \rightsquigarrow \overline{x}$, where all the literals on the two paths are pairwise strictly distinct (except for x and $\overline{x}$). Suppose that the pattern p has total length ℓ. Then, among 2^m equivalent sum-representations of G, $2^{m-\ell}$ sum-representations contain the pattern p unaltered. Among 2^ℓ possible combinations of edge rotations of one of the ℓ edges of p only, there are exactly 2 possibilities that result in the same type of pattern, the second one obtained by converting the path $x \rightsquigarrow \overline{x}$ into its complementary.

Therefore, the number of sum-representation digraphs with a distinguished pattern $p = y \rightsquigarrow x \rightsquigarrow \overline{x}$ of length ℓ counted with multiplicity 2^ℓ enumerates (with multiplicity 2-to-1) implication digraphs with a distinguished minimal spinal path. This allows us to rewrite the first summand of $\mathbb{E}\sigma$ as

$$\frac{\left|\left\{(G,y,p) \mid G \in \mathcal{D}^\circ(2n,m);\ p \text{ is a minimal spinal path } y \rightsquigarrow \overline{y} \text{ in } G + \overline{G}\right\}\right|}{|\mathcal{D}^\circ(2n,m)|}$$
$$= \frac{1}{2} \cdot \frac{\sum_\ell 2^\ell \left|\{(G,y,x,p) \mid G \in \mathcal{D}^\circ(2n,m);\ p \text{ is a distinguished pattern } y \rightsquigarrow x \rightsquigarrow \overline{x} \text{ of length } \ell \text{ in } G\}\right|}{|\mathcal{D}^\circ(2n,m)|}.$$

All the subsequent summands of $\mathbb{E}\sigma$ can be rewritten in the same manner as well, though it requires considering several cases for the mutual configuration of p_1 and p_2 and using possibly different factors instead of $\frac{1}{2}$ for the multiplicities.

7.2.2 Classifying the contradictory components according to their excesses

Recall that contradictory circuits and variables are defined in Definition 2.2.3. If a variable x is contradictory, it can belong to multiple contradictory circuits at the same time. When the complexity of a contradictory component increases, the number of circuits simultaneously containing a given variable can grow exponentially on its excess. However, we show that the excess grows very slowly in the subcritical phase of the 2-SAT transition, so this technique allows to obtain some structural properties before the number of multiplicities blows up.

Similarly to the excess of a complex component in a simple graph which equals to the number of its edges minus the number of its vertices, we introduce the excess of a contradictory graph. (See Figures 7.5 to 7.8 for different kernels of contradictory components of excess 1 to 2.)

Definition 7.2.3 (Contradictory component and its excess)**.** We call a digraph D with $2n$ conventionally labelled vertices a *contradictory digraph* or a *contradictory component* if for its every edge $e \in D$ its complementary edge $\overline{e}$ is also in D, and every vertex $x \in D$ implies its complementary: $x \rightsquigarrow \overline{x}$. The *excess* of a contradictory component is defined to be equal to the difference between the number of its edges and vertices divided by 2. The excess of an empty graph is defined to be zero.

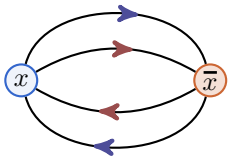


Figure 7.5: First possible contradictory component of excess 1.

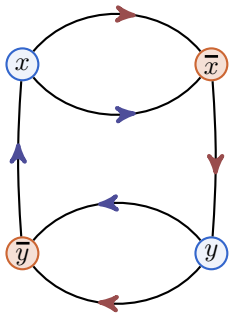


Figure 7.6: Second possible contradictory component of excess 1.

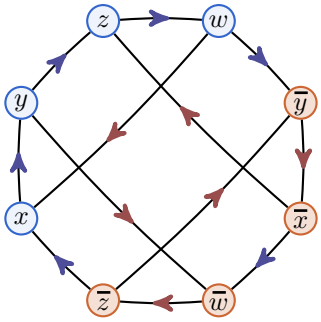


Figure 7.7: A minimal contradictory component of excess $\frac{12-8}{2} = 2$.

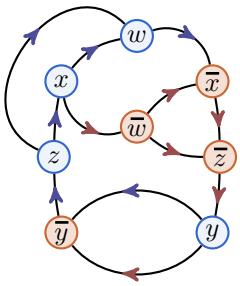


Figure 7.8: Contradictory component of excess 2 which is not minimal.

In contrast with the phase transition of simple graphs, where the whole structure of the graph is known with high probability, no such information is available for the phase transition of 2-SAT yet. On a positive side, Kim has obtained the number of variables and clauses in a core of a random formula. In this work, we focus on the contradictory component only.

In order to access the probability of satisfiability, we use the *inclusion-exclusion method*. We consider *minimal* contradictory components which are contradictory implication digraphs that do not have any proper contradictory subgraphs. Two examples of minimal contradictory components of excess 1 and 2 are given in Figures 7.6 and 7.7, the arrows play the role of strictly distinct paths. Figure 7.8 represents a contradictory strongly component of excess 2 which is not minimal.

Lemma 7.2.2. Let ξ denote the number of minimal contradictory components in a random implication digraph corresponding to a random formula $F \in \mathcal{F}(n,m)$, counted with multiplicities and with possible overlappings. Then, the probability of satisfiability of a random formula $F \in \mathcal{F}(n,m)$ can be expressed using the principle of inclusion-exclusion:

$$\mathbb{P}(F \in \mathcal{F}(n,m) \text{ is SAT}) = 1 - \mathbb{E}_{n,m}\xi + \frac{1}{2!}\mathbb{E}_{n,m}\xi(\xi-1) - \frac{1}{3!}\mathbb{E}_{n,m}\xi(\xi-1)(\xi-2) + \cdots . \tag{7.2.1}$$

Proof. A formula is satisfiable if and only if the contradictory component is empty, i.e. the number of minimal contradictory components is equal to zero. Let N denote the total possible number of subgraphs in a digraph with n vertices, and let E_i denote the event that ith subgraph forms a minimal contradictory component. Using de Morgan's rule and the inclusion-exclusion principle, we obtain

$$\mathbb{P}(\xi = 0) = \mathbb{P}(\overline{E_1} \wedge \cdots \wedge \overline{E_N}) = 1 - \mathbb{P}(E_1 \vee \cdots \vee E_N) = 1 - \sum_{i=1}^{N} \mathbb{P}(E_i) + \sum_{i<j} \mathbb{P}(E_i \wedge E_j) - \cdots .$$

Finally, we note that

$$\sum_{i_1<\ldots<i_k} \mathbb{P}(E_{i_1} \wedge \cdots E_{i_k}) = \binom{N}{k} \mathbb{P}(E_1 \wedge \cdots \wedge E_k) = \frac{1}{k!}\mathbb{E}_{n,m}\xi(\xi-1)\cdots(\xi-k+1),$$

which finishes the proof. □

The term $\mathbb{E}\xi(\xi-1)$ denotes the expected number of pairs of minimal contradictory components, and requires going through different possible cases of their mutual configuration. Further terms will provide even more complicated combinatorial structures, but on the asymptotic level, they will all appear to be negligible in the subcritical phase of the 2-SAT phase transition.

Example 7.2.1. Figure 7.8 representing a contradictory component of excess 2 can be also considered as a pair of contradictory components each of excess 1: the first one being $x \rightsquigarrow w \rightsquigarrow \overline{x} \rightsquigarrow \overline{z} \rightsquigarrow y \rightsquigarrow \overline{y} \rightsquigarrow z \rightsquigarrow x$ with a double sequence $y \rightsquigarrow \overline{y}$ and a mirror path $x \rightsquigarrow \overline{w} \rightsquigarrow \overline{x}$, and the second one being described as $z \rightsquigarrow w \rightsquigarrow \overline{x} \rightsquigarrow \overline{z} \rightsquigarrow y \rightsquigarrow \overline{y} \rightsquigarrow z$ with a double path $y \rightsquigarrow \overline{y}$ and a complementary sequence $z \rightsquigarrow x \rightsquigarrow \overline{w} \rightsquigarrow \overline{z}$. In such a way, the first component of excess 1 is obtained by dropping the pair $x \rightsquigarrow w$ and $\overline{w} \rightsquigarrow \overline{z}$, and the second – by dropping $x \rightsquigarrow w$ and $\overline{w} \rightsquigarrow \overline{x}$.

7.2.3 The case of the simplest minimal contradictory component

It is convenient to represent ξ from Lemma 7.2.2 as the sum $\xi = \sum_{r \geqslant 1} \xi_r$ where ξ_r is the number of distinguished minimal contradictory components of excess r in a random implication digraph.

We focus on the case $r = 1$ first. The expected value of ξ_1 is equal to the number of implication digraphs with a distinguished contradictory component of excess 1 (e.g. Figure 7.6) divided by the total number of implication digraphs. In order to count the implication digraphs with distinguished contradictory components, we are going to count their respective sum-representations and then divide by 2^m which is the number of sum-representations of a single implication digraph with m edges. Both numerator and denominator of the total fraction are then divided by the same factor 2^m which cancels out. In other words, $\mathbb{E}\xi_1$ can be expressed as

$$\mathbb{E}\xi_r = \frac{\left|\left\{(G,p) \;\middle|\; \begin{smallmatrix} G \text{ is a sum-representation digraph with } 2n \text{ vertices and } m \text{ edges} \\ p \text{ is a minimal contradictory pattern of excess } r \text{ in } G + \overline{G} \end{smallmatrix}\right\}\right|}{\left|\{G \mid G \text{ is a sum-representation digraph with } 2n \text{ vertices and } m \text{ edges}\}\right|}. \tag{7.2.2}$$

We further note that distinguishing a contradictory pattern in $G + \overline{G}$ can be replaced with distinguishing a different contradictory pattern in a sum-representation digraph G if a proper multiplicity factor is used.

There are only two possible multigraphs which can result in cancelling of a contradictory digraph of excess 1, depicted, respectively, in Figures 7.5 and 7.6. Accordingly, we consider two different cases.

First case. Consider a sum-representation digraph with a distinguished pattern of type depicted in Figure 7.9. Each arrow there represents a sequence of strictly distinct literals, and it is also required that all the literals on both sequences (except x and $\overline{x}$) are pairwise strictly distinct, and that the two distinguished literals marked x and $\overline{x}$ are complementary. Suppose that the pattern has ℓ oriented edges. Then, among 2^ℓ graphs obtained by edge rotations of the pattern, there are exactly 4 which belong to the same pattern: it is possible to take a complementary of the path $x \rightsquigarrow \overline{x}$ which gives 2 possibilities, and the complementary of the path $\overline{x} \rightsquigarrow x$.

Second case. Consider a sum-representation digraph with a distinguished pattern of type depicted in Figure 7.10. Again, each arrow corresponds to a sequence, and all the literals on the three sequences $x \rightsquigarrow \overline{x}$, $\overline{x} \rightsquigarrow y$, $y \rightsquigarrow \overline{y}$ are pairwise strictly distinct, except for x and $\overline{x}$, and y and $\overline{y}$, which are required to be complementary. The literals x and y are also required to be strictly distinct. Again, if the total length of the pattern is ℓ, there are 8 sum-representations among 2^ℓ equivalent ones, such that these sum-representations form the same pattern. Indeed, 4 options are given by replacing either of the two paths $x \rightsquigarrow \overline{x}$ and $y \rightsquigarrow \overline{y}$ by its complementary. If the path $\overline{x} \rightsquigarrow y$ is replaced by its complementary, then we obtain a different graph constructed of the three sequences $x \rightsquigarrow \overline{x}$, $\overline{y} \rightsquigarrow x$, $y \rightsquigarrow \overline{y}$ which belongs to the same pattern if we swap the variables x and y. No other combinations of edge rotations give the same pattern because it is required that all the literals are pairwise strictly distinct.

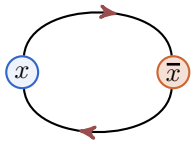


Figure 7.9: First sum-representation contradictory pattern corresponding to Figure 7.5.

Figure 7.10: Second sum-representation contradictory pattern corresponding to Figure 7.6.

Suppose that a pattern (either from Figure 7.9 or from Figure 7.10) has ℓ edges and is a distinguished subgraph of a sum-representation digraph containing m edges in total. Then, among 2^m equivalent, $2^{m-\ell}$ contain the pattern unaltered. Therefore, the number of sum-representation digraphs with a distinguished contradictory pattern of length ℓ from Figures 7.9 and 7.10 counted with multiplicity 2^ℓ enumerates (with respective multiplicities 1-to-4 and 1-to-8) implication digraphs with a distinguished contradictory digraph Figures 7.5 and 7.6 of excess 1.

Combining the two cases, we obtain a new expression for $\mathbb{E}\xi_1$:

$$\mathbb{E}\xi_1 = \frac{1}{4}\cdot\frac{\sum_{\ell=0}^{\infty} 2^\ell \left|\left\{(G,p_\ell) \mid \substack{G \text{ is a sum-representation digraph with } 2n \text{ vertices and } m \text{ edges} \\ p_\ell \text{ is a pattern from Figure 7.9 of length } \ell \text{ in } G}\right\}\right|}{\left|\{G \mid G \text{ is a sum-representation digraph with } 2n \text{ vertices and } m \text{ edges}\}\right|} + \frac{1}{8}\cdot\frac{\sum_{\ell=0}^{\infty} 2^\ell \left|\left\{(G,p_\ell) \mid \substack{G \text{ is a sum-representation digraph with } 2n \text{ vertices and } m \text{ edges} \\ p_\ell \text{ is a pattern from Figure 7.10 of length } \ell \text{ in } G}\right\}\right|}{\left|\{G \mid G \text{ is a sum-representation digraph with } 2n \text{ vertices and } m \text{ edges}\}\right|}.$$

As we shall see in Section 7.3, the contribution of the first summand is negligible. An intuitive way to see why this happens is the following: the shape from Figure 7.5 can be obtained by contracting a path $\overline{y} \rightsquigarrow x$ in Figure 7.6. Therefore, the first case is a degenerate case when the length of the corresponding path is equal to zero and the literal x coincides with the literal $\overline{y}$. In general, only the structures with cubic kernels have dominant contribution.

A similar decomposition can be obtained for $\mathbb{E}\xi_r$ for any finite r. Instead the two presented cases, the sum should be taken over all the possible minimal contradictory components of given excess r, divided by a corresponding multiplicity factor, which plays the analog of a compensation factor for simple graphs. A detailed explanation of the nature of such compensation factors is given in Section 2.2.3.

7.2.4 Incorporating relations between labels of vertices

We continue to use the labelling convention for digraphs and implication digraphs introduced in Section 7.2, so that for digraphs having $2n$ vertices, the labels of the vertices are partitioned into the set of positive literals $\{1, \cdots, n\}$ and negative ones $\{\overline{1}, \cdots, \overline{n}\}$.

Definition 7.2.4 (Literal linking construction)**.** Suppose that in a given family $\mathcal{A}$ of digraphs, each graph $G \in \mathcal{A}$ contains two distinguished *empty nodes*, i.e. vertices for which we do not assign any labels and which do not contribute to the total size of the graph. We define the family $\Lambda[\mathcal{A}]$ as the family obtained from $\mathcal{A}$ by replacing the two empty nodes by complementary literals.

Lemma 7.2.3. If $\mathrm{A}(z)$ is the EGF of a given family $\mathcal{A}$ of digraphs with even number of nodes, two distinguished empty nodes (of size 0), then the EGF of $\Lambda[\mathcal{A}]$ is given by

$$\Lambda[\mathrm{A}](z) := z\int_0^z \mathrm{A}(t)dt. \tag{7.2.3}$$

Proof. The number of graphs of size $(2n-2)$ from $\mathcal{A}$ is $(2n-2)![z^{2n-2}]\,\mathrm{A}(z)$. Then, the number of ways to insert the complementary labels into two distinguished positions is $2n$. The total number of graphs of size $2n$ from the family $\Lambda[\mathcal{A}]$ is therefore $2n\cdot(2n-2)![z^{2n}]z^2A(z) = \frac{1}{2n-1}(2n)![z^{2n-2}]\,\mathrm{A}(z)$. We conclude by noting that $(2n)$-th coefficient of $z\int_0^z \mathrm{A}(t)dt$ equals to $\frac{1}{2n-1}[z^{2n-2}]\,\mathrm{A}(z)$ and multiplying by $(2n)!$ gives the total number of graphs of size $2n$ from $\Lambda[\mathcal{A}]$. □

Remark 7.2.1. It is possible to insert more than one pair of complementary literals, in this case more than one application of Λ is required. Naturally, in this case, it is required that the number of unlabelled empty slots should be equal to two times the number of inserted complementary literals.

7.3 Extracting the asymptotics

7.3.1 Contradictory components in simple digraphs

Before constructing the sum-representation digraphs with marked directed subgraphs, we start by constructing the subcritical *simple* digraphs first. We are going to mark the same contradictory patterns, but without an additional assumption that the paths are strictly distinct and that the digraph is conflict free, and without excluding the edges $x \to \overline{x}$.

As we will see later in Section 7.3.2, this set has a very similar structure to the set of sum-representation digraphs $\mathcal{D}^{\circ}(2n, m)$: the number of edge conflicts in a random digraph $D \in \mathcal{D}(2n, m)$ asymptotically follows a Poisson distribution with parameter $1/8$. Excluding edge conflicts is equivalent to conditioning on the value of this Poisson variable being zero.

We start by demonstrating a result directly related to the probability of satisfiability in the subcritical phase, which is shown by Kim in [Kim08] to be asymptotically

$$\mathbb{P}(F \in \mathcal{F}(n, m) \text{ is SAT}) = 1 - \frac{1 + o(1)}{16|\mu|^3}$$

for $m = n(1 + \mu n^{-1/3})$, as $\mu \to -\infty$ with n, and $|\mu| = o(n^{1/3})$. In this section we give a new explanation of the factor $1/16$ and show how to extend this result to obtain a complete asymptotic distribution in powers of $|\mu|^{-3}$.

The contradictory pattern that we are going to identify first, takes a form of three sequences $x \rightsquigarrow \overline{x} \rightsquigarrow y \rightsquigarrow \overline{y}$ (Figure 7.10). We start by considering the case when this pattern is a part of a weakly connected component which is a tree (Figure 7.11).

Lemma 7.3.1. Consider simple digraphs $G \in \mathcal{D}(2n, m)$ with a distinguished contradictory pattern $x \rightsquigarrow \overline{x} \rightsquigarrow y \rightsquigarrow \overline{y}$ such that the weakly connected component containing this pattern is a tree, and there are no weakly connected components whose non-oriented projections have positive excess. The proportion of such digraphs among $\mathcal{D}(2n, m)$ where each such digraph is taken with weight 2^{ℓ}, where ℓ is the length of the pattern, is

$$\frac{(2n)!}{|\mathcal{D}(2n, m)|}[z^{2n}]z \int z \int \left[\frac{U_{\to}(z)^{2n-m-1}}{(2n-m-1)!} e^{V_{\to}(z)} \frac{8T_{\to}(z)^4 z^{-4}}{(1-2T_{\to}(z))^3}\right] dz^2 = \frac{1}{2|\mu|^3} + O(|\mu|^{-6}). \tag{7.3.1}$$

Proof. The weakly connected component containing a distinguished pattern $x \rightsquigarrow \overline{x} \rightsquigarrow y \rightsquigarrow \overline{y}$ can be represented as three sequences of trees, each sequence of length at least one, and an additional tree (Figure 7.11). The generating function for one such sequence, equipped with a weight 2^{ℓ}, where ℓ is the length of this sequence, is

$$\sum_{\ell \geqslant 1} T_{\to}(z)^{\ell} 2^{\ell} = \frac{2T_{\to}(z)}{1 - 2T_{\to}(z)}.$$

Taking three such sequences and adding a triple multiple 2 corresponding to linking three directed edges between the sequences, and by adding the last tree, we get an EGF $\frac{8T_{\to}(z)^4}{(1-2T_{\to}(z))^3}$. The next step is to divide by z^4 which corresponds to erasing the labels of the four distinguished nodes (which will later become the labels x, $\overline{x}$, y, and $\overline{y}$).

The digraphs required in the current lemma are obtained as product of the family of directed trees and unicycles and the constructed contradictory pattern, through literal linking construction (see Definition 7.2.4). By applying the complementary label insertion operation $\Lambda[A](z) = z\int A(z)dz$ from Lemma 7.2.3 twice, we obtain the desired EGF. The (weighted) number of digraphs is obtained by taking the $(2n)$th coefficient of the EGF.

Since the number of trees in the forest of directed trees (without the distinguished one) is $(2n - m - 1)$, the EGF of digraphs containing this forest and set of directed unicycles is then

$$\frac{U_{\to}(z)^{2n-m-1}}{(2n-m-1)!} e^{V_{\to}(z)},$$

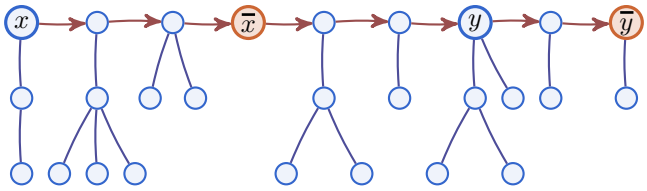


Figure 7.11: The case when the weakly connected component of the contradictory pattern is a tree.

and multiplication by the EGF of the distinguished tree with removed nodes $\frac{8T_{\to}(z)^4 z^{-4}}{(1-2T_{\to}(z))^3}$ and double application of the complementary label insertion operator finishes the construction.

The asymptotic analysis of the expression is done in two steps. Firstly, we get rid of the operator Λ by using the properties

$$[z^n]zF(z) = [z^{n-1}]F(z), \quad [z^n]\int F(z)dz = \frac{1}{n}[z^{n-1}]F(z).$$

We then obtain

$$\begin{aligned}
&[z^{2n}]z\int z\int\left[\frac{U_{\to}(z)^{2n-m-1}}{(2n-m-1)!}e^{V_{\to}(z)}\frac{8T_{\to}(z)^4z^{-4}}{(1-2T_{\to}(z))^3}\right]dz^2\\
&=\frac{1}{(2n-1)(2n-3)}[z^{2n-4}]\frac{U_{\to}(z)^{2n-m-1}}{(2n-m-1)!}e^{V_{\to}(z)}\frac{8T_{\to}(z)^4z^{-4}}{(1-2T_{\to}(z))^3}\\
&=\frac{1}{(2n-1)(2n-3)}[z^{2n}]\frac{U_{\to}(z)^{2n-m-1}}{(2n-m-1)!}e^{V_{\to}(z)}\frac{8T_{\to}(z)^4}{(1-2T_{\to}(z))^3}.
\end{aligned}$$

The second step is to apply Lemma 3.2.2 taking $H(t)$ such that $H(T_{\to}(z)) = 8T_{\to}(z)^4(2n-m)U_{\to}(z)^{-1}$, $y = 3$. Since $U_{\to}(z) = T_{\to}(z) - T_{\to}(z)^2$, the resulting value $H(1/2)$ will be $8/16 \cdot (2n-m)\cdot 4 \sim 2n$ when $m/n \to 1$. An application of the lemma gives

$$\begin{aligned}
&\frac{(2n)!}{|\mathcal{D}(2n,m)|}\frac{1}{(2n-1)(2n-3)}[z^{2n}]\frac{U_{\to}(z)^{2n-m}}{(2n-m)!}e^{V_{\to}(z)}\frac{H(T_{\to}(z))}{(1-2T_{\to}(z))^3}\\
&=\frac{2n(1+\mu n^{-1/3})}{(2n-1)(2n-3)}\left(\frac{n^{1/3}}{|\mu|}\right)^3(1+O(|\mu|^{-3})) = \frac{1}{2|\mu|^3}+O(|\mu|^{-6}).
\end{aligned}$$

□

The weakly connected component containing the pattern $x \rightsquigarrow \overline{x} \rightsquigarrow y \rightsquigarrow \overline{y}$ may happen to be a unicycle (see Figure 7.12), or a complex component of excess at least 1 (see Figure 7.13). The constructions analogous to Lemma 7.3.1 counting simple digraphs with a distinguished contradictory pattern weighted as 2^ℓ, where ℓ is the length of the pattern, yield asymptotics of order $O(|\mu|^{-6})$ or smaller. This concept can be illustrated by the unicyclic case.

Instead of the three sequences with the generating function $\frac{1}{(1-2T_{\to}(z))^3}$ we consider 6 sequences (possibly empty). There are several different mutual configurations of the distinguished contradictory pattern and the unicycle which contains this pattern, but all can be described by 6 sequences. For one of these sequences the directions of the edges are not fixed, but its length is not accounted in the weight of the graph. Therefore, the generating function for such a sequence is also proportional to $\frac{1}{1-2T_{\to}(z)}$, where a factor 2 in front of $T_{\to}(z)$ stands for the choice of orientation of each edge. The number of trees in such a graph remains $(2n-m)$, and so, the contribution obtained from Lemma 3.2.2 has order $\Theta\left(\frac{1}{(2n-1)(2n-3)}\cdot\frac{n^2}{|\mu|^6}\right) = \Theta(|\mu|^{-6})$.

Remark 7.3.1. The same type of reasoning is used to show that the terms $\mathbb{E}\xi_r$ with $r \geqslant 2$ and $\mathbb{E}\xi(\xi-1)\cdots(\xi-k+1)$ for $k \geqslant 2$ do not contribute to the term of order $\Theta(|\mu|^{-3})$ and start contributing from $\Theta(|\mu|^{-6})$. Refinement of this technique using the complete asymptotic expansion of the saddle point lemma (see the

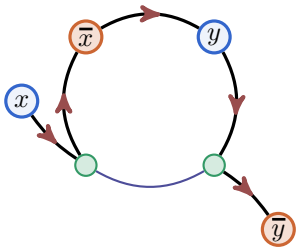


Figure 7.12: Contradictory path in component of excess 0.

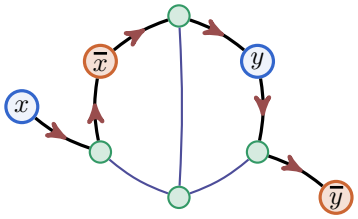


Figure 7.13: Contradictory path in component of excess 1.

refinement of Lemma 3.2.1 in Section 3.3) can be used to obtain the complete asymptotic expansions of the probability of satisfiability in powers of $|\mu|^{-3}$. The problem of enumeration of the mutual combinations of tuples of minimal contradictory components inside an implication digraph, in order to compute the factorial moments of ξ, seems to be a challenging task.

7.3.2 From simple digraphs to sum-representations

The two details peculiar to sum-representations that were not treated in the previous section are the following: firstly, all the literals of the paths $x \rightsquigarrow \overline{x}$, $\overline{x} \rightsquigarrow y$, $y \rightsquigarrow \overline{y}$ of the distinguished pattern, except x and $\overline{x}$, and y and $\overline{y}$, should be pairwise strictly distinct; secondly, there should be no edge conflicts, i.e. pairs of complementary edges. Note that a presence of an edge $x \to \overline{x}$ automatically induces a conflict, so there should be no such edges as well.

Excluding these cases requires inclusion-exclusion: in order to count the instances not containing conflicts and complementary literal pairs, we count the instances with distinguished conflicts and distinguished complementary literal pairs, and then take the alternating sum over them. The two inclusions-exclusions can be done independently.

Lemma 7.3.2 (Excluding complementary literals on the paths)**.** Among the digraphs with a distinguished pattern $x \rightsquigarrow \overline{x} \rightsquigarrow y \rightsquigarrow \overline{y}$, taken with weight 2 to the power of the length of the pattern, the asymptotic proportion of digraphs in which there exist two literals on the pattern which are not strictly distinct (except the pairs x and $\overline{x}$, and y and $\overline{y}$), is only $O(n^{-1/3}|\mu|^{-2})$.

Proof. Let a random variable X denote the number of pairs of complementary literals on the distinguished pattern, not counting x and $\overline{x}$, and y and $\overline{y}$. Using the inclusion-exclusion principle, we can express the probability of the event $[X = 0]$ as

$$\mathbb{P}(X = 0) = 1 - \mathbb{E}X + \frac{1}{2!}\mathbb{E}X(X-1) - \cdots ,$$

where $\mathbb{E}X$ corresponds to the proportion of digraphs having a distinguished pattern and a distinguished additional pair of complementary literals z and $\overline{z}$. Further terms correspond to distinguishing several pairs of complementary literals, etc.

By marking two complementary literals in the case when a contradictory pattern is a tree (see Lemma 7.3.1 and Figure 7.11), we obtain 5 sequences instead of 3, and we need three applications of the label insertion operator Λ instead of just two. We then obtain a generating function

$$F(z) = \Lambda^3\left(\frac{U_{\to}(z)^{2n-m-1}}{(2n-m-1)!}e^{V_{\to}(z)}\frac{2^5 T_{\to}(z)^6 z^{-6}}{(1-2T_{\to}(z))^5}\right) \tag{7.3.2}$$

and after extracting the asymptotics using Lemma 3.2.2, we obtain

$$\mathbb{E}X = \frac{[z^{2n}]\Lambda^3\left(\frac{U_{\to}(z)^{2n-m-1}}{(2n-m-1)!}e^{V_{\to}(z)}\frac{2^5T_{\to}(z)^6z^{-6}}{(1-2T_{\to}(z))^5}\right)}{[z^{2n}]\Lambda^2\left(\frac{U_{\to}(z)^{2n-m-1}}{(2n-m-1)!}e^{V_{\to}(z)}\frac{2^3T_{\to}(z)^4z^{-4}}{(1-2T_{\to}(z))^3}\right)} = \Theta(n^{-1/3}|\mu|^{-2})$$

The same factor $n^{-1/3}$ multiplied by a polynomial of $|\mu|^{-1}$ and $n^{-1/3}$ appears when the weakly connected component containing the contradictory pattern is not necessarily a tree. Further factorial moments of X will have respective contributions of order $n^{-k/3}$, $k \geqslant 2$. From [Jan+93] it is known that the components with infinite excess appear with an asymptotic zero probability, therefore, the interiors of the distinguished paths have pairwise strictly distinct literals with a probability of $1 - O(n^{-1/3}|\mu|^{-2})$. □

As opposed to the first inclusion-exclusion on the number of complementary literals in the distinguished pattern, it can be shown that the number of complementary edge pairs in the whole graph follows a limiting Poisson distribution with parameter $1/8$, so the absence of conflicting edges comes at a positive probability $e^{-1/8}$. In this paper, we focus only on the probability of not having edge conflicts.

Lemma 7.3.3 (Excluding edge conflicts)**.** With an asymptotic probability $e^{-1/8}$, the digraphs from $\mathcal{D}(2n, m)$ with a distinguished pattern $x \rightsquigarrow \overline{x} \rightsquigarrow y \rightsquigarrow \overline{y}$ weighted according to 2 to the power of its length, are conflict-free.

Proof. Following the same inclusion-exclusion principle, we introduce a random variable X denoting the number of conflicting pairs of edges. The probability of the event $[X = 0]$ can be expressed using the inclusion-exclusion principle.

Let us start by computing $\mathbb{E}X$, i.e. the expected number of the conflicting edges. According to the linearity of the mathematical expectation, $\mathbb{E}X$ equals to the total number of possible complementary edge pairs times the probability that one distinguished edge pair, say $1 \to 2$, $\overline{2} \to \overline{1}$ is involved in a conflict.

Using Lemma 7.2.3, we construct the EGF for the family of digraphs containing a distinguished conflict $x \to y$, $\overline{y} \to \overline{x}$ by inserting two pairs of labels $x, \overline{x}$ and $y, \overline{y}$ at the free slots. This insertion can happen in several different ways: each of the edges can belong to the forest of directed trees, to the set of unicycles, and to the complex component (or to the component containing the distinguished pattern $x \rightsquigarrow \overline{x} \rightsquigarrow y \rightsquigarrow \overline{y}$).

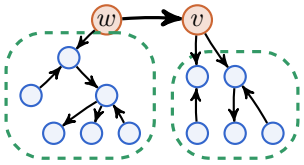


Figure 7.14: Digraph tree with a marked edge and two empty slots.

First, consider the case when each of the distinguished edges belongs to a distinct tree in a forest of directed trees. We denote the random variable corresponding to the number of such occurrences as X_1 which we will show to be asymptotically equivalent to X. A tree containing a distinguished edge $v \to w$ can be represented as a pair of trees with removed roots (which will be later filled by certain labels), see Figure 7.14. The corresponding EGF for digraphs with two distinguished trees containing edges $v \to w$, $\overline{w} \to \overline{v}$, and containing a distinguished pattern $x \rightsquigarrow \overline{x} \rightsquigarrow y \rightsquigarrow \overline{y}$ weighted as 2 to the power of the length of the pattern, is then equal to

$$\Lambda^4\left[\frac{1}{2!}\left(\frac{T_{\to}(z)}{z}\right)^4\frac{U_{\to}(z)^{2n-m-3}}{(2n-m-3)!}e^{V_{\to}(z)}\frac{8T_{\to}(z)^4z^{-4}}{(1-2T_{\to}(z))^3}\right], \tag{7.3.3}$$

where two of the applications of the operator Λ stand for the insertion of the label pairs v and $\overline{v}$, and w and $\overline{w}$ for the distinguished conflict edge pair, and another two applications for the insertion of complementary

literals x and $\overline{x}$, and y and $\overline{y}$ in a distinguished pattern. The term $1/2 \cdot T_{\to}(z)^4 z^{-4}$ is an EGF of the set of two trees with a distinguished edge. The cardinality $(2n-m)$ of the forest of trees is then reduced by 3.

The expected value of X_1 is then expressed as

$$\mathbb{E}X_1 = \frac{[z^{2n}]\Lambda^4 \left(\frac{1}{2!} \left(\frac{T_{\to}(z)}{z} \right)^4 \frac{U_{\to}(z)^{2n-m-3}}{(2n-m-3)!} e^{V_{\to}(z)} \frac{8T_{\to}(z)^4 z^{-4}}{(1-2T_{\to}(z))^3} \right)}{[z^{2n}]\Lambda^2 \left(\frac{U_{\to}(z)^{2n-m-1}}{(2n-m-1)!} e^{V_{\to}(z)} \frac{8T_{\to}(z)^4 z^{-4}}{(1-2T_{\to}(z))^3} \right)}.$$

This expression can be readily analysed by Lemma 7.2.3. Using the property that the exchange of the operator $[z^{2n}]$ with Λ is done according to the rule $[z^{2n}]\Lambda F(z) = \frac{1}{2n-1}[z^{2n}]z^2 F(z)$, we conclude that the unique multiples in the numerator of $\mathbb{E}X_1$, if carefully counted, appear as follows: (i) a multiple asymptotically equivalent to $(2n)^{-2}$ from the double application of Λ and its exchange with $[z^{2n}]$; (ii) a multiple $1/2$ because the two trees with distinguished conflict form a set; (iii) a multiple 2^{-4} appear from $T_{\to}(z)^4$; (iv) a multiple 4^2 appears from $U_{\to}(z)^2$; (v) a multiple n^2 from the change in the factorial $(2n-m-1)!$ of the EGF of forest; (vi) all the occurrences of z cancel out. Collecting the powers of two, we finally obtain

$$\mathbb{E}X_1 \sim 2^{-2-1-4+4} = \frac{1}{8}.$$

So far, we have considered only one particular case, but most of the work has been actually already done.

Without caring too much for the exact coefficients in the asymptotics, let us count the expected values of the following random variables: (i) X_2 for conflicting pairs of edges which are located in the same directed tree; (ii) X_3 for conflicting pairs, one located in a tree, another located in a unicycle (or a complex component); (iii) X_4 for conflicting pairs, neither is located in the forest. For $\mathbb{E}X_2$, for the first conflicting edge we need to fix a pair of rooted trees, and assuming without loss of generality (but with a loss of the exact multiplicative constant) that the second conflicting edge is inside the first tree, we obtain a path from this edge to the root. Therefore, the EGF for such objects is proportional, up to a constant, to $\frac{1}{1-2T_{\to}(z)}(T_{\to}(z)/z)^4$, which corresponds to four trees without roots and a sequence of trees, each term of the sequence equipped with a weight 2 corresponding to the choice of the orientation of the edge in the sequence. Plugging it into the EGF of counted graphs yields

$$\mathbb{E}X_2 \sim \frac{C_2}{\Theta(|\mu|^{-3})}[z^{2n}]\Lambda^4 \left(\frac{1}{1-2T_{\to}(z)} \left(\frac{T_{\to}(z)}{z} \right)^4 \frac{U_{\to}(z)^{2n-m-2}}{(2n-m-2)!} e^{V_{\to}(z)} \frac{8T_{\to}(z)^4 z^{-4}}{(1-2T_{\to}(z))^3} \right). \tag{7.3.4}$$

We immediately notice that using fewer than two trees costs an asymptotic multiple of n because of the factorial in the denominator for the EGF of forest $U_{\to}(z)^{2n-m}/(2n-m)!$, while we gain only a multiple $n^{1/3}$ for an additional sequence constructor. Thus, $\mathbb{E}X_2$ is negligible by a factor $n^{2/3}$. The same story happens again for X_3 and X_4 since in all these cases fewer than two trees are used, and for each lost multiple n it is possible to obtain at most two sequence constructions, contributing each $n^{1/3}$.

Again, the same principle is applicable for computing the higher moments: the most probable situation is to have all the conflicting edges separately in distinct trees (otherwise the contribution will be negligible by a factor at least $n^{2/3}$), and in this case,

$$\frac{1}{k!}\mathbb{E}X(X-1)\cdots(X-k+1) \sim \frac{1}{8^k k!},$$

and therefore, as $n \to \infty$, $\mathbb{P}(X=0) \to \sum_{k \geqslant 0} \frac{(-1)^k}{8^k k!} = e^{-1/8}$. □

Corollary 7.3.1 (Formulation of the counting technique)**.** Note that, by simple asymptotic analysis, $\frac{|\mathcal{D}(2n,m)|}{|\mathcal{D}^{\circ}(2n,m)|} \to e^{1/8}$. Hence, by following Lemma 7.3.2, we obtain that the asymptotic probabilities obtained by counting the (weighted) patterns in *simple digraphs*, without excluding conflicting edges and without taking care of strict distinctness of the paths, are equal to the asymptotic probabilities that sum-representation digraphs contain such (weighted) patterns.

Remark 7.3.2. As can be seen from the proof, the same technique can be applied to the situations when instead of the pattern $x \rightsquigarrow \overline{x} \rightsquigarrow y \rightsquigarrow \overline{y}$ a different pattern is distinguished. This can be either used for the case when the expectations of ξ_r are counted, or when further factorial moments of ξ are being considered. In all such cases, the edge conflicts will most probably appear in the forest component, and excluding such conflicts will always give a multiple $e^{-1/8}$.

Remark 7.3.3. Using the same techniques and the inclusion-exclusion method, it is possible to consider the 2-SAT model where some of the conditions (C1)–(C3) are violated. Accordingly, in such models the loops or multiple edges may appear. The number of the counted graphs will then be multiplied by a certain constant appearing from inclusion-exclusion or from a different symbolic construction, while the total number of graphs in the denominator will be coincidentally multiplied by the same constant. Therefore, the probability of satisfiability and its asymptotic expansion will remain unchanged under the different models.

7.3.3 The structure of contradictory components

Let us recall that the contradictory component in the implication digraph, defined as the set of contradictory variables $x \rightsquigarrow \overline{x} \rightsquigarrow x$, forms a set of strongly connected components, see Remark 2.2.1. The following theorem gives a description of the contradictory component in the subcritical phase.

Theorem 7.3.1. Suppose that $m = n(1 + \mu n^{-1/3})$, and $\mu \to -\infty$ with n while remaining $|\mu| \leqslant n^{1/12}$. The contradictory component of an implication digraph corresponding to a random formula $F \in \mathcal{F}(n, m)$ has an excess r with probability

$$\mathbb{P}(\text{excess of the contradictory component} = r) = C_r |\mu|^{-3r}(1 + O(|\mu|^{-3})). \tag{7.3.5}$$

Moreover, for every finite r, the kernel of this contradictory component is cubic with probability $1-O(n^{-1/3})$. The coefficient C_r is equal to the sum of $\sum_M 2^{-r}\varkappa(M)/(2r)!$ taken over all possible labelled cubic contradictory components of excess r.

Proof. Let ξ_r denote the random variable which equals to the number of contradictory components of excess r (each component is not necessarily connected and the different components may possibly overlap). Using a manipulation with formal power series that can be interpreted as a variation of the inclusion-exclusion principle, we can express the probability that ξ_r equals 1.

Let $F(x) := \sum_{k\geqslant 0} \mathbb{P}(\xi_r = k)x^k$ be the probability generating function (PGF) of ξ_r. Then, $F^{(k)}(z) = \mathbb{E}\xi_r(\xi_r - 1)\cdots(\xi_r - k + 1)$. Using Taylor series expansion at $x = 1$, provided that $F(x)$ is analytic in a circle of radius greater than 1, we obtain

$$F(x) = \sum_{k\geqslant 0} \frac{\mathbb{E}\xi_r(\xi_r - 1)\cdots(\xi_r - k + 1)}{k!}(x - 1)^k. \tag{7.3.6}$$

Using the fact that $\mathbb{P}(\xi_r = 1) = F'(0)$, we obtain the expression

$$\mathbb{P}(\xi_r = 1) = \frac{\mathbb{E}\xi_r}{0!} - \frac{\mathbb{E}\xi_r(\xi_r - 1)}{1!} + \frac{\mathbb{E}\xi_r(\xi_r - 1)(\xi_r - 2)}{2!} - \cdots . \tag{7.3.7}$$

Here, $\mathbb{E}\xi_r$ can be rewritten as the number of implication digraphs with a distinguished contradictory component of excess r, $\mathbb{E}\xi_r(\xi_r - 1)$ is the number of implication digraphs with a distinguished pair of contradictory components, etc. The distinguished pair of contradictory components forms a contradictory component by itself, of excess at least $(r + 1)$, as e.g. in Example 7.2.1.

Let us choose a contradictory component $\mathcal{C}$ of excess r. We choose a *contradictory pattern* π which is chosen by taking an arbitrary sum-representation of the kernel of $\mathcal{C}$, so that no two complementary edges are chosen. Assume that all the isolated vertices are not included into the pattern, so that there might appear distinguished vertices that do not have their complementaries in π. Finally, every edge of π is replaced by a sequence of directed trees to obtain a directed weakly connected component $\mathcal{P}$. By combining the reasoning

of Section 7.2.3 and using Lemma 2.2.1, we conclude that $\mathbb{E}\xi_r$ is asymptotically worth the number of all sum-representation digraphs $\mathcal{D}^\circ(2n,m)$ containing $\mathcal{P}$, counted with weight 2^ℓ, where ℓ is the number of edges of π, and divided by the compensation factor of $\mathcal{C}$. Note that such a weakly connected component $\mathcal{P}$ may be a part of a larger weakly connected component in a sum-representation digraph, so it is required to take the sum over all possible weakly connected components containing $\mathcal{P}$.

Let us define the following quantities: (i) $\tau(\pi)$ equal to the number of directed edges of π; (ii) $\nu(\pi)$ equal to the number of pairs of complementary variables in π; (iii) $\varphi(\pi)$ equal to the number of literals that do not have their complementaries in π. Then, the number of Boolean variables in $\mathcal{C}$ equals $\nu(\pi)+\varphi(\pi)$, and the excess r then equals $r=\tau(\pi)-\varphi(\pi)-\nu(\pi)$. At the same time, the excess of the unoriented projection of π (in the sense of simple graphs) is equal to $\tau(\pi)-\varphi(\pi)-2\nu(\pi)=r-\nu(\pi)$.

The EGF $f_{\mathcal{P}}(z)$ for digraphs $D\in\mathcal{D}(2n,m)$ containing $\mathcal{P}$ as a separate weakly connected components, can be expressed as

$$f_{\mathcal{P}}(z)=\frac{1}{\varkappa(\pi)}\Lambda^{\nu}\left(\frac{U_{\to}(z)^{2m-n+(r-\nu(\pi))}}{(2m-n+r-\nu(\pi))!}e^{V_{\to}(z)}\frac{(T_{\to}(z))^{\varphi(\pi)+2\nu(\pi)}z^{-2\nu(\pi)}2^{\tau(\pi)}}{(1-2T_{\to}(z))^{\tau}}\right). \tag{7.3.8}$$

In the case above, each of the literals of π is coloured into a separate distinguished colour, and only the compensation factor of π is considered.

By analysing the asymptotics of $\dfrac{(2n)!}{\mathcal{D}(2n,m)}[z^{2n}]f_{\mathcal{P}}(z)$ similarly to Lemma 7.3.1, we obtain

$$\frac{(2n)!}{\mathcal{D}(2n,m)}[z^{2n}]f_{\mathcal{P}}(z)\sim\frac{1}{\varkappa(\pi)}\frac{1}{(2n)^{\nu(\pi)}}4^{\nu(\pi)-r}n^{\nu(\pi)-r}2^{-\varphi(\pi)-2\nu(\pi)}2^{\tau(\pi)}n^{\tau(\pi)/3}|\mu|^{-\tau(\pi)}.$$

By using the relation $r=\tau(\pi)-\varphi(\pi)-\nu(\pi)$, we obtain

$$\frac{(2n)!}{\mathcal{D}(2n,m)}[z^{2n}]f_{\mathcal{P}}(z)\sim n^{\tau(\pi)/3-r}|\mu|^{-\tau(\pi)}\frac{1}{\varkappa(\pi)}2^{-r}.$$

If the kernel of the component $\mathcal{C}$ is not a cubic multigraph, then $r>\tau(\pi)/3$, and the contributions of such terms are negligible. Otherwise, $\tau(\pi)=3r$. In the case when $\mathcal{P}$ is a part of a larger weakly connected component of higher excess, this results of asymptotic of order $|\mu|^{-3r-3}$ which is negligible compared to $|\mu|^{-3r}$.

Finally, let us obtain the asymptotics of $\mathbb{E}\xi_r$. By Corollary 7.3.1, enumeration inside simple digraphs gives the same asymptotics as the probability in sum-representations.

The expected value $\mathbb{E}\xi_r$ can be obtained by adding up the contributions of all possible contradictory components $\mathcal{C}$ of excess r whose kernels are cubic. We denote such a contribution by $\mathbb{E}\xi_{\mathcal{C}}$, where $\xi_{\mathcal{C}}$ is the corresponding random variable. Then, recalling the reasoning from Section 7.2.3, and by choosing a corresponding sum-representation π of the kernel of $\mathcal{C}$, by using Lemma 2.2.1 and the fact that a cubic multigraph of excess r contains $2r$ vertices, we express $\mathbb{E}\xi_{\mathcal{C}}$ as

$$\mathbb{E}\xi_{\mathcal{C}}\sim\frac{\varkappa(\mathcal{C})}{(2r)!\varkappa(\pi)}\frac{\sum_{\ell=0}^{\infty}2^{\ell}\left|\left\{(G,p_\ell)\mid {\scriptstyle G\in\mathcal{D}^\circ(2n,m),\ p_\ell\subset G \text{ is obtained from } \pi \text{ by} \atop \text{inserting sequences of trees of total length } \ell}\right\}\right|}{\left|\{G\mid G\in\mathcal{D}^\circ(2n,m)\}\right|}. \tag{7.3.9}$$

Taking the sum over all such $\mathcal{C}$ we obtain the dominant contribution of $\mathbb{E}\xi_r$. Further terms of the inclusion-exclusion for $\mathbb{P}(\xi_r=1)$ have order at most $|\mu|^{-3r-3}$ and are, therefore, negligible by a factor $|\mu|^3$. Collecting the dominant contributions, we conclude that

$$\mathbb{P}(\xi_r=1)\sim\sum_{\substack{\mathcal{C}\text{ of excess }r\\ \text{with cubic}\\ \text{kernels}}}\mathbb{E}\xi_{\mathcal{C}}\sim\sum_{\mathcal{C}}\frac{\varkappa(\mathcal{C})}{2^r(2r)!}|\mu|^{-3r}. \tag{7.3.10}$$

□

We present a different proof of a theorem from [Kim08] using the compensation factors of the contradictory components which comes as a corollary of the above theorem.

Corollary 7.3.2. For a random formula $F \in \mathcal{F}(n,m)$, when $m = n(1+\mu n^{-1/3})$ and $\mu \to -\infty$ with n while remaining $|\mu| \leqslant n^{1/12}$,

$$\mathbb{P}(F \text{ is satisfiable}) = \left(1 - \frac{1}{16|\mu|^3}\right)(1 + O(\mu n^{-1/3} + \mu^{-3})). \tag{7.3.11}$$

Proof. In the subcritical phase, the compensation factor of the only possible cubic contradictory implication multidigraph of excess 1 (viz. Figure 7.6) is equal to $1/4$. Therefore, the probability of having a contradictory component of excess 1 is $\frac{2^{-2}}{2!} \cdot \frac{1}{4}|\mu|^{-3} = \frac{1}{16}|\mu|^{-3}$. The probability of having a contradictory component of higher excess is then $\Theta(|\mu|^{-6})$, and so, is negligible.

□

7.3.4 Number of contradictory variables

Theorem 7.3.2. Let $m = n(1+\mu n^{-1/3})$, $\mu \to -\infty$, $|\mu| \leqslant n^{1/12}$. Assuming that the excess of the contradictory component is r, and this component has a cubic kernel, the number of contradictory variables V_n in a random formula $F \in \mathcal{F}(n,m)$ follows asymptotically a Gamma law with shape parameter $3r$ and scale parameter $n^{1/3}|\mu|^{-1}$, so that

$$\lim_{n\to\infty} \mathbb{P}\Big(V_n = x n^{1/3}|\mu|^{-1} \;\Big|\; \text{excess } = r\Big) = \frac{x^{3r-1}}{\Gamma(3r)} e^{-x}. \tag{7.3.12}$$

Proof. Fix a contradictory component $\mathcal{C}$ of excess r with a cubic kernel. Construct a contradictory pattern π by taking an arbitrary sum-representation of the kernel of $\mathcal{C}$, and replace every oriented edge of π by a sequence of directed trees, thus obtaining a digraph $\mathcal{P}$. Consider an EGF $F_{\mathcal{P}}(z,u)$ for directed graphs with a distinguished weakly connected component $\mathcal{P}$ counted with weight 2^ℓ where ℓ denotes the length of the 2-core of $\mathcal{P}$. The variable u then marks all the contradictory variables on $\mathcal{C}$. Then, using the similar constructions as in Theorem 7.3.1, we express $F_{\mathcal{P}}(z,u)$:

$$F_{\mathcal{P}}(z,u) = \frac{1}{\varkappa(\pi)} \Lambda^{2r} \left(\frac{U_{\to}(z)^{2m-n-r}}{(2m-n-r)!} e^{V_{\to}(z)} \frac{(T_{\to}(z))^{4r} z^{-4r} 2^{3r}}{(1-2uT_{\to}(z))^{3r}} \right). \tag{7.3.13}$$

The expected value of the number of contradictory variables conditioned on this pattern $\mathcal{P}$ is then

$$\mathbb{E}[V_n \mid \mathcal{P}] = \frac{\partial_u [z^{2n}] F_{\mathcal{P}}(z,u)|_{u=1}}{[z^{2n}] F_{\mathcal{P}}(z,1)} \sim 3r \cdot n^{1/3}|\mu|^{-1}, \tag{7.3.14}$$

and more generally, k-th factorial moment can be expressed as

$$\mathbb{E}[V_n \cdots (V_n - k + 1) \mid \mathcal{P}] = \frac{\partial_u^k [z^{2n}] F_{\mathcal{P}}(z,u)|_{u=1}}{[z^{2n}] F_{\mathcal{P}}(z,1)} \sim \frac{\Gamma(3r+k)}{\Gamma(3r)!} (n^{1/3}|\mu|^{-1})^k. \tag{7.3.15}$$

Note that the resulting factorial moments do not depend on the choice of $\mathcal{C}$, therefore, these are also the asymptotic moments of the unconditioned variable V_n.

The sequence of moments of the scaled random variable $\widehat{V}_n := V_n n^{-1/3}|\mu|$ coincides with the sequence of moments of the Gamma distribution with shape parameter $3r$: if its density is $f(x) = x^{3r}e^{-x}/\Gamma(3r)$, then k-th moment is calculated as

$$\int_0^\infty \frac{x^{3r+k-1}}{\Gamma(3r)} e^{-x} dx = \frac{\Gamma(3r+k)}{\Gamma(3r)!}. \tag{7.3.16}$$

By checking Carleman's condition for Stieltjes moment problem on $(0,+\infty)$, we conclude that this distribution is uniquely defined by its moments, which finishes the proof. □

Corollary 7.3.3. Since in the subcritical phase a random 2-CNF has a contradictory component of excess 1 with probability $\frac{1}{16|\mu|^3}$, and components of higher excess with a negligible probability, the distribution of the number of contradictory variables can be approximated by a mixture of a deterministic value 0 with probability $\mathbb{P}(V = 0) = 1 - \frac{1}{16|\mu|^3}$ and of Gamma distribution with parameter 3 and scale $n^{1/3}|\mu|^{-1}$ with probability $\frac{1}{16|\mu|^3}$.

7.3.5 Structure of the spine

In the paper [Bol+01] it is proven that the expected size of the spine in the subcritical phase, i.e. for $m = n(1 + \mu n^{-1/3})$ when $\mu \to -\infty$ with n, is asymptotically $\frac{1}{2}|\mu|^{-2}n^{2/3}$. As proven in Lemma 7.2.1, for every literal y from the spine of a formula $F \in \mathcal{F}(n, m)$ there exists a minimal spinal path of the form $y \rightsquigarrow x \rightsquigarrow \overline{x}$, such that all the internal nodes of the paths $y \rightsquigarrow x$ and $x \rightsquigarrow \overline{x}$ are pairwise strictly distinct. We show that for almost all literals y from the spine such a minimal spinal path is unique.

Theorem 7.3.3. Consider random formulae $F \in \mathcal{F}(n, m)$ in the subcritical phase, $m = n(1 + \mu n^{-1/3})$, $\mu \to -\infty$, $|\mu| \leqslant n^{1/12}$. The expected number of spine variables y that have exactly k unique paths from y to $\overline{y}$ is asymptotically equal to $C_k n^{2/3}|\mu|^{-2-3k}$, where C_k is some algorithmically computable constant. In particular, the expected proportion of spine variables y having a unique path $y \rightsquigarrow \overline{y}$ is $1 - O(|\mu|^{-3})$.

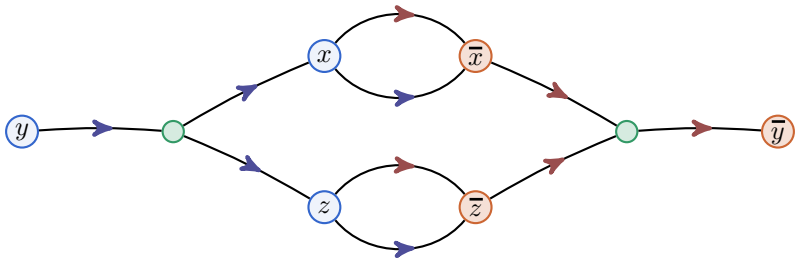


Figure 7.15: One possible configuration of two distinct paths $y \rightsquigarrow \overline{y}$.

Proof. Let a random variable P_n denote the number of spine literals y of a random formula $F \in \mathcal{F}(n, m)$, each variable counted with a multiplicity of the number of paths $y \rightsquigarrow \overline{y}$. We also define $P_n^{(2)}$ which counts the number of spine literals y counted with the multiplicities of the pairs of distinct paths $y \rightsquigarrow \overline{y}$, and similarly $P_n^{(\ell)}$ for ℓ-tuples of distinct paths.

The cardinality of the spine $\mathcal{S}(F)$, i.e. the number of literals y for which there exists at least one such path, can be then counted using the variant of the inclusion-exclusion approach:

$$\mathbb{E}\mathcal{S}(F) = \mathbb{E}P_n - \mathbb{E}P_n^{(2)} + \frac{1}{2!}\mathbb{E}P_n^{(3)} - \frac{1}{3!}\mathbb{E}P_n^{(4)} + \cdots . \tag{7.3.17}$$

In order to prove the theorem, we show that $\mathbb{E}P_n = \Theta(n^{2/3}\mu^{-2})$, $\mathbb{E}P_n^{(2)} = \Theta(n^{2/3}\mu^{-5})$, and, generically, $\mathbb{E}P_n^{(k)} = \Theta(n^{2/3}\mu^{-3k+1})$.

The expected value of minimal spinal paths can be again counted using sum-representations. By distinguishing a pattern $y \rightsquigarrow x \rightsquigarrow \overline{x}$ and counting such graphs with weight 2^ℓ where ℓ is the length of the pattern, we are counting the minimal spinal paths with multiplicity 2, as there are 2 distinct paths $x \rightsquigarrow \overline{x}$.

Knowing that exclusion of the conflicting edges gives the same result as counting the proportions of graphs with a distinguished pattern in simple digraphs (see Section 7.3.2), we pass directly to the counting in simple digraphs. The corresponding EGF for simple digraphs $D \in \mathcal{D}(2n, m)$ with distinguished pattern which forms a weakly connected component which is a tree, is then (taking into account the compensating factor 1/2 arising from the multiplicity mentioned above)

$$\frac{1}{2}\Lambda\left(\frac{U_\to(z)^{2n-m-1}}{(2n-m-1)!}e^{V_\to(z)}4T_\to(z)^3 z^{-2}(1-2T_\to(z))^2\right)$$

and therefore, by applying Lemma 3.2.2, we obtain the expected value of the dominant term

$$\mathbb{E}P_n \sim \frac{1}{2} \cdot \frac{1}{2n} \cdot 4 \cdot n \cdot 4 \cdot \frac{1}{8} n^{2/3} \mu^{-2} = \frac{1}{2} n^{2/3} \mu^{-2}.$$

The above expectation also includes the cases when the excess of the weakly connected component containing the pattern $y \rightsquigarrow x \rightsquigarrow \overline{x}$ is greater than -1, but such cases give contribution at most $O(n^{2/3}\mu^{-5})$ and therefore are negligible.

Similarly, by constructing all possible cases of having two distinct paths between y and $\overline{y}$ in the implication digraph, we note that the dominant contributions come only from the case when vertices of degree 3 are inserted into the core of the pattern, and therefore the pattern is almost cubic (except for the nodes y, x and $\overline{x}$, and the pairs of complementary literals, where each such pair has a summary degree 3). It is easy to show that such components contribute a factor $O(n^{2/3}\mu^{-5})$ (one example of such configuration is given in Figure 7.15). Considering spine literals of further complexity gives the next orders of asymptotics. This observation finishes the proof. □

Corollary 7.3.4. By removing on average a $\Theta(|\mu|^{-3})$ proportion of the spine literals, the spine breaks down into non-intersecting tree-like components, viz. Figure 7.4. For each component there exists a distinct literal x such that every literal y from this component has a unique path to x, and the path $x \rightsquigarrow \overline{x}$ is unique and strictly distinct.

7.4 Conclusions and open problems

While the analysis of random graphs and the curve of their phase transition have been described in details in [Jan+93], there is a certain obstacle which doesn't immediately allow to go inside the critical phase of 2-SAT phase transition using the inclusion-exclusion approach. In [Col+18], a notion of a *patchwork* has been specifically designed for a very similar reason. However, their approach also doesn't suggest any explicit way for obtaining such patchworks. These considerations give rise to the following questions.

Problem 7.4.1. What is the number of cubic strongly connected contradictory multigraph components with given excess? How many minimal contradictory (cubic) multigraphs implication multigraphs are there?

A variation of the above question, directly related to the inclusion-exclusion method, leads to the following question.

Problem 7.4.2. Using a version of [Jan+93, Lemma 3] when $|\mu| = \Theta(1)$, is it possible to express the probability of satisfiability as a *converging* sum of Airy functions $A(y, \mu)$?

Some of the simulations suggested that the introduced notion of the excess correctly captures the discrete nature of the phase transition. According to the results, the distribution of the excess is discrete for finite μ and doesn't depend on n in the limit.

Conjecture 7.4.1. Let ξ denote the total excess of the strongly connected contradictory component in a random implication digraph corresponding to a formula F chosen uniformly at random from $\mathcal{F}(n, m)$. If μ is constant and $m = n(1 + \mu n^{-1/3})$, the distribution of ξ is discrete. Moreover, when μ is constant, the number of strongly connected components of the contradictory component digraph follows a discrete limiting law; when $\mu \to +\infty$, the contradictory component is strongly connected with high probability.

When $\mu \to +\infty$, the expected value of the excess of the complex component in simple graphs is known and is $\frac{2}{3}\mu^3$. Some simulations suggest that exactly the same asymptotics may hold when $\mu \to \infty$ for the expected excess of the contradictory component.

Problem 7.4.3. When $m = n(1 + \mu n^{-1/3})$ and $\mu \to +\infty$, what is the expected excess of the contradictory component of a random 2-CNF formula?

One of the motivations to study the phase transition in 2-SAT is its similarity to the phase transition of the appearance of the strongly connected component in directed graphs. The papers [LS09; PP17] seem to come the closest to resolving the question, but is it possible to give the exact description?

Problem 7.4.4. It is possible to describe a "giant strong component" of a critical directed graph having n vertices and $m = n(1 + \mu n^{-1/3})$ oriented edges in the terms of cubic components and their excesses? Is it then possible to express the probability of having a strong component of excess r in terms of Airy function depending on r?

One of the applications of analysis with the help of generating functions, is the study of 2-SAT with a given set of degree constraints, similarly to [dPR16] and [DR18]. It is clear that the same analysis can be done for the case of formulae with literal set degree constraints, however the present analysis is done only for the subcritical phase.

Conjecture 7.4.2. The point r of the phase transition in the 2-SAT model with literal degree constraints from a set $\Delta = \{d_1, d_2, \ldots\}$ (possibly weighted) can be computed from the system of equations

$$z\frac{\omega''(z)}{\omega'(z)} = 1, \quad z\frac{\omega'(z)}{\omega(z)} = r, \tag{7.4.1}$$

where $\omega(z) := \sum_{d\in\Delta} \frac{z^d}{d!}$ and Δ satisfies the condition $1 \in \Delta$.

Chapter 8

Statistics of closed lambda terms

Contents

8.1 Preliminaries

8.1.1 Lambda calculus

λ-calculus is a theoretical formalism famously equivalent in expressiveness to Turing machines, see [Bar84]. In this calculus, computations are represented as λ-*terms* defined by the formal grammar $T ::= x \mid (\lambda x.T) \mid (T\ T)$ in which x belongs to the countable, infinite alphabet of *variables*; $(\lambda x.T)$ is an *abstraction* of variable x in T; and $(T\ T)$ denotes an *application* of two λ-terms. Given an abstraction $(\lambda x.T)$, occurrences of x in T are said to be *bound*. Unbound variable occurrences are said to occur *freely*.

Lambda terms, intended to represent anonymous functions, are executed by means of the iterated process of β-*reduction*. First, an arbitrary β-*redex* subterm in form of $(\lambda x.N)M$ is selected (if no such subterm exists, computations are terminated). Next, the selected β-redex is replaced with $N[x := M]$, i.e. N in which each occurrence of x is substituted, in a *capture-avoiding* manner, by M. While substituting M for x in N we

have to avoid the unintended situation in which free variable occurrences in M get bound, in other words *captured*, by some abstractions occurring in N. For instance, let $N = (\lambda y.x)$ and $M = y$. The term $(\lambda x.N)M$ should not be reduced to $\lambda y.y$ as, by doing so, the free variable occurrence y gets bound due to a coincidental *clash* with the inner abstraction variable name. Certainly, the arbitrary choice of the formal variable name y should not influence the intended semantics of the represented computation. Following this motivation, λ-terms differing only in bound variable names are considered equivalent (in other words α-*convertible*). In order to avoid potential name clashes, we can therefore *rename* bound variable occurrences before proceeding with β-reduction. Since there is an infinite supply of available variable names, it is always possible to avoid variable captures. Consequently, we can equivalently α-convert $(\lambda x.\lambda y.x)y$ into, say, $(\lambda x.\lambda w.x)y$ and proceed with $(\lambda x.\lambda w.x)y \to_\beta (\lambda w.x)[x := y] = \lambda w.y$.

Though intuitive, explicit variable names pose considerable conceptual and implementation problems. For instance, consider the terms $\lambda x.x$ and $\lambda y.y$. Although syntactically different, both semantically represent the same anonymous identity function as $(\lambda x.x)T \to_\beta T$ and $(\lambda y.y)T \to_\beta T$ for arbitrary T. In order to facilitate automatic computations, de Bruijn proposed a different notation for λ-terms eliminating in effect the troublesome variable names [dBru72]. In his notation, variable occurrences are replaced with *indices* represented as non-negative integers. The intention is to view λ-terms as natural tree-like structures and encode variable occurrences as indices denoting their relative distance to respective variable binders – each index $\underline{\mathsf{n}}$ denotes a variable occurrence x whose relative distance to its binder (in terms of passed lambda symbols) is equal to $n + 1$. For instance, $\underline{\mathsf{0}}$ corresponds to a variable occurrence bound to the nearest abstraction on its unique path to the root in the associated tree-like representation of the considered λ-term. Consequently, α-convertible λ-terms have the same de Bruijn representation. In effect, each λ-term in the de Bruijn notation represents an entire α-equivalence class of λ-terms in the classic variable notation. For instance, both $\lambda x.x$ and $\lambda y.y$, being α-convertible, are represented as $\lambda\underline{\mathsf{0}}$ in the de Bruijn notation. The use of de Bruijn indices significantly simplifies the automatic substitution operation. Due to the lack of explicit variable names, variable captures and name clashes do not pose implementation issues.

Remark 8.1.1. There exists a disagreement in the literature whether to start de Bruijn indices with $\underline{\mathsf{0}}$ or $\underline{\mathsf{1}}$. Although de Bruijn himself assumed the latter [dBru72], some authors follow his footsteps, see e.g [GL13; GL15] whereas others do not, including $\underline{\mathsf{0}}$ in the set of admissible indices, see e.g. [GG16; Ben+17; BT17]. Certainly, neither convention is better than the other. In the current work, we follow the convention of starting de Bruijn indices with $\underline{\mathsf{0}}$ so the keep consistent with the most recent literature.

Definition 8.1.1. Let $\{\underline{\mathsf{0}}, \underline{\mathsf{1}}, \ldots\}$ be an infinite, denumerable set of available indices. Then, the set $\mathcal{L}_\infty$ of *λ-terms in the de Bruijn notation* is defined inductively as follows:

1. Each index $\underline{\mathsf{n}}$ is a λ-term;
2. If N and M are λ-terms, then (NM) is a λ-term;
3. If N is a λ-term, then (λN) is a λ-term.

Following usual notational conventions, we omit outermost parentheses and drop parentheses from left-associated λ-terms. For instance, $\lambda x.\lambda y.\lambda z.((xy)z)$ in the classical variable notation is written as $\lambda\lambda\lambda\underline{\mathsf{210}}$.

An index occurrence $\underline{\mathsf{n}}$ is said to be *bound* in the term N if there exist at least $n + 1$ lambda symbols on the unique path from $\underline{\mathsf{n}}$ to the root of the associated tree-like representation of N, see e.g. Figure 8.1. Otherwise, $\underline{\mathsf{n}}$ is said to be occurring *freely* in N and hence corresponds to a free variable in the classical λ-calculus notation. For convenience, we refer to de Bruijn indices both as indices and variables. If each index occurrence in N is bound, then N is said to be *closed*. Otherwise, it is said to be *open*. And so, $\lambda\lambda\lambda\underline{\mathsf{210}}$ is closed whereas $\lambda\lambda\underline{\mathsf{21}}$ is not as here $\underline{\mathsf{2}}$ is not bound. If prepending N with m lambdas turns it into a closed λ-term, then N is said to be *m-open*. Certainly, if N is m-open, then it is also $(m + 1)$-open. Moreover, 0-open λ-terms correspond exactly to closed λ-terms. Hence, though $\lambda\lambda\underline{\mathsf{21}}$ is not closed, it is 1-open as $\lambda\lambda\lambda\underline{\mathsf{21}}$ is a closed λ-term. Finally, we write that a λ-term is *plain* if we mean to denote that it is either open or closed, without specifying which case holds.

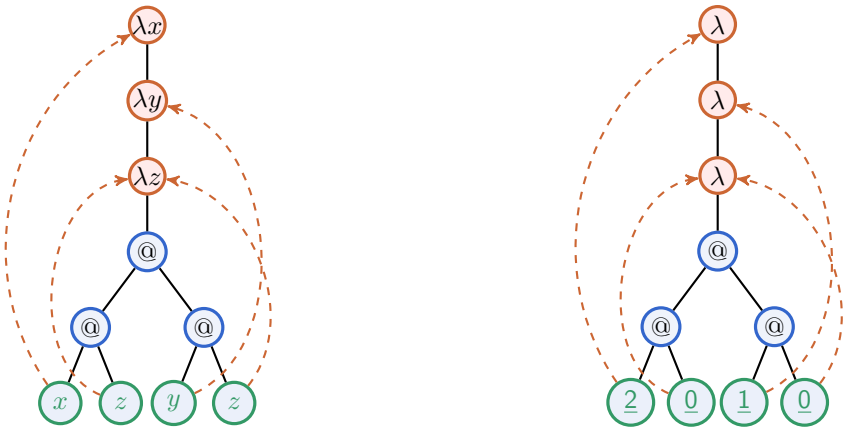


Figure 8.1: Tree-like representations associated with the same example λ-term $\lambda x.\lambda y.\lambda z.xz(yz)$ and its de Bruijn notation variant $\lambda\lambda\lambda\underline{2}(\underline{1}\underline{0})$. Back pointers to abstractions are included for illustrative purposes only.

Enumeration

In the current work we follow [Ben+16; Ben+17; GG16; BGG17] and investigate the statistical properties of random λ-terms in the de Bruijn representation. We assume a unary base encoding of indices, i.e. an encoding in which $\underline{\mathsf{n}}$ is identified with an n-fold application of the *successor* operator S to zero. Formally, the set $\mathcal{L}_\infty$ of λ-terms is described by the following formal grammar:

$$\begin{aligned} \mathcal{L}_\infty &::= \underline{\mathsf{n}} \mid (\lambda\mathcal{L}_\infty) \mid (\mathcal{L}_\infty\mathcal{L}_\infty) \\ \underline{\mathsf{n}} &::= \mathsf{0} \mid \mathsf{S}\ \underline{\mathsf{n}}. \end{aligned} \tag{8.1.1}$$

In order to enumerate λ-terms, we have to assign a formal notion of *size* to each term in such a manner that for each available size n the number of terms of size n is finite. Though various size measures are considered in the literature, most notably the general size model framework of Gittenberger and Gołębiewski [GG16], we assume the simple *natural size notion* [Ben+16] in which the size of T is equal to the total number of abstractions, applications, successors and zeros occurring in T. Formally, we define the size of T inductively as follows:

$$\begin{aligned} |\mathsf{0}| &= 1 & |M\,N| &= |M| + |N| + 1 \\ |\mathsf{S}\ \underline{\mathsf{n}}| &= |\underline{\mathsf{n}}| + 1 & |\lambda M| &= |M| + 1. \end{aligned} \tag{8.1.2}$$

Example 8.1.1. Note that, in general, $\underline{\mathsf{n}}$ is of size $n+1$ as it consists of n successors applied to zero. Consequently, the term $\lambda\lambda\lambda\underline{2}\underline{1}\underline{0}$ is of size 11 as it consists of three λ symbols, two applications between $\underline{2}, \underline{1}$ and $\underline{0}$, by convention omitted in writing, and indices $\underline{0}, \underline{1}, \underline{2}$ of total size six.

Remark 8.1.2. It is worth noticing that, with some minor technical overhead, the analysis presented in the current work extends onto the more general size model framework of Gittenberger and Gołębiewski [GG16] including the assumed natural size notion as a special case. We prefer to avoid technicalities related to the general size notion and so, for the reader's convenience, favour a lucid presentation of the key arguments.

Let $\mathcal{L}_m$ denote the set of m-open λ-terms, see Definition 8.1.1 (plain terms can be viewed as "infinitely" open, hence the ∞ symbol in the subscript of $\mathcal{L}_\infty$). Like plain λ-terms (8.1.1), $\mathcal{L}_m$ can be described in terms of a formal, though now infinite, grammar as follows:

$$\begin{aligned} \mathcal{L}_m &::= (\lambda\mathcal{L}_{m+1}) \mid (\mathcal{L}_m\mathcal{L}_m) \mid \underline{0}, \underline{1}, \ldots, \underline{\mathsf{m}\text{ - }1} \\ \mathcal{L}_{m+1} &::= (\lambda\mathcal{L}_{m+2}) \mid (\mathcal{L}_{m+1}\mathcal{L}_{m+1}) \mid \underline{0}, \underline{1}, \ldots, \underline{\mathsf{m}} \\ &\ldots \qquad \ldots \end{aligned} \tag{8.1.3}$$

An m-open λ-term T can take one of the three forms. Either T is in the form of abstraction followed by an $(m+1)$-open λ-term; or it is an application of two m-open λ-terms; or, finally, T is one of the indices $\underline{0}, \underline{1}, \ldots, \underline{\mathsf{m\text{-}1}}$.

Due to the infinite combinatorial specification (8.1.3) for $\mathcal{L}_m$ standard analytic combinatorics techniques are not readily applicable. Consequently, enumerating closed λ-terms poses a considerable challenge. In [GG16] a partial solution bounding the asymptotic growth rate of the number of m-open λ-terms of size n was proposed. Although both the lower and upper bounds were of the form $C\rho^n n^{-3/2}$, a typical trait of various tree-like structures, the asymptotic growth rate of m-open terms remained open. Remarkably, some time later in their joint paper [BGG17] Bodini, Gittenberger and Gołębiewski closed the remaining gap and confirmed the conjectured $C\rho^n n^{-3/2}$ form of the asymptotic growth of m-open λ-terms. Furthermore, two combinatorial problems related to random closed λ-terms were studied. Specifically, the number of terms with an *a priori* fixed number of abstractions and the number of terms in β-normal form, i.e. without β-redexes. In this context, our contribution is a natural continuation of their work. In addition, we offer a different proof of the asymptotic growth rate of m-open λ-terms.

8.1.2 Analytic tools

We use the standard tools from analytic combinatorics discussed in Part I and also novel machinery for infinite systems of algebraic equations Chapter 4. For our purposes, combinatorial parameter analysis outlines as follows:

- Let $a_{n,k}$ denote the number of plain (closed) lambda terms of size n for which the investigated parameter takes value k. Note that we do not assume that the numbers $a_{n,k}$ are *a priori* known. With such a two-dimensional sequence of numbers we associate a bivariate generating function

$$A(z,u) := \sum_{n,k\geqslant 0} a_{n,k} z^n u^k; \tag{8.1.4}$$

  In order to simultaneously study several different parameters of interest, we introduce *multivariate generating functions* in form of

$$A(z,\boldsymbol{u}) = \sum_{n,\boldsymbol{k}\geqslant 0} a_{n,\boldsymbol{k}} z^n \boldsymbol{u}^{\boldsymbol{k}} \tag{8.1.5}$$

  where $\boldsymbol{u} = (u_1, \ldots, u_d)$ is a d-dimensional variable, $\boldsymbol{k} = (k_1, \ldots, k_d)$ is a d-dimensional index satisfying $k_i \geqslant 0$, $\boldsymbol{u}^{\boldsymbol{k}} := u_1^{k_1} u_2^{k_2} \cdots u_d^{k_d}$, and $a_{n,\boldsymbol{k}}$ denotes the number of plain (closed) lambda terms of size n for which the investigated parameter values equal $k_1, k_2, \ldots, k_d$, respectively;

- Considered combinatorial parameters (patterns) inside plain or closed λ-terms are described in terms of admissible combinatorial specifications (sometimes infinite, as in the case of closed terms);

- So obtained specifications are then converted into systems of equations involving multivariate generating functions where additional variables $\boldsymbol{u} = (u_1, u_2, \ldots, u_d)$ *mark* associated combinatorial patterns;

- In the case of plain lambda terms, the resulting systems of equations are solved, usually approximately, in terms of standard complex-valued functions like $f(z) = \sqrt{1-z}$. The coefficients of associated generating functions depend on the marking variables $\boldsymbol{u} = (u_1, u_2, \ldots, u_d)$. In the case of closed lambda terms, novel tools developed in Chapter 4 are applied;

- Finally, an application of Flajolet and Odlyzko's transfer theorem provides access to probability generating functions of the limiting probability distributions. In consequence, properties of investigated combinatorial parameters become readily available.

8.2 Statistics of plain lambda terms

In this section we investigate several basic combinatorial parameters related to random plain λ-terms. Let us start with invoking the combinatorial specification (8.1.1) describing the set $\mathcal{L}_\infty$ of plain λ-terms. Recall that $\mathcal{L}_\infty$ is specified as

$$\begin{aligned} \mathcal{L}_\infty &::= \underline{\mathsf{n}} \mid \lambda\mathcal{L}_\infty \mid (\mathcal{L}_\infty\mathcal{L}_\infty) \\ \underline{\mathsf{n}} &::= \mathsf{0} \mid \mathsf{S}\ \underline{\mathsf{n}}. \end{aligned} \tag{8.2.1}$$

Equivalently, the set $\mathcal{L}_\infty$ of λ-terms can be specified using a pictorial tree grammar, see Figure 8.2 (note the explicit @ symbol for term application).

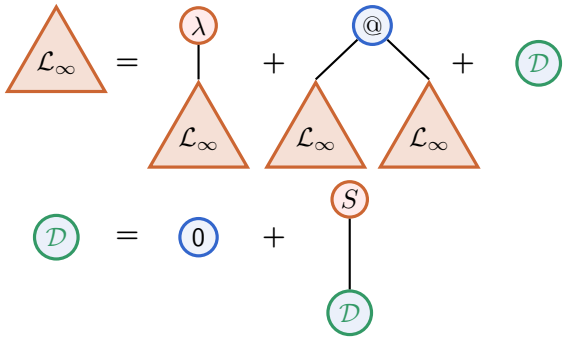


Figure 8.2: Combinatorial specification for plain λ-terms.

Following symbolic methods [FS09, Part A: Symbolic Methods] we note that the generating function $D(z)$ corresponding to de Bruijn indices takes the form $D(z) = \dfrac{z}{1-z}$ and so the generating function $L_\infty(z)$ associated with plain λ-terms satisfies the following functional equation:

$$L_\infty(z) = zL_\infty(z) + zL_\infty(z)^2 + \frac{z}{1-z}. \tag{8.2.2}$$

Solving (8.2.2) for $L_\infty(z)$ we obtain two formal solutions. Since we know *a priori* that the resulting generating function has non-negative coefficients $[z^n]L_\infty(z)$ we conclude that

$$L_\infty(z) = \frac{1}{2z}\left(1 - z - \sqrt{(1-z)^2 - \frac{4z^2}{1-z}}\right). \tag{8.2.3}$$

In this form, we can easily verify that the radicand expression $(1-z)^2 - \dfrac{4z^2}{1-z}$ carries the single dominant square-root type singularity ρ of $L_\infty(z)$. At this point, a straightforward application of the transfer theorem (see Proposition 3.1.1) gives us access to the asymptotic growth rate of the counting sequence corresponding to plain λ-terms.

Proposition 8.2.1 (see [Ben+16])**.** Let $L_\infty(z)$ be the generating function associated with plain λ-terms (8.2.3). Then, the number $[z^n]L_\infty(z)$ of plain terms of size n admits the following asymptotic approximation:

$$[z^n]L_\infty(z) \xrightarrow[n\to\infty]{} C\rho^{-n}n^{-3/2} \tag{8.2.4}$$

where

$$\rho \doteq 0.29559774 \quad \text{and} \quad C \doteq 0.606767. \tag{8.2.5}$$

More specifically, ρ is the positive real root of the polynomial $z^3+z^2+3z-1=0$ whereas $C = \frac{1}{2(1-\rho)}\sqrt{\frac{\rho+2}{\pi}}$.

8.2.1 Variables in plain lambda terms

We start our investigations with the variable distribution in plain λ-terms.

Proposition 8.2.2. Let X_n be a random variable corresponding to the number of variables in a random plain λ-term of size n. Then, after standardisation, X_n converges in law to a Gaussian distribution. Its expectation μ_n and variance σ_n^2 satisfy (up to numerical approximation)

$$\mu_n \xrightarrow[n\to\infty]{} 0.306849n \quad \text{and} \quad \sigma_n^2 \xrightarrow[n\to\infty]{} 0.0516364n. \tag{8.2.6}$$

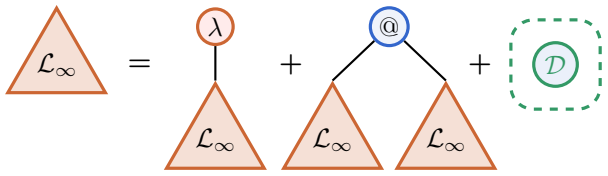


Figure 8.3: Marking variables in plain terms.

Proof. Let us consider a bivariate generating function $L_\infty(z,u)$ in which $[z^n u^k]L_\infty(z,u)$, i.e. the coefficient standing by $z^n u^k$, denotes the number of plain λ-terms of size n with k variables (equivalently k occurrences of 0). Marking all occurrences of 0 in the defining equation (8.2.2) of $L_\infty(z)$, see Figure 8.3, we obtain the following combinatorial specification for $L_\infty(z,u)$:

$$L_\infty(z,u) = zL_\infty(z,u) + zL_\infty(z,u)^2 + \frac{uz}{1-z} \tag{8.2.7}$$

and hence

$$L_\infty(z,u) = \frac{1}{2z}\left(1 - z - \sqrt{(1-z)^2 - \frac{4uz^2}{1-z}}\right) \tag{8.2.8}$$

as $L_\infty(z,1) = L_\infty(z)$.

The dominant singularity $\rho(u)$ of $L_\infty(z,u)$, is the real positive root of the radicand expression $F(z,u) = (1-z)^2 - \frac{4uz^2}{1-z}$. Moreover, the singularity has a non-zero derivative at $u=1$. According to Remark 4.1.4 (moving singularity framework) this yields a Gaussian limit law.

The mean and the variance of the resulting normal distribution can be computed by Proposition 4.1.4 using the values $\rho'(1)$ and $\rho''(1)$ from the partial derivatives of $F(z,u)$. Since $F(\rho(u),u) = 0$, after taking the derivative with respect to u we obtain

$$\rho'(1) = -\frac{\partial_u F(\rho,1)}{\partial_z F(\rho,1)} \quad \text{and} \quad \rho''(1) = -\frac{\partial_u^2 F(\rho,u) + 2\rho'(1)\partial_z\partial_u F(\rho,1) + \rho'(1)^2\partial_z^2 F(\rho,1)}{\partial_z F(\rho,1)}. \tag{8.2.9}$$

□

Remark 8.2.1. Let us note that, in general, $\rho(u)$ is a root of the polynomial $(1-z^b)F(z,u)$ whose degree depends on the specific size model parameters a,b,c,d denoting the weights of zero, successor, abstraction and application, respectively, cf. [GG16]. Specifically,

$$(1-z^b)F(z,u) = (1-z^b)(1-z^c)^2 - 4uz^{a+d}. \tag{8.2.10}$$

Consequently, for most admissible size notion parameters we cannot explicitly obtain analytic expression of $\rho(u)$. Instead, in order to check the premises of the multivariate central limit theorem we have to work with the implicit equation (8.2.10).

The main technical obstacle lies in the verification of the requested variability condition $B''(1) + B'(1) - B'(1)^2 \neq 0$, see (4.1.13). The remaining argumentation is virtually identical to the one presented for the specific case of $a=b=c=d=1$.

8.2.2 Redexes in plain lambda terms

Basic marking techniques allow us also to investigate limiting distributions of various sub-patterns in plain λ-terms. In what follows we study the fundamental pattern of β-redexes. Recall that a β-redex is a λ-term in form of $(\lambda N)M$ where N and M arbitrary λ-terms. In other words, a sub-pattern which can be depicted as  .

Proposition 8.2.3. Let X_n be a random variable denoting the number of β-redexes in a random plain λ-terms of size n. Then, after standardisation, X_n converges in law to a Gaussian distribution with expectation μ_n and variance σ_n^2 satisfying (up to numerical approximation)

$$\mu_n \xrightarrow[n\to\infty]{} 0.0907039n \quad \text{and} \quad \sigma_n^2 \xrightarrow[n\to\infty]{} 0.0519495n. \tag{8.2.11}$$

Proof. We start with establishing a formal specification for β-redexes in plain λ-terms. For that purpose, we introduce an auxiliary class $\mathcal{N}$ consisting of de Bruijn indices and terms in application form.

Note that if N is in application form, then either $N = (\lambda M)P$, i.e. N is a β-redex, or $N = MP$ where M belongs itself to class $\mathcal{N}$. Furthermore, with $\mathcal{N}$ at hand, we notice that each plain λ-term N takes either the form $N = \lambda M$ for some λ-term M, or is an element of $\mathcal{N}$. We can therefore write down the following joint combinatorial specification (8.2.12) for $L_\infty(z,u)$ and $N(z,u)$ corresponding to $\mathcal{L}_\infty$ and $\mathcal{N}$, respectively, using the variable u to mark the redex occurrences in $\mathcal{N}$, see Figure 8.4:

$$\begin{aligned} L_\infty(z,u) &= zL_\infty(z,u) + N(z,u) \\ N(z,u) &= \frac{z}{1-z} + uz^2 L_\infty(z,u)^2 + zN(z,u)L_\infty(z,u). \end{aligned} \tag{8.2.12}$$

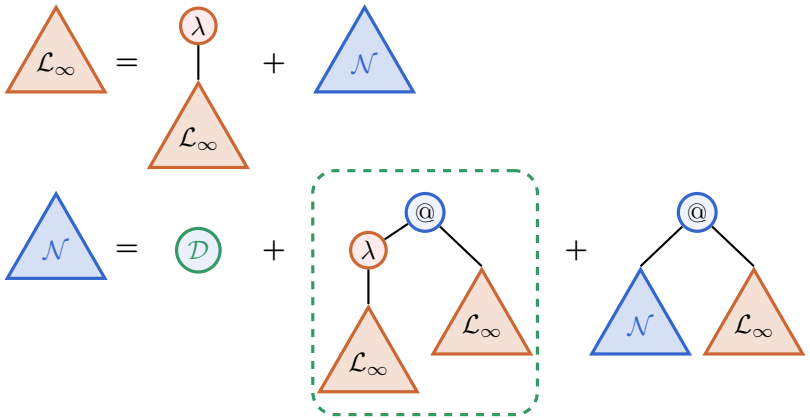


Figure 8.4: Marking redexes in plain terms.

Given that $L_\infty(z,1) = L_\infty(z)$ we solve (8.2.12) for $L_\infty(z)$ we finally obtain

$$L_\infty(z,u) = \frac{1-z-\sqrt{(1-z)^2 - \frac{4z^2(1+z(u-1))}{1-z}}}{2z\,(1+z(u-1))} \tag{8.2.13}$$

With the closed-form formula (8.2.13) we can now easily access the dominant singularity $\rho(u)$ of $L_\infty(z,u)$ carried by the radicand expression $(1-z)^2 - \frac{4z^2(1+z(u-1))}{1-z}$. Consequently, a straightforward application of the multivariate central limit theorem finishes the proof, see Proposition 4.1.4. □

8.2.3 Joint distribution of variables, abstractions, successors and redexes

Proposition 8.2.4. Let $\boldsymbol{X}_n = (X_{n(\mathsf{var})}, X_{n(\mathsf{red})}, X_{n(\mathsf{suc})}, X_{n(\mathsf{abs})})$ be a random vector denoting

1. the number $X_{n(\mathsf{var})}$ of variables;
2. the number $X_{n(\mathsf{red})}$ of β-redexes;
3. the number $X_{n(\mathsf{suc})}$ of successors, and
4. the number $X_{n(\mathsf{abs})}$ of abstractions

in a random plane λ-terms of size n. Then, after standardisation, the random vector $\boldsymbol{X}_n$ converges in law to a multivariate Gaussian distribution satisfying (up to numerical approximation)

$$X_n \xrightarrow{d} \mathcal{N}\left(n \begin{pmatrix} 0.307 \\ 0.091 \\ 0.129 \\ 0.258 \end{pmatrix}, n \begin{pmatrix} 0.052 & -0.013 & -0.034 & -0.069 \\ -0.013 & 0.052 & -0.022 & 0.047 \\ -0.034 & -0.022 & 0.145 & -0.076 \\ -0.069 & 0.047 & -0.076 & 0.214 \end{pmatrix} \right). \tag{8.2.14}$$

Proof. Like in the corresponding proofs for single parameters (see Proposition 8.2.2 and Proposition 8.2.3) we base our proof on the multivariate central limit theorem (see Proposition 4.1.4). We start with a joint multivariate specification for plain λ-terms including the investigated combinatorial parameters marked using auxiliary variable vector $\boldsymbol{u} = (u_{(\mathsf{var})}, u_{(\mathsf{red})}, u_{(\mathsf{suc})}, u_{(\mathsf{abs})})$ corresponding to respective components of $\boldsymbol{X}_n$, see Figure 8.5 (cf. Figure 8.4).

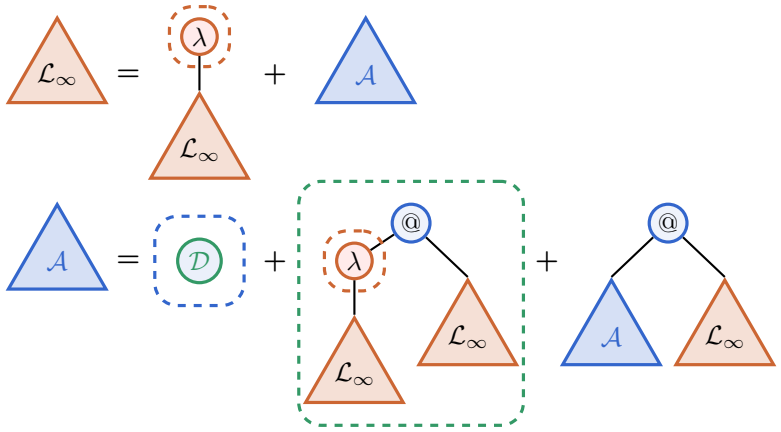


Figure 8.5: Marking abstractions, variables, successors and redexes in plain λ-terms.

Let $\boldsymbol{u}$ denote the vector of considered marking variables. Such a specification, when converted into a system of functional equations involving the generating functions $L_\infty(z, \boldsymbol{u})$ and $A(z, \boldsymbol{u})$ associated with $\mathcal{L}_\infty$ and $\mathcal{A}$, respectively, yields

$$\begin{aligned} L_\infty(z, \boldsymbol{u}) &= u_{(\mathsf{abs})} z L_\infty(z, \boldsymbol{u}) + A(z, \boldsymbol{u}) \\ A(z, \boldsymbol{u}) &= \frac{u_{(\mathsf{var})} z}{1 - u_{(\mathsf{suc})} z} + u_{(\mathsf{red})} u_{(\mathsf{abs})} z^2 L_\infty(z, \boldsymbol{u})^2 + z A(z, \boldsymbol{u}) L_\infty(z, \boldsymbol{u}). \end{aligned} \tag{8.2.15}$$

Furthermore, reformulating (8.2.15) we find that

$$(1 - u_{(\mathsf{abs})} z) L_\infty(z, \boldsymbol{u}) = \left(z u_{(\mathsf{abs})} (u_{(\mathsf{red})} - 1) + 1 \right) z L_\infty(z, \boldsymbol{u})^2 + \frac{u_{(\mathsf{var})} z}{1 - u_{(\mathsf{suc})} z}. \tag{8.2.16}$$

Let $\Delta(z, \boldsymbol{u})$ be the discriminant of the quadratic equation (8.2.16) defining $L_\infty(z, \boldsymbol{u})$.

Note that $\Delta(z, \boldsymbol{u})$ satisfies

$$\Delta(z, \boldsymbol{u}) = (1 - u_{(\mathsf{abs})} z)^2 - 4 \left(z u_{(\mathsf{abs})} (u_{(\mathsf{red})} - 1) + 1 \right) \frac{u_{(\mathsf{var})} z^2}{1 - u_{(\mathsf{suc})} z}. \tag{8.2.17}$$

In this form, we can access the dominant singularity $\rho(\boldsymbol{u})$ of $L_\infty(z, \boldsymbol{u})$ solving $\Delta(z, \boldsymbol{u}) = 0$ for z as a function of $\boldsymbol{u}$. Since (8.2.17) is a cubic equation in $z(\boldsymbol{u})$ we have access to the analytic form of its roots. We can therefore easily check that only one solution $z(\boldsymbol{u})$ of (8.2.17) coincides at $\boldsymbol{u} = \mathbf{1}$ with the dominant singularity ρ corresponding to plain terms (8.2.5). Consequently, the generating function $L_\infty(z, \boldsymbol{u})$ admits a corresponding Puiseux expansion in form of

$$L_\infty(z, \boldsymbol{u}) = \alpha(z, \boldsymbol{u}) - \beta(z, \boldsymbol{u}) \sqrt{1 - \frac{z}{\rho(\boldsymbol{u})}} + O\left(\left| 1 - \frac{z}{\rho(\boldsymbol{u})} \right| \right) \tag{8.2.18}$$

where both $\alpha(z, \boldsymbol{u})$ and $\beta(z, \boldsymbol{u})$ are analytic and non-vanishing near $(z, \boldsymbol{u}) = (\rho, \mathbf{1})$.

The required variability condition can be directly verified once the analytic form of $\rho(\boldsymbol{u})$ is calculated. At this point, the multivariate central limit theorem is readily applicable yielding the asserted convergence. A direct computation gives the vector of corresponding means and covariance matrix, see (8.2.14). □

Remark 8.2.2. Arguably, the most interesting part of the covariance matrix (8.2.14) is the sign of the correlations and the absolute values of associated variances. Note that in the natural size notion, the number of abstractions has greater variance than other constructors. Interestingly, the number of β-redexes is positively correlated with the number of abstractions. Not surprisingly, all other parameters are negatively correlated.

8.2.4 Head abstractions in plain lambda terms

In this section we turn to head abstractions in plain λ-terms showing that the corresponding random variable converges to a discrete geometric distribution.

Proposition 8.2.5. Let X_n be a random variable denoting the number of head abstractions in a random plain λ-term of size n. Then, X_n converges in law to a geometric distribution $\textsc{Geom}(\rho)$ with parameter ρ. Specifically,

$$\mathbb{P}(X_n = h) \xrightarrow[n \to \infty]{} \mathbb{P}(\textsc{Geom}(\rho) = h) = (1 - \rho)\rho^h. \tag{8.2.19}$$

Proof. Note that each λ-term starts with a (perhaps empty) sequence of consecutive head abstractions followed either by a de Bruijn index or an application of two terms (recall that abstractions therein are no longer considered to be head abstractions). Consequently, the set $\mathcal{L}_\infty$ of plain λ-terms can be specified using the auxiliary class $\mathcal{H}$ of head abstractions as depicted in Figure 8.6.

The bivariate generating function $L_\infty(z, u)$ corresponding to plain λ-terms with marked head abstractions satisfies therefore

$$L_\infty(z, u) = \frac{1}{1 - zu} \left(\frac{z}{1 - z} + z L_\infty(z, 1)^2 \right). \tag{8.2.20}$$

In such a form we immediately note that the dominant singularity $\rho(u)$ of $L_\infty(z, u)$ does not depend on u. In fact, it is constant and equal to ρ (i.e. the dominant singularity of $L_\infty(z)$, see (8.2.5)) as, by construction, $L_\infty(z, 1) = L_\infty(z)$.

Following the fact that $L_\infty(z)$ admits a Puiseux expansion near $z = \rho$ we can represent $L_\infty(z, 1)^2$ as a corresponding Puiseux series in form of

$$L_\infty(z, 1)^2 = \alpha - \beta \sqrt{1 - \frac{z}{\rho}} + O\left(\left| 1 - \frac{z}{\rho} \right| \right). \tag{8.2.21}$$

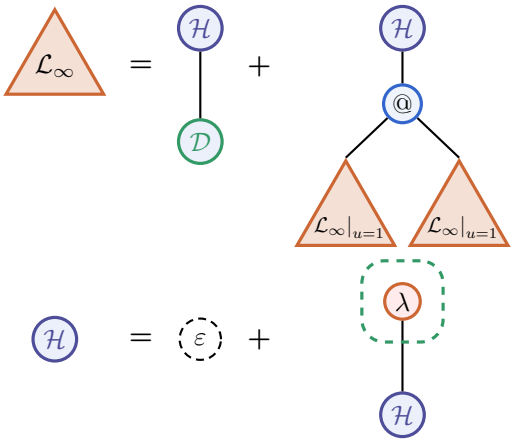


Figure 8.6: Marking head abstractions in plain terms.

Given (8.2.20) we further note that

$$L_\infty(z,u) = \widetilde{\alpha}(u) - \left(\frac{\rho}{1-\rho u}\right)\beta\sqrt{1-\frac{z}{\rho}} + O\left(\left|1-\frac{z}{\rho}\right|\right) \tag{8.2.22}$$

for u fixed sufficiently close to 1.

At this point, we apply Proposition 4.1.3 and find that the limit probability generating function $p(u)$ associated with $L_\infty(z,u)$ satisfies

$$p(u) = \frac{1-\rho}{1-\rho u} \tag{8.2.23}$$

which indeed corresponds to the asserted limit geometric distribution (8.2.19) of X_n. □

Remark 8.2.3. The mean number μ_n of head abstractions in a random λ-term of size n satisfies

$$\mu_n \xrightarrow[n\to\infty]{} \frac{\rho}{1-\rho}. \tag{8.2.24}$$

The limit mean (8.2.24) is close to 0.4196. Consequently, sufficiently large plain terms have, on average, less than one head abstraction. Such a result stands in sharp contrast to the canonical representation of David et al. [Dav+13] where the number of head abstractions in a random (closed) λ-term of size n is at least of order $\sqrt{n/\log n}$; in particular, it is a moderately growing function of n.

8.2.5 De Bruijn index values in plain lambda terms

In the current subsection we focus on the distribution of de Bruijn index values in random λ-terms.

Proposition 8.2.6. Let X_n be a random variable denoting the de Bruijn index value m of an index $\underline{\mathsf{m}}$ taken uniformly at random from a random plain λ-term of size n. Then, X_n converges in law to a geometric distribution $\textsc{Geom}(\rho)$ with parameter ρ. Specifically,

$$\mathbb{P}(X_n = m) \xrightarrow[n\to\infty]{} \mathbb{P}(\textsc{Geom}(\rho) = m) = (1-\rho)\rho^m. \tag{8.2.25}$$

Proof. Let $V_n^{(m)}$ be a random variable denoting the number of de Bruijn indices $\underline{\mathsf{m}}$ in a plain λ-term of size n. Marking the index $\underline{\mathsf{m}}$ in the specification for plain terms, see Figure 8.7, we note that the bivariate generating function $L_\infty^{(m)}(z,u)$ associated with $V_n^{(m)}$ satisfies

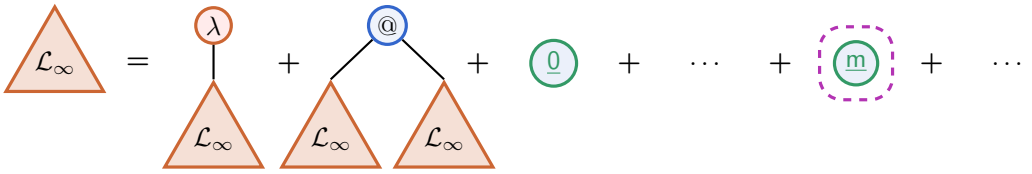


Figure 8.7: Marking the index $\underline{\mathsf{m}}$ in plain λ-terms.

$$L_\infty^{(m)}(z,u) = zL_\infty^{(m)}(z,u) + zL_\infty^{(m)}(z,u)^2 + \frac{z}{1-z} + (u-1)z^{m+1}. \tag{8.2.26}$$

Denote $\frac{\partial}{\partial u}L_\infty^{(m)}(z,u)|_{u=1}$ as $D_\infty^{(m)}(z)$. Then, taking the partial derivative ∂u at $u = 1$ of both sides of (8.2.26) we arrive at

$$D_\infty^{(m)}(z) = zD_\infty^{(m)}(z) + 2zD_\infty^{(m)}(z)L_\infty(z) + z^{m+1} \tag{8.2.27}$$

as $L_\infty^{(m)}(z,1) = L_\infty(z)$ for each $m \geqslant 0$, cf. (8.2.26) and (8.2.3). Note that $[z^n]D^{(m)}(z)$ corresponds to the weighted sum over all plain λ-terms of size n where each term comes with weight equal to the total number of occurrences of index $\underline{\mathsf{m}}$ in it.

Let $S_\infty(z,w) = \sum_{m\geqslant 0} D_\infty^{(m)}(z)w^m$. Taking the weighted sum over all $m \geqslant 0$ of both sides of (8.2.27) such that weight corresponding to m is w^m we obtain

$$S_\infty(z,w) = zS_\infty(z,w) + 2zS_\infty(z,w)L_\infty(z) + \frac{z}{1-zw}. \tag{8.2.28}$$

Consequently, $[z^n w^m]S_\infty(z,w)$ stands for $[z^n]D^{(m)}(z)$ whereas $[z^n]S_\infty(z,1)$ denotes the weighted sum of all λ-terms of size n where each term has weight equal to its total number of variables. In other words, variable w in $S_\infty(z,w)$ marks the probability mass function corresponding to X_n. Solving (8.2.28) we find that

$$S_\infty(z,w) = \frac{1}{1-zw}\left(\frac{z}{1-z-2zL_\infty(z)}\right) = \frac{z}{1-zw}\left(\sqrt{(1-z)^2 - \frac{4z^2}{1-z}}\right)^{-1} \tag{8.2.29}$$

where the latter equality follows from the formula (8.2.3) for $L_\infty(z)$.

Furthermore, given the known Puiseux expansion of the right-hand side square-root expression (see, e.g. Proposition 8.2.1) we easily note that $S_\infty(z,w)$ admits a Puiseux series expansion required for Proposition 4.1.3. Finally, a routine computation verifies the asserted geometric limit distribution $\textsc{Geom}(p)$ corresponding to the variable X_n. □

Remark 8.2.4. The mean value μ_n of a random index with a random plain terms satisfies

$$\mu_n \xrightarrow[n\to\infty]{} \frac{\rho}{1-\rho}. \tag{8.2.30}$$

The limit value (8.2.30), coinciding in the natural size notion with the corresponding mean for head abstractions (8.2.24), is close to 0.4196. This result stands, again, is sharp contrast to the canonical model of David et al. [Dav+13] in which variables (in closed terms) tend to be arbitrarily far from their binding abstractions.

Let us point out that such a disparity is a consequence of the different combinatorial models for λ-terms. In the canonical representation, the distance from a variable to its binding abstraction does not contribute to the weight of the corresponding variable (all variables have weight zero). On the other hand, in the de Bruijn representation weights of bound indices are proportional to the distances to their binding abstractions. Consequently, de Bruijn indices in large random λ-terms tend to be, on average, shallow. This central difference of both combinatorial models leads to remarkably contrasting asymptotic properties, including normalisation of large random λ-terms, cf. [Dav+13; Ben+16; Ben17].

8.2.6 Leftmost-outermost redex search

In order to evaluate an expression represented by a λ-term M we iteratively choose a β-redex (i.e. a subterm in form of $(\lambda P)Q$) in M and contract it using β-reduction, see e.g. Figure 8.8 (cf. Section 8.1.1).

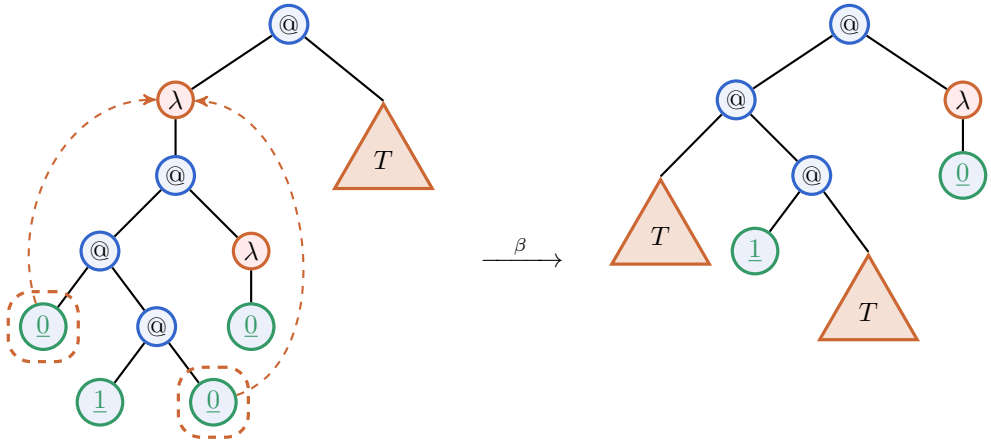


Figure 8.8: An example of β-reduction $(\lambda\underline{0}(\underline{10})(\lambda\underline{0}))T \to_\beta T(\underline{1}T)(\lambda\underline{0})$.

The order of evaluation (i.e. the order in which redexes are contracted) has a crucial impact on the computational effect of repeated β-reduction. Consider λ-terms $\Omega := (\lambda\underline{00})(\lambda\underline{00})$ and $P := (\lambda\lambda\underline{0})\Omega$. Note that Ω can be evaluated *ad infinitum* as $\Omega \to_\beta \Omega$. If we choose the Ω redex over the main $(\lambda\lambda\underline{0})\Omega$ redex in P, then $P \to_\beta P$. Otherwise, if we choose the *leftmost-outermost* redex $(\lambda\lambda\underline{0})\Omega$ instead of Ω we note that $P \to_\beta \lambda\underline{0}$ since the topmost abstraction in P has no associated indices. We cannot continue β-reducing $\lambda\underline{0}$ as it contains no more redexes (such terms are in so-called *(β-)normal form*) and so we terminate the evaluation process.

Terms from which it is possible to reach a β-normal form using repeated β-reduction are said to be *normalisable*. Since it is possible to emulate computations of arbitrary Turing machines in λ-calculus by means of normalisable terms, it is undecidable to determine whether a given λ-term is normalisable or not, see [Bar84]. Remarkably, the following classical result provides a normalising evaluation strategy guaranteed to find normal forms of normalisable λ-terms.

Theorem 8.2.1 (Standardisation theorem, see e.g [Bar84])**.** Let N be a normalising λ-term. Then, the iterated process of applying β-reduction to the leftmost-outermost redex in N leads to the (unique) normal form of N.

When searching for the leftmost-outermost redex in a given λ-term, we traverse the associated λ-tree in a depth-first manner favouring left branches of application nodes. Both cases of handling abstractions and indices are trivial – when visiting an abstraction node, we immediately recurse to its subterm looking for the leftmost-outermost redex; indices cannot contain β-redexes hence after arriving at an index, we terminate the traversal.

The most interesting part of the traversal algorithm lies in visiting application nodes. Suppose that we are currently visiting an application node. If its left branch starts with an abstraction node, we terminate the traversal as we have just found the leftmost-outermost redex. Otherwise, we have two possibilities based on whether the left branch is in β-normal form or not. If it is, we move into the left branch and, as we cannot find a β-redex, return from the recursion moving to the corresponding right branch. Otherwise if the left branch is not in β-normal form, we handle it recursively, however since it contains a β-redex, we eventually terminate the search before ever returning to visit the right branch.

Let us note that such a traversal algorithm, henceforth abbreviated to LO, introduces some *a priori* non-trivial computational overhead to the execution cost of β-reduction in λ-calculus. If carried out on a

β-redex, LO takes constant time to run. In contrast, when carried out on a λ-term in β-normal form, LO traverses nearly the whole λ-term. Such a varying traversal cost poses the natural question of the average-case performance of LO. In what follows, we show that the execution cost of LO, when viewed as a random variable ranging over random λ-terms, tends to a discrete limit law with constant expectation. Consequently, finding the leftmost-outermost redex introduces, on average, only a constant overhead to the cost of carrying out a single β-reduction.

Proposition 8.2.7. Let X_n be a random variable denoting the number of nodes visited by the LO traversal algorithm while searching for the leftmost-outermost β-redex in a random plain λ-term of size n. Then, X_n converges in law to a discrete limiting distribution with computable probability generating function and constant expectation. The corresponding means μ_n satisfy (up to numerical approximation)

$$\mu_n \xrightarrow[n\to\infty]{} 6.222262521. \tag{8.2.31}$$

Proof. We start with providing a combinatorial specification for plain λ-terms marking all nodes that are visited by the leftmost-outermost redex traversal algorithm LO. For that purpose we introduce the following three auxiliary classes:

- $\mathcal{N}$ for the class of β-normal forms;
- $\mathcal{M}$ for the class of so-called *neutral terms*, and
- $\mathcal{A}$ for the class of de Bruijn indices and plain λ-terms starting with an application.

In order to give their combinatorial specification we follow the presentation of [Ben+16] and give a joint description for the class $\mathcal{N}$ of β-normal forms and the associated class $\mathcal{M}$ of neutral terms. A β-normal form is either a plain λ-term starting with an abstraction followed by another β-normal form, or a neutral term. A neutral term, in turn, is either a de Bruijn index, or an application of a neutral term to a β-normal form, see Figure 8.9.

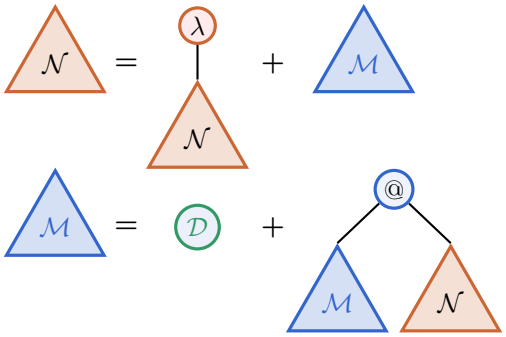


Figure 8.9: Joint specification for β-normal forms $\mathcal{N}$ and neutral terms $\mathcal{M}$.

Such a specification provides the following system of functional equations defining the generating functions $N(z)$ and $M(z)$ corresponding to the class of β-normal forms and neutral terms, respectively:

$$\begin{aligned} N(z) &= zN(z) + M(z) \\ M(z) &= zM(z)N(z) + \frac{z}{1-z}. \end{aligned} \tag{8.2.32}$$

Solving (8.2.32) we note that $M(z)$ and $N(z)$ satisfy

$$N(z) = \frac{M(z)}{1-z} \quad \text{and} \quad M(z) = \frac{1 - z - \sqrt{(1+z)(1-3z)}}{2z}. \tag{8.2.33}$$

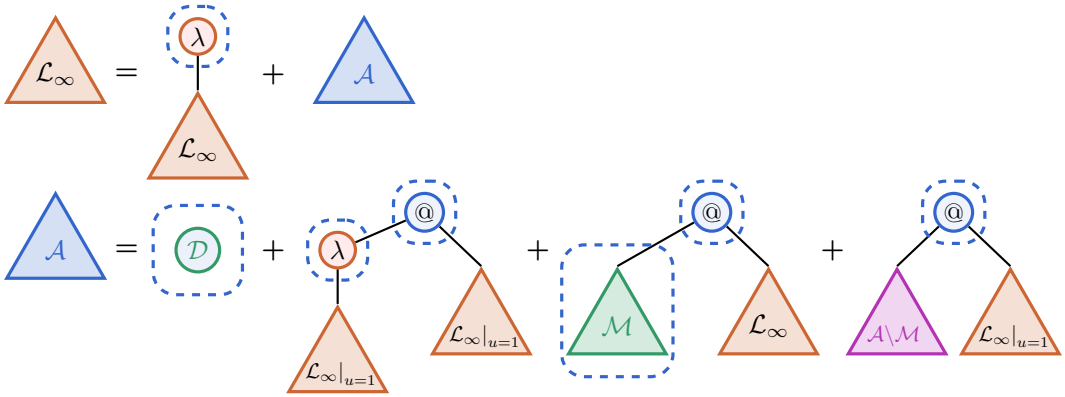


Figure 8.10: Specification for plain λ-terms with marked nodes following the execution of the redex finding traversal algorithm LO.

With both $N(z)$ and $M(z)$ at hand, we can now proceed and give the announced specification for plain λ-terms with marked nodes visited during the execution of LO, see Figure 8.10.

Let T be a plain λ-term in the class $\mathcal{A}$. Certainly, if T is a de Bruijn index, we mark only its topmost atom (i.e. the topmost successor or 0 if T is equal to $\underline{0}$). Otherwise, T starts with an application. If T is a β-redex, we mark two atoms – the topmost application and the abstraction starting the left branch of T. Remaining nodes are left unmarked. If T is not a redex however its left branch is a neutral term, we mark the entire left branch as well as the topmost application. Finally, if the left branch of T is not a neutral term, we take the difference class $\mathcal{A} \setminus \mathcal{M}$ of $\mathcal{A}$ and (marked) neutral terms $\mathcal{M}$ for the left branch. The right branch remains unmarked.

Such a specification yields the following system of functional equations defining the generating functions $L_\infty(z,u)$ and $A(z,u)$ corresponding to classes $\mathcal{L}_\infty$ and $\mathcal{A}$, respectively:

$$\begin{aligned} L_\infty(z,u) &= zuL_\infty(z,u) + A(z,u) \\ A(z,u) &= \frac{zu}{1-z} + z^2u^2L_\infty(z,1)^2 + zuM(zu)L_\infty(z,u) \\ &\quad + zu\left(A(z,u) - M(zu)\right)L_\infty(z,1). \end{aligned} \tag{8.2.34}$$

Knowing *a priori* that $L_\infty(z,1)$ corresponds to the generating function for plain λ-terms (8.2.3) we solve system (8.2.34) and find that $L_\infty(z,u)$ satisfies

$$L_\infty(z,u) = \frac{zuM(zu)L_\infty(z,1) - \left(zuL_\infty(z,1)\right)^2 - \frac{zu}{1-z}}{zuM(zu) - \left(1 - zuL_\infty(z,1)\right)(1-zu)}. \tag{8.2.35}$$

What remains to finish the proof is to check that $L_\infty(z,u)$ meets the premises of Proposition 4.1.3. Specifically, it admits a single, fixed dominant singularity $\rho(u) = \rho$ and a corresponding Puiseux expansion in form of

$$L_\infty(z,u) = \alpha(u) - \beta(u)\sqrt{1 - \frac{z}{\rho}} + O\left(\left|1 - \frac{z}{\rho}\right|\right). \tag{8.2.36}$$

Denote the denominator expression of (8.2.35) as $F(z,u)$. Note that $F(\rho,1) > 0$ and hence in a fixed neighbourhood of $u = 1$ the denominator $F(z,u)$ is non-zero. Consequently, $L_\infty(z,u)$ shares its (fixed) dominant singularity with $L_\infty(z,1)$. The required form of the Puiseux expansion of $L_\infty(z,u)$ follows as a consequence of the Puiseux expansions of both $L_\infty(z,1)$ and its power $L_\infty(z,1)^2$, see (8.2.21). A direct computation gives access to the corresponding probability generating function (omitted for brevity) and also the specific limit mean (8.2.31) of X_n. □

Remark 8.2.5. The generating function $M(z)$ associated with neutral terms, see (8.2.32), also corresponds to the well-known class of Motzkin numbers enumerating, *inter alia*, plane unary-binary trees, see e.g. [FS09, Note I.39, p.68]. We refer the curious reader to [Ben+16] for a size-preserving correspondence between neutral terms of size n and Motzkin trees with n nodes.

8.2.7 Height profile in plain lambda terms

The goal of this section is to obtain some insight into the mean height profile of plain λ-terms. We distinguish essentially two different notions of height. The first notion which we call *unary height*, takes into account only the number of abstractions above the considered node. The second notion concerns the *natural height* of a tree, i.e. the number of predecessors of a node which can be either abstractions or applications. In both situations, the semi-large powers theorem (see Proposition 3.1.2) can be applied. Consequently, the mean profile is always related to the Rayleigh distribution.

We are interested in mean profile of different types of nodes. In what follows we consider three types of mean profiles:

- the mean (unary or natural) profile of leaves;
- the mean (unary or natural) profile of abstractions, and
- the mean (unary or natural) profile of applications.

Proposition 8.2.8. Let H_n be a random variable denoting the unary (respectively natural) height of a randomly chosen variable (application or abstraction) in a random plain λ-term. Then, with x in any compact subinterval of $(0, +\infty)$, H_n admits a limiting Rayleigh distribution

$$\mathbb{P}(H_n = k) \sim \frac{C}{\sqrt{n}} \cdot \frac{x}{2} e^{-x^2/4} \quad \text{where} \quad x = \frac{k}{\sqrt{n}} \cdot C \tag{8.2.37}$$

with $C \doteq 4.30187$ for unary height, and $C \doteq 1.27162$ for the natural height. The average value of mean height is $\sqrt{\pi n}/C$ whereas the peak value is attained at $k^* = \sqrt{2n}/C$.

More specifically, the average number of

- variables at unary height k is asymptotically equal to $2.839\, k e^{-C^2k^2/4n}$;
- variables at natural height k is asymptotically equal to $0.248\, k e^{-C^2k^2/4n}$;
- abstractions at unary height k is asymptotically equal to $2.383\, k e^{-C^2k^2/4n}$;
- abstractions at natural height k is asymptotically equal to $0.208\, k e^{-C^2k^2/4n}$;
- applications at unary height k is asymptotically equal to $2.839\, k e^{-C^2k^2/4n}$;
- applications at natural height k is asymptotically equal to $0.248\, k e^{-C^2k^2/4n}$.

Proof. We start with the unary height profile of variables. Consider generating functions $C_k(z, u)$ corresponding to plain λ-terms with u marking the variables at the unary height k. These functions satisfy the following system of equations:

$$\begin{cases} C_0(z,u) = \dfrac{uz}{1-z} + zL_\infty(z) + zC_0(z,u)^2, & \text{if } k = 0; \\ C_k(z,u) = \dfrac{z}{1-z} + zC_{k-1}(z,u) + zC_k(z,u)^2, & \text{if } k > 0. \end{cases} \tag{8.2.38}$$

Taking partial derivatives of each equation in (8.2.38) with respect to u, we obtain a linear system for derivatives of generating functions. Setting $u = 1$ we can solve this linear system and obtain

$$\frac{\partial}{\partial u} C_k(z,u)\Big|_{u=1} = \frac{1}{1-z} \left(\frac{z}{1 - 2zL_\infty(z)} \right)^{k+1}. \tag{8.2.39}$$

Furthermore, a direct computation provides the following Puiseux series expansions as $z \to \rho$:

$$\frac{z}{1-2zL_\infty(z)} \sim 1 - \beta\sqrt{1-\frac{z}{\rho}} \quad \text{and} \quad L_\infty(z) \sim \frac{1-\rho}{2\rho} - b_\infty\sqrt{1-\frac{z}{\rho}} \tag{8.2.40}$$

where $\beta = 2b_\infty \doteq 4.301868701457$. Consequently, the numbers $M_{n,k}$ of variables at unary level k in a random plain lambda term of size n satisfy

$$M_{n,k} = \frac{[z^n]\dfrac{1}{1-z}\left(\dfrac{z}{1-2zL_\infty(z)}\right)^{k+1}}{[z^n]L_\infty(z)}. \tag{8.2.41}$$

An application of the semi-large powers theorem (see Proposition 3.1.2) and the transfer theorem (see Proposition 3.1.1) to the numerator and denominator, respectively, result in the following asymptotic estimate:

$$M_{n,k} \sim \frac{2k}{1-\rho}e^{-\beta^2k^2/4n}. \tag{8.2.42}$$

After normalising by the total sum $\sum_{k=0}^{n} M_{n,k}$ we obtain the declared limiting distribution.

Next, we turn to the case of natural height profile of variables. Consider generating functions $C_k(z,u)$ where now u marks the variables at the natural height k, instead of the unary height. As in the previous case, we obtain a system of equations

$$\begin{cases} C_0(z,u) = \dfrac{uz}{1-z} + zL_\infty(z) + zL_\infty(z)^2, & \text{if } k=0; \\ C_k(z,u) = \dfrac{z}{1-z} + zC_{k-1}(z,u) + zC_{k-1}(z,u)^2, & \text{if } k>0. \end{cases} \tag{8.2.43}$$

Again, taking partial derivatives ∂u at $u=1$ we can solve the resulting system and find that

$$\frac{\partial}{\partial u}C_k(z,u)\bigg|_{u=1} = \frac{z}{1-z}\left(z+2zL_\infty(z)\right)^k. \tag{8.2.44}$$

In this case, a direct computation verifies that the function $z+2zL_\infty(z)$ admits a Puiseux series expansion in form of $1-\gamma\sqrt{1-z/\rho}+O\left(|1-z/\rho|\right)$ where $\gamma = \beta\rho \doteq 1.27162265120953$. Consequently, this estimate yields a Rayleigh distribution with parameter 1.27162265120953. In particular, the average number $M_{n,k}$ of variables at natural height k in a random plain λ-term of size n satisfies

$$M_{n,k} \sim \frac{2\rho^2}{1-\rho}ke^{-\gamma^2k^2/4n}. \tag{8.2.45}$$

In order to mark remaining nodes, i.e. abstractions and applications, it is sufficient to change the first equation of the system (8.2.38). Accordingly, only the constant multiple behind the mean tree width at level k changes. For abstractions, we obtain, respectively,

$$C_0(z,u) = \frac{z}{1-z} + zuL_\infty(z) + zC_0(z,u)^2 \tag{8.2.46}$$

for unary height whereas

$$C_0(z,u) = \frac{z}{1-z} + zuL_\infty(z) + zL_\infty(z)^2 \tag{8.2.47}$$

for natural height. This change gives the constants $2L_\infty(\rho) = \frac{(1-\rho)}{\rho} \doteq 2.383$ for unary height, and $2\rho^2L_\infty(\rho) = \rho(1-\rho) \doteq 0.208$ for the natural height, respectively.

Similarly, marking applications yields a change in the first equation for the generating function

$$C_0(z,u) = \frac{z}{1-z} + zL_\infty(z) + zuC_0(z,u)^2 \tag{8.2.48}$$

for unary height, and

$$C_0(z,u) = \frac{z}{1-z} + zL_\infty(z) + zuL_\infty(z)^2 \tag{8.2.49}$$

for natural height. We obtain the constants $2L_\infty^2(\rho) = \frac{(1-\rho)^2}{2\rho^2} \doteq 2.839$ for unary height, and $2\rho^2 L_\infty^2(\rho) = \frac{(1-\rho)^2}{2} \doteq 0.248$ for natural height, respectively.

The mean value is obtained by using the integral approximation for the ratio of sums $kM_{n,k}/\sum_{k=0}^{n} M_{n,k}$ whereas the peak value is obtained by finding the maximum value of $M_{n,k}$ as a function of k. □

8.3 Parameters in closed lambda terms

In the following section we investigate more parameters related to plain and closed λ-terms. In particular, we consider:

- Several parameters related to closed λ-terms, resulting in Gaussian limit laws;
- Further parameters whose limiting distributions are discrete, including the leftmost-outermost redex search time in closed terms, the number of free variables in plain terms, the number of head abstractions in closed terms, and mean degree profile in closed terms;
- Finally, the mean height profile of closed terms for several different notions of height.

8.3.1 m-openness and the enumeration of closed terms

Recall that a term is said to be m-open (see Section 8.1.1) if by prepending it with m head abstractions we obtain a closed λ-term as a result. Following this natural, hierarchical notion, the set $\mathcal{L}_m$ of m-open λ-terms can be specified as

$$\begin{aligned} \mathcal{L}_m &::= \lambda\mathcal{L}_{m+1} \mid (\mathcal{L}_m\mathcal{L}_m) \mid \underline{0}, \underline{1}, \ldots, \underline{\mathsf{m}\text{ - }1} \\ \mathcal{L}_{m+1} &::= \lambda\mathcal{L}_{m+2} \mid (\mathcal{L}_{m+1}\mathcal{L}_{m+1}) \mid \underline{0}, \underline{1}, \ldots, \underline{\mathsf{m}} \\ \ldots &\qquad \ldots \end{aligned} \tag{8.3.1}$$

Let $L_m(z)$ denote the generating function associated with the set of m-open λ-terms, i.e. $L_m(z) = \sum_{n\geqslant 0} a_{n,m} z^n$ where $a_{n,m}$ stands for the number of m-open lambda terms of size n. Using (8.3.1) we obtain a corresponding infinite system for the functions $L_m(z)$:

$$\begin{aligned} L_0(z) &= zL_1(z) + zL_0(z)^2, \\ L_1(z) &= zL_2(z) + zL_1(z)^2 + z, \\ &\cdots, \\ L_m(z) &= zL_{m+1}(z) + zL_m(z)^2 + z\frac{1-z^m}{1-z}, \\ &\cdots \end{aligned} \tag{8.3.2}$$

In [BGG17, Lemma 8] the authors prove that for each $m \geqslant 0$ the generating functions for m-open λ-terms $L_m(z)$ admit Puiseux expansions in form of

$$L_m(z) \sim a_m - b_m\sqrt{1-\frac{z}{\rho}}. \tag{8.3.3}$$

Moreover, by the virtue of their proof, we obtain an suitable approximation procedure for the coefficients a_m and b_m by truncating the system (8.3.2) and replacing the function $L_m(z)$ with $L_\infty(z)$. Furthermore, the estimated coefficients $\widetilde{a}_m$ and $\widetilde{b}_m$ tend to their respective limits with an error of order $O(\frac{1}{\sqrt{m}})$. Using Theorem 4.4.1 it is possible to prove that $\widetilde{a}_m$ and $\widetilde{b}_m$ converge to their respective limits exponentially

fast. Consequently, the approximation procedure proposed in [BGG17] convergences exponentially fast, as well.

Immediately, this implies that the probability that a random plain λ-terms is m-open, but not $(m-1)$-open is $\dfrac{b_m - b_{m-1}}{b_\infty}$. Certainly, the limiting distribution associated with m-openness is discrete.

Note that the probability distribution function of m-openness is proportional to the coefficient at z^n in the bivariate generating function

$$L(z,u) = \sum_{k\geqslant 1} u^k \big(L_k(z) - L_{k-1}(z)\big) \sim \sum_{k\geqslant 1} u^k (a_k - a_{k-1}) - \sum_{k\geqslant 1} u^k (b_k - b_{k-1}) \sqrt{1 - \frac{z}{\rho}}. \tag{8.3.4}$$

The mean value corresponding to m-openness of plain terms can be calculated as

$$\frac{[z^n] \dfrac{\partial}{\partial u} L(z,u)\Big|_{u=1}}{[z^n] L_\infty(z)} \sim \frac{\sum_{k\geqslant 1} k(b_k - b_{k-1})}{\sum_{k\geqslant 1}(b_k - b_{k-1})} = \frac{\sum_{k\geqslant 0}(b_\infty - b_k)}{b_\infty} = \sum_{k\geqslant 0} \left(1 - \frac{b_k}{b_\infty}\right). \tag{8.3.5}$$

In order to compute this expectation we use the approximation procedure discussed above. Using the (aptly truncated) recurrence for the coefficients a_m and b_m

$$a_m = \frac{1}{2\rho}\left(1 - \sqrt{1 - 4\rho^2 \frac{1-\rho^m}{1-\rho} - 4\rho^2 a_{m+1}}\right), \quad b_m = \frac{\rho b_{m+1}}{\sqrt{1 - 4\rho^2 \frac{1-\rho^m}{1-\rho} - 4\rho^2 a_{m+1}}} \tag{8.3.6}$$

we obtain the numerical approximation for the mean value corresponding to m-openness. Numerical approximation yields an estimate 2.01922912627.

8.3.2 Variables, abstractions, successors and redexes in closed terms

In the following section we investigate the joint distribution of several parameters in closed λ-terms, utilising the novel Theorem 4.4.1.

Proposition 8.3.1. Let $\boldsymbol{X}_n = (X_{n(\mathsf{var})}, X_{n(\mathsf{red})}, X_{n(\mathsf{suc})}, X_{n(\mathsf{abs})})$ denote a vector of random variables denoting the number of variables, redexes, successors and abstractions in a random closed λ-term of size n, respectively. Then, after standardisation, the random vector $\boldsymbol{X}_n$ converges in law to a multivariate Gaussian distribution with identical parameters as plain terms.

Proof. Let us recall that the system of equations from Proposition 8.2.4 associated with the four parameters that we consider is, in the general class of plain terms, of the form

$$\begin{aligned} L(z,\boldsymbol{u}) &= u_{(\mathsf{abs})} z L(z,\boldsymbol{u}) + A(z,\boldsymbol{u}),\\ A(z,\boldsymbol{u}) &= \frac{u_{(\mathsf{var})} z}{1 - u_{(\mathsf{suc})} z} + u_{(\mathsf{red})} u_{(\mathsf{abs})} z^2 L(z,\boldsymbol{u})^2 + z A(z,\boldsymbol{u}) L(z,\boldsymbol{u}) \end{aligned} \tag{8.3.7}$$

with $\boldsymbol{u} = (u_{(\mathsf{var})}, u_{(\mathsf{red})}, u_{(\mathsf{suc})}, u_{(\mathsf{abs})})$ corresponding to respective components of $\boldsymbol{X}_n$.

In order to compose a similar, infinite system for m-open terms, we index respective generating functions in accordance with the natural combinatorial interpretation of m-openness; if an abstraction stands before an occurrence of $L(z,\boldsymbol{u})$, its respective index should be increased by one. This leads us to the following system:

$$\begin{aligned} L_m(z,\boldsymbol{u}) &= u_{(\mathsf{abs})} z L_{m+1}(z,\boldsymbol{u}) + A_m(z,\boldsymbol{u}),\\ A_m(z,\boldsymbol{u}) &= u_{(\mathsf{var})} \frac{z(1 - (u_{(\mathsf{suc})} z)^m)}{1 - u_{(\mathsf{suc})} z} + u_{(\mathsf{red})} u_{(\mathsf{abs})} z^2 L_m(z,\boldsymbol{u}) L_{m+1}(z,\boldsymbol{u}) + z A_m(z,\boldsymbol{u}) L_m(z,\boldsymbol{u}). \end{aligned} \tag{8.3.8}$$

Equivalently, we can represent (8.3.8) as

$$\begin{pmatrix} L_m(z,\boldsymbol{u}) \\ A_m(z,\boldsymbol{u}) \end{pmatrix} = \boldsymbol{\mathcal{K}}_m(L_m(z,\boldsymbol{u}), L_{m+1}(z,\boldsymbol{u}), A_m(z,\boldsymbol{u}), z, \boldsymbol{u}), \quad m = 0,1,2,\dots \tag{8.3.9}$$

It is straightforward to check that all the conditions of Theorem 4.4.1 are satisfied. Consequently, the function $L_0(z,\boldsymbol{u})$ admits a Puiseux expansion in form of

$$L_0(z,\boldsymbol{u}) \sim a_0(\boldsymbol{u}) - b_0(\boldsymbol{u})\sqrt{1-\frac{z}{\rho(\boldsymbol{u})}} \tag{8.3.10}$$

with the same $\rho(\boldsymbol{u})$ as in Proposition 8.2.4. Therefore, the limiting distribution, after standardisation, is Gaussian with the mean vector and the covariance matrix completely determined by the behaviour of the singularity $\rho(\boldsymbol{u})$ near the point $\boldsymbol{u} = \mathbf{1}$. □

8.3.3 Free variables in plain terms

Proposition 8.3.2. Let X_n be a random variable denoting the number of free variables in a random plain lambda term of size n. Then, X_n converges in law to a computable, discrete limiting distribution.

Proof. Consider the infinite system of functional equations $(L_m(z,u))_{m=0}^{\infty}$ where $L_m(z,u)$ corresponds to the generating function for plain λ-terms in which each de Bruijn index whose value k exceeds its unary height at least by m, is marked. For example, $L_0(z,u)$ corresponds to plain λ-terms with marked free variables. Note that

$$\begin{aligned} L_0(z) &= zL_1(z) + zL_0(z)^2 + uz + uz^2 + \cdots, \\ L_1(z) &= zL_2(z) + zL_1(z)^2 + z + uz^2 + \cdots, \\ &\cdots, \\ L_m(z) &= zL_{m+1}(z) + zL_m(z)^2 + z\frac{1-z^m}{1-z} + uz\frac{z^m}{1-z}, \\ &\cdots \end{aligned} \tag{8.3.11}$$

Let us apply Theorem 4.4.1 to this system taking $L_\infty(z)$ as the limiting equation. Certainly, the limiting equation does not depend on the marking variable u. Therefore, the singular point z^* also does not depend on u. An application of Proposition 4.1.3 finishes the proof. □

In order to compute the mean value we set $L_m^{\square}(z) := \dfrac{\partial}{\partial u} L_m(z,u)\Big|_{u=1}$ and represent the respective derivative as

$$L_m^{\square}(z) \sim c_m - d_m\sqrt{1-\frac{z}{\rho}}. \tag{8.3.12}$$

Based on (8.3.11) we note that

$$L_m^{\square}(z) = \frac{z}{1-2zL_\infty(z)}\left(L_{m+1}^{\square}(z) + \frac{z^m}{1-z}\right). \tag{8.3.13}$$

Since $\frac{z}{1-2zL_\infty(z)} \sim 1 - 2b\sqrt{1-z/\rho}$ as $z \to \rho$ we establish the following recurrence relation for the coefficients c_m and d_m:

$$c_m = c_{m+1} + \frac{\rho^m}{1-\rho} \quad \text{and} \quad d_m = d_{m+1} + 2b_\infty c_{m+1} + 2b_\infty\frac{\rho^m}{1-\rho}. \tag{8.3.14}$$

Once solved, this implies

$$c_m = \frac{\rho^m}{(1-\rho)^2} \quad \text{and} \quad d_m = \frac{2b_\infty\rho^m}{(1-\rho)^3}. \tag{8.3.15}$$

Consequently, the mean value corresponding the number of free variables in a random plain lambda term is equal to $\frac{d_0}{b_\infty} = \frac{2}{(1-\rho)^3} \doteq 5.7222625231204$.

8.3.4 Head abstractions in closed terms

Proposition 8.3.3. Let X_n be a random variable denoting the number of head abstractions in a closed λ-term of size n, chosen uniformly at random. Then, X_n converges in law to a computable, discrete limiting distribution. The corresponding expectation is close to 1.447.

Proof. Let $L_m(z,u)$ be the bivariate generating function associated with m-open lambda where u marks head abstractions. Then, the system $(L_m(z,u))_{m=0}^{\infty}$ satisfies

$$\begin{aligned} L_0(z,u) &= zuL_1(z,u) + zL_0(z,1)^2, \\ L_1(z,u) &= zuL_2(z,u) + zL_1(z,1)^2 + z, \\ &\dots \\ L_m(z,u) &= zuL_{m+1}(z,u) + zL_m(z,1)^2 + z\frac{1-z^m}{1-z}, \\ &\dots \end{aligned} \tag{8.3.16}$$

If a λ-term starts with a head abstraction, then after its removal, the openness of the respective subterm increases by one. Consequently, we include the expression $uzL_{m+1}(z,u)$ in the equation for $L_m(z,u)$. On the other hand, if the λ-term does not start with a head abstraction, i.e. starts with an application or is itself a de Bruijn index, we do not mark remaining abstractions as they are no longer head abstractions. Hence, we also include expressions $zL_m^2(z,1)$ and $z\frac{1-z^m}{1-z}$ in the equation corresponding to $L_m(z,u)$.

Having established the system (8.3.16) we note that $L_0(z,u)$ can be obtained as a limit of the solutions of truncated systems (see the description of the approximation procedure in Theorem 4.4.1) and this limit is equal to the sum

$$\begin{aligned} L_0(z,u) &= zL_0(z,1)^2 + zu\left(zuL_2(z,u) + zL_1(z,1)^2 + z\right) \\ &= zL_0(z,1)^2 + z^2uL_1(z,1)^2 + z^2u + (zu)^2\left(zuL_3(z,u) + zL_2(z,1)^2 + z + z^2\right) \\ &= \dots \end{aligned} \tag{8.3.17}$$

and so

$$\begin{aligned} L_0(z,u) &= z\sum_{m\geqslant 0}(uz)^m L_m(z,1)^2 + \sum_{m\geqslant 1}u^m(z^{m+1}+\dots+z^{2m}) \\ &= z\sum_{m\geqslant 0}(uz)^m L_m(z,1)^2 + \sum_{m\geqslant 1}(uz)^m z\frac{1-z^m}{1-z} \\ &= z\sum_{m\geqslant 0}(uz)^m L_m(z,1)^2 + \frac{uz^2}{(1-z)(1-uz)} - \frac{uz^3}{(1-z)(1-uz^2)}. \end{aligned} \tag{8.3.18}$$

Denote the final sum in (8.3.18) as $S(z,u)$. Since for each m the function $L_m(z,1)$ admits a Puiseux series expansion $L_m(z,1) \sim a_m - b_m\sqrt{1-z/\rho}$, near $z=\rho$ it holds

$$S(z,u) \sim c(u) + z\sum_{m\geqslant 0}(uz)^m\left(a_m^2 - 2a_mb_m\sqrt{1-\frac{z}{\rho}}\right) \tag{8.3.19}$$

where $c(u)$ comes from the last two summands of the previous expression. Since $a_m^2 \leqslant a_\infty^2$ and $a_mb_m \leqslant a_\infty b_\infty$, $S(z,u)$ is convergent near $(z,u)=(\rho,1)$ and the function $L_0(z,u)$ admits a Puiseux series expansion in form of

$$L_0(z,u) \sim a_0(u) - b_0(u)\sqrt{1-\frac{z}{\rho}}. \tag{8.3.20}$$

Consequently, $p(u) = b_0(u)/b_0(1)$ is the limiting probability generating function corresponding to the number of head abstractions in closed λ-terms. The function $b_0(u)$ satisfies

$$b_0(u) = 2\rho\sum_{m\geqslant 0}(u\rho)^m a_mb_m. \tag{8.3.21}$$

□

8.3.5 De Bruijn index values in closed lambda terms

Proposition 8.3.4. Let X_n be a random variable denoting the de Bruijn index value m of a random index $\underline{\mathsf{m}}$ in a random closed λ-term of size n. Then, X_n converges in law to a geometric distribution $\textsc{Geom}(\rho)$ with parameter ρ. Specifically,

$$\mathbb{P}(X_n = h) \xrightarrow[n\to\infty]{} \mathbb{P}(\textsc{Geom}(\rho) = h) = (1-\rho)\rho^h. \tag{8.3.22}$$

Proof. Let $L_{m,k}(z,u)$ denote the generating function for m-open λ-terms with u marking the number of occurrences of de Bruijn index $\underline{\mathsf{k}}$. Note that $L_{m,k}(z)$ satisfies a functional equation

$$L_{m,k}(z,u) = zL_{m+1,k}(z,u) + zL_{m,k}(z,u)^2 + z\frac{1-z^m}{1-z} + (u-1)z^{k+1}\mathbf{1}_{[k<m]} \tag{8.3.23}$$

where $\mathbf{1}_{[\cdot]}$ stands for the Iverson bracket notation.

Taking the partial derivative of (8.3.23) with respect to u and assigning $u = 1$, we obtain the generating function corresponding to λ-terms weighted by the number of occurrences of de Bruijn index $\underline{\mathsf{k}}$. Denote $\frac{\partial}{\partial u}L_{m,k}(z,u)|_{u=1}$ as $L^{\square}_{m,k}(z)$. Then, taking into account that $L_{m,k}(z,1) = L_m(z)$ we arrive at

$$L^{\square}_{m,k}(z) = zL^{\square}_{m+1,k}(z) + 2zL_m(z)L^{\square}_{m,k}(z) + z^{k+1}\mathbf{1}_{[k<m]}. \tag{8.3.24}$$

Consider the generating function

$$E_m(z,w) = \sum_{k\geqslant 0} L^{\square}_{m,k}(z)w^k. \tag{8.3.25}$$

Note that $[z^n]E_m(z,w)$ denotes the probability generating function associated with the distribution of variables in m-open λ-terms (cf. Proposition 8.2.6). Consequently, summing (8.3.24) over k we obtain

$$E_m(z,w) = zE_{m+1}(z,w) + 2zL_m(z)E_m(z,w) + z\frac{1-(wz)^m}{1-wz}. \tag{8.3.26}$$

These equations generate an infinite system for which Theorem 4.4.1 with a small modification is applicable. Each of the equations of the infinite system is linear, and the generating functions $L_m(z)$ enter the equations as coefficients. This yields the desired behaviour of the Puiseux expansions, because the non-linearity of the components is used only to provide the Puiseux expansion of the limiting equation, which is given in our case by construction.

The condition of exponential convergence holds because the difference between the limiting system and the mth equation of the system is equal to

$$2zE_\infty(z,w)\left(L_\infty(z) - L_m(z)\right) + \frac{z}{1-wz}(wz)^m \tag{8.3.27}$$

and decreases at exponential speed. Hence, the limiting distribution of de Bruijn index value is identical to the respective parameter in plain λ-terms. □

8.3.6 Leftmost-outermost redex search time in closed terms

Proposition 8.3.5. Let X_n denote the number of vertices visited by depth-first traversal algorithm searching for the leftmost-outermost β-redex in a random closed λ-term of size n (see Section 8.2.6). Then, the random variable X_n converges in law to a computable, discrete limiting distribution.

Proof. Recall that the system (8.2.34) defining the generating function $L_\infty(z,u)$ corresponding to plain terms with u marking visited nodes is written as

$$\begin{aligned} L_\infty(z,u) &= uzL_\infty(z,u) + A(z,u), \\ A(z,u) &= u\frac{z}{1-z} + z^2u^2L_\infty(z,1)^2 + zuM(zu)L_\infty(z,u) + uz(A(z,u) - M(zu))L_\infty(z,1) \end{aligned} \tag{8.3.28}$$

with $M(z)$ being the generating function associated with so-called neutral terms and also Motzkin numbers, see Remark 8.2.5:

$$M(z) = \frac{1 - z - \sqrt{(1+z)(1-3z)}}{2z}. \tag{8.3.29}$$

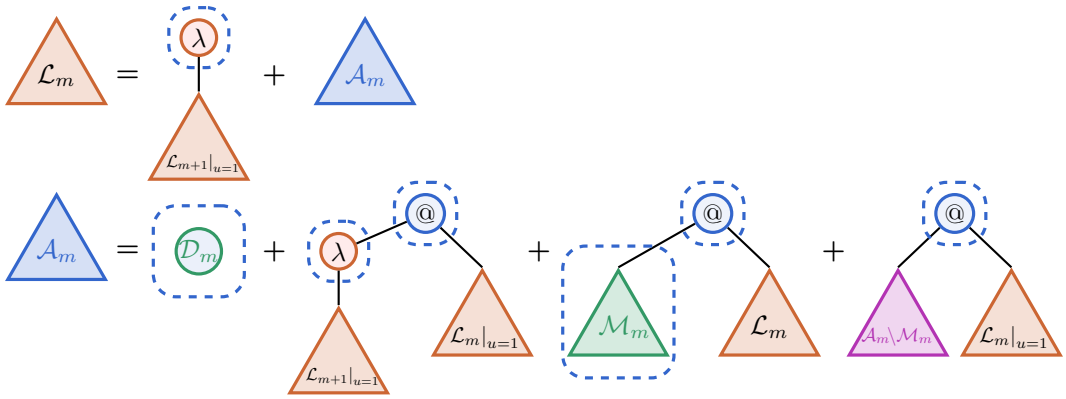


Figure 8.11: Specification corresponding to redex search time in closed lambda terms.

Note that including indices in (8.3.28) according to m-openness we obtain

$$\begin{aligned} L_m(z,u) &= uzL_{m+1}(z,u) + A_m(z,u), \\ A_m(z,u) &= uz\frac{1-z^m}{1-z} + z^2u^2L_m(z,1)L_{m+1}(z,1) \\ &\quad + zuM_m(zu)L_m(z,u) + uz(A_m(z,u) - M_m(zu))L_m(z,1) \end{aligned} \tag{8.3.30}$$

where $M_m(z)$ is the generating function for m-open neutral lambda terms. The sequence of functions $(M_m(z))_{m=0}^\infty$ can be obtained from the system of equations

$$\begin{aligned} N_m(z) &= zN_{m+1}(z) + M_m(z), \\ M_m(z) &= zM_m(z)N_m(z) + z\frac{1-z^m}{1-z}. \end{aligned} \tag{8.3.31}$$

Comparing (8.3.31) with its limiting counterpart (8.2.32) we note that the $M_\infty(z) - M_m(z)$ decays at exponential speed as $m \to \infty$ by virtue of Theorem 4.4.1. As additionally follows from the theorem, the functions $(M_m(z))_{m=0}^\infty$ share the same singularity $\rho = 1/3$.

Next, the system of equations (8.3.31) can be represented in the form

$$\begin{pmatrix} L_m(z,u) \\ A_m(z,u) \end{pmatrix} = \boldsymbol{\mathcal{K}}_m(L_m(z,u), L_{m+1}(z,u), A_m(z,u), z, u). \tag{8.3.32}$$

In order to apply Theorem 4.4.1 to (8.3.32) we need to replace the condition that the limiting system satisfies the premises of Drmota–Lalley–Woods theorem by an assumption that the limiting system admits Puiseux expansion. It was proven in Proposition 8.2.7 that $L_\infty(z,u)$ has a fixed singularity z^* which is independent of u. Therefore, all the functions $L_m(z,u)$ have a fixed singularity z^* and by applying Proposition 4.1.3, we obtain that the limiting distribution of the redex search time is discrete. □

8.3.7 Node height profile in closed terms

Like in Section 8.2.7, in the current section we consider unary and natural height profile of variables, abstractions and applications in closed lambda terms. For this purpose, we provide a variation of the semi-large powers theorem (see Proposition 3.1.2).

Theorem 8.3.1. Let $(f_k(\rho z))_{k\geqslant 0}$ be a sequence of functions analytic in delta-domain $\Delta(R)$ (see Proposition 3.1.1) for some $R > \rho$ admitting Puiseux series expansions in form of

$$f_k(z) \sim \sigma_k - a_k\sqrt{1-\frac{z}{\rho}} \tag{8.3.33}$$

as $z \to \rho$. Assume there exist β and $\widehat{\sigma}$ such that the sequences $(\sigma_k)_{k\geqslant 0}$ and $(a_k)_{k\geqslant 0}$ satisfy

$$\sum_{j=0}^{k}\frac{a_j}{\sigma_j} \sim \beta k \quad \text{and} \quad \lim_{k\to\infty}\prod_{j=0}^{k}\sigma_k \to \widehat{\sigma}. \tag{8.3.34}$$

Then, for x in any compact subinterval of $(0,+\infty)$, as $n \to \infty$, it holds

$$[z^n]\prod_{j=0}^{k} f_j(z) \sim \widehat{\sigma}\frac{\rho^{-n}}{n}S(\beta x) \quad \text{and} \quad x = \frac{k}{\sqrt{n}} \tag{8.3.35}$$

where $S(x)$ is the Rayleigh function defined in Proposition 3.1.2.

Proof. We recall that in the course of the proof of the semi-large power theorem, see [FS09, Theorem IX.16], the coefficient $[z^n]f(z)^k$ is expressed as the following complex contour integral with the help of Cauchy's integral theorem:

$$[z^n]f(z)^k = \frac{1}{2\pi i}\oint f(z)^k\frac{dz}{z^{n+1}} = \frac{1}{2\pi i}\oint e^{h_{n,k}(z)}\frac{dz}{z}, \quad h_{n,k}(z) = k\log f(z) - n\log z. \tag{8.3.36}$$

With the change of variables $z = \rho(1-t/n)$ the coefficient $[z^n]f(z)^k$ can be accordingly approximated by the following real integral:

$$[z^n]f(z)^k \sim -\frac{\rho^n}{n}\frac{1}{2\pi i}\int_0^\infty e^{t-ax\sqrt{t}}dt. \tag{8.3.37}$$

As proven in the referenced literature, this yields the Rayleigh approximation. In the statement of the current theorem, the function $h_{n,k}(z)$, i.e. the logarithm of the sub-integral expression, is replaced by

$$\widetilde{h}_{n,k} = \sum_{j=0}^{k}\log f_j(z) - n\log z. \tag{8.3.38}$$

Accordingly, with the variable change $z = \rho(1-t/n)$ the coefficient $[z^n]\prod_{j=0}^{k} f_j(z)$ becomes

$$[z^n]\prod_{j=0}^{k} f_j(z) = \prod_{j=0}^{k}\sigma_k \cdot [z^n]\prod_{j=0}^{k}\left(1-\frac{a_k}{\sigma_k}\sqrt{1-\frac{z}{\rho}}\right) \sim -\frac{\rho^n\prod_{j=0}^{k}\sigma_k}{n}\frac{1}{2\pi i}\int_0^\infty e^{t-\beta x\sqrt{t}}dt \tag{8.3.39}$$

which has the same form as (8.3.37), finishing the proof. □

Proposition 8.3.6. Let H_n be a random variable denoting the unary (respectively natural) height of a uniformly random variable in a random closed lambda term. Then, with x in any compact subinterval of $(0,+\infty)$, H_n follows the Rayleigh limiting distribution

$$\mathbb{P}(H_n = k) \sim \frac{C}{\sqrt{n}}\cdot\frac{x}{2}e^{-x^2/4}, \quad \text{where} \quad x = \frac{k}{\sqrt{n}}\cdot C \tag{8.3.40}$$

with $C \doteq 4.30187$ for unary height and $C \doteq 1.27162$ for the natural height.

Proof. Let $C_{m,k}(z,u)$ denote the bivariate generating function corresponding to m-open λ-terms where variable u marks de Bruijn indices at unary height $k-m$. Certainly, $C_{m,k}(z,1) = L_m(z)$ for each m and k. Note that, the functions $(C_{m,k}(z,u))_{m=0}^{\infty}$ satisfy jointly

$$\begin{cases} C_{m,k}(z,u) = z\dfrac{1-z^m}{1-z} + zC_{m+1,k}(z,u) + zC_{m,k}(z,u)^2 & \text{if } m < k, \\ C_{m,k}(z,u) = zu\dfrac{1-z^m}{1-z} + zL_{m+1}(z) + zC_{m,k}(z,u)^2 & \text{if } m = k, \\ C_{m,k}(z,u) = L_m(z) & \text{if } m > k. \end{cases} \tag{8.3.41}$$

A straightforward induction yields

$$\frac{\partial}{\partial u} C_{0,k}(z,u)\Big|_{u=1} = \prod_{j=0}^{k} \frac{z}{1-2zL_j(z)} \cdot \frac{1-z^k}{1-z}. \tag{8.3.42}$$

This function is amenable to asymptotic analysis of their coefficients by Theorem 8.3.1. First show that in the respective Puiseux expansions of the functions

$$\frac{z}{1-2zL_j(z)} \sim \sigma_j - c_j\sqrt{1-z/\rho}$$

the sequence σ_j tends to 1 at exponential speed, and the sequence c_j tends to a limit $2b_\infty \doteq 4.30187$ again at exponential speed. This holds because in the course of the proof of Theorem 4.4.1 we have shown that the sequences of coefficients of the Puiseux expansion of $(L_j(z))_{j=0}^{\infty}$ (respectively, the the sequence of first coefficients $L_j(\rho)$, and the sequence of the second coefficients) tend to their respective limits, i.e. to the coefficients of the Puiseux expansion of $L_\infty(z)$ exponentially fast. Comparing with the Puiseux expansion of $\frac{z}{1-2zL_\infty(z)}$ given in the proof of Proposition 8.2.8 we obtain the limiting values of the sequences $(\sigma_j)_{j=0}^{\infty}$ and $(c_j)_{j=0}^{\infty}$. Since the speed of convergence is exponential, the product $\prod_{j=0}^{k}\sigma_j$ converges to some $\widehat{\sigma}$, and the sum of the ratios c_j/σ_j tends to a linear function $\beta k = 2b_\infty k$.

Note that up to a normalising constant, the height profile of other parameters, namely the height profile distribution of abstractions and applications, remains asymptotically the same because from the generating function viewpoint only the multiple in front of the product $\prod_{j=0}^{k} f_j(z)$ changes (see Proposition 8.2.8).

In the same manner, there can be obtained Rayleigh distribution for natural height profile of different parameters. For example, in the case of variable height profile, we obtain the system of equations for the family of generating functions $C_{m,k}(z,u)$ for m-open lambda terms with variable u marking de Bruijn indices at unary height $k-m$:

$$\begin{cases} C_{m,k}(z,u) = z\dfrac{1-z^m}{1-z} + zC_{m+1,k}(z,u) + zC_{m+1,k}(z,u)^2, & 0 \leqslant m < k; \\ C_{m,k}(z,u) = uz\dfrac{1-z^m}{1-z} + zL_{m+1}(z) + zL_m(z)^2, & m = k; \\ C_{m,k}(z,u) = L_m(z), & m > k. \end{cases}$$

This implies

$$\frac{\partial}{\partial u} C_{0,k}(z,u)\Big|_{u=1} = \prod_{j=1}^{k} (z + 2zL_j(z)) \cdot z\frac{1-z^k}{1-z}.$$

Using the same argument as in the previous case, and taking into account two first terms of Puiseux expansion of $(z + 2zL_\infty(z))$ (see proof of Proposition 8.2.8), we obtain again Rayleigh distribution, with the same parameter as for plain lambda terms. □

Chapter 9

Statistical properties of random maps

Contents

This chapter follows [Bod+18a].

9.1 Introduction

9.1.1 Motivation for our work

Rooted maps form a ubiquitous family of combinatorial objects, of considerable importance in combinatorics, in theoretical physics, and in image processing. They describe the possible ways to embed graphs into compact oriented surfaces [LZ04].

The present work focuses on asymptotic enumeration of basic parameters in rooted maps with no restriction on genus. From a generating function viewpoint, if the genus of the maps is not fixed, then the generating function of rooted maps is non-analytic (namely, convergent only at zero) and often satisfies a Riccati differential equation, in contrast to *planar maps* for which analytic (convergent) generating functions abound. The divergent Riccati equations appear frequently in enumerative combinatorics. For example, at least 39 entries in Sloane's OEIS [Slo] were found containing sequences whose generating functions satisfy Riccati equations, including some entries related to the families of indecomposable combinatorial objects, moments of probability distributions, chord diagrams [CY17; CYZ16; FN00], Feynman diagrams [CLP78], etc. Some of these are closely connected to maps. Indeed, it is known that rooted maps with no genus restriction also encode different combinatorial families such as chord diagrams and Feynman diagrams on the one hand, and different fragments of lambda calculus [BGJ13; ZG15] on the other hand. Thus most asymptotic information obtained on maps can often be transferred to the aforementioned objects and lead to a better understanding of them in the corresponding domains.

While the asymptotics and stochastics on planar maps have been extensively studied (see for example [Ban+01; BB17; BR86; DP13; Lis99]), those on rooted maps with no genus restriction have received comparatively much less attention in the literature. Of closest connection to our study here is the paper by

Arquès and Béraud [AB00], which contains several characterisations of the number of rooted maps and their generating functions. In particular, they give an explicit formula for the number of maps, expressed as an infinite sum, from which the asymptotic number of maps with n edges can be deduced (which is $(2n+1)!!$). Recently, Carrance [Car17] obtained the distribution of genus in bipartite random maps. To our knowledge, no other asymptotic distribution properties of map statistics have been properly examined so far. Along a different direction, Flajolet and Noy [FN00] investigated basic statistics on chord diagrams, and Courtiel and Yeats [CY17] studied the distribution of *terminal chords*.

From an asymptotic point of view, for planar enumeration, as Bender and Richmond put it in [BR86]: "*The two most successful techniques for obtaining asymptotic information from functional equations of the sort arising in planar enumeration are Lagrange inversion and the use of contour integration.*" An equally useful analytic technique is the saddle-point method as large powers of generating functions are ubiquitous in map asymptotics; see [Ban+01; FS09] for more detailed information. In contrast, for divergent series, Odlyzko writes in his survey [Odl95]: "*There are few methods for dealing with asymptotics of formal power series, at least when compared to the wealth of techniques available for studying analytic generating functions.*" We show however that a few simple linearizing techniques are very helpful in deriving the diverse limit laws mentioned in the Abstract; the approaches we use may also be of potential application to other closely related problems.

9.1.2 Definitions

For a rigorous definition of a rooted combinatorial map we refer, for example, to [LZ04; AB00]. For our purposes in this extended abstract we use a less formal but more intuitive definition.

Definition 9.1.1 (Maps)**.** A *map* is a connected multigraph endowed with a cyclic ordering of consecutive half-edges incident to each vertex. Multiple edges and loops are allowed. Around each vertex, each pair of adjacent half-edges is said to form a *corner*. If there is only one half-edge, there is only one corner. A *rooted map* is a map with a distinguished corner.

Figure 9.1 shows some examples of rooted maps. Observe that the first two maps are different since the cyclic ordering is not the same: in the first map, the pendant edge follows counterclockwise the edge after the root (the node pointed to by an arrow), while in the second map it precedes in counterclockwise order. In contrast, the last two maps are equal: although the leaves are at different positions, one can find an isomorphism between the two maps preserving the vertices, the root and the cyclic orderings around each vertex. The corners of the leftmost map are displayed in Figure 9.2 (left), showing all the possible rootings of this map.

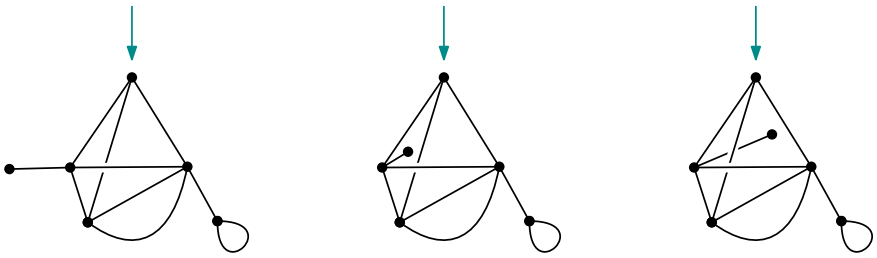


Figure 9.1: Three rooted maps. Each root is marked by an arrow. The two last maps are equal.

Definition 9.1.2 (Map features)**.** A *face* can be obtained by starting at some corner, moving along an incident half-edge, then switching to the next clockwise half-edge and repeating the procedure until the starting corner is met. A *loop* is an edge that connects the same vertex. An *isthmus* is an edge such that the deletion of this edge increases the number of connected components of the underlying graph. The *degree* of a vertex is the number of half-edges incident to this vertex.

These definitions are illustrated in Figure 9.2 (right).

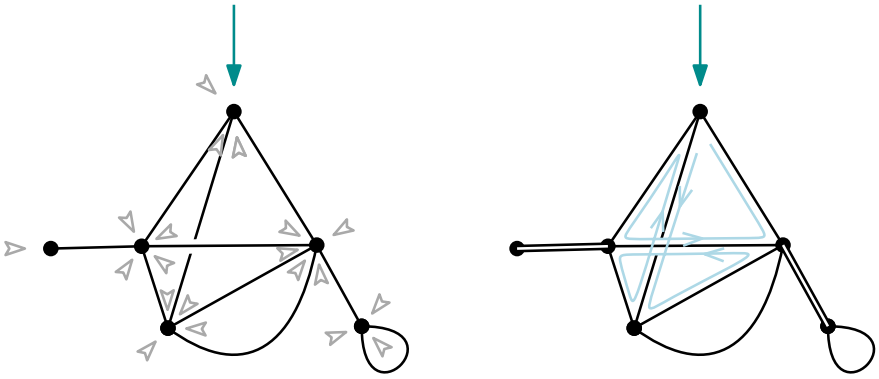


Figure 9.2: *Left:* The small triangles point at every corner of the map. *Right:* The light-blue line marks the contour of one face of the map. The double-lined edges are the isthmii of the map. The only loop of the map is adjacent to the rightmost isthmus, and the vertex incident to this loop has degree 3.

Arquès and Béraud [AB00] prove that the generating function of maps $M(z) := \sum_{n\geqslant 0} m_n z^n$, where m_n enumerates the number of maps with n edges, satisfies

$$2z^2M'(z) = (1-z)M(z) - 1 - zM(z)^2, \tag{9.1.1}$$

a typical Riccati equation whose first few Taylor coefficients read $M(z) = 1+2z+20z^2+444z^3+16944z^4+\cdots$.

Table 9.1: The six map statistics and their limit laws studied in this extended abstract.

Statistics	**Differential equation**	**Mean**	**Limit law**
leaves	$L = v + (2-u)zL + zL^2 + 2z^2\partial_z L + z(1-v)\partial_v L$	1	Poisson(1)
root isthmic parts	$C = 1+zC+vzC\vert_{v=1}C+2z^2\partial_z C$	2	Geom($\frac{1}{2}$)
vertices	$X = v + zX + zX^2 + 2z^2\partial_z X$	$\log n$	$\mathcal{N}(\log n, \log n)$
loops	$Y = v + vzY + vzY\vert_{v=1}Y + 2vz^2\partial_z Y + v^2z(vw-1)\partial_v Y$	$\frac{1}{2}n$	A new law*
root edges	$E = 1 + vzE + vzE\vert_{v=1}E + 2vz^2\partial_z E$	$\frac{2}{3}n$	Beta$(1, \frac{1}{2})$
root degree	$D = 1 + v^2zD + vzD\vert_{v=1}D + 2vz^2\partial_z D - v^2(1-v)z\partial_v D$	n	Uniform$[0, 2]$

9.1.3 Results and methods

We address in this work the analysis of the extended equations of (9.1.1) for bivariate (and in one case, trivariate) generating functions $M(z, v) := \sum_{n,k\geqslant 0} m_{n,k} z^n v^k$, where $m_{n,k}$ stands for the number of maps with n edges and the value of the shape parameter equal to k. We obtain limit laws for the distributions of six different parameters (see Figures 9.3 to 9.5).

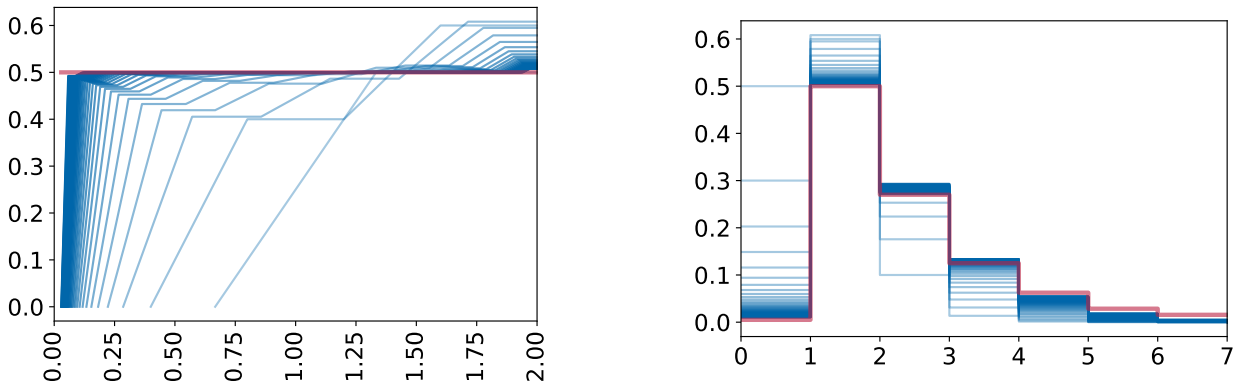


Figure 9.3: *Left:* Root vertex degree. *Right:* Number of root isthmic parts.

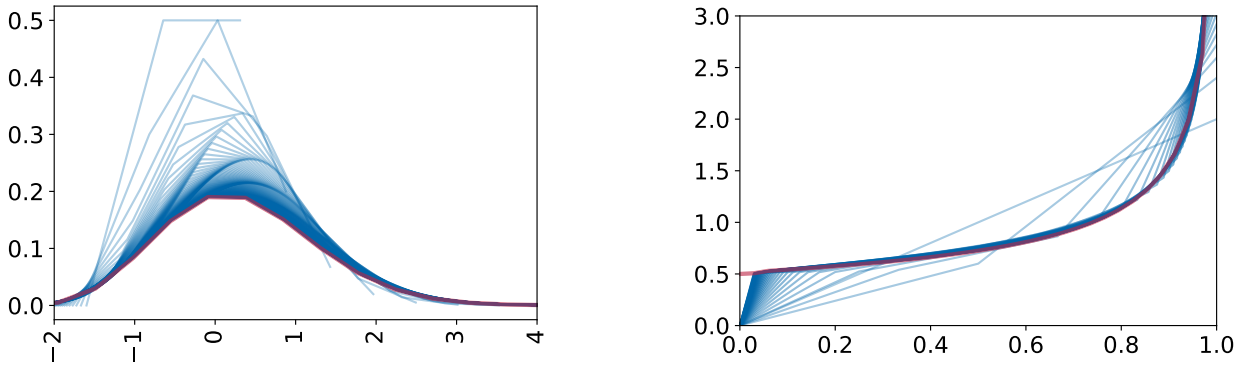


Figure 9.4: *Left:* Number of vertices. *Right:* Number of root edges.

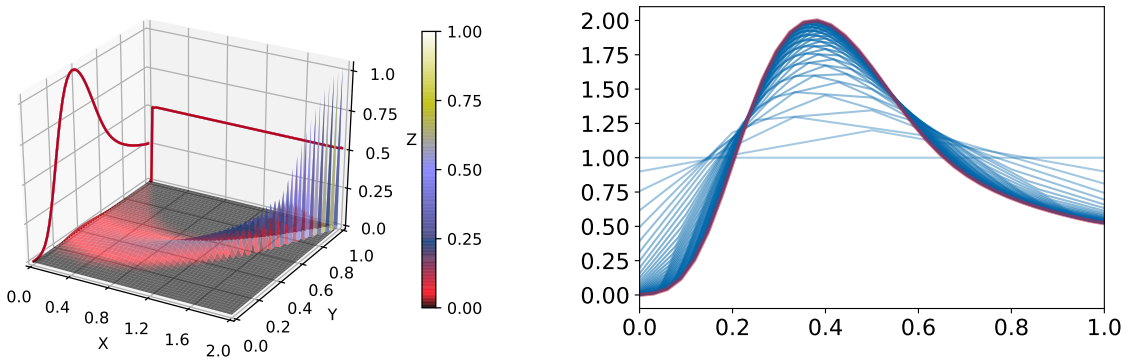


Figure 9.5: *Left:* Joint distribution of root vertex degree and the number of loops. *Right:* Number of loops.

We collect the statistics and their limit laws studied here in Table 9.1 for comparison. We see that some of the limit laws are discrete (Poisson and Geometric), one of them (the number of vertices) is Gaussian with a logarithmic mean, which we denote by $\mathcal{N}(\log n, \log n)$, and the others are continuous. For the number of root edges, root degree and loops, the corresponding limit laws are normalized by n, the total number of

edges. The distribution of the number of loops follows a rather unusual limit law (see Figure 9.5) in the sense that we can only characterise the limit law by its moment sequence, η_l, which satisfies $\eta_l = \eta_{0,l}$ with $\eta_{k,l}$ computable only through a recurrence involving $\eta_{k-1,l}$ and $\eta_{k+1,l-1}$. The corresponding probability density function of this law remains unknown and does not have an explicit expression at this stage (see Figure 9.5). Finally, by the bijection from [CYZ16] and a known property of chord diagrams in [FN00], it is possible to deduce the limit laws for the number of leaves.

One technique we use several times in our proofs consists in linearizing the differential equations satisfied by the generating functions, by choosing a suitable transformation, inspired from the resolution of Riccati equations. Once the dominant term is identified, the analysis for the limit law becomes more or less straightforward. When such a technique fails, we rely then on the method of moments, which establishes weak convergence by computing all higher derivatives of $M(z, v)$ at $v = 1$ and by examining asymptotically the ratios $[z^n]\partial_v^k M(z, v)|_{v=1}/[z^n]M(z, 1)$ (which correspond to the moments of random variable). Such a procedure also linearises to some extent the more complicated bivariate nature of the differential equations and facilitates the resolution complexity of the asymptotic problem.

Structure of the Paper. In Section 9.2 we derive the nonlinear differential equations satisfied by the generating functions of the map statistics. Then in Section 9.3 we sketch the proofs for the limit laws of five statistics based on generating functions. The Poisson law for the number of leaves (together with the root face degree and the number of trivial loops) will be proved by a direct combinatorial approach in the last section.

9.2 Differential equations for maps

In this section, we derive the differential equations satisfied by the bivariate or trivariate generating functions with the additional variable(s) counting the shape statistics.

Univariate generating function of maps. Since the Riccati equation (9.1.1) lies at the basis of all other extended equations in Table 9.1, we give a quick proof of it via the recurrence satisfied by m_n, the number of maps with n edges (see Figure 9.6):

$$m_n = \mathbf{1}_{[n=0]} + \sum_{0 \leqslant k < n} m_k m_{n-1-k} + (2n-1)m_{n-1}, \tag{9.2.1}$$

which then implies the Riccati equation (9.1.1).

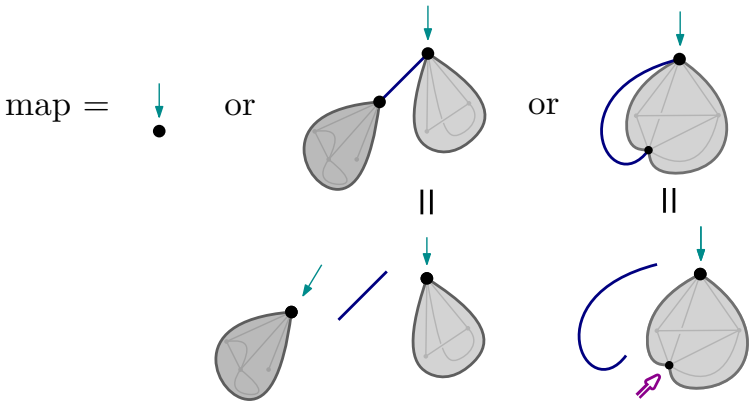


Figure 9.6: A symbolic construction of rooted maps.

First, $m_0 = 1$ because there is only one map with 0 edges. Then a map with n edges can be formed either by connecting the roots of two maps (with k and $n-k-1$ edges, respectively) with an isthmus, or by adding an edge to a map with $n-1$ edges, connecting the root and a corner. The number of possible ways to insert an edge in this way is equal to $2n-1$, because there are $2n-2$ corners in a map of size $n-1$, and there are two possible ways to insert a new edge at the root corner (either before, or after the root). This proves (9.2.1).

Vertices. Consider now the bivariate generating function $X(z,v) = \sum_{n,k\geqslant 0} x_{n,k} z^n v^k$, where $x_{n,k}$ is equal to the number of rooted maps with n edges and k vertices. Arquès and Béraud [AB00] showed that

$$X(z,v) = v + zX(z,v) + zX(z,v)^2 + 2z^2\partial_z X(z,v). \tag{9.2.2}$$

This recurrence can be obtained from (9.2.1) by noticing that no new vertex is created when we connect two maps with an isthmus, nor when we add a new root edge to a map. Note that $X(z,v)$ satisfies another functional equation (see [AB00])

$$X(z,v) = v + zX(z,v)X(z,v+1),$$

which seems less useful from an asymptotic point of view.

Root isthmic parts. We count here the *root isthmic parts*, which are the number of isthmic constructions used at the root vertex. Note that an isthmic part may not be a bridge because the additional edge constructor may induce additional connections.

We show that the bivariate generating function $C(z,v) = \sum_{n,k\geqslant 0} c_{n,k} z^n v^k$, where $c_{n,k}$ enumerates the number of maps with n edges and k root isthmic parts, satisfies

$$C(z,v) = 1 + zC(z,v) + vzC(z,v)C(z,1) + 2z^2\partial_z C(z,v). \tag{9.2.3}$$

In Figure 9.6, the number of root isthmic parts only changes whenever two maps are connected by an isthmus. This yields $vzC(z,v)C(z,1)$ instead of zC^2.

Root edges. Similarly, consider $E(z,v) = \sum_{n,k\geqslant 0} e_{n,k} z^n v^k$, where $e_{n,k}$ counts the number of rooted maps with n edges and k root edges. We show that $E(z,v)$ satisfies

$$E = 1 + vzE + vzE|_{v=1}E + 2vz^2\partial_z E. \tag{9.2.4}$$

This again results from the recurrence (9.2.1) and from Figure 9.6: the non-root edges come from the bottom map in the isthmic construction, yielding the term $vzE(z,v)E(z,1)$.

Root Degree. Consider the degree of the root vertex. Note that this may be different from the number of root edges because for the root degree, each loop edge is counted twice, therefore the degree of the root vertex varies from 0 to $2n$. By duality, the distribution of the root face degree is the same as the distribution of the root vertex degree.

Let $D(z,v) = \sum_{n,k\geqslant 0} d_{n,k} z^n v^k$ denote the bivariate generating function for maps with variable v marking root degree. Then

$$D = 1 + v^2 zD + vzD|_{v=1}D + 2vz^2\partial_z D - v^2(1-v)z\partial_v D. \tag{9.2.5}$$

In this case, the original construction in Figure 9.6 is insufficient, and we need to consider further cases in Figure 9.7. When an additional edge becomes a loop, it increases the degree of the root vertex by 2; otherwise, the root degree is increased merely by 1. Note that the equation (9.2.5) is now a *bona fide* partial differential equation, making the analysis more difficult.

Leaves. The differential equation for the bivariate generating function of maps with variable v marking leaves (see Table 9.1) can be obtained in a similar way by considering different cases in the new edge constructor. The number of special leaf corners is equal to the number of leaves.

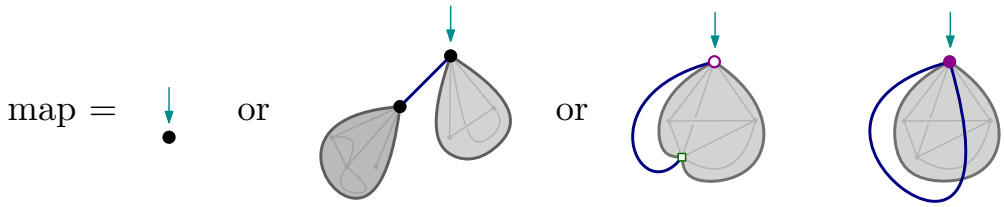


Figure 9.7: Symbolic method to count root degree and loops in rooted maps.

Loops. Finally, we look at the number of loops whose enumeration necessitates the consideration of the joint distribution of the number of loops and the number of root edges, namely, we consider the trivariate generating function $Y(z,v,w) = \sum_{n,k,m} y_{n,k,m} z^n v^k w^m$, where $y_{n,k,m}$ denotes the number of rooted maps with n edges, root degree equal to k, and m loops. We show that $Y(z,v,w)$ satisfies a partial differential equation

$$Y = 1 + zvY + zvY|_{v=1} Y + 2z^2 v\partial_z Y + zv^2(vw-1)\partial_v Y. \tag{9.2.6}$$

As in the symbolic construction of Figure 9.7, a new edge becomes a loop only if it is attached to one of the corners incident to the root vertex. The differential equation (9.2.6) is then a modification of (9.2.5) with an additional variable counting the number of loops.

Note that Equation (9.2.6) is *catalytic* with respect to the variable v, i.e. putting $v = 1$ introduces a new unknown object $\partial_v Y|_{v=1}$ to the differential equation. One of the strategies for dealing with catalytic equations was developed by Bousquet-Mélou and Jehanne [BJ06], generalising the so-called kernel method and quadratic method. However, their method does not work in our case because our equation is differentially algebraic.

9.3 Limit laws

This section describes the techniques we employ to establish the limit laws.

From now on, by a random map (with n edges) we assume that all rooted map with n edges are equally likely. For notational convention, we use $X' = \partial_z X$ to denote derivative with respect to z. Due to space limit, we give only the sketches of the proofs.

9.3.1 Transformation into a linear differential equation

For most of the equations in the previous section, it turns out that a transformation similar to that used for Riccati equations largely simplifies the resolution and leads to solvable recurrences, which are then suitable for our asymptotic purposes. We begin by solving the standard Riccati equation (9.1.1) and see how a similar idea extends to other differential equations.

Proposition 9.3.1. The number m_n of maps with n edges satisfies

$$\frac{m_n}{\phi_n} = 2n - 1 + O\big(n^{-1}\big), \quad \text{where} \quad \phi_n = \frac{(2n)!}{2^n n!} = (2n-1)!!. \tag{9.3.1}$$

Proof. We solve the Riccati equation (9.1.1) by considering the transformation

$$M(z) = 1 + \frac{2z\phi'(z)}{\phi(z)}, \tag{9.3.2}$$

for some function $\phi(z)$ with $\phi(0) = 1$. Substituting this form into the equation (9.2.1), we get the second-order differential equation $2z^2\phi'' + (5z-1)\phi' + \phi = 0$. From this equation, the coefficients $\phi_n := [z^n]\phi(z)$ satisfy the recurrence $\phi_{n+1} = (2n+1)\phi_n$, which implies the double factorial form of ϕ_n by $\phi_0 = 1$.

Moreover, by extracting the coefficient of z^n in (9.3.2), we obtain a relation between the coefficients m_k and ϕ_ℓ. By the inequality $m_n \geqslant (2n-1)m_{n-1}$ (see (9.2.1)), we then deduce the asymptotic relation (9.3.1). □

Theorem 9.3.1. Let X_n denote the number of vertices in a random rooted map with n edges. Then X_n follows a central limit theorem with logarithmic mean and logarithmic variance:

$$\frac{X_n - \mathbb{E}(X_n)}{\sqrt{\mathbb{V}(X_n)}} \xrightarrow{d} \mathcal{N}(0,1), \quad \mathbb{E}(X_n) \sim \log n, \quad \mathbb{V}(X_n) \sim \log n. \tag{9.3.3}$$

Proof. Similar to (9.3.2), we define a bivariate generating function $S(z,v) = \sum_{n\geqslant 0} s_n(v) z^n$ such that

$$X(z,v) = v + \frac{2zS'}{S}, \quad S(0) = 1.$$

Substituting this $X(z,v)$ into (9.2.2) leads to a linear differential equation from which one can extract the recurrence

$$s_n(v) = \frac{(2n+v-2)(2n+v-1)}{2n}\, s_{n-1}(v).$$

We then get an explicit expression for $s_n(v)$, from which we deduce, by singularity analysis, that

$$\mathbb{E}(v^{X_n}) = \frac{2^{v-1}}{\Gamma(v)}\, n^{v-1}(1 + O(n^{-1})),$$

and conclude by applying the Quasi-Powers Theorem [FS09; Hwa98]. □

A finer Poisson($\log n + c$) approximation, for a suitably chosen c, is also possible, which results in a better convergence rate $O(\log n)^{-1}$ instead of $(\log n)^{-\frac{1}{2}}$; see [Hwa99] for details.

Theorem 9.3.2. Let C_n denote the number of root isthmic parts in a random rooted map with n edges. Then,

$$C_n \xrightarrow{d} \textsc{Geom}\big(\tfrac{1}{2}\big).$$

Proof. Since $C(1,z) = M(z)$, we use again the substitution (9.3.2) and apply it to (9.2.3):

$$2z^2(\phi C' + v\phi' C) = (1 - (1+v)z)\phi C - \phi.$$

The trick here is to multiply both sides by $\phi(z)^{v-1}$ and set $Q(z,v) = \phi(z)^v C(z,v)$. We then obtain

$$2z^2 Q' = (1 - (1+v)z)Q - \phi^v.$$

Using the recurrence for the normalised coefficients $\widehat{q}_n(v) := q_n(v)/\phi_n$ and dominant-term approximations, we find that the n-th coefficient of Q is proportional to

$$\widehat{q}_n(v) = \frac{v}{2n} \sum_{1\leqslant k\leqslant n} \Big(\frac{n}{k}\Big)^{v/2} + O(n^{-1/2}) = \frac{v}{2-v} + O(n^{-\frac{1}{2}}).$$

This corresponds to a (shifted by 1) geometric distribution with parameter $\frac{1}{2}$. By the definition $Q(z,v) = \phi(z)^v C(z,v)$, we deduce that the limiting distribution of C_n is also geometric with parameter $\frac{1}{2}$. □

Theorem 9.3.3. Let E_n denote the number of edges incident to the root vertex in a random rooted map with n edges. Then E_n follows asymptotically a Beta distribution:

$$\frac{E_n}{n} \xrightarrow{d} \mathrm{Beta}\big(1, \tfrac{1}{2}\big), \tag{9.3.4}$$

with the density function $\frac{1}{2}(1-t)^{-\frac{1}{2}}$ for $t \in [0,1)$.

Proof. We use again the substitution $E(z,1) = M(z) = 1 + 2z\frac{\phi'}{\phi}$ in (9.2.4), giving

$$2vz^2(\phi E' + \phi' E) = (1 - 2vz)\phi E - \phi.$$

With $Q(z,v) = \phi(z)E(z,v)$, we then obtain

$$2vz^2 Q' = (1 - 2vz)Q - \phi. \tag{9.3.5}$$

This linear differential equation translates into a recurrence for the coefficients $q_n(v)$ of $Q(z,v)$, which yields the closed-form expression

$$q_n(v) = 2^n n! \sum_{0 \leqslant j \leqslant n} \binom{2j}{j} 4^{-j} v^{n-j}. \tag{9.3.6}$$

Returning to $E(z,v)$, we see that its coefficients behave asymptotically like $q_n(v)$. This implies the Beta limit law (9.3.4) for the random variable E_n/n since $\binom{2j}{j}4^{-j} \sim (\pi j)^{-1/2}$ for large j. □

Theorem 9.3.4. Let D_n denote the degree of the root vertex in a random rooted map with n edges. Then, D_n, divided by the number of edges, converges in law to the uniform distribution on $[0,2]$:

$$\frac{D_n}{n} \xrightarrow{d} \text{Uniform}\,[0,2]\,. \tag{9.3.7}$$

Proof. The substitutions

$$D(z,1) = M(z) = 1 + \frac{2z\phi'}{\phi}, \quad \text{and} \quad D(z,v) = \frac{Q(z,v)}{\phi(z)}$$

lead to a partial differential equation, which in turn yields the recurrence for the coefficients $q_n(v) := [z^n]Q(z,v)$:

$$q_n(v) = v(2n - 1 + v)q_{n-1} - v^2(1 - v)q'_{n-1}(v) + \phi_n.$$

We then get the exact solution $q_n(v) = \phi_n(1 + v + \cdots + v^{2n})$.

In order to proceed, we show the following curious result:

$$m_n = -[z^{k+1}]\frac{1}{\phi(z)}. \tag{9.3.8}$$

From the definition of $\phi(z)$ we have two recurrences, following from the respective convolution identities, first for $\phi(z)M(z)$ and the second for $z\phi'(z) \times \dfrac{1}{\phi(z)}$:

$$\sum_{k=0}^{n} m_k \phi_{n-k} = (2n+1)\phi_n,$$

$$m_n = \mathbf{1}_{[n=0]} + 2\sum_{k=0}^{n}(n-k)\phi_{n-k}[z^k]\frac{1}{\phi(z)}.$$

By combining these two recurrences, we obtain (9.3.8).

Accordingly, (9.3.8) implies that $d_n(v) := [z^n]D(z,v) \sim q_n(v)$. This, in turn, implies the uniform limit law (9.3.7). □

A more intuitive interpretation of this uniform limit law is given in the next section.

9.3.2 Approximation and method of moments

Unlike all previous proofs, we use the method of moments to establish the limiting distribution of the number of loops. The situation is complicated by the presence of the term involving $\partial_v Y$ in (9.2.6), which introduces higher order derivatives with respect to v at $v = 1$ when computing the asymptotic of the moments.

Theorem 9.3.5. Let Y_n denote the total number of loops in a random rooted map with n edges. Then

$$\frac{Y_n}{n} \xrightarrow{d} \mathcal{L}, \tag{9.3.9}$$

where $\mathcal{L}$ is a probability measure on $[0, 1]$.

Proof. First, we show by induction that there exist constants $\eta_{k,\ell}$, such that as $n \to \infty$,

$$[z^n]\partial_v^k \partial_w^\ell Y(z, v, w)\big|_{v=w=1} \sim \eta_{k,\ell} n^{k+\ell+1} \phi_n, \quad k, \ell \geqslant 0. \tag{9.3.10}$$

For $k = \ell = 0$ the statement clearly holds. Let $y_n^{(k,\ell)} := [z^n]\partial_v^k \partial_w^\ell Y(z, v, w)\big|_{v=w=1}$ for larger $k, \ell \geqslant 0$. By translating (9.2.6) into the corresponding recurrence for the coefficients and by collecting the dominant terms (using the induction hypothesis (9.3.10)), we deduce that

$$y_n^{(k,\ell)} \sim (2n + k) y_{n-1}^{(k,\ell)} + \ell y_{n-1}^{(k+1,\ell-1)} + (2kn - 2k) y_{n-1}^{(k-1,\ell)} + \mathbf{1}_{[k=0]} y_{n-1}^{(k,\ell)}.$$

Accordingly, we are led to the recurrence

$$\eta_{k,\ell} = \frac{1}{k + 2\ell + \mathbf{1}_{[k>0]}} (2k\eta_{k-1,\ell} + \ell\eta_{k+1,\ell-1}), \tag{9.3.11}$$

for $k + \ell > 0$ (provided that we interpret $\eta_{k,\ell} = 0$ when any index becomes negative). In particular, when $\ell = 0$, we obtain the moments of the random variable E_n, the number of root edges: $\eta_{k,0} = \frac{2^{k+1}}{k+1}$, which coincides with the moments of the uniform random variable Uniform$[0, 2]$.

Let us present several first coefficients:

ℓ \ k	**0**	1	2	3
0	**2**	2	8/3	4
1	**1**	7/6	26/15	14/5
2	**7/12**	139/180	391/315	887/420
3	**139/360**	1133/2016	21631/22680	63559/37800
4	**1133/4032**			
5	0.2188..			
6	0.1787..			
7	0.1510..			
8	0.1308..			
9	0.1155..			

Note that $\eta_{0,\ell} \cdot (2\ell + 1)!$ are integer numbers:

$$(2, 6, 70, 1946, 101970, 8735782, \ldots)$$

as well as

$$\eta_{0,\ell} \cdot (2\ell - 1)! \cdot \ell \colon (1, 7, 139, 5665, 397081, 42817831, \ldots)$$

Neither of them is present in OEIS.

Let us show that the coefficients $\eta_{0,\ell}$ satisfy Carleman's condition

$$S_n := \sum_{\ell=1}^{n} \eta_{0,\ell}^{1/2\ell} \underset{n\to\infty}{\longrightarrow} \infty \ .$$

Indeed, as follows from the recurrence (9.3.11),

$$\eta_{0,\ell} \geqslant \frac{\ell\eta_{1,\ell-1}}{2\ell} = \frac{1}{2}\eta_{1,\ell-1} \geqslant \frac{1}{2} \cdot \frac{2\eta_{0,\ell-1}}{2+2\ell-2} = \frac{1}{\ell}\eta_{0,\ell-1} \ ,$$

in what follows that $\eta_{0,\ell} \geqslant \frac{2}{\ell!}$. Therefore, since $(\frac{1}{\ell!})^{1/2\ell} \sim e^{1-\log \ell} \sim \frac{e}{\ell}$, it follows that partial sums grow at least as fast as harmonic sum, and therefore, the sequence of partial sums is divergent. In particular, this implies that the condition of Hausdorff moment problem is satisfied, i.e. $\eta_{0,\ell}$ uniquely determine the limiting random variable defined on $[0,1]$. □

9.4 Combinatorics of map statistics

We examine briefly the combinatorial aspect of the map statistics, relying our arguments on the close connection between maps and chord diagrams (see [Cor09]).

Recall that a *chord diagram* [FN00] with n chords is a set of vertices labelled with the numbers $\{1, 2, \ldots, 2n\}$ equipped with a perfect matching. A chord diagram is *indecomposable* if it cannot be expressed as a concatenation of two smaller diagrams.

Why the root degree follows a uniform law? We begin with Cori's bijection [Cor09] between rooted maps and indecomposable diagrams. In this bijection, each chord connecting labels i and j corresponds to matching of the half-edges with labels i and j. The set of half-edges incident to each vertex of the resulting map corresponds to the set of nodes to the right of the starting points of the so-called *outer chords*, i.e. chords that do not lie under any other chord.

Proposition 9.4.1. There exists a bijection between rooted maps of root degree d with n edges, and indecomposable diagrams with $n+1$ chords such that the vertex $k-2$ is matched with vertex 1.

Once this proposition is available, it leads to a simpler and more intuitive proof of Theorem 9.3.4 as follows. In a (not necessarily indecomposable) diagram, the label of the vertex matched with 1 follows exactly a uniform law on $\{2, \ldots, 2n\}$. But a diagram is almost surely an indecomposable diagram (because its cardinality is asymptotically the same); thus the label of the vertex matched with 1 divided by $2n$ obeys asymptotically a uniform law on $[0,1]$ (or Uniform$[0,2]$ if divided by n as in Theorem 9.3.4).

Uniform random generation. Cori's bijection is also useful for generating random rooted maps. Uniformly sampling a random diagram can be achieved by adding the chords sequentially one after another. If this procedure results in an indecomposable diagram, it is rejected (which occurs with asymptotic probability 0). A successful sampled diagram is then transformed into a map using Cori's bijection [Cor09]. Figure 9.8 shows two instances of random maps thus generated.

The number of leaves. Another bijection in [CYZ16] is useful in proving the Poisson limit law of the number of leaves. This bijection sends leaves of a map into the *isolated* chords (namely, edges connecting vertices k and $k+1$) of an indecomposable chord diagram. According to [FN00, Theorem 2], the number of isolated edges in a random chord diagram has a Poisson distribution with parameter 1. We can then deduce the following theorem.

Theorem 9.4.1. The number of leaves in a random map with n edges follows asymptotically a Poisson law with parameter 1.

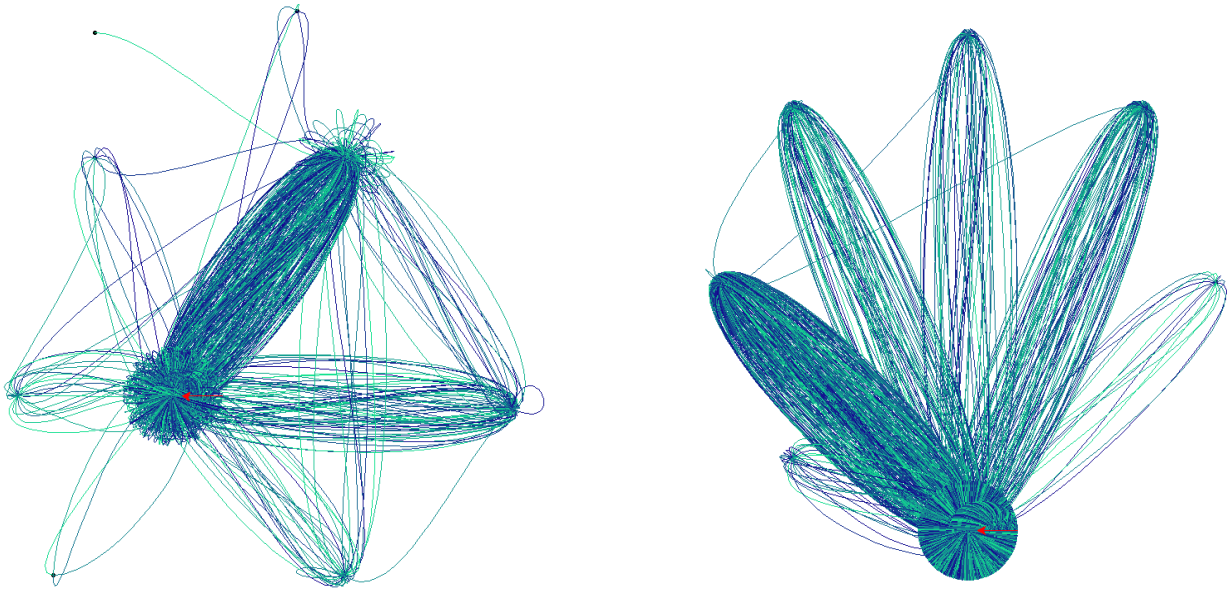


Figure 9.8: Random rooted maps, respectively with 1000 and 20000 edges.

Two dual parameters. We briefly remark that two other parameters, namely *root face degree* and the number of *trivial loops* do not seem easily dealt with by the method of generating functions because marking them requires additional nested information such as the degrees of all the faces. However, such parameters can be easily marked in their corresponding dual maps. Their limit distributions are uniform and Poisson, respectively.

Chapter 10

Applications of random sampling

Contents

10.1 Software verification

The tactics of program verification range from experimentation to formal verification techniques which prove the correctness of a program. While formal verification is a promising field, most of the industry-scaled applications are too heavy to be fully rigorously described.

An example of a very general application is an *optimising compiler*. Intrinsically, a compiler of any programming language typically takes an input and generates a so-called *intermediate representation* (abstract syntax tree) which is essentially a tree-like structure which keeps the abstract nature of the code written and is converted to the *byte-code* later. Most optimisations are done at the stage of the intermediate representation then. This standardised representation allows to use well-established tools like GNU compiler collection and LLVM (low-level virtual machine) to process a variety of programming languages in a uniform manner and for different target architectures.

To have a picture how optimising compilers actually work in reality, we may consider a very widely-used example which is an *compiler's optimisation level* in GCC. A variety of `-O` flags such as

`-O0, -O1, -O2, -O3, -Ofast`

provide different levels of speed optimisation. While the nature of such optimisations may be highly technical, involving how the float numbers are multiplied, the optimisations regarding the intermediate representation can be considered as combinatorial ones. At the same time, optimisation above `-O2` is typically *not recommended* as it yields some *undefined behaviour* and breaks some packages included.

Similar optimisation flags exist for different compilers, and external code optimisers can be used as well. In order to avoid undefined behaviour, testing or verification is required. One prominent example of an optimising compiler is CompCert for C99 written by a team led by Xavier Leroy [Ler+12]. This optimising compiler is fully programmed and *proved* in Coq, an interactive theorem prover programming language associated with Thierry Coquand and Gérard Pierre Huet [Bar+99].

Another approach, discussed in [Pał12] consists in generating many random input instances (programs) and verifying statistical properties of these instances before and after optimisations. This approach does not provide rigorous evidence, but rather a heuristic one. For example, a tool CSmith generating random C programs [Yan+11], allowed to detect as many as 325 previously unreported bugs in GCC, LLVM, some commercial C compilers and even CompCert whose core is formally verified. The technique suggested in [Pał12] is to generate *simply typed lambda terms* in order to test Haskell's optimising compiler.

One of the new applications of the multiparametric Boltzmann tuner Chapter 5 is the generation of *untyped lambda terms* also known as *closed lambda terms*, with a lot of flexibility and control. It is known that random generation with a *uniform distribution* corresponding to univariate Boltzmann samplers is not sufficient to quickly find the bugs in the compilers, while the latter fail only at some very specific *corner cases*, when one subset of the parameters behaves normally, and the other part takes very large or very small values. What is also important, is the ability to find relatively small counterexamples, since the size of the generated object can also be tuned.

As we show in Chapter 8, specifications originating from unambiguous context-free grammars allow to mark several parameters in plain and closed lambda terms allowing skewed random generation. A few possible examples of parameters might include:

(*i*) number of atomic nodes of distinguished colors;

(*ii*) number of redexes;

(*iii*) number of head abstractions;

(*iv*) number of closed subterms;

(*v*) number of any tree-like patterns.

While generation of simply typed lambda terms stays an open problem for now, the flexible generation of corresponding closed terms can potentially lead to advances in testing techniques.

10.2 Belief propagation for RNA design

One of the practical problems requiring a proper multiparametric Boltzmann tuning was communicated to me by Yann Ponty and his coauthor Sebastian Will. In this section, the pictures are re-drawn, as the original pictures are copyrighted by Springer. The concepts of the pictures and the description of the presented application come entirely from [Ham+19], while the method used in the aforementioned paper is different. On a general level, the problem is formulated as follows.

Problem 10.2.1. Generate, uniformly at random, RNA structures of given length N, such that, given a list of secondary structures $(s_1, s_2, \ldots, s_m)$ and corresponding energies $(E_1, E_2, \ldots, E_m)$, each secondary structure s_i is possible for generated RNA chain and has an (average) energy E_i.

This section is intended to give an intuitive explanation of all the terms introduced and guide through the process of random generation.

An RNA chain is a sequence of symbols from the alphabet $\{\mathsf{A}, \mathsf{U}, \mathsf{G}, \mathsf{C}\}$, corresponding to nucleotides *adenine*, *uracil*, *guanine* and *cytosine*. The fifth element, *thymine*, conventionally denoted by letter T, replaces uracil in DNA. The *secondary structure* of an RNA is a *chord diagram* with non-intersecting chords constructed on a subset of symbols of the sequence such that each chord may connect only three possible combinations $\mathsf{C}-\mathsf{G}$, $\mathsf{G}-\mathsf{U}$ and $\mathsf{U}-\mathsf{A}$. Sampling tree alignments from Boltzmann distribution is also considered in [CCP16].

In a basic setting, each of the three types of chords receives its own energy value determined from chemical properties, and the total energy of the secondary structure configuration is then obtained by adding up all the compound energies of the links.

Remark 10.2.1. In reality, the interactions of the adjacent pairs of the secondary configuration are also taken into account, but they are negligible. Their contribution can be investigated by considering a table of interactions of adjacent pairs. The next refinement requires taking triples, etc.

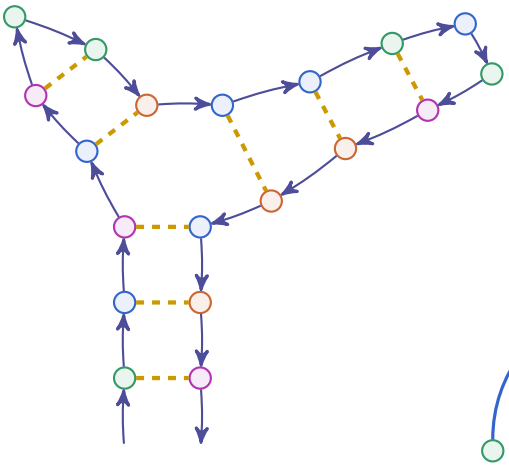


Figure 10.1: An example of a secondary structure

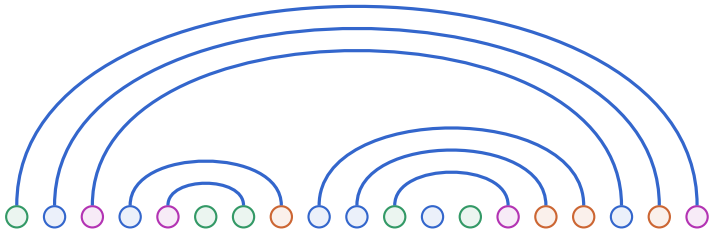


Figure 10.2: RNA secondary structure viewed as a chord diagram

The first step is to construct a graph whose edges are the union of the chords of the chord diagrams corresponding to the secondary structures.

Lemma 10.2.1. There exists a nucleotide assignment if and only if the resulting graph is bipartite.

Proof. According to a given nucleotide assignment, the vertices can be partitioned in such a way that the nodes with labels C and U belong to the first part, and the nodes with labels G and A belong to the second. The edges can exist only between the parts, but not inside them, therefore a graph is bipartite. Conversely, if a graph is bipartite, it is possible to label all the nodes from the first part into G, and from the second part into U. □

Curiously, the number of RNA satisfying all the secondary structures, can be expressed the number of *independent sets*, that is, subsets of nodes having no pairwise edges between them.

Lemma 10.2.2. The number of nucleotide assignments complying the secondary structures forms a 2^{ℓ}-to-1 bijection with distinguished independent sets of the graph formed of chords of the secondary structures, where ℓ is the number of connected components.

Proof. Consider one connected component in a graph of chords. The goal is to choose nucleotide labellings in such a way that the nodes from the distinguished independent set have only labels C and A, and all other

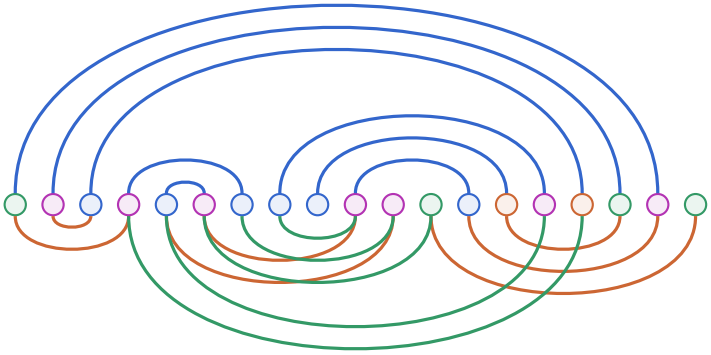


Figure 10.3: Several superimposed secondary structures

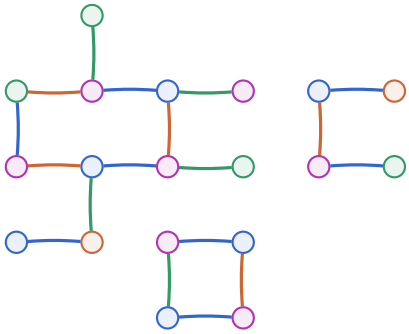


Figure 10.4: Graph of nucleotide connections from the secondary structures

nodes can only have labels G and U. It is clear that once one of the labels in a graph is fixed, all other labels in this connected component can be reconstructed uniquely. Therefore, for every connected component, there are exactly two label choices. □

The problem of enumeration of independent sets is known to be #P-complete, so no hope for a polynomial algorithm exists, and the presence of the energies and the tuning only makes the problem even more complex.

Such problems are known to be *fixed parameter tractable*, by analogy with independent sets in sparse random graphs with bounded excess of the complex component. A special tool called *tree-decomposition* is introduced, and the complexity of enumeration becomes a polynomial of the input, with a constant multiple possibly exponential in a special parameter called *tree-width*.

Definition 10.2.1 (Tree decomposition)**.** A *tree decomposition* of G is a rooted tree T such that

(i) every node of T represents a subset of nodes of G;

(ii) the root represents an empty set of nodes;

(iii) if an edge connects a *parent* and a *child* node, the contents of the parent node can be formed by removing an arbitrary subset of vertices from the child node and by adding at most one new node;

(iv) every edge $u-v$ of G is present in T, that is, connects two nodes t_1 and t_2 such that $u \in t_1$ and $v \in t_2$;

(v) for every node $v \in G$ the subgraph corresponding to all nodes $t \in T$ containing v forms a tree.

The *treewidth* of the tree decomposition is equal to the maximum cardinality over nodes $t \in T$. It can be shown that if a graph has excess r, it has a tree-decomposition with a tree-width not exceeding $r+3$, constructible by a dfs.

A physical notion of the partition function is synonymic to the notion of the generating function where the size of the object corresponds to the energy of a physical state, and the value of the argument of the generating function is related to the temperature of the system.

Definition 10.2.2. The *partition function* corresponding to a given list of secondary structures $(s_1, \ldots, s_m)$ is defined as

$$P(\beta) = \sum_S \exp\left(-\beta E(S; s_1, \ldots, s_m)\right) \tag{10.2.1}$$

where the sum is taken over all admissible RNA sequences S and $E(S; s_1, \ldots, s_m)$ is the energy of S taken as a sum of energies corresponding to each of the energies of the secondary structures.

To each node of the tree-width we assign a function $m(\beta; x_1, \ldots, x_e)$ whose arguments $(x_1, \ldots, x_e)$ are the nodes that are not removed while passing to the parent of this node, the values that each of the arguments may take is a set of possible nucleotide assignments. The value β corresponds to the argument of the partition function. The value of the function $m(x_1, \ldots, x_e)$ can be then interpreted as the partition function of a partial assignment corresponding to the subtree of the given vertex.

10.2.1 Belief propagation

Belief propagation, also known as *sum-product message passing* algorithm [K+01] is designed to compute the marginal distributions of discrete random variables. Let $g(x_1, \ldots, x_n)$ be some real-valued function whose arguments take the values in a finite alphabet Σ. We are interested in computing *marginal functions* in the form

$$g_i(x_i) := \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} g(x_1, \ldots, x_n). \tag{10.2.2}$$

This seems to be an easy task if $g(\boldsymbol{x})$ represents a joint probability function of independent random variables, that is, when this function is expressed as a product of individual arguments, dependent on a single argument each. It turns out that in some more general situations, a fast computation algorithm is possible.

We consider situations when

$$g(x_1, \ldots, x_n) = \prod_{j \in J} f_j(X_j) \tag{10.2.3}$$

where each X_j is a subset of $\{x_1, \ldots, x_n\}$ and J is a discrete index set. For example,

$$g(x_1, x_2, x_3, x_4, x_5) = f_A(x_1) f_B(x_2) f_C(x_1, x_2, x_3) f_D(x_3, x_4) f_E(x_3, x_5). \tag{10.2.4}$$

A *factor graph* is then defined as a bipartite graph whose nodes are of two types: the *variable nodes* and *argument nodes*. The sum-product algorithm is initially defined for the case when the factor graph is a tree, but in modern applications related to physics and machine learning, it is applied to situations when the underlying graphs contains cycles.

In order to directly see how the marginal distributions of a tree-like function can be simplified straight away, using the example above, we use the distributive property and rooting a tree at the corresponding vertex to obtain

$$g_1(x_1) = f_A(x_1) \sum_{\sim\{x_1\}} \Big(f_B(x_2) f_C(x_1, x_2, x_3) \Big(\sum_{\sim\{x_3\}} f_D(x_3, x_4) \Big) \Big(\sum_{\sim\{x_3\}} f_E(x_3, x_5) \Big) \Big), \tag{10.2.5}$$

where $\sum_{\sim\{x_i\}}$ denotes a sum taken over all arguments except x_i.

However, a *message-passing procedure* makes the process even more efficient by de-symmetrising the tree and starting a computation simultaneously at all its leaves. The messages are passed from variables to functions, and from functions to variables, to all of their neighbours. They are defined as follows:

$$\mu_{x \to f}(x) := \prod_{h \in n(x) \setminus \{f\}} \mu_{h \to x}(x), \tag{10.2.6}$$

$$\mu_{f \to x}(x) := \sum_{\sim\{x\}} \left(f(X) \prod_{y \in n(f) \setminus \{x\}} \mu_{y \to f}(y) \right), \tag{10.2.7}$$

where $n(v)$ denotes the set of neighbours of a given node v, and $n(f)$ denotes the set of arguments of the function f. The meaning of x in the index of $\mu_{x \to f}$ and $\mu_{f \to x}$ is the corresponding index of the variable x in the list of arguments $(x_1, \ldots, x_n)$, while the meaning of x in the argument of $\mu_{x \to f}(x)$ is the value of x which is taken from the alphabet Σ where x is allowed to take values.

It is shown in [K+01] that after initialising the message functions at leaves (both variable leaves and function leaves) as follows:

$$\mu_{f\to x}(x) := f(x), \tag{10.2.8}$$
$$\mu_{x\to f}(x) := 1, \tag{10.2.9}$$

and running the message-passing until every edge is initialised.

10.2.2 Message passing for tree-decomposition of RNA

Define a message function $m_{t_1\to t_2}$ from a child node $t_1 \in T$ to a parent node $t_2 \in T$ as follows: it takes a set of arguments $t_1 \cap t_2$ and

$$m_{t_1\to t_2}(x_{t_1\cap t_2}) := \sum_{\text{allowed } x_{t_1\setminus t_2}} \prod_{t\to t_1} m_{t\to t_1}(x_{t\cap t_1}). \tag{10.2.10}$$

The *allowed* arguments x_t mean that inserting nucleotides at the positions corresponding to the nodes of t induce only complementary "allowed" pairs.

Starting the message passing at leaves and finishing at $m_{\varnothing}$ it is then possible to compute the total number of independent sets. What is more important, we can modify the message function to add an argument which is a variable marking the energy of the configuration in a respective secondary structure, to obtain a partition function. The adjustment is seen as

$$m_{t_1\to t_2}(x_{t_1\cap t_2}) := \sum_{\text{allowed } x_{t_1\setminus t_2}} \prod_{t\to t_1} m_{t\to t_1}(x_{t\cap t_1}) u_c^{-\text{energy of added edge}}. \tag{10.2.11}$$

where u_c is a marking variable with an index c representing the index of the secondary structure to which an edge belongs to. (If there is an edge belonging to several secondary structures, a product should be then taken.)

This approach builds an *algebraic grammar* of multivariate generating functions. By setting the values of the variables u_c it is possible to do Boltzmann generation from this grammar to obtain RNA sequences with expected values of energies. In the paper [Ham+19] the weight optimisation is done using the backtrack, while using the current approach a direct optimisation is possible.

10.3 Bose–Einstein condensate in quantum harmonic oscillator

Bose–Einstein condensation is a phenomenon occurring in quantum physics and can be described as follows: an gas of N particles (typically, *bosons*, like photons or any other particle with an integer spine, without going into too much details) is confined in a magnetic trap at a certain temperature; each of the bosons is only allowed to take a discrete set of energies and the joint probability of energies is a well-defined function, which is also a function of temperature; then, as the temperature lowers, below a certain critical point $t < t_c$ most of the bosons with high probability appear in the lowest possible energy state.

So, essentially, Bose–Einstein condensation can be regarded as a purely probabilistic phenomenon, if rigorously formulated. More generally, this phenomenon occurs when the particles live in dimension d where d is not necessarily equal to 3. One of the modern observations in network science, in the study of real-world social networks by Bianconi and Barabási [BB01] in 2001 was the explanation of the uneven degree distributions in such networks. This explanation suggested that an evolving graph can be regarded as a process of adding particles into a bosonic gas. Each particle then corresponds to a half-edge in the graph, and Bose–Einstein condensation can be viewed as a topological phase transition.

One of possible applications of multidimensional tuning is weighted generation of integer partitions. In fact, integer partitions (such as $16 = 1 + 3 + 3 + 4 + 5$) can be represented as the states of one-dimensional quantum harmonic oscillator. Each summand of the integer partition corresponds to a particle, the value of the summand corresponds to the energy of the partition. The generalisation to multidimensional case is done by further partitioning of each of the summands of the partition into at most d summands, so that a particle

Bose gas	network evolution
temperature	temperature
energy	energy
particle	half-edge
number of energy levels	$\leqslant$ number of nodes
Bose–Einstein condensation	topological phase transition

Table 10.1: Comparison of Bose gas and network evolution

assembly is then viewed as *partition of partitions*. In fact, the Bose–Einstein condensation is nothing else but the limit shape of the partition, assuming that the partitions are taken from Boltzmann distribution.

Weighted partition	Random particle assembly
Sum of numbers	Total energy
Number of colours	Dimension (d)
Row of Young table	Particle
Number of rows	Number of particles
Number of squares in the row	Energy of a particle (λ)
Partition limit shape	Bose–Einstein condensation
$\binom{d+\lambda-1}{\lambda}$	Number of particle states

Table 10.2: Correspondence between weighted integer partitions and quantum oscillator states

10.4 Multiclass queueing networks

Queueing theory is an important field of industrial engineering. A queueing network can be described as a directed graph whose nodes are *queues* (also called *servers*) and are capable of processing *jobs* (sometimes called *clients* or *customers*). The rules according to which the jobs are processed, differ over a large variety of queue types and network types. The choice of the design of a queueing system for particular practical purposes is dictated by statistical behaviour of each of the systems under process arrival assumptions. Therefore, a detailed understanding of such behaviours is crucial for applications.

One of the possible approaches to queueing theory is a detailed analysis of the case when there are two processors by representing the capacity vector of the two queues by an integer point and by considering the evolution of the system as a random walk on an integer plane. This approach and its generalisations is extensively covered in a very recent book [Fay+99].

Another approach to multiparametric queueing systems is to compute the probability generating function of the stationary distribution and study the limiting properties using the multidimensional complex integration techniques.

10.4.1 Gordon–Newell queueing network

In Gordon–Newell network, all the customers are indistinguishable. A queueing network is represented by a complete graph with m nodes whose behaviour is defined by the following rules:

1. The total population of the graph is $K = k_1 + k_2 + \ldots + k_m$ where k_i is the number of customers in the queue number i;

2. Each queue is designed by principle *first in, first out*;
3. Service time at queue i is exponentially distributed with parameter μ_i; after a client is served, he jumps from queue number i to queue number j with probability p_{ij} where $(p_{ij})_{i,j=1}^m$ is a given transition matrix;
4. No client enters and no client leaves the network.

It is proven in [GN67] that this queueing network has a stationary distribution whose probability generating function is

$$p(z) = \frac{1}{(1-\rho_1 z)\cdots(1-\rho_m z)}, \tag{10.4.1}$$

where $\rho_i = \lambda_i/\mu_i$, and λ_i is the stationary vector of the transition matrix

$$\lambda_i = \sum_{j=1}^{m} \lambda_j p_{ji}.$$

The Boltzmann generation from the stationary distribution of Gordon–Newell network is described in [Rov17] and sampling techniques from network with bounded capacities, such as fast computation of the *partition function* (here: the synonym of the probability generating function) is given in [BBR14].

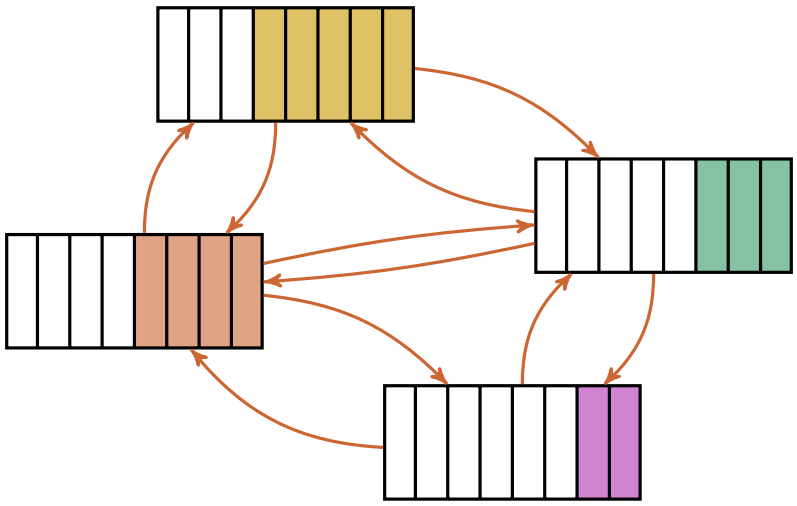


Figure 10.5: An example of a state of Gordon–Newell network

10.4.2 Introducing customer classes

The product form that the generation function Equation (10.4.1) has, suggests that there should be an analog of the symbolic approach for queueing networks: the generating function specifies the *sequence* of m *sequences*, while this description is explicitly corresponding the definition of the network.

A multiclass generalisation, where instead of indistinguishable customers, J classes of customers are introduced, is considered in [Kog02; PW08; BM93]. In these papers, the authors use the techniques of multivariate saddle-point integration to obtain the asymptotic properties of such networks when the number of nodes and customer types is possibly large.

The first example of a multiclass queueing network with m *processor sharing* stations and J customer types has a limiting distribution with a probability generating function

$$G(z_1,\ldots,z_J) = \frac{1}{\prod_{k=1}^{m}(1-\sum_{j=1}^{J}\rho_{ji}z_j))} \tag{10.4.2}$$

where the matrix $(\rho_{ji})_{j,i}$ is determined by network properties. A slightly modified case includes one infinite server (such that every arriving job is served immediately and this the waiting time is zero) and m processor sharing stations. The corresponding probability generating function of the stationary distribution then becomes

$$G(z_1, \dots, z_J) = e^{z_1 + \dots z_J} \frac{1}{\prod_{k=1}^{m} (1 - \sum_{j=1}^{J} \rho_{ji} z_j))}. \tag{10.4.3}$$

The values ρ_{ji} are determined by network properties.

Other variations are possible. One common thing about these networks is that the stationary distribution has product form which yields a simple expression for the multivariate generating function. In such a case, sampling from a weighted distribution is possible, with additional control over the expected queue capacities.

10.5 Combinatorial learning and Maximum Likelihood

The method of maximum likelihood estimation plays a central role in mathematical statistics. It allows to construct asymptotically effective[1] estimates for the unknown parameters using a finite amount of data available, assuming that the data comes from a known family of distributions with fixed values of parameters.

Consider the following basic example. If $X_1, \dots, X_n \in \mathbb{R}^d$ are independent random variables from distribution $\mathbb{P}_\theta$, and $\theta \in \Theta \subset \mathbb{R}^r$ is an r-dimensional *parameter*, the *log-likelihood function* is defined as

$$L(X_1, \dots, X_n; \theta) := \log \mathbb{P}_\theta(X_1 \wedge \dots \wedge X_n) = \sum_{i=1}^{n} \log \mathbb{P}_\theta(X_i). \tag{10.5.1}$$

Without *a priori* knowing the value θ, the function can be computed for all possible values of this (vector) parameter. The value of this parameter is then estimated as

$$\widehat{\theta} = \widehat{\theta}(X_1, \dots, X_n) := \arg\max_{\theta \in \Theta} L(X_1, \dots, X_n; \theta). \tag{10.5.2}$$

The obtained estimate is then a random variable, since it is dependent on random quantities $X_1, \dots, X_n$.

Under certain smoothness assumptions on the distribution family, the obtained estimate $\widehat{\theta}_n$ can be proven to converge to a normal random variable at speed $\sqrt{n}$ with the mean value equal the target parameter, and the variance equal to so-called *asymptotic information matrix*, or *Fisher information matrix*.

The idea inside the maximum likelihood method extends further than just parameter estimation and can be applied in the following settings:

1. regression;
2. model selection;
3. density estimation
4. hypothesis testing;
5. generalised linear models;

10.5.1 Multiparametric tuning and maximum likelihood approach

The procedure of multiparametric tuning described in Chapter 5, can be viewed as likelihood optimisation. Consider a multivariate generating function $F(z)$ for a combinatorial class $\mathcal{F}$:

$$F(z) := \sum_{\boldsymbol{n} \succeq \boldsymbol{0}} a_{\boldsymbol{n}} \boldsymbol{z}^{\boldsymbol{n}}.$$

[1] Maximum likelihood estimator achieves the Cramér–Rao bound for asymptotically large sample sizes; when the size of the sample is finite, non-asymptotic improvements upon maximum likelihood estimation method are possible, such as unbiased variance estimator for normal variable.

Suppose that a sequence of combinatorial objects $X_1, \ldots, X_N$ has been sampled. It is convenient to assume that the whole generated objects are not available, but rather the corresponding values of their parameters $\boldsymbol{n}_1, \ldots, \boldsymbol{n}_N$. Then, the log-likelihood function for the sample is written as

$$L(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_N; \boldsymbol{z}) = \sum_{i=1}^{N} \mathbb{P}_{\boldsymbol{z}}(\chi(X) = \boldsymbol{n}_i) = \sum_{i=1}^{N} \log \frac{a_{\boldsymbol{n}_i} \boldsymbol{z}^{\boldsymbol{n}_i}}{F(\boldsymbol{z})}, \tag{10.5.3}$$

where $\chi(X)$ is a function returning the size parameter vector of X.

Taking the equation for maximum likelihood estimate, we obtain

$$\nabla_{\boldsymbol{z}} L(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_N; \boldsymbol{z}) = 0$$

which is equivalent to

$$\frac{\sum_{i=1}^{N} \boldsymbol{n}_i}{N} = \boldsymbol{z} \frac{\nabla F(\boldsymbol{z})}{F(\boldsymbol{z})}. \tag{10.5.4}$$

Interesting things happen when we use the logarithmic transform $\boldsymbol{z} \mapsto e^{\boldsymbol{\xi}}$. From statistical viewpoint, it corresponds to passing to an *exponential family in a canonical form*. In particular, this representation gives an alternative proof that $\log F(e^{\boldsymbol{\xi}})$ is convex.

Lemma 10.5.1. Let Ξ denote the random vector yielding the parameter value $\boldsymbol{n}$ for Boltzmann distribution with parameter $\boldsymbol{z}$. Let $\boldsymbol{z} = e^{\boldsymbol{\xi}}$. Then,

$$\nabla^2_{\boldsymbol{\xi}} \log F(e^{\boldsymbol{\xi}}) = \operatorname{Cov}_{\boldsymbol{z}}[\Xi]. \tag{10.5.5}$$

Since the covariance matrix is non-negative definite, the function $\log F(e^{\boldsymbol{\xi}})$ is convex.

Proof. Since the variables are identically distributed, it is sufficient to prove the lemma for a single random variable Ξ. We shall use the fact that for any family of distributions, Fisher information at the is equal to the curvature matrix, defined as follows:

$$\operatorname{Cov}_{\theta^*} \nabla L(X, \theta)|_{\theta=\theta^*} = -\mathbb{E}_{\theta^*} \nabla^2 L(X, \theta)|_{\theta=\theta^*}. \tag{10.5.6}$$

The first gradient of the likelihood function can be computed as

$$\nabla_{\boldsymbol{\xi}} L(\Xi, e^{\boldsymbol{\xi}}) = \nabla_{\boldsymbol{\xi}} [\log a_{\Xi} + \Xi^{\top} \boldsymbol{\xi} - \log F(e^{\boldsymbol{\xi}})] = \Xi - \nabla_{\boldsymbol{\xi}} \log F(e^{\boldsymbol{\xi}}).$$

The variance of this gradient under the true value $\boldsymbol{\xi} = \boldsymbol{\xi}^*$ is equal to

$$\operatorname{Cov}_{\boldsymbol{\xi}^*} [\Xi - \nabla_{\boldsymbol{\xi}} \log F(e^{\boldsymbol{\xi}})] = \operatorname{Cov} \Xi.$$

Next, the expected value of the log-likelihood can be expressed as

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\xi}^*} L(\Xi, e^{\boldsymbol{\xi}}) &= \sum_{\boldsymbol{n}} L(\boldsymbol{n}, e^{\boldsymbol{\xi}}) \frac{a_{\boldsymbol{n}} e^{\boldsymbol{n}^{\top} \boldsymbol{\xi}^*}}{F(e^{\boldsymbol{\xi}})} \\ &= \frac{1}{F(e^{\boldsymbol{\xi}})} \sum_{\boldsymbol{n}} \left[e^{\boldsymbol{n}^{\top} \boldsymbol{\xi}^*} a_{\boldsymbol{n}} \log a_{\boldsymbol{n}} + \boldsymbol{n}^{\top} \boldsymbol{\xi} \cdot a_{\boldsymbol{n}} e^{\boldsymbol{n}^{\top} \boldsymbol{\xi}^*} - a_{\boldsymbol{n}} e^{\boldsymbol{n}^{\top} \boldsymbol{\xi}^*} F(e^{\boldsymbol{\xi}}) \right]. \end{aligned}$$

Taking the second derivative, we obtain

$$\nabla^2_{\boldsymbol{\xi}} \mathbb{E}_{\boldsymbol{\xi}^*} L(\Xi, e^{\boldsymbol{\xi}}) = -\nabla^2_{\boldsymbol{\xi}} \log F(e^{\boldsymbol{\xi}}) \frac{\sum_{\boldsymbol{n}} a_{\boldsymbol{n}} e^{\boldsymbol{n}^{\top} \boldsymbol{\xi}^*}}{F(e^{\boldsymbol{\xi}})}. \tag{10.5.7}$$

By using the fact that Fisher information is equal to the curvature matrix, we conclude the proof. □

10.5.2 Hidden parameter estimation

As we seen in the previous section, there are several intriguing connections between the likelihood estimation method and multiparametric Boltzmann tuning. However, if written explicitly, the maximum likelihood estimation does not give any new non-trivial parameter estimator, as the estimation of the expected parameter is just the average value among the observed parameters.

Nonetheless, such an analogy gives rise to a new approach of hidden parameter estimation, which we show to be #P-complete. Suppose that the objects are sampled from Boltzmann distribution with parameter

$$\boldsymbol{z} = (z_1, \dots, z_k)$$

but we observe only some part of the parameters $(n_1, \dots, n_\ell)$, $\ell < k$. The problem is to estimate the whole vector $\boldsymbol{z}$ using the available data.

Let us assume that the full vector of Boltzmann parameters is $(\boldsymbol{z}, \boldsymbol{u})$, while the observable part is only corresponding to $\boldsymbol{z}$. We need to estimate the whole vector $(\boldsymbol{z}, \boldsymbol{u})$. After expanding the likelihood maximiser equations, we obtain

$$\sum_{i=1}^{N} \boldsymbol{n}_i - N \frac{\nabla_{\boldsymbol{z}} F(\boldsymbol{z}, \boldsymbol{u})}{F(\boldsymbol{z})} = 0, \tag{10.5.8}$$

$$\sum_{i=1}^{N} \frac{\nabla_{\boldsymbol{u}} [\boldsymbol{z}^{\boldsymbol{n}_i}] F(\boldsymbol{z}, \boldsymbol{u})}{[\boldsymbol{z}^{\boldsymbol{n}_i}] F(\boldsymbol{z}, \boldsymbol{u})} - N \frac{\nabla_{\boldsymbol{u}} F(\boldsymbol{z}, \boldsymbol{u})}{F(\boldsymbol{z}, \boldsymbol{u})} = 0. \tag{10.5.9}$$

Since the coefficient extraction operator is known to be #P-hard for generating functions from algebraic grammars Chapter 5, the problem of hidden parameter estimation is also #P-hard.

At the same time, an efficient relaxation of this problem can be constructed and solved as a series of N individual tuning problems. The principle of Boltzmann sampling itself relaxates the size by superimposing Boltzmann distribution, so that

$$z^n [z^n] F(z) \approx F(z^*(n)), \quad z^*(n) \frac{F'(z^*(n))}{F(z^*(n))} = n. \tag{10.5.10}$$

Multiparametric tuning provides an oracle for computing $z^*(n)$ numerically. By using the same replacement in the coefficient extraction problem, namely,

$$\frac{\nabla_{\boldsymbol{u}} [\boldsymbol{z}^{\boldsymbol{n}_i}] F(\boldsymbol{z}, \boldsymbol{u})}{[\boldsymbol{z}^{\boldsymbol{n}_i}] F(\boldsymbol{z}, \boldsymbol{u})} \approx \frac{\nabla_{\boldsymbol{u}} F(\boldsymbol{z}^*(\boldsymbol{n}_i), \boldsymbol{u})}{F(\boldsymbol{z}^*(\boldsymbol{n}_i), \boldsymbol{u})}, \tag{10.5.11}$$

and after finding the vectors $\boldsymbol{z}^*(\boldsymbol{n}_i)$ by solving N tuning problems, we are able to solve the relaxed approximation of the hidden parameter estimation problem.

To conclude, the combination of maximum likelihood approach, and combinatorial Boltzmann approach can lead to some new developments and problem formulations on the intersection of mathematical statistics and analytic combinatorics. The main contribution lies in the possibility of taking advantage of the combinatorial source of the observed scalar parameters $\boldsymbol{n}_i$ that are not anymore just numbers, but some meaningful quantities like the number of nodes in a tree, or its height, or any other parameter. All this allows to do a more meaningful data analysis assuming that a corresponding model generates the data. Further tools like regression, model selection, are readily available and can be combined in order to produce further refinements.

10.6 Practical benchmarks

In this section we present several examples illustrating the wide range of applications of our tuning techniques. Afterwards, we briefly discuss our prototype sampler generator and its implementation details.

10.6.1 Polyomino tilings.

We start with a benchmark example of a rational specification defining $n \times 7$ rectangular tilings using up to 126 different tile variants (a toy example of so-called transfer matrix models, cf. [FS09, Chapter V.6, Transfer matrix models]).

Figure 10.6: Examples of admissible tiles

We begin the construction with defining the set T of admissible tiles. Each tile $t \in T$ consists of two horizontal layers. The base layer is a single connected block of width $w_t \leqslant 6$. The second layer, placed on top of the base one, is a subset (possibly empty) of w_t blocks, see Figure 10.6. For presentation purposes each tile is given a unique, distinguishable colour.

Next, we construct the asserted rational specification following the general construction method of defining a deterministic automaton with one state per each possible partial tiling configuration using the set T of available tiles. Tracking the evolution of attainable configurations while new tiles arrive, we connect relevant configurations by suitable transition rules in the automaton. Finally, we (partially) minimise the constructed automaton removing states unreachable from the initial empty configuration. Once the automaton is created, we tune the tiling sampler such that the target colour frequencies are uniform, i.e. each colour occupies, on average, approximately $\frac{1}{126} \approx 0.7936\%$ of the outcome tiling area. Figure 10.7 depicts an exemplary tiling generated by our sampler.

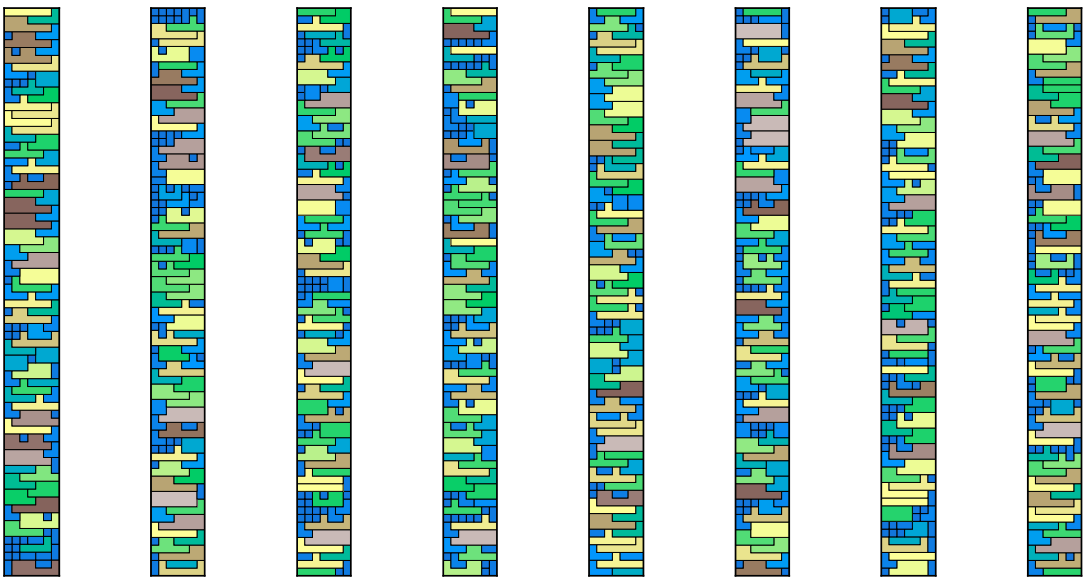


Figure 10.7: Eight random $n \times 7$ tilings of areas in the interval $[500; 520]$ using in total 95 different tiles.

The automaton corresponding to our tiling sampler consists of more than 2000 states and 28,000 transitions. We remark that this example is a notable improvement over the work of Bodini and Ponty [BP10] who were able to sample $n \times 6$ tilings using 7 different tiles (we handle 126) with a corresponding automaton consisting of roughly 1500 states and 3200 transitions.

10.6.2 Simply-generated trees with node degree constraints.

Next, we give an example of simple varieties of plane trees with fixed sets of admissible node degrees, satisfying the general equation

$$y(z) = z\phi(y(z)) \quad \text{for some polynomial} \quad \phi\colon \mathbb{C} \to \mathbb{C}\,.$$

Let us consider the case of plane trees where nodes have degrees in the set $D = \{0, \ldots, 9\}$, i.e. $\phi(y(z)) = a_0 + a_1 y(z) + a_2 y(z)^2 + \cdots + a_9 y(z)^9$. Here, the numbers $a_0, a_1, a_2, \ldots, a_9$ are non-negative real coefficients. We tune the corresponding algebraic specification so to achieve a target frequency of 1% for all nodes of degrees $d \geqslant 2$. Frequencies of nodes with degrees $d \leqslant 1$ are left undistorted. For presentation purposes all nodes with equal degree are given the same unique, distinguishable colour. Figure 10.8 depicts two exemplary trees generated in this manner.

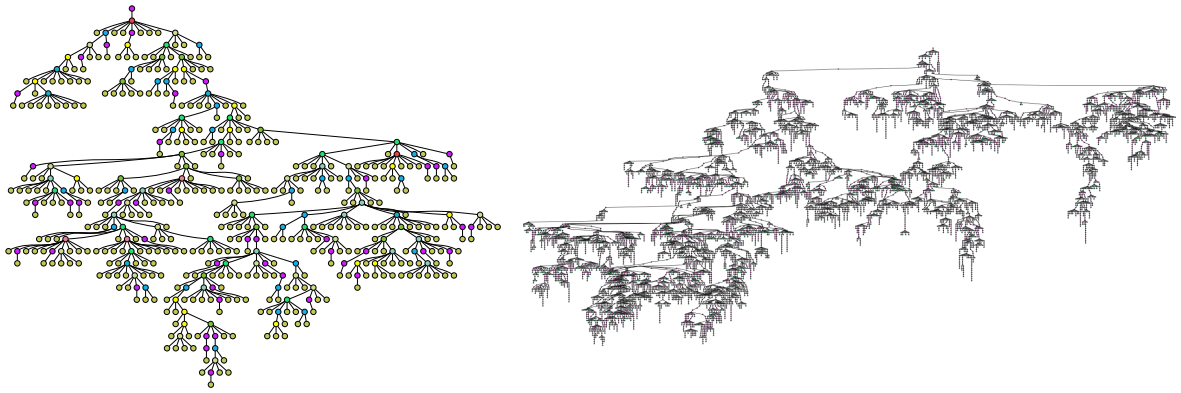


Figure 10.8: Two random plane trees with degrees in the set $D = \{0, \ldots, 9\}$. On the left, a tree of size in between 500 and 550; on the right, a tree of size in the interval [10 000; 10 050].

Empirical frequencies for the right tree of Figure 10.8 and a simply-generated tree of size in between $10,000$ and $10,050$ with default node degree frequencies are included in Table 10.3.

Node degree	0	1	2	3	4	5	6	7	8	9
Tuned frequency	- - -	- - -	1.00%	1.00%	1.00%	1.00%	1.00%	1.00%	1.00%	1.00%
Observed frequency	35.925%	56.168%	0.928%	0.898%	1.098%	0.818%	1.247%	0.938%	1.058%	0.918%
Default frequency	50.004%	24.952%	12.356%	6.322%	2.882%	1.984%	0.877%	0.378%	0.169%	0.069%

Table 10.3: Empirical frequencies of the node degree distribution.

We briefly remark that for this particular problem, Bodini, David and Marchal proposed a different, bit-optimal sampling procedure for random trees with given partition of node degrees [BDM16].

10.6.3 Variable distribution in plain λ-terms.

To exhibit the benefits of distorting the intrinsic distribution of various structural patterns in algebraic data types, we present an example specification defining so-called plain λ-terms with explicit control over the distribution of de Bruijn indices.

In their nameless representation due to de Bruijn [dBru72] λ-terms are defined by the formal grammar $L ::= \lambda L \mid (LL) \mid D$ where $D = \{\underline{0}, \underline{1}, \underline{2}, \ldots\}$ is an infinite denumerable set of so-called indices (cf. [Ben+17; GG16]). Assuming that we encode de Bruijn indices as a sequence of successors of zero (i.e. use a unary base representation), the class $\mathcal{L}$ of plain λ-terms can be specified as $\mathcal{L} = \mathcal{Z}\mathcal{L} + \mathcal{Z}\mathcal{L}^2 + \mathcal{D}$ where $\mathcal{D} = \mathcal{Z}\,\mathsf{Seq}(\mathcal{Z})$.

In order to control the distribution of de Bruijn indices we need a more explicit specification for de Bruijn indices. For instance:

$$\mathcal{D} = \mathcal{U}_0\mathcal{Z} + \mathcal{U}_1\mathcal{Z}^2 + \cdots + \mathcal{U}_k\mathcal{Z}^{k+1} + \mathcal{Z}^{k+2}\,\mathsf{Seq}(\mathcal{Z})\,.$$

Here, we roll out the $k+1$ initial indices and assign distinct marking variables to each one of them, leaving the remainder sequence intact. In doing so, we are in a position to construct a sampler tuned to enforce a uniform distribution of 8% among all marked indices, i.e. indices $\underline{0}, \underline{1}, \ldots, \underline{8}$, distorting in effect their intrinsic geometric distribution.

Figure 10.9 illustrates two random λ-terms with such a new distribution of indices. For presentation purposes, each index in the left picture is given a distinct colour.

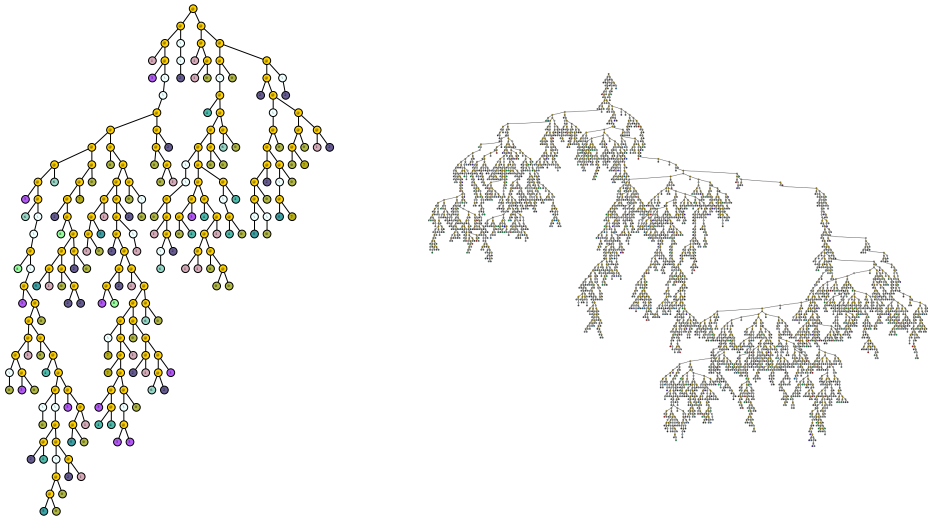


Figure 10.9: On the left, a random λ-term of size in the interval $[500; 550]$; on the right, a larger example of a random λ-term of size between $10,000$ and $10,050$.

Empirical frequencies for the right term of Figure 10.9 and a plain λ-term of size in between $10,000$ and $10,050$ with default de Bruijn index frequencies are included in Table 10.4.

Index	$\underline{0}$	$\underline{1}$	$\underline{2}$	$\underline{3}$	$\underline{4}$	$\underline{5}$	$\underline{6}$	$\underline{7}$	$\underline{8}$
Tuned frequency	8.00%	8.00%	8.00%	8.00%	8.00%	8.00%	8.00%	8.00%	8.00%
Observed frequency	7.50%	7.77%	8.00%	8.23%	8.04%	7.61%	8.53%	7.43%	9.08%
Default frequency	21.91%	12.51%	5.68%	2.31%	0.74%	0.17%	0.20%	0.07%	- - -

Table 10.4: Empirical frequencies (with respect to the term size) of index distribution.

Let us note that algebraic data types, an essential conceptual ingredient of various functional programming languages such as Haskell or OCaml, and the random generation of their inhabitants satisfying additional structural or semantic properties is one of the central problems present in the field of property-based software testing (see, e.g. [CH00; Pał12]). In such an approach to software quality assurance, programmer-declared function invariants (so-called properties) are checked using random inputs, generated accordingly to some predetermined, though usually not rigorously controlled, distribution. In this context, our techniques provide a novel and effective approach to generating random algebraic data types with fixed average frequencies of type constructors. In particular, using our methods it is possible to *boost* the intrinsic frequencies of certain desired subpatterns or *diminish* those which are unwanted.

10.6.4 Weighted partitions.

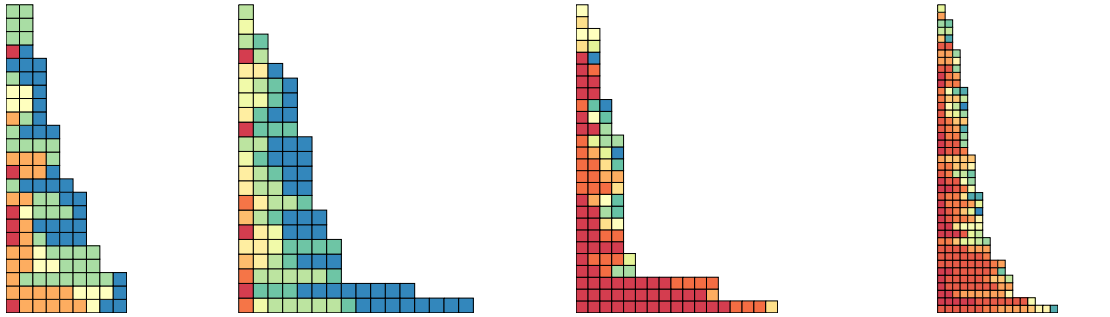
Integer partitions are one of the most intensively studied objects in number theory, algebraic combinatorics and statistical physics. Hardy and Ramanujan obtained the famous asymptotics which has later been refined by Rademacher [FS09, Chapter VIII]. In his article [Ver96], Vershik considers several combinatorial examples related to statistical mechanics and obtains the limit shape for a random integer partition of size n with $\alpha\sqrt{n}$ parts and summands bounded by $\theta\sqrt{n}$. Let us remark that Bernstein, Fahrbach, and Randall [BFR17] have recently analysed the complexity of exact-size Boltzmann sampler for weighted partitions. In the model of ideal gas, there are several particles (bosons) which form a so-called assembly of particles. The overall energy of the system is the sum of the energies $\Lambda = \sum_{i=1}^{N} \lambda_i$ where λ_i denotes the energy of i-th particle. We assume that energies are positive integers. Depending on the energy level λ there are $j(\lambda)$ possible available states for each particle; the function $j(\lambda)$ depends on the physical model. Since all the particles are indistinguishable, the generating function $P(z)$ for the number of assemblies $p(\Lambda)$ with energy Λ takes the form

$$P(z) = \sum_{\Lambda=0}^{\infty} p(\Lambda) z^{\Lambda} = \prod_{\lambda>0} \frac{1}{(1-z^{\lambda})^{j(\lambda)}} \ . \tag{10.6.1}$$

In the model of d-dimensional harmonic trap (also known as the Bose-Einstein condensation) according to [CMZ99; HHA97; LR08] the number of states for a particle with energy λ is $\binom{d+\lambda-1}{\lambda}$ so that each state can be represented as a multiset with λ elements having d different colours. Accordingly, an assembly is a multiset of particles (since they are bosons and hence indistinguishable) therefore the generating function for the number of assemblies takes the form

$$P(z) = \mathsf{MSet}(\mathsf{MSet}_{\geqslant 1}(\mathcal{Z}_1 + \cdots + \mathcal{Z}_d)) \ . \tag{10.6.2}$$

It is possible to control the expected frequencies of colours using our tuning procedure and sample resulting assemblies as Young tableaux. Each row corresponds to a particle whereas the colouring of the row displays the multiset of included colours, see Figure 10.10. We also generated weighted partitions of expected size 1000 (which are too large to display) with tuned frequencies of 5 colours, see Table 10.5.



(a) [5, 10, 15, 20, 25] (b) [4,4,4,4, 10, 20, 30, 40] (c) [80, 40, 20, 10, 9, 8, 7, 6, 5] (d) $[20, 60, 30, 20, 10, 5^9]$

Figure 10.10: Young tableaux corresponding to Bose–Einstein condensates with expected numbers of different colours. Notation $[c_1, c_2, \ldots, c_k]$ provides the expected number c_j of the j-th colour, c_k^m is a shortcut for m occurrences of c_k.

Let us briefly explain our generation procedure. Boltzmann sampling for the outer MSet operator is described in **??**, **??**. The sampling of inner $\mathsf{MSet}_{\geqslant 1}(\mathcal{Z}_1 + \ldots + \mathcal{Z}_d)$ is more delicate. The generating function for this multiset can be written as

$$\mathsf{MSet}_{\geqslant 1}(z_1 + \cdots + z_d) = \prod_{i=1}^{d} \frac{1}{1-z_i} - 1 \ . \tag{10.6.3}$$

Colour index	1	2	3	4	5	size
Tuned frequency	0.03	0.07	0.1	0.3	0.5	1000
Observed frequency	0.03	0.08	0.07	0.33	0.49	957
	0.03	0.06	0.09	0.28	0.54	1099
	0.03	0.08	0.09	0.34	0.46	992
	0.04	0.07	0.1	0.31	0.49	932
	0.04	0.09	0.1	0.25	0.52	1067

Table 10.5: Empirical frequencies of colours observed in random partition.

In order to correctly calculate the branching probabilities, we introduce slack variables $s_1, \ldots, s_d$ satisfying $(1+s_i) = (1-z_i)^{-1}$. Boltzmann samplers for the newly determined combinatorial classes $\Gamma\mathcal{S}_i$ are essentially Boltzmann samplers for $\mathsf{Seq}_{\geqslant 1}(\mathcal{Z}_i)$. Let us note that after expanding brackets the expression becomes

$$\mathsf{MSet}_{\geqslant 1}(z_1 + \cdots + z_d) = (s_1 + \cdots + s_d) + (s_1 s_2 + \cdots + s_{d-1} s_d) + \cdots + s_1 s_2 \ldots s_d.$$

The total number of summands is $2^d - 1$ where each summand corresponds to choosing some subset of colours. Finally, let us explain how to precompute all the symmetric polynomials and efficiently handle the branching process in quadratic time using a dynamic programming approach. We can recursively define two arrays of real numbers $p_{k,j}$ and $q_{k,j}$ satisfying

$$\begin{cases} p_{1,j} = s_j, \quad j \in \{1, \ldots, d\}; \\ q_{k,d} = p_{k,d}, \quad k \in \{1, \ldots, d\}; \\ q_{k,j} = p_{k,j} + q_{k,j+1}, \quad j \in \{k, \ldots, d-1\}, \quad k \in \{1, \ldots, d\}; \\ p_{k,j} = s_{j-k+1} \cdot q_{k-1,j}, \quad j \in \{k, \ldots, d-1\}, \quad k \in \{2, \ldots d\}; \end{cases} \tag{10.6.4}$$

Arrays $(p_{k,j})_{j=k}^{d}$ contain the branching probabilities determining the next colour inside the k-th symmetric polynomial. Arrays $(q_{k,j})_{j=k}^{d}$ contain partial sums for the k-th symmetric polynomial and are required in intermediate steps. Numbers $q_{k,k}$ are equal to the total values of symmetric polynomials $(s_1 + \cdots + s_2), (s_1 s_2 + \cdots + s_{d-1} s_d), \ldots, s_1 s_2, \ldots, s_d$ and they define initial branching probabilities to choose the number of colours.

10.7 Prototype sampler generator.

Consider the following example of an input file for Boltzmann Brain:

```
-- Motzkin trees
Motzkin = Leaf (3)
        | Unary Motzkin
        | Binary Motzkin Motzkin (2) [0.3].
```

Here, a `Motzkin` algebraic data type is defined. It consists of three constructors: a constant `Leaf` of weight three, a `Unary` constructor of weight one (default value if not explicitly annotated) and a constructor `Binary` of weight two together with an explicit tuning frequency of 30%. Such a definition corresponds to the combinatorial specification $\mathcal{M} = \mathcal{Z}^3 + \mathcal{Z}\mathcal{M} + \mathcal{U}\mathcal{Z}^2\mathcal{M}^2$ where the objective is to obtain the mean proportion of $\mathcal{U}\mathcal{Z}^2\mathcal{M}^2$ equal 30% of the total structure size. All the terms `Leaf`, `Unary`, `Motzkin`, `Binary` are user-defined keywords. Given such a specification on input, bb builds a corresponding singular Boltzmann sampler implemented in form of a self-contained Haskell module.

Bibliography

[AB00] Didier Arquès and Jean-François Béraud. "Rooted maps on orientable surfaces, Riccati's equation and continued fractions". In: *Discrete mathematics* 215.1-3 (2000), pp. 1–12.

[ABG12] Louigi Addario-Berry, Nicolas Broutin, and Christina Goldschmidt. "The continuum limit of critical random graphs". In: *Probability Theory and Related Fields* 152.3-4 (2012), pp. 367–406.

[AC08] Dimitris Achlioptas and Amin Coja-Oghlan. "Algorithmic barriers from phase transitions". In: *2008 49th Annual IEEE Symposium on Foundations of Computer Science*. IEEE. 2008, pp. 793–802.

[Ald97] David Aldous. "Brownian excursions, critical random graphs and the multiplicative coalescent". In: *The Annals of Probability* (1997), pp. 812–854.

[APT82] Bengt Aspvall, Michael F Plass, and Robert Endre Tarjan. "A linear-time algorithm for testing the truth of certain quantified boolean formulas". In: *Information Processing Letters* 14.4 (1982), p. 195.

[Ayr+17] Peter Ayre, Amin Coja-Oghlan, Pu Gao, and Noëla Müller. "The satisfiability threshold for random linear equations". In: *arXiv preprint arXiv:1710.07497* (2017).

[Ban+01] Cyril Banderier, Philippe Flajolet, Gilles Schaeffer, and Michele Soria. "Random maps, coalescing saddles, singularity analysis, and Airy phenomena". In: *Random Structures & Algorithms* 19.3-4 (2001), pp. 194–246.

[Ban+12] Cyril Banderier, Olivier Bodini, Yann Ponty, and Hanane Tafat Bouzid. "On the diversity of pattern distributions in rational language". In: *Proceedings of the Ninth Workshop on Analytic Alg. and Combinatorics*. 2012, pp. 107–115.

[Bar+99] Bruno Barras, Samuel Boutin, Cristina Cornes, Judicaël Courant, Yann Coscoy, David Delahaye, Daniel de Rauglaudre, Jean-Christophe Filliâtre, Eduardo Giménez, Hugo Herbelin, et al. "The Coq proof assistant reference manual". In: *INRIA, version* 6.11 (1999).

[Bar84] Henk P. Barendregt. *The Lambda Calculus: Its Syntax and Semantics*. Revised. Vol. 103. North Holland, 1984.

[BB01] Ginestra Bianconi and Albert-László Barabási. "Bose-Einstein condensation in complex networks". In: *Physical review letters* 86.24 (2001), p. 5632.

[BB17] Olivier Bernardi and Mireille Bousquet-Mélou. "Counting coloured planar maps: differential equations". In: *Communications in Mathematical Physics* 354.1 (2017), pp. 31–84.

[BBD18a] 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)*. 2018, pp. 92–106. DOI: 10.1137/1.9781611975062.9.

[BBD18b] Maciej Bendkowski, Olivier Bodini, and Sergey Dovgal. "Statistical properties of lambda terms". In: *arXiv preprint arXiv:1805.09419* (2018).

[BBJ13] Axel Bacher, Olivier Bodini, and Alice Jacquot. "Exact-size sampling for Motzkin trees in linear time via Boltzmann samplers and holonomic specification". In: *Proceedings of the Meeting on Analytic Algorithmics and Combinatorics*. 2013, pp. 52–61.

[BBR14] Anne Bouillard, Ana Bušić, and Christelle Rovetta. "Perfect sampling for closed queueing networks". In: *Performance Evaluation* 79 (2014), pp. 146–159.

[BBY10] Jason P. Bell, Stanley N. Burris, and Karen A. Yeats. "Characteristic points of recursive systems". In: *The Electronic Journal of Combinatorics* 17.1 (2010), p. 121.

[BCM02] Giulio Biroli, Simona Cocco, and Rémi Monasson. "Phase transitions and complexity in computer science: an overview of the statistical physics approach to the random satisfiability problem". In: *Physica A: Statistical Mechanics and its Applications* 306 (2002), pp. 381–394.

[BDM16] Olivier Bodini, Julien David, and Philippe Marchal. "Random-bit optimal uniform sampling for rooted planar trees with given sequence of degrees and applications". In: *Conference on Algorithms and Discrete Applied Mathematics*. 2016, pp. 97–114.

[Bea+10] Laura Beaudin, Joanna Ellis-Monaghan, Greta Pangborn, and Robert Shrock. "A little statistical mechanics for the graph theorist". In: *Discrete Mathematics* 310.13-14 (2010), pp. 2037–2053.

[Ben+16] Maciej Bendkowski, Katarzyna Grygiel, Pierre Lescanne, and Marek Zaionc. "A Natural Counting of Lambda Terms". In: *SOFSEM 2016: Theory and Practice of Computer Science - 42nd International Conference on Current Trends in Theory and Practice of Computer Science, Harrachov, Czech Republic, January 23-28, 2016, Proceedings*. Ed. by Rusins Martins Freivalds, Gregor Engels, and Barbara Catania. Vol. 9587. Lecture Notes in Computer Science. Springer, 2016, pp. 183–194. ISBN: 978-3-662-49191-1. DOI: 10.1007/978-3-662-49192-8_15.

[Ben+17] Maciej Bendkowski, Katarzyna Grygiel, Pierre Lescanne, and Marek Zaionc. "Combinatorics of λ-terms: a natural approach". In: *Journal of Logic and Computation* (2017). DOI: https://doi.org/10.1093/logcom/exx018.

[Ben17] Maciej Bendkowski. "Quantitative aspects and generation of random lambda and combinatory logic terms". PhD thesis. Kraków, Poland: Jagiellonian University, May 2017.

[BF01] Tom Bohman and Alan Freize. "Avoiding a giant component". In: *Random Structures and Algorithms* 19.1 (2001), pp. 75–85.

[BFP10] Olivier Bodini, Éric Fusy, and Carine Pivoteau. "Random sampling of plane partitions". In: *Combinatorics, Probability and Computing* 19.2 (2010), pp. 201–226.

[BFR17] Megan Bernstein, Matthew Fahrbach, and Dana Randall. "Analyzing Boltzmann Samplers for Bose-Einstein Condensates with Dirichlet Generating Functions". In: *arXiv preprint arXiv:1708.02266* (2017).

[BG12] Olivier Bernardi and Omer Giménez. "A linear algorithm for the random sampling from regular languages". In: *Algorithmica* 62.1 (2012), pp. 130–145.

[BGG17] Olivier Bodini, Bernhard Gittenberger, and Zbigniew Gołębiewski. "Enumerating Lambda Terms by Weighted Length of Their De Bruijn Representation". In: *CoRR* abs/1707.02101 (2017). URL: https://arxiv.org/abs/1707.02101.

[BGG18] Olivier Bodini, Antoine Genitrini, and Bernhard Gittenberger. "On the number of increasing trees with label repetitions". In: *arXiv preprint arXiv:1809.04314* (2018).

[BGJ13] Olivier Bodini, Danielle Gardy, and Alice Jacquot. "Asymptotics and random sampling for BCI and BCK lambda terms". In: *Theoretical Computer Science* 502 (2013), pp. 227–238.

[BGN19] Olivier Bodini, Antoine Genitrini, and Mehdi Naima. "Ranked Schröder Trees". In: *2019 Proceedings of the Sixteenth Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*. SIAM. 2019, pp. 13–26.

[BGR15] Olivier Bodini, Antoine Genitrini, and Nicolas Rolin. "Pointed versus singular Boltzmann samplers: a comparative analysis". In: *Pure Mathematics and Application* 25.2 (2015), pp. 115–131.

[Bha+17] Prateek Bhakta, Ben Cousins, Matthew Fahrbach, and Dana Randall. "Approximately sampling elements with fixed rank in graded posets". In: *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms SODA*. 2017, pp. 1828–1838.

[BJ06] Mireille Bousquet-Mélou and Arnaud Jehanne. "Polynomial equations with one catalytic variable, algebraic series and map enumeration". In: *Journal of Combinatorial Theory, Series B* 96.5 (2006), pp. 623–672.

[BLL98] François Bergeron, Gilbert Labelle, and Pierre Leroux. *Combinatorial species and tree-like structures*. Vol. 67. Cambridge University Press, 1998.

[BLR15] Olivier Bodini, Jérémie Lumbroso, and Nicolas Rolin. "Analytic samplers and the combinatorial rejection method". In: *Proceedings of the Meeting on Analytic Algorithmics and Combinatorics*. 2015, pp. 40–50.

[BM93] Andrea Bertozzi and James McKenna. "Multidimensional residues, generating functions, and their application to queueing networks". In: *SIAM review* 35.2 (1993), pp. 239–268.

[Bod+11] Manuel Bodirsky, Éric Fusy, Mihyun Kang, and Stefan Vigerske. "Boltzmann samplers, Pólya theory, and cycle pointing". In: *SIAM J. Comp.* 40.3 (2011), pp. 721–769.

[Bod+16] Olivier Bodini, Matthieu Dien, Xavier Fontaine, Antoine Genitrini, and Hsien-Kuei Hwang. "Increasing Diamonds". In: *Latin American Symposium on Theoretical Informatics*. 2016, pp. 207–219.

[Bod+18a] Olivier Bodini, Julien Courtiel, Sergey Dovgal, and Hsien-Kuei Hwang. "Asymptotic distribution of parameters in random maps". In: *arXiv preprint arXiv:1802.07112* (2018).

[Bod+18b] Olivier Bodini, Matthieu Dien, Antoine Genitrini, and Alfredo Viola. "Beyond series-parallel concurrent systems: the case of arch processes". In: *arXiv preprint arXiv:1803.00843* (2018).

[Bod10] Olivier Bodini. "Autour de la génération aléatoire sous modèle de Boltzmann [On random generation under Boltzmann models]". habilitation. Université Pierre et Marie Curie, Paris, 2010.

[Bol+01] Béla Bollobás, Christian Borgs, Jennifer T Chayes, Jeong Han Kim, and David B Wilson. "The scaling window of the 2-SAT transition". In: *Random Structures & Algorithms* 18.3 (2001), pp. 201–256.

[Bol80] Béla Bollobás. "A probabilistic proof of an asymptotic formula for the number of labelled regular graphs". In: *European Journal of Combinatorics* 1 (1980), pp. 311–316.

[Bol85] Béla Bollobás. *Random graphs*. Academic Press, Inc., London, 1985.

[BP10] Olivier Bodini and Yann Ponty. "Multi-dimensional Boltzmann sampling of context-free languages". In: *21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA'10)*. Vol. AM. 2010.

[BR83] Edward A. Bender and L. Bruce Richmond. "Central and local limit theorems applied to asymptotic enumeration II: Multivariate generating functions". In: *Journal of Combinatorial Theory, Series A* 34.3 (1983), pp. 255–265.

[BR86] Edward A Bender and L.Bruce Richmond. "A survey of the asymptotic behaviour of maps". In: *Journal of Combinatorial Theory, Series B* 40.3 (1986), pp. 297–329. ISSN: 0095-8956. DOI: https://doi.org/10.1016/0095-8956(86)90086-9. URL: http://www.sciencedirect.com/science/article/pii/0095895686900869.

[BRS12] Olivier Bodini, Olivier Roussel, and Michèle Soria. "Boltzmann samplers for first-order differential specifications". In: *Disc. App. Math.* 160.18 (2012), pp. 2563–2572.

[BT17] Olivier Bodini and Paul Tarau. "On Uniquely Closable and Uniquely Typable Skeletons of Lambda Terms". In: *CoRR* abs/1709.04302 (2017). URL: http://arxiv.org/abs/1709.04302.

[Car17] Ariane Carrance. "Uniform random colored complexes". In: *arXiv preprint arXiv:1705.11103* (2017).

[CCP16] Cedric Chauve, Julien Courtiel, and Yann Ponty. "Counting, generating and sampling tree alignments". In: *International Conference on Algorithms for Computational Biology*. Springer. 2016, pp. 53–64.

[CH00] Koen Claessen and John Hughes. "QuickCheck: a lightweight tool for random testing of Haskell programs". In: *Proceedings of the Fifth ACM SIGPLAN International Conference on Functional Programming*. 2000, pp. 268–279.

[Cha] Guillaume Chapuy. "Rencontres autour de la combinatoire des cartes [Encounters around the combinatorics of maps]". habilitation.

[CKP18] Oliver Cooley, Mihyun Kang, and Yury Person. "Largest components in random hypergraphs". In: *Combinatorics, Probability and Computing* 27.5 (2018), pp. 741–762.

[CLP78] Predrag Cvitanovi ć, B. Lautrup, and Robert B. Pearson. "Number and weights of Feynman diagrams". In: *Phys. Rev. D* 18 (6 Sept. 1978), pp. 1939–1949. DOI: 10.1103/PhysRevD.18.1939. URL: https://link.aps.org/doi/10.1103/PhysRevD.18.1939.

[CMZ99] K. C. Chase, A. Z. Mekjian, and L. Zamick. "Canonical and microcanonical ensemble approaches to Bose-Einstein condensation: The thermodynamics of particles in harmonic traps". In: *The European Physical Journal B-Condensed Matter and Complex Systems* 8.2 (1999), pp. 281–285.

[Col+18] Gwendal Collet, Elie De Panafieu, Danièle Gardy, Bernhard Gittenberger, and Vlady Ravelomanana. "Threshold functions for small subgraphs in simple graphs and multigraphs". In: *arXiv preprint arXiv:1807.05772* (2018).

[Coo+18] Oliver Cooley, Nicola Del Giudice, Mihyun Kang, and Philipp Sprüssel. "Vanishing of cohomology groups of random simplicial complexes". In: *arXiv preprint arXiv:1806.04566* (2018).

[Cop+04] Don Coppersmith, David Gamarnik, Mohammad Taghi Hajiaghayi, and Gregory B Sorkin. "Random MAX SAT, random MAX CUT, and their phase transitions". In: *Random Structures & Algorithms* 24.4 (2004), pp. 502–545.

[Cor09] Robert Cori. "Indecomposable permutations, hypermaps and labeled Dyck paths". In: *Journal of Combinatorial Theory, Series A* 116.8 (2009), pp. 1326–1343.

[CP16] Amin Coja-Oghlan and Konstantinos Panagiotou. "The asymptotic k-SAT threshold". In: *Advances in Mathematics* 288 (2016), pp. 985–1068.

[CV13] Amin Coja-Oghlan and Dan Vilenchik. "Chasing the k-colorability threshold". In: *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. IEEE. 2013, pp. 380–389.

[CY17] Julien Courtiel and Karen Yeats. "Terminal chords in connected chord diagrams". In: *Annales de l'Institut Henri Poincaré D* 4.4 (2017), pp. 417–452.

[CYZ16] Julien Courtiel, Karen Yeats, and Noam Zeilberger. "Connected chord diagrams and bridgeless maps". In: *arXiv preprint arXiv:1611.04611* (2016).

[Dav+13] René David, Katarzyna Grygiel, Jakub Kozik, Christophe Raffalli, Guillaume Theyssier, and Marek Zaionc. "Asymptotically almost all λ-terms are strongly normalizing". In: *Logical Methods in Computer Science* 9 (2013), pp. 1–30.

[DB16] Steven Diamond and Stephen Boyd. "CVXPY: A Python-embedded bodeling language for convex optimization". In: *J. Mach. Learn. Research* 17.83 (2016), pp. 1–5.

[dBru72] 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.

[DCB13] Alexander Domahidi, Eric Chu, and Stephen Boyd. "ECOS: An SOCP solver for embedded systems". In: *Control Conference (ECC), 2013 European*. IEEE. 2013, pp. 3071–3076.

[DGM12] Michael Drmota, Bernhard Gittenberger, and Johannes F. Morgenbesser. "Infinite Systems of Functional Equations and Gaussian Limiting Distributions". In: *23rd International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA'12)*. Vol. DMTCS Proceedings vol. AQ, 23rd Intern. Meeting on Probabilistic, Combinatorial, and Asymptotic Methods for the Analysis of Algorithms (AofA'12). DMTCS Proceedings. Montreal, Canada: Discrete Mathematics and Theoretical Computer Science, 2012, pp. 453–478.

[Die+10] Martin Dietzfelbinger, Andreas Goerdt, Michael Mitzenmacher, Andrea Montanari, Rasmus Pagh, and Michael Rink. "Tight thresholds for cuckoo hashing via XORSAT". In: *International Colloquium on Automata, Languages, and Programming*. Springer. 2010, pp. 213–225.

[DM03] Luc Devroye and Pat Morin. "Cuckoo hashing: further analysis". In: *Information Processing Letters* 86.4 (2003), pp. 215–219.

[Dom13] Sander Dommers. "Spin models on random graphs". PhD thesis. Technische Universiteit Eindhoven, 2013.

[Dov19] Sergey Dovgal. "The birth of the contradictory component in random 2-SAT". In: *arXiv preprint arXiv:1904.10266* (2019).

[DP13] Michael Drmota and Konstantinos Panagiotou. "A central limit theorem for the number of degree-k vertices in random maps". In: *Algorithmica* 66.4 (2013), pp. 741–761.

[dPan15] Élie de Panafieu. "Phase transition of random non-uniform hypergraphs". In: *Journal of Discrete Algorithms* 31 (2015), pp. 26–39.

[dPan16] Élie de Panafieu. "Analytic combinatorics of connected graphs". In: *Random Structures & Algorithms* (2016).

[dPD19] Élie de Panafieu and Sergey Dovgal. "Symbolic method and directed graph enumeration". In: *Acta Mathematica Universitatis Comenianae* 88.3 (2019), pp. 989–996.

[dPR16] Élie de Panafieu and Lander Ramos. "Graphs with degree constraints". In: *2016 Proceedings of the Thirteenth Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*. SIAM. 2016, pp. 34–45.

[DPT10] Alain Denise, Yann Ponty, and Michel Termier. "Controlled non-uniform random generation of decomposable structures". In: *Theoretical Computer Science* 411.40-42 (2010), pp. 3527–3552.

[DR18] Sergey Dovgal and Vlady Ravelomanana. "Shifting the Phase Transition Threshold for Random Graphs Using Degree Set Constraints". In: *Latin American Symposium on Theoretical Informatics*. Springer. 2018, pp. 399–412.

[Drm09] Michael Drmota. *Random Trees: An Interplay between Combinatorics and Probability*. Springer Science & Business Media, 2009.

[Drm97] Michael Drmota. "Systems of functional equations". In: *Rand. Struct. & Alg.* 10.1-2 (1997), pp. 103–124.

[Duc+04] Philippe Duchon, Philippe Flajolet, Guy Louchard, and Gilles Schaeffer. "Boltzmann samplers for the random generation of combinatorial structures". In: *Combinatorics, Probability & Computing* 13.4-5 (2004), pp. 577–625.

[Dum17] Hugo Duminil-Copin. "Lectures on the Ising and Potts models on the hypercubic lattice". In: *arXiv preprint arXiv:1707.00520* (2017).

[DZ99] Alain Denise and Paul Zimmermann. "Uniform random generation of decomposable structures using floating-point arithmetic". In: *Theoretical Computer Science* 218.2 (1999), pp. 233–248.

[ER60] Paul Erdős and Alfred Rényi. "On the evolution of random graphs". In: *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei* 5 (1960), pp. 17–61.

[Fay+99] Guy Fayolle, VA Malyshev, Roudolf Iasnogorodski, and Guy Fayolle. *Random walks in the quarter-plane*. Vol. 40. Springer, 1999.

[FFP07] Philippe Flajolet, Éric Fusy, and Carine Pivoteau. "Boltzmann sampling of unlabelled structures". In: *Proceedings of the Meeting on Analytic Algorithmics and Combinatorics*. 2007, pp. 201–211.

[FKP89] Philippe Flajolet, Donald E Knuth, and Boris Pittel. "The first cycles in an evolving graph". In: *Discrete Mathematics* 75.1-3 (1989), pp. 167–215.

[FN00] Philippe Flajolet and Marc Noy. "Analytic combinatorics of chord diagrams". In: *Formal Power Series and Algebraic Combinatorics*. Springer. 2000, pp. 191–201.

[FO82] Philippe Flajolet and Andrew M. Odlyzko. "The average height of binary trees and other simple trees". In: *Journal of Computer and System Sciences* 25 (2 1982), pp. 171–213.

[FPK89] Philippe Flajolet, Boris Pittel, and Donald E. Knuth. "The First Cycles in an Evolving Graph". In: *Discrete Mathematics* 75 (1–3 1989), pp. 167–215.

[FS09] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. 1st ed. Cambridge University Press, 2009. ISBN: 978-0-521-89806-5.

[FSS04] Philippe Flajolet, Bruno Salvy, and Gilles Schaeffer. "Airy phenomena and analytic combinatorics of connected graphs". In: *the electronic journal of combinatorics* 11.1 (2004), p. 34.

[Fus05] Éric Fusy. "Quadratic exact size and linear approximate size random generation of planar graphs". In: *Discr. Math. & Theor. Comp. Sc.* 2005, pp. 125–138.

[FZC94] Philippe Flajolet, Paul Zimmermann, and Bernard Van Cutsem. "A calculus for the random generation of labelled combinatorial structures". In: *Theoretical Computer Science* 132.1 (1994), pp. 1–35.

[GBY06] Michael Grant, Stephen Boyd, and Yinyu Ye. "Disciplined convex programming". In: *Global optimization*. 2006, pp. 155–210.

[Ges95] Ira Martin Gessel. "Enumerative applications of a decomposition for graphs and digraphs". In: *Discrete Mathematics* 139.1 (1995), pp. 257–271. ISSN: 0012-365X.

[Ges96] Ira Martin Gessel. "Counting acyclic digraphs by sources and sinks". In: *Discrete Mathematics* 160.1–3 (1996), pp. 253–258.

[GG16] Bernhard Gittenberger and Zbigniew Gołębiewski. "On the Number of Lambda Terms With Prescribed Size of Their De Bruijn Representation". In: *33rd Symposium on Theoretical Aspects of Computer Science (STACS 2016)*. Ed. by Nicolas Ollinger and Heribert Vollmer. Vol. 47. Leibniz International Proceedings in Informatics (LIPIcs). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2016, 40:1–40:13. ISBN: 978-3-95977-001-9.

[GJ08] Ian P Goulden and David M Jackson. "The KP hierarchy, branched covers, and triangulations". In: *Advances in Mathematics* 219.3 (2008), pp. 932–951.

[GL13] Katarzyna Grygiel and Pierre Lescanne. "Counting and generating lambda terms". In: *Journal of Functional Programming* 23.5 (2013), pp. 594–628. DOI: 10.1017/S0956796813000178.

[GL15] Katarzyna Grygiel and Pierre Lescanne. "Counting and generating terms in the binary lambda calculus". In: *Journal of Functional Programming* 25 (2015), e24. DOI: 10.1017/S0956796815000271.

[GN67] William J Gordon and Gordon F Newell. "Closed queuing systems with exponential servers". In: *Operations research* 15.2 (1967), pp. 254–265.

[GS96] Ira Martin Gessel and Bruce Eli Sagan. "The Tutte polynomial of a graph, depth-first search, and simplicial complex partitions". In: *Electron. J. Combin* 3.2 (1996), R9.

[Ham+19] Stefan Hammer, Wei Wang, Sebastian Will, and Yann Ponty. "Fixed-parameter tractable sampling for RNA design with multiple target structures". In: *BMC bioinformatics* 20.1 (2019), p. 209.

[Han81] Phil Hanlon. "Algebras of acyclic type". In: *Canadian Journal of Mathematics* 33.1 (1981), pp. 129–141.

[Heu07] Clemens Heuberger. "Hwang's quasi-power-theorem in dimension two". In: *Quaestiones Mathematicae* 30.4 (2007), pp. 507–512.

[HHA97] Tor Haugset, Harek Haugerud, and Jens O. Andersen. "Bose-Einstein condensation in anisotropic harmonic traps". In: *Physical Review A* 55.4 (1997), p. 2922.

[HK16] Clemens Heuberger and Sara Kropf. "On the Higher Dimensional Quasi-Power Theorem and a Berry-Esseen Inequality". In: *27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms Kraków, Poland, July 4–8, 2016*. 2016.

[HM12] Hamed Hatami and Michael Molloy. "The scaling window for a random graph with a given degree sequence". In: *Random Structures and Algorithms* 41.1 (2012), pp. 99–123. ISSN: 1098-2418.

[HR11] Daudé Hervé and Vlady Ravelomanana. "Random 2 XORSAT phase transition". In: *Algorithmica* 59.1 (2011), pp. 48–65.

[Hwa98] 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.

[Hwa99] Hsien-Kuei Hwang. "Asymptotics of Poisson approximation to random discrete distributions: an analytic approach". In: *Advances in Applied Probability* 31.2 (1999), pp. 448–491.

[IBP05] Gabriel Istrate, Stefan Boettcher, and Allon G Percus. "Spines of random constraint satisfaction problems: definition and connection with computational complexity". In: *Annals of Mathematics and Artificial Intelligence* 44.4 (2005), pp. 353–372.

[Jan+93] Svante Janson, Donald E Knuth, Tomasz Łuczak, and Boris Pittel. "The birth of the giant component". In: *Random Structures & Algorithms* 4.3 (1993), pp. 233–358.

[JŁR11] Svante Janson, Tomasz Łuczak, and Andrzej Rucinski. *Random graphs*. John Wiley & Sons, 2011.

[Joo+16] Felix Joos, Guillem Perarnau, Dieter Rautenbach, and Bruce Reed. "How to determine if a random graph with a fixed degree sequence has a giant component". In: *2016 IEEE 57th Ann. Symposium on Found. of Comp. Sc. (FOCS)* (2016), pp. 695–703.

[Joy81] André Joyal. "Une théorie combinatoire des séries formelles". In: *Advances in mathematics* 42.1 (1981), pp. 1–82.

[JVV86] Mark R Jerrum, Leslie G Valiant, and Vijay V Vazirani. "Random generation of combinatorial structures from a uniform distribution". In: *Theoretical Computer Science* 43 (1986), pp. 169–188.

[K+01] Frank R Kschischang, Brendan J Frey, Hans-Andrea Loeliger, et al. "Factor graphs and the sum-product algorithm". In: *IEEE Transactions on information theory* 47.2 (2001), pp. 498–519.

[Kei91] Michael Keith. *From polychords to Pólya; adventures in musical combinatorics*. Vinculum Press, Princeton, New Jersey, 1991. ISBN: 0963009702.

[Kim08] Jeong Han Kim. "Finding cores of random 2-SAT formulae via Poisson cloning". In: *arXiv preprint arXiv:0808.1599* (2008).

[Knu15] Donald E Knuth. *The Art of Computer Programming, Volume 4, Fascicle 6: Satisfiability*. Addison-Wesley Professional, 2015.

[Knu92] Donald E Knuth. “Convolution polynomials”. In: *arXiv preprint math/9207221* (1992).

[Kog02] Yaakov Kogan. “Asymptotic expansions for large closed and loss queueing networks”. In: *Mathematical problems in engineering* 8.4-5 (2002), pp. 323–348.

[Kon92] Maxim Kontsevich. “Intersection theory on the moduli space of curves and the matrix Airy function”. In: *Communications in Mathematical Physics* 147.1 (1992), pp. 1–23.

[Kra07] David Kravitz. “RANDOM 2-SAT Does not depend on a giant”. In: *SIAM Journal on Discrete Mathematics* 21.2 (2007), pp. 408–422.

[KS60] J.G. Kemeny and J.L. Snell. *Finite Markov chains*. Springer, 1960.

[Kud91] NA Kudryashov. “On types of nonlinear nonintegrable equations with exact solutions”. In: *Physics Letters A* 155.4-5 (1991), pp. 269–275.

[Lal93] Steven P. Lalley. “Finite range random walk on free groups and homogeneous trees”. In: *The Annals of Probability* (1993), pp. 2087–2130.

[Ler+12] Xavier Leroy et al. “The CompCert verified compiler”. In: *Documentation and user’s manual. INRIA Paris-Rocquencourt* 53 (2012).

[Lis00] Valery Anisimovich Liskovets. “Some easily derivable integer sequences”. In: *Journal of Integer Sequences* 3.2 (2000), p. 3.

[Lis69a] Valery Anisimovich Liskovets. “Лисковец Валерий Анисимович. Об одном рекуррентном методе подсчета графов с отмеченными вершинами”. Russian. In: *Доклады Академии наук*. Vol. 184. 6. [On one recurrent method of counting graphs with marked vertices, Doklady Akademii Nauk]. 1969, pp. 1284–1287.

[Lis69b] Valery Anisimovich Liskovets. “Лисковец Валерий Анисимович. Подсчет корневых инициально связных ориентированных графов”. Russian. In: *Известия АН БССР*. 5. [Enumeration of rooted initially connected oriented graphs, Izv. Akad. Nauk BSSR]. 1969, pp. 23–32.

[Lis70] Valery Anisimovich Liskovets. “The number of strongly connected directed graphs”. In: *Mathematical notes of the Academy of Sciences of the USSR* 8.6 (1970), pp. 877–882.

[Lis73] Valery Anisimovich Liskovets. “Лисковец Валерий Анисимович. К перечислению сильно связных ориентированных графов.” In: *ДАН БССР* 17 (1973). [A contribution to the enumeration of strongly connected digraphs, Dokl. Akad. Nauk BSSR], pp. 1077–1080.

[Lis99] Valery A. Liskovets. “A pattern of asymptotic vertex valency distributions in planar maps”. In: *Journal of Combinatorial Theory, Series B* 75.1 (1999), pp. 116–133.

[LR08] James Lucietti and Mukund Rangamani. “Asymptotic counting of BPS operators in superconformal field theories”. In: *J. Math. Phys.* 49.8 (2008), p. 082301.

[ŁS09] Tomasz Łuczak and Taral Guldahl Seierstad. “The critical behavior of random digraphs”. In: *Random Structures & Algorithms* 35.3 (2009), pp. 271–293.

[LZ04] Sergei K. Lando and Alexander K. Zvonkin. *Graphs on Surfaces and Their Applications*. Encyclopaedia of Mathematical Sciences 141. Springer-Verlag, 2004.

[Mol08] Michael Molloy. “When does the giant component bring unsatisfiability?” In: *Combinatorica* 28.6 (2008), pp. 693–734.

[MPZ02] Marc Mézard, Giorgio Parisi, and Riccardo Zecchina. “Analytic and algorithmic solution of random satisfiability problems”. In: *Science* 297.5582 (2002), pp. 812–815.

[MR95] Michael Molloy and Bruce A. Reed. “A critical point for random graphs with a given degree sequence”. In: *Random Structures and Algorithms* 6.2/3 (1995), pp. 161–180.

[MZ97] Rémi Monasson and Riccardo Zecchina. “Statistical mechanics of the random K-satisfiability model”. In: *Physical Review E* 56.2 (1997), p. 1357.

[NN94] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

[NP08] Asaf Nachmias and Yuval Peres. “Critical random graphs: Diameter and mixing time”. In: *The Annals of Probability* 36.4 (2008), pp. 1267–1286.

[NR02] Marek Nowakowski and Haret C Rosu. "Newton's laws of motion in the form of a Riccati equation". In: *Physical Review E* 65.4 (2002), p. 047602.

[NW78] Albert Nijenhuis and Herbert S. Wilf. *Combinatorial Algorithms*. 2nd ed. Academic Press, 1978.

[Odl95] Andrew M. Odlyzko. "Asymptotic enumeration methods". In: *Handbook of combinatorics* 2.1063-1229 (1995), p. 1229.

[ODo+16] Brendan O'Donoghue, Eric Chu, Neal Parikh, and Stephen Boyd. "Conic optimization via operator splitting and homogeneous self-dual embedding". In: *J. Opt. Th. & App.* 169.3 (2016), pp. 1042–1068.

[Pał12] Michał H. Pałka. "Random structured test data generation for black-box testing". PhD thesis. Chalmers University of Technology, 2012.

[Pet16] Fedor Petrov. *Analytic Combinatorics: upper bound for sum of absolute values of two complex functions:* $|zf'(z)| + |2f(z) - zf'(z)| \leqslant 2f(|z|)$. 2016. URL: http://mathoverflow.net/q/242106.

[Pól36] George Pólya. "Algebraische Berechnung der Anzahl der Isomeren einiger organischer Verbindungen". In: *Zeitschrift für Kristallographie-Crystalline Materials* 93.1-6 (1936), pp. 415–443.

[PP17] BG Pittel and DJ Poole. "Birth of a giant (k1, k2)-core in the random digraph". In: *Advances in Applied Mathematics* 86 (2017), pp. 132–174.

[PS16] Boris Pittel and Gregory B Sorkin. "The satisfiability threshold for k-XORSAT". In: *Combinatorics, Probability and Computing* 25.2 (2016), pp. 236–268.

[PSS12] Carine Pivoteau, Bruno Salvy, and Michele Soria. "Algorithms for combinatorial structures: Well-founded systems and Newton iterations". In: *Journal of Combinatorial Theory, Series A* 119.8 (2012), pp. 1711–1773.

[Puy04] Vincent Puyhaubert. "Generating functions and the satisfiability threshold". In: *Discrete Mathematics and Theoretical Computer Science* 6.2 (2004).

[PW08] Robin Pemantle and Mark C Wilson. "Twenty combinatorial examples of asymptotics derived from multivariate generating functions". In: *Siam Review* 50.2 (2008), pp. 199–272.

[PW13] Robin Pemantle and Mark C. Wilson. *Analytic Combinatorics in Several Variables*. Cambridge Studies in Advanced Mathematics, 2013.

[Rio12] Oliver Riordan. "The Phase Transition in the Configuration Model". In: *Combinatorics, Probability & Computing* 21.1-2 (2012), pp. 265–299.

[Rob77a] Robert W. Robinson. "Counting strong digraphs". In: *Journal of Graph Theory* 1.2 (1977), pp. 189–190.

[Rob77b] Robert W Robinson. "Counting unlabeled acyclic digraphs". In: *Combinatorial mathematics V*. Springer, 1977, pp. 28–43.

[Rov17] Christelle Rovetta. "Simulation parfaite de réseaux fermés de files d'attente et génération aléatoire de structures combinatoires". PhD thesis. Paris Sciences et Lettres, 2017.

[RRN13] Juanjo Rué, Vlady Ravelomanana, and Marc Noy. "The probability of planarity of a random graph near the critical point". In: *Disc. Math. & Theor. Comp. Sc.* (2013).

[RW12] Oliver Riordan and Lutz Warnke. "Achlioptas process phase transitions are continuous". In: *Annals of Applied Probability* 22.4 (2012), pp. 1450–1464.

[RW17] Oliver Riordan and Lutz Warnke. "The phase transition in bounded-size Achlioptas processes". In: *arXiv preprint arXiv:1704.08714* (2017).

[Slo] N. J. A. Sloane. *The On-Line Encyclopedia of Integer Sequences*. URL: http://oeis.org.

[Sol] David Solymosi. "On the spine of 3-XORSAT". In: *Master Thesis* ().

[Spe13] Joel Spencer. *The strange logic of random graphs*. Vol. 22. Springer Science & Business Media, 2013.

[SS88] Saharon Shelah and Joel Spencer. "Zero-one laws for sparse random graphs". In: *Journal of the American Mathematical Society* 1.1 (1988), pp. 97–115.

[Sta73] Richard P Stanley. "Acyclic orientations of graphs". In: *Discrete Mathematics* 5.2 (1973), pp. 171–178.

[Van16] Remco Van Der Hofstad. *Random graphs and complex networks*. Cambridge university press, 2016.

[Ver96] Anatolii Moiseevich Vershik. "Statistical mechanics of combinatorial partitions, and their limit shapes". In: *Functional Analysis and Its Applications* 30.2 (1996), pp. 90–105.

[WG01] Dominic Welsh and Amy Gale. "The complexity of counting problems". In: *Aspects of Complexity, de Gruyter Series in Logic and Its Applications* (2001), pp. 115–154.

[Woo97] Alan R. Woods. "Coloring rules for finite trees, and probabilities of monadic second order sentences". In: *Random structures and algorithms* 10.4 (1997), pp. 453–485.

[Wri71] Edward Maitland Wright. "The number of strong digraphs". In: *Bulletin of the London Mathematical Society* 3.3 (1971), pp. 348–350.

[Yan+11] 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.

[ZG15] Noam Zeilberger and Alain Giorgetti. "A correspondence between rooted planar maps and normal planar lambda terms". In: *Logical Methods in Computer Science* 11.3:22 (2015).

List of Figures

List of Tables

Alphabetical Index