On counting untyped lambda terms

Pierre Lescanne
University of Lyon, ENS de Lyon,
46 allée d'Italie, 69364 Lyon, France

Abstract

We present several results on counting untyped lambda terms, i.e., on telling how many terms belong to such or such class, according to the size of the terms and/or to the number of free variables.

1 Introduction

This paper presents several results on counting untyped lambda terms, i.e., on telling how many terms belong to such or such class, according to the size of the terms and/or to the number of free variables. In particular, it gives results on the distribution of terms according to their number of free variables and explores properties of these distributions, especially through generating functions. In addition to the inherent interest of these results from the mathematical point of view, we expect that a knowledge on the distribution of terms will improve the implementation of reduction and that results on asymptotic distributions of terms will give a better insight of the lambda calculus. We show strong evidence for the claim that when the number m of free variables tends to infinity, the ratio "number of normal forms of size n over number of terms of size n" decreases and tends to 1, meaning that when m grows, more and more terms are normal forms. The results presented here are preliminary, but we expect them to be a milestone in describing probabilistic properties of lambda terms with answers to questions like: How does a random lambda term looks like? How does a random normal form look like? How to generate a random lambda term (a random normal form)?

Related works

Previous unpublished works on counting lambda terms were by R. David et al. [1] and J. Wang [9]. Related works are on counting types and/or counting tautologies [10, 6, 2, 7].

2 Untyped lambda terms with de Bruijn indices

I am dedicating this book to N. G. "Dick" de Bruijn, because his influence can be felt on every page. Ever since the 1960s he has been my chief mentor, the main person who would answer my question when I was stuck on a problem that I had not been taught how to solve.

Donald Knuth in preface of [8]

In this paper we represent terms by De Bruijn indices [4], this means that variables are represented by numbers $\underline{1}, \underline{2}, ..., \underline{m}, ...,$ where an index, for instance $\underline{k}$, is the number of λ, above the location of the index and below the λ that binds the variable, in a representation of λ-terms by trees. For instance, the term with variables $\lambda x.\lambda y.x\, y$ is represented by the term with de Bruijn indices $\lambda\lambda\underline{2}\underline{1}$. The variable x is bound by the top λ. Above the occurrence of x, there are two λ's, therefore x is represented by $\underline{2}$ and from the occurrence of y, one counts just the λ that binds y; so y is represented by $\underline{1}$. In what follows we will call *terms*, the untyped terms with de Bruijn indices. Let us call $\mathcal{T}_{n,m}$, the set of terms of size n, with m de Bruijn indices, i.e., with indices in $\mathcal{I}(m) = \{\underline{1}, \underline{2}, ..., \underline{m}\}$. A term in $\mathcal{T}_{n,m}$, is either a de Bruijn index or an abstraction on a term with $m+1$ indices i.e., a term in $\mathcal{T}_{n,m+1}$ or an application of a term in $\mathcal{T}_{n,m}$ on a term in $\mathcal{T}_{n,m}$. We can write, using @ as the application symbol,

$$\mathcal{T}_{n+1,m} = \mathcal{I}(m) \uplus \lambda\mathcal{T}_{n,m+1} \uplus \biguplus_{k=0}^{n} \mathcal{T}_{n-k,m} @ \mathcal{T}_{k,m}.$$

Moreover terms of size 1 are only made of de Bruijn indices, therefore

$$\mathcal{T}_{1,m} = \mathcal{I}(m).$$

From this one gets:

$$\begin{aligned} T_{n+1,m} &= T_{n,m+1} + \sum_{k=1}^{n} T_{n-k,m}.T_{k,m} \qquad \text{for } n > 1 \\ T_{1,m} &= m \end{aligned}$$

$T_{n,0}$ is the set of closed terms (terms with no non bound indices). Let us illustrate this result by the array of closed terms up to size 5:

n	terms	$T_{n,0}$
1	none	0
2	$\lambda\underline{1}$	1
3	$\lambda\lambda\underline{1}, \lambda\lambda\underline{2},$	2
4	$\lambda\lambda\lambda\underline{1}, \lambda\lambda\lambda\underline{2}, \lambda\lambda\lambda\underline{3}, \lambda(\underline{1}.\underline{1})$	4
5	$\lambda\lambda\lambda\lambda\underline{1}, \lambda\lambda\lambda\lambda\underline{2}, \lambda\lambda\lambda\lambda\underline{3}, \lambda\lambda\lambda\lambda\underline{4}, \lambda\lambda(\underline{1}.\underline{1}), \lambda\lambda(\underline{1}.\underline{2}), \lambda\lambda(\underline{2}.\underline{1}), \lambda\lambda(\underline{2}.\underline{2}),$ $\lambda(\underline{1}.\lambda\underline{1}), \lambda(\underline{1}.\lambda\underline{2}), \lambda((\lambda\underline{1}).\underline{1}), \lambda((\lambda\underline{2}).\underline{1}), \lambda\underline{1}.\lambda\underline{1}$	13

The equation that defines $T_{n,m}$ allows us to compute it, since it relies on entities $T_{i,j}$ where either $i < n$ or $j < m$. Figure 1 is a table of the first values of $T_{n,m}$

up to $T_{18,7}$. We are mostly interested by the sequence of sizes of the closed terms, namely $T_{n,0}$, in other words the first column of the table. The values of $T_{n,0}$ correspond to sequence **A135501** (see `http://www.research.att.com/~nudges/sequences/A135501`) due to Christophe Raffalli, which is the *number of closed lambda-terms of size n*. His formula for those numbers is more complex. He considers the values of the double sequence $f_{n,m}$. $T_{n,m}$ and $f_{n,m}$ coincide for $m = 0$, i.e., $T_{n,0} = f_{n,0}$.

$$\begin{aligned}
f_{1,1} &= 1 \\
f_{0,m} &= 0 \\
f_{n,m} &= 0 \text{ if } m > 2n-1 \\
f_{n,m} &= f_{n-1,m} + f_{n-1,m+1} + \sum_{p=1}^{n-2}\sum_{c=0}^{m}\sum_{l=0}^{m-c} \binom{c}{m}\binom{l}{m-c} f_{p,l+c} f_{n-p-1,m-l}.
\end{aligned}$$

He adds

> The last term is for the application where c is the number of common variables in both subterms. $f_{n,m}$ can be computed only using $f_{n',m'}$ with $n' < n$ and $m' \leq m + n - n'$.

Notice that he deals only with sequence $T_{n,0}$, whereas we consider the values for any value of m, from which we can expect to extract interesting informations. The main interesting statement we can draw from this is that considering lambda terms with explicit variables or considering lambda terms with de Bruijn indices makes no difference, at least when no β-reduction is taken into account. We feel that considering lambda terms with de Bruijn indices makes the task easier and produces more results.

3 The functions $m \mapsto T_{n,m}$

Due to properties of the generating function (see Section 5) we are not able to give a simple expression for the function $n \mapsto T_{n,m}$, so we focus on the function $m \mapsto T_{n,m}$. We have found experimentally that this function is a polynomial P_n^T and we give the first 18 ones in Figure 2. From this table, we collected a few facts on the coefficients of the polynomials, with the aim that we will be able to find the constant term of P_n^T, which is $T_{n,0}$.

- The degree of P_{2q-1}^T and of P_{2q}^T is q.
- Leading coefficients $p_{2q+1,q+1}^T$ of P_{2q+1}^T are Catalan numbers $\frac{1}{q+1}\binom{2q}{q}$.
- Leading coefficients $p_{2q,q}^T$ of P_{2q}^T are $\binom{2q-1}{q} = (2q-1)\, p_{2q-1,q}^T$.
- Second leading coefficients $p_{2q+1,q}^T$ of P_{2q+1}^T are $q\binom{2q-1}{q} = q\, p_{2q,q}^T$.

- Hence typically

$$\begin{aligned} P^T_{2q+1}(m) &= p^T_{2q+1,q+1} m^{q+1} + p^T_{2q+1,q} m^q + \ldots + p^T_{2q+1,0} \\ &= \frac{1}{q+1}\binom{2q}{q} m^{q+1} + p\binom{2q-1}{q} m^q + \ldots + T_{2q,0}. \end{aligned}$$

- Hence we get the induction relations for the leading coefficients:

$$\begin{aligned} p^T_{1,1} &= 1 \qquad & p^T_{2q+1,q+1} &= \tfrac{2(2q-1)}{q+1} p^T_{2q-1,q} \\ p^T_{3,1} &= 1 \qquad & p^T_{2q+1,q} &= \tfrac{2(2q-1)}{q-1} p^T_{2q-1,q-1} \\ p^T_{2,1} &= 1 \qquad & p^T_{2q,q} &= \tfrac{2(2q-1)}{q} p^T_{2q-2,q-1} \end{aligned}$$

Recall also the well-known formula on Catalan numbers applied to $p^T_{2q+1,q+1}$:

$$p^T_{2q+1,q+1} = \sum_{k=0}^{2q} p^T_{2(q-k)+1,q-k+1} \cdot p^T_{2k+1,k}$$

We can summarize the coefficients $p^T_{n,i}$ in the table:

$n\backslash m^i$	m^7	m^6	m^5	m^4	m^3	m^2	m	1
1							1	0
2							*1*	*1*
3						1	1	2
4						*3*	*5*	*4*
5					2	6	17	13
6					*10*	*26*	*49*	*42*
7				5	30	111	179	139
8				*35*	*134*	*405*	*683*	*506*
9			14	140	652	1658	2629	1915
10			*126*	*676*	*2812*	*7122*	*10725*	*7558*
11		42	630	3610	12760	30783	45195	31092
12		*462*	*3334*	*17670*	*60240*	*138033*	*196355*	*132170*
13	132	2772	19218	87850	285982	635178	880379	580466

Open questions

The table raises a bunch of open questions.

- What is the induction relation between $p^T_{2q+1,q-1}$ and $p^T_{2q-1,q-2}$ (we know the first numbers of the sequence, namely $2, 17, 111, 652, 3610, 19218$) and more generally between $p^T_{2q+1,q-j}$ and $p^T_{2q-1,q-j-1}$?
- Can we describe the second coefficient $p^T_{2q,q-1}$ of P^T_{2q}?
- Can we describe the coefficients of the P^T_n's? Very likely they are described by binomials, close to Catalan numbers.

- Can we describe the last coefficient of P_n^T? If yes, we know $T_{n,0}$. Actually since the sequence is increasing, we can tell the asymptotic behavior of $T_{n,0}$ if we know only $T_{2q+1,0}$, i.e., $p_{2q+1,0}^T$.

4 Normal forms

Lest us call $\mathcal{F}_m$ the set of normal forms with m de Bruijn indices and $\mathcal{G}_m$ the sets of normal forms with no head λ. The combinatorial structure equations are

$$
\begin{aligned}
\mathcal{G}_m &= \mathcal{I}(m) \uplus \mathcal{G}_m @ \mathcal{F}_m \\
\mathcal{F}_m &= \lambda\, \mathcal{F}_{m+1} \uplus \mathcal{G}_m
\end{aligned}
$$

Let $G_{n,m}$ be the number of normal forms of size n with no head λ and with de Bruijn indices in $\mathcal{I}(m)$ and $F_{n,m}$ be the number of normal forms of size n with de Bruijn indices in $\mathcal{I}(m)$. The relations between $G_{n,m}$ and $F_{n,m}$ are

$$
\begin{aligned}
G_{1,m} &= m \\
G_{n+1,m} &= \sum_{k=0}^{n} G_{n-k,m} F_{k,m} \\
\\
F_{1,m} &= m \quad = \quad G_{1,m} \\
F_{n+1,m} &= F_{n,m+1} + G_{n+1,m}
\end{aligned}
$$

whereas the relations between generating functions are

$$
\begin{aligned}
G_m(z) &= m\, z + z\, G_m(z)\, F_m(z) \\
F_m(z) &= z\, F_{m+1}(z) + G_m(z).
\end{aligned}
$$

The coefficients $F_{n,m}$ are given in Figure 3.

The functions $m \mapsto F_{n,m}$

Like for $m \mapsto T_{n,m}$, the functions $m \mapsto F_{n,m}$ are polynomials of degree $n+1 \div 2$, which we write P_n^{NF} and which are given in Figure 4. Conjectured facts about the coefficients of polynomials P_n^{NF} are somewhat similar to those conjectured for polynomials P_n^T.

- Like for P_{2q+1}^T the coefficients of P_{2q+1}^{NF} are Catalan numbers.
- The coefficients of P_{2q+2}^{NF} are $2\binom{2q-1}{q}$, i.e., $P_{2q+2}^{NF} = 2P_{2q}^T$.

If this is provable, this means that the leading coefficients of P_n^T and P_n^{NF} have asymptotically the same behavior, so this shows the following fact: when the number m of free variables grows, $P_{2q+1}^T(m) \sim P_{2q+1}^{NF}(m)$ and since $P_{2q-1}^T < P_{2q}^T < P_{2q+1}^T$ and $P_{2q-1}^{NF} < P_{2q}^{NF} < P_{2q+1}^{NF}$ then when m grows the set of normal forms tends to be the same as the sets of terms and the set of normal forms tends to be no more negligible.

5 Generating functions for terms

We consider the bivariate generating function for terms which coefficients are the $T_{n,m}$ and the vertical generating functions which gives the $T_{n,m}$ for each value of m.

The bivariate generating function for terms

Let $T(z,u)$ be the bivariate generating function of $T_{n,m}$.

$$\begin{aligned}
T(0,u) &= 0 \\
&\text{and} \\
T(z,u) &= \sum_{n,m\geq 0} T_{n,m} z^n u^m \\
&= z\sum_{m\geq 0} m u^m + \sum_{n\geq 1}\sum_{m\geq 0} T_{n-1,m+1} z^n u^m + zT^2(z,u) \\
&= \frac{z\,u}{(1-u)^2} + \sum_{n\geq 0}\sum_{m\geq 1} T_{n,m} z^{n+1} u^{m-1} + zT^2(z,u) \\
&= \frac{z\,u}{(1-u)^2} + \frac{z}{u}(T(z,u) - T(z,0)) + zT^2(z,u)
\end{aligned}$$

Hence

$$zuT^2(z,u) + (z-u)T(z,u) + \frac{z\,u^2}{(1-u)^2} - zT(z,0) = 0$$

and

$$\boxed{T(z,u) \quad = \quad \frac{-(z-u)+\sqrt{(z-u)^2+4zu(\frac{z\,u^2}{(1-u)^2}-zT(z,0))}}{2zu}.}$$

We notice that

$$\lim_{u\to 0} T(z,u) = T(z,0).$$

On the other hand we see the presence of $T(z,0)$ in the lefthand side, which shows that the equation is not a definition of T but rather an equation which characterizes T as a fixed point. Presented like that, the equation is not amenable to the same treatment as those presented in Flajolet and Sedgewick [5] and specialists consider it as difficult and no result is known about the asymptotic behavior of its coefficient.

Vertical generating functions

We see that

$$T_{n,m+1} = T_{n+1,m} - \sum_{k=0}^{n} T_{n-k,m} T_{k,m}.$$

Hence

$$T^{\langle m\rangle}(0) \quad = \quad 0$$

and

$$\begin{aligned}
T^{\langle m+1\rangle}(z) &= \sum_{n=0}^{\infty} T_{n,m+1} z^n \\
&= \sum_{n=0}^{\infty} T_{n+1,m} z^n - \sum_{n=0}^{\infty}\sum_{k=0}^{n} T_{n-k,m} T_{k,m} z^n \\
&= \frac{T^{\langle m\rangle}(z)}{z} - (T^{\langle m\rangle}(z))^2
\end{aligned}$$

In other words

$$z(T^{\langle m\rangle}(z))^2 - T^{\langle m\rangle}(z) + zT^{\langle m+1\rangle}(z) = 0.$$

Hence

$$T^{\langle m\rangle}(z) = \frac{1 - \sqrt{1 - 4z^2 T^{\langle m+1\rangle}(z)}}{2z}.$$

One sees that $T^{\langle m\rangle}$ is defined from $T^{\langle m+1\rangle}$. Like the bivariate generating function $T(z,u)$, $T^{\langle m\rangle}(z)$ is also difficult to study, because we have $T^{\langle m\rangle}$ defined in term of $T^{\langle m+1\rangle}$.

6 Conclusion

We have given several parameters on numbers of untyped lambda terms and untyped normal forms and proved or conjectured facts about them. Further tracks can lead to compute moments which will allow us to characterize precisely properties of large random lambda terms which hold with high probability. On another direction, it could be worth to study typed lambda terms, whereas we have only analyzed untyped lambda terms in this paper.

References

[1] René David, Christophe Raffalli, Guillaume Theyssier, Katarzyna Grygiel, Jakub Kozik, and Marek Zaionc. Some properties of random lambda terms. *CoRR*, abs/0903.5505, 2009.

[2] René David and Marek Zaionc. Counting proofs in propositional logic. *Arch. Math. Log.*, 48(2):185–199, 2009.

[3] Nicolaas Govert de Bruijn. *Asymptotic Methods in Analysis*. North-Holland, Amsterdam, 1958.

[4] Nicolaas Govert de Bruijn. Lambda calculus with nameless dummies, a tool for automatic formula manipulation, with application to the Church-Rosser theorem. *Proc. Koninkl. Nederl. Akademie van Wetenschappen*, 75(5):381–392, 1972.

[5] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics.* Cambridge University Press, 2008.

[6] Hervé Fournier, Danièle Gardy, Antoine Genitrini, and Marek Zaionc. Classical and intuitionistic logic are asymptotically identical. In Jacques Duparc and Thomas A. Henzinger, editors, *CSL*, volume 4646 of *Lecture Notes in Computer Science*, pages 177–193. Springer, 2007.

[7] Hervé Fournier, Danièle Gardy, Antoine Genitrini, and Marek Zaionc. Tautologies over implication with negative literals. *Math. Log. Q.*, 56(4):388–396, 2010.

[8] Donald E. Knuth. *Selected Papers on Analysis of Algorithms*, volume 102 of *CSLI Lecture Notes.* Stanford, California: Center for the Study of Language and Information, 2000.

[9] Jue Wang. Generating random lambda calculus terms. Technical report, Citeseer, 2005.

[10] Marek Zaionc. Probability distribution for simple tautologies. *Theor. Comput. Sci.*, 355(2):243–260, 2006.

$n\backslash m$	0	1	2	3	4	5	6	7
1	0	1	2	3	4	5	6	7
2	1	2	3	4	5	6	7	8
3	2	4	8	14	22	32	44	58
4	4	12	26	46	72	104	142	186
5	13	38	87	172	305	498	763	1112
6	42	127	324	693	1294	2187	3432	5089
7	139	464	1261	2890	5831	10684	18169	29126
8	506	1763	5124	12653	27254	52671	93488	155129
9	1915	7008	21709	57070	130863	269260	508513	896634
10	7558	29019	94840	265129	646458	1406983	2791564	5136885
11	31092	124112	427302	1264362	3262352	7502892	15703602	30429782
12	132170	548264	1977908	6168242	16811366	40776020	89671904	181746638
13	580466	2491977	9384672	30755015	88253310	225197061	520076012	1104714147
14	2624545	11629836	45585471	156409882	471315501	1263116040	3058077451	6789961206
15	12190623	55647539	226272369	810506769	2558249963	7184911623	18208806189	42244969589
16	58083923	272486289	1146515237	4275219191	14098296495	41417170373	109721440529	265618096347
17	283346273	1363838742	5923639803	22933607180	78832280277	241776779298	668513708207	1686996660888
18	1413449148	6968881025	31177380822	125027527671	446961983408	1428444131853	4116538065930	10816530842627

Figure 1: Values of $T_{n,m}$ up to $(18, 7)$

n	$P_n^T(m)$
1	m
2	$m+1$
3	m^2+m+2
4	$3m^2+5m+4$
5	$2m^3+6m^2+17m+13$
6	$10m^3+26m^2+49m+42$
7	$5m^4+30m^3+111m^2+179m+139$
8	$35m^4+134m^3+405m^2+683m+506$
9	$14m^5+140m^4+652m^3+1658m^2+2629m+1915$
10	$126m^5+676m^4+2812m^3+7122m^2+10725m+7558$
11	$42m^6+630m^5+3610m^4+12760m^3+30783m^2+45195m+31092$
12	$462m^6+3334m^5+17670m^4+60240m^3+138033m^2+196355m+132170$
13	$132m^7+2772m^6+19218m^5+87850m^4+285982m^3+635178m^2+880379m+580466$
14	$1716m^7+16108m^6+104034m^5+449290m^4+1390246m^3+2991438m^2+4052459m+2624545$
15	$429x^8+12012x^7+99386x^6+560854x^5+2308173x^4+6895122x^3+14436365x^2+19144575x+12190623$
16	$6435x^8+76444x^7+584878x^6+3076878x^5+12039895x^4+34815210x^3+71170791x^2+92631835x+58083923$
17	$1430x^9+51480x^8+502384x^7+3389148x^6+16925916x^5+63753310x^4+179178860x^3+358339416x^2+458350525x+283346273$
18	$24310x^9+357256x^8+3176112x^7+19799164x^6+93981244x^5+342274990x^4+938333964x^3+1840448776x^2+2317036061x+1413449148$

Figure 2: The polynomials P_n^T for the function $m \mapsto T_{n,m}$

$n\backslash m$	0	1	2	3	4	5	6	7	8
1	0	1	2	3	4	5	6	7	8
2	1	2	3	4	5	6	7	8	9
3	2	4	8	14	22	32	44	58	74
4	4	10	20	34	52	74	100	130	164
5	10	25	58	121	226	385	610	913	1306
6	25	72	185	400	753	1280	2017	3000	4265
7	72	223	614	1497	3244	6347	11418	19189	30512
8	223	728	2195	5716	12863	25688	46723	78980	125951
9	728	2549	8108	22745	56360	125093	253004	473753	832280
10	2549	9254	31253	93734	244997	564854	1173029	2237558	3983189
11	9254	35168	124778	395720	1109222	2770904	6261818	12999728	25130630
12	35168	138606	512898	1720040	5097660	13347978	31308206	66902388	132274680
13	138606	563907	2174894	7645095	23948550	66818531	167837142	384821079	816168830
14	563907	2369982	9459993	34771380	114618495	335857722	880524117	2092596528	4571548155
15	2369982	10231830	42221886	161568762	558056526	1723895502	4785906510	12073186866	28016723742
16	10231830	45381558	192944940	765787548	2764390146	8947158690	25962816408	68135021640	163627733358
17	45381558	206266797	901441688	3701763855	13912595562	47127027713	143678500332	397091138883	1005324501470
18	206266797	959283300	4302919895	18223902654	71123969121	251343711032	799893538635	2302171013970	6046781201429

Figure 3: Values of $F_{n,m}$ up to $(18, 8)$

n	$P_n^{NF}(m)$
1	m
2	$m+1$
3	m^2+m+2
4	$2m^2+4m+4$
5	$2m^3+3m^2+10m+10$
6	$6m^3+15m^2+26m+25$
7	$5m^4+12m^3+49m^2+85m+72$
8	$20m^4+62m^3+155m^2+268m+223$
9	$14m^5+50m^4+240m^3+589m^2+928m+728$
10	$70m^5+263m^4+870m^3+2146m^2+3356m+2549$
11	$42m^6+210m^5+1153m^4+3658m^3+8351m^2+12500m+9254$
12	$252m^6+1128m^5+4658m^4+14838m^3+33575m^2+48987m+35168$
13	$132m^7+882m^6+5446m^5+21198m^4+63138m^3+137695m^2+196810m+138606$
14	$924m^7+4862m^6+24086m^5+93748m^4+275898m^3+587814m^2+818743m+563907$
15	$429m^8+3696m^7+25372m^6+117120m^5+429435m^4+1223102m^3+2558090m^2+3504604m+2369982$
16	$3432m^8+20996m^7+121286m^6+556920m^5+2011411m^4+5601948m^3+11448828m^2+15384907m+10231830$
17	$1430m^9+15444m^8+116892m^7+624768m^6+2717670m^5+9524196m^4+26064412m^3+52459126m^2+69361301m+45381558$
18	$12870m^9+90683m^8+598120m^7+3162562m^6+13513606m^5+46329205m^4+124109404m^3+245453736m^2+319746317m+206266797$

Figure 4: The polynomials P_n^{NF} for the function $m \mapsto F_{n,m}$