Keywords:
Limit Distribution
Catalan Trees
Singularity Analysis

Distribution of the Size of Simplified or Reduced Trees

Michel Nguyễn Thế
Laboratoire d'Informatique de l'X
École Polytechnique
Palaiseau, France

mnguyen@lix.polytechnique.fr

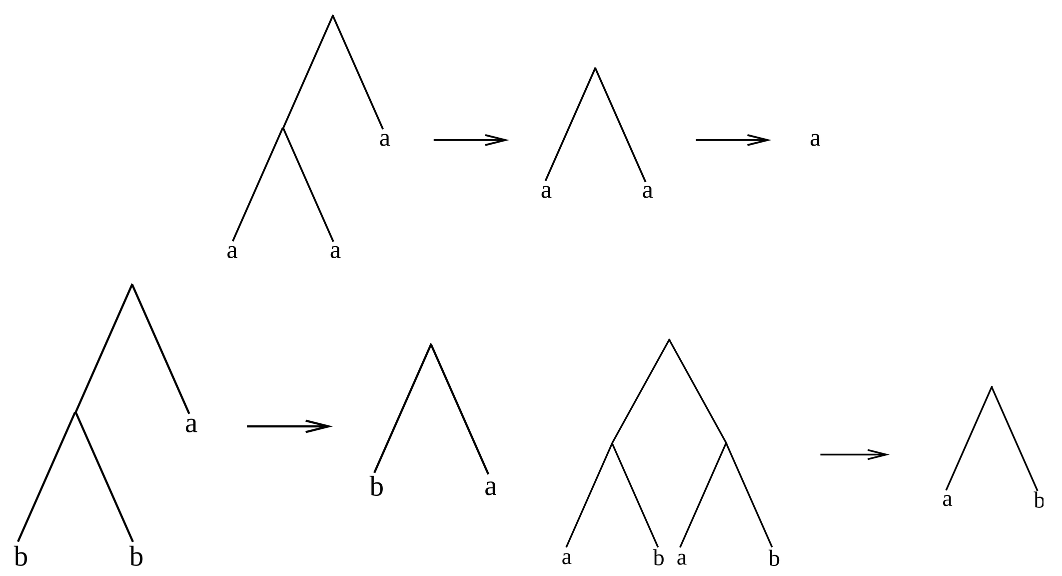
Laboratoire d'Informatique de l'X
École Polytechnique

Abstract

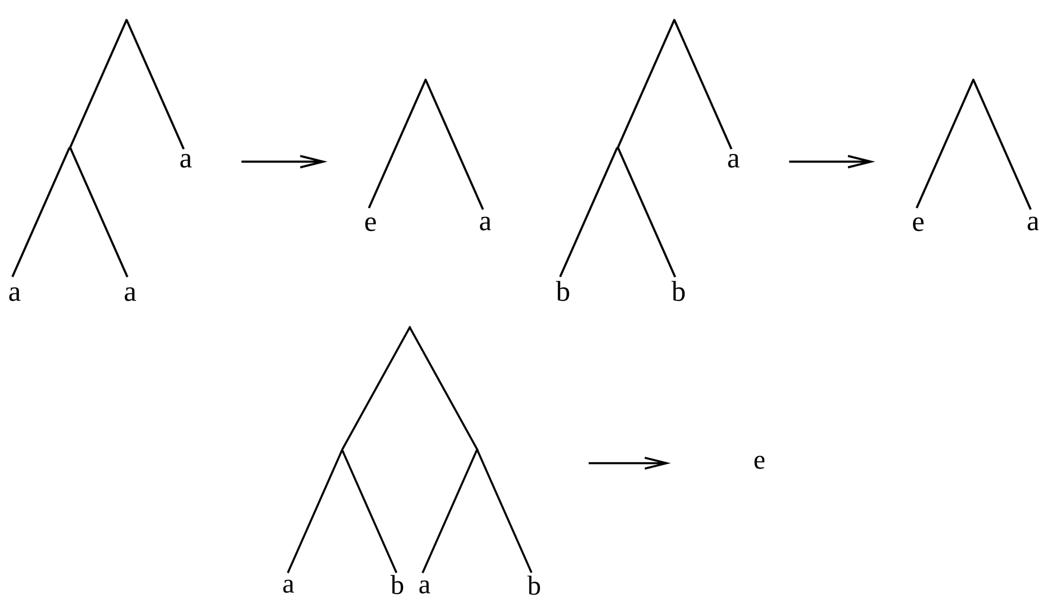
Previous works by Casas, Fernández-Camacho and Steyaert have determined the mean and variance of the size of simplified or reduced trees of 2-ary trees. Using the implicit function and the Quasi Powers theorems, we show that the limit distribution of that size is Gaussian.

1. Motivation

This poster deals with the analysis of a number of simplifying rewriting rules, as used in computer algebra systems. Instances of such rewriting rules commonly encountered in mathematics are for example $x^2 = x$ in Boolean rings, $p \cdot x = 0$ and $x^p = x$ in fields of characteristic p, $x - x = 0$ over the reals, and so on.
We consider expressions generated by a finite algebra starting from a finite number $q \geq 2$ of constants $e_1, \cdots, e_q$, appearing with probabilities π_{e_i}.
We will consider two types of simplifying rewriting rules. The first is defined by an idempotent 2-ary law $\wedge$ such that $f \wedge f \longrightarrow f$. We will call this operation *simplification*.



The second simplifying rewriting rule is defined by a nilpotent 2-ary law $\wedge$ such that $f \wedge f \longrightarrow e$, where $e = e_1$ is a constant reminiscent of a neutral element ($x - x = 0$).



We want to know the gain in space brought by these two types of operation. The study of that problem was developed in [FC88] and [CFS90], where the average size and the variance of the size of simplified and reduced trees were computed and were proved to be linear in the size of the initial tree, using the Darboux-Pólya method. In this poster, we will show that these new tools allow us to establish that the limit distribution of the size of simplified and reduced trees is Gaussian in the case of binary trees, but it also holds for m-ary trees, at least for equiprobable constants [Ngu02].

2. Simplification

Generating functions

An *irreducible* is a tree that cannot be further simplified. We denote $\mathcal{Q}$ the family of all considered trees, $\mathcal{I}$ the family of irreducible trees, π_t the probability of a tree t, $s(u)$ the irreducible tree to which u simplifies. The generating functions $Q_t(z) := \sum_{s(u)=t} \pi_t z^{|u|}$ for $t \in \mathcal{I}$, satisfy, for all $u \neq v \in \mathcal{I}$,

$$Q_a(z) = \pi_a + zQ_a^2(z), \qquad Q_{u\circ v}(z) = zQ_u(z)Q_v(z) + zQ_t^2(z),$$

so that, if we define $d(z,y) = \sum_{t\in\mathcal{I}} \left(y^{2|t|+1} - y^{|t|}\right) Q_t^2(z)$, the bivariate generating function $X(z,y) := \sum_{t\in\mathcal{Q}} y^{|s(t)|} z^{|t|}$ satisfy

$$X(z,y) \;=\; 1 + yzX(z,y)^2 - z\sum_{t\in\mathcal{I}} \left(y^{2|t|+1} - y^{|t|}\right) Q_t^p(z)$$

$$X(z,y) \;=\; \frac{1-\sqrt{1-4yzd(z,y)^2}}{2yz}$$

An algebraic representation

Let $\rho = 1/4$ and $\rho_I > \rho$ be the dominant singularities of $z \mapsto X(z,1)$ and $z \mapsto \sum_{t\in\mathcal{I}} \pi_t z^{|t|}$. We want to show that $X(z,y)$ admits a representation of the form

$$X(z,y) = A(z,y) + B(z,y)\sqrt{C(z,y)},$$

in a domain of the form

$$\mathcal{D} = \{\ (z,y)\ \big|\ |z| < z_1,\ |y| < y_1\},$$

with analytic functions A, B, C and reals $z_1 > \rho$, $y_1 > 1$.
This amounts to establishing that d converges beyond the domain of convergence of X. Indeed, using elementary computations, we can show by induction that there exists $z_0, \gamma, c > 0$ such that $\rho < z_0$, $\gamma^2 < \rho_I$ and $\boxed{Q_t(z_0) < c\sqrt{\pi_t}\gamma^{|t|}}$,
from which we deduce that the series $\sum_{t\in\mathcal{I}} y^{2|t|+1}Q_t^2(z)$ and $\sum_{t\in\mathcal{I}} y^{|t|}Q_t^2(z)$ converge beyond $(\rho, 1)$, and their difference d too.

Mean and variance

Using singularity analysis, we show that the expectation

$$\mu_n = \frac{[z^n]X_y(z,1)}{[z^n]Q(z)} = \mu n + O(1) \text{ with } \mu > 0,$$

and the variance

$$v_n = \frac{[z^n]\left(X_{y^2}(z,1) + X_y(z,1)\right)}{[z^n]Q(z)} - \mu_n^2 = \overline{v}n + O(1).$$

of the size of the simplified trees are linear in the size of the initial trees. To establish that $\underline{\lim} v_n > 0$, we have to use the majorization $[z^n]Q_t(z) \geq cz_0^{-n}$ for all n and t, and Chebyshev's inequality.

Limit Distribution

From X's representation and the variability condition, we deduce thanks to the Quasi Powers Theorem that we have a Gaussian limit.
For $\pi_{e_1} = \pi_{e_2} = 1/2$, we obtain [CFS90]: $\mu = 0.8196...$ and $\overline{v} = 0.2166...$

3. Reduction

We have the same notion of irreducible tree, and if $r(u)$ denotes the irreducible tree to which u reduces, the generating functions $R_t(z) := \sum_{r(u)=t} z^{|u|}$ for $t \in \mathcal{I}$ satisfy

$$\forall a \neq e, R_a(z) = \pi_a, \quad R_e(z) = \pi_e + z\sum_{t\in\mathcal{I}} R_t(z)^2, \quad R_{u\circ v}(z) = zR_u(z)R_v(z).$$

Studying the bivariate generating function

$$H(z,y) := \sum_{t\in\mathcal{Q}} y^{|r(t)|} z^{|t|}$$

along the same lines as for X, we establish the Gaussian limit for reduction.
A crucial point consists in finding c, z_0 and γ such that $R_t(z_0) < c\sqrt{\pi_t}\gamma^{|t|}$ for all t. We take $c = z_0^{-1/2}$ and $\gamma = \sqrt{z_0}$. These values prove adapted to show by induction the desired inequality for $t \neq e$, but the main difficulty consists in showing it for $t = e$, that is $R_e(z_0) < c\pi_e$. We proceed by using the family of trees without the label e, of generating function S. This family of trees can be strictly injected in the family of trees that do not reduce to e, so that $R_e(\rho) < R(\rho) - S(\rho)$. This inequality remains true for z_0 slightly greater than ρ, and as $S(\rho) = 1 - \pi_e + \rho S^2(\rho) = R(\rho) - 2\sqrt{\pi(e)}$, we deduce the desired inequality $R_e(z_0) < 2\sqrt{\pi_e}$. The values of c, z_0 and γ prove adapted to show by induction that $R_t(z_0) < c\sqrt{\pi_t}\gamma^{|t|}$ for all t.
For $\pi_{e_1} = \pi_{e_2} = 1/2$, we obtain [CFS90]: $\mu = 0.8162...$ et $\overline{v} = 0.2479...$

4. A Third rule

There exist a third type of rule of reduction used in computer algebra.

Absorption (4 rules)

$f \wedge F \to F$ (False)
$F \wedge f \to F$
$f \vee T \to T$ (True)
$T \vee f \to T$

Neutral element (4 rules)

$f \wedge T \to f$
$T \wedge f \to f$
$f \vee F \to f$
$F \vee f \to f$

From now on we will denote $\wedge$ et $\vee$ by $\odot$ and $\oplus$. We have begun the study of that rule.
The model is the following: we put operators $\odot$ and $\oplus$ with probability $1/2$ on internal nodes, operands T and F with probability $p/2$, and another operand X with probability $q = 1 - p$.
This model is simpler than the former ones, in so far as we can find an explicit expression for the bivariate generating function

$$G(z,y) \;=\; 1 - (1-y)p(1-\sqrt{1-4z}) \\ -\sqrt{1-4z+4(1-y)\big((1-p^2)z + (p^2-p)(1-\sqrt{1-4z})/2\big)}$$

but we have the following singularity diagram:

y	0	1	$+\infty$
$\rho(y)$	$1/4$	$1/4$	z_1

with

$$z_1 \;=\; 1/4 - 1/4\,(p-1)^2\,t - 1/2p\,(p-1)^2\,t^{3/2} - 1/16\left(-4+12p^2\right)(p-1)^2\,t^2 \\ -1/16\left(4p + 8\left(-2+2p^2\right)p\right)(p-1)^2\,t^{5/2} + O(t^3)$$

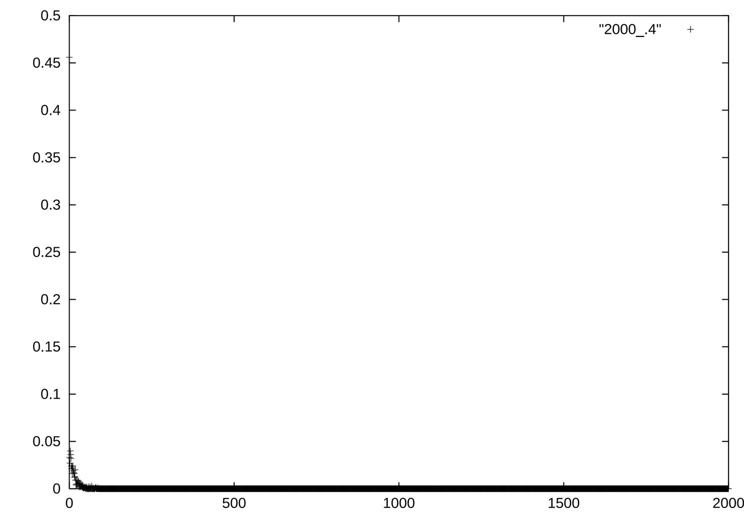
if $t = 1 - y$.
We found an expectation and a variance respectively given by:

$$\mu_n = \frac{[z^n]G_y(z,1)}{C_{n-1}} \underset{n\to\infty}{=} q^2 n\left(1 + \frac{1}{4n} + O\left(\frac{1}{n^2}\right)\right)$$

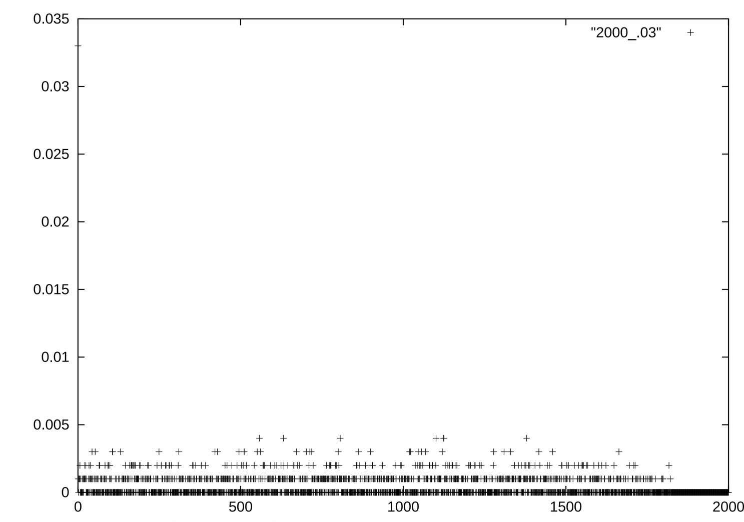
and

$$\sigma_n^2 = \frac{[z^n]G_{y^2}(z,1)}{C_{n-1}} + \mu_n - \mu_n^2 = (8(1-q^2) - p)q^2\sqrt{\pi} \cdot n^{3/2} + O(n),$$

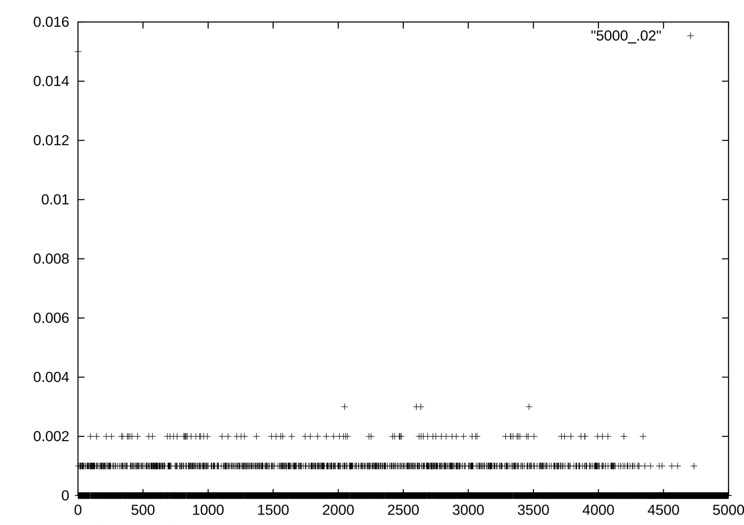
where C_n is the nth Catalan number.
Simulations for $p = 0.4, 0.03, 0.02, 0.01$ indicate that there exist no limit law.



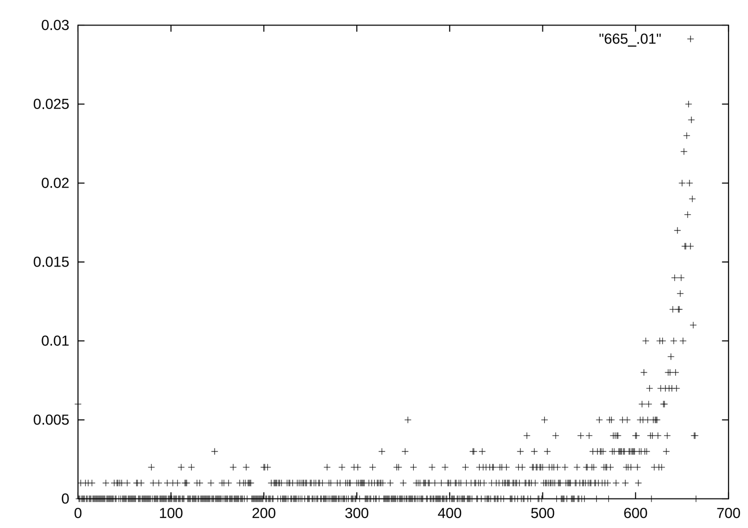
Simulation for $p = 0.4$ and $n = 2000$.



Simulation for $p = 0.03$ and $n = 2000$.



Simulation for $p = 0.02$ and $n = 5000$.



Simulation for $p = 0.01$ and $n = 665$.

5. Conclusion

For the rules of simplification and reduction, we show a Gaussian limit, and a mix of these two operators probably also leads to a Gaussian limit. For the third rule of reduction, with absorption and neutral element, there seems to be no limit law.

References

[CFS90] Rafael Casas, María-Inés Fernández Camacho, and Jean-Marc Steyaert. Algebraic simplification in computer algebra: An analysis of bottom-up algorithms. *Theoretical Computer Science*, 74(74):273–298, 1990.

[FC88] María-Inés Fernández-Camacho. *Analisis Medio de Algoritmos de reducción sobre árboles*. Tesis, Universidad Complutense de Madrid, 1988.

[Ngu02] Michel Nguyễn Thế. Distribution of the size of simplified or reduced trees. In *Colloquium on Mathematics and Computer Science: Algorithms, Trees, Combinatorics and Probabilities*, Trends in Mathematics, pages 341–354. Birkhäuser, 2002.