Le paradoxe des facteurs (des mots de Dyck généralisés)

Cyril Banderier*

7 février 2001 (Soumission Poster à Discrete Models...)

Résumé

L'ubiquité des mots de Dyck en combinatoire et en théorie des langages est bien connue. Une généralisation simple consiste à considérer un alphabet à plus de 2 lettres, ayant chacune un poids dans $\mathbf{Z}$: on obtient alors les "langages de Dyck généralisés", qui s'avèrent être des langages algébriques. Nous montrons qu'il est possible d'établir un comportement typique pour les mots de Dyck généralisés, indépendamment du nombre de lettres et des poids choisis. Nous utilisons pour ce faire les méthodes de la combinatoire analytique (séries génératrices et analyse de singularités), notamment pour étudier le nombre de facteurs d'un mot de Dyck généralisé (en mots "primitifs"). Par exemple, un mot de Dyck typique se factorise en un produit de 3 mots de Dyck. L'étude de nombreux autres paramètres fondamentaux (aire, hauteur) des mots de Dyck généralisés devrait suivre la piste proposée ici. C'est un équivalent discret des marches aléatoires continues (mouvement Brownien), mais le modèle probabiliste uniforme (le plus naturel en combinatoire) amène des résultats asymptotiques parfois contre-intuitifs car différents du cas continu.

1 Langages de Dyck généralisés

Considérons un alphabet à deux lettres (a de poids $+1$ et b de poids -1). Un mot de Dyck est un mot dont la somme de toutes ses lettres est nulle et dont la somme des lettres de chacun des préfixes gauches est ≥ 0. On obtient ainsi le langage $\mathcal{D} = \{\epsilon, ab, abab, aabb, aaabbb, ababab, aababb, abaabb, aabbab, \ldots\}$ des mots de Dyck (en l'honneur de Walther von Dyck [1856-1934] qui a introduit les représentations de groupes). Les "langages de Dyck généralisés" sont bâtis similairement à partir d'un alphabet fini : on donne à chaque lettre a_i une valeur entière positive ou négative v_i et on forme un langage avec des conditions de positivité similaires à celles des mots de Dyck (valeur totale nulle, valeur positive pour tous les préfixes gauches).

Théorème 1 *Chaque langage de Dyck généralisé est engendré par une grammaire. La série génératrice des chemins de Dyck généralisés est algébrique.*

Preuve. Labelle et Yeh [7] l'ont démontré en déterminant une grammaire génératrice, ceci a l'avantage de permettre l'utilisation de grammaires attribuées (comme dans [5] pour des alphabets à deux lettres). Nous proposons une nouvelle preuve, analytique, qui a le mérite de capter une structure analytique remarquable, restant invisible avec les grammaires génératrices, et qui ouvre ainsi la porte à diverses études énumératives et asymptotiques. Notre preuve repose sur une méthode aussi astucieuse que simple, baptisée *méthode du noyau*, qui trouve ses origines dans le folklore combinatoire (Tutte, Knuth, Cori...), mais dont la généralité n'a été que récemment dégagée [1, 2, 3, 4].

Chaque mot de Dyck peut être considéré comme une marche sur $\mathbf{N}$ (partant et finissant en 0) dont $\{v_i\}$ est exactement l'ensemble des sauts possibles. Notons $P(u) = \sum u^{v_i}$ et soit $f_n(u)$ l'ensemble des positions atteintes en n sauts (la hauteur est codée par u). La récurrence $f_{n+1}(u) = \{u^{\geq 0}\}P(u)f_n(u)$ (où $\{u^{\geq 0}\}$ signifie que l'on ne prend en compte que les monômes de degré positif), en multipliant par

*Cyril.Banderier@inria.fr Projet Algorithmes, Inria-Rocquencourt htpp://algo.inria.fr/banderier

z^{n+1} et en sommant pour $n \geq 0$, aboutit à l'équation fonctionnelle pour $F(z,u) = \sum_{n\geq 0} f_n(u) z^n$:

$$(1 - zP(u))F(z,u) = 1 - z \sum_{j=0}^{c-1} q_j(u) \partial_u^j F(z,0)\,,$$

où les q_j sont des polynômes de Laurent, où $c = -\min v_i$ et où les dérivées en u de F (évaluées en $u = 0$) sont de nouvelles inconnues. Toutefois le "noyau" $1 - zP(u) = 0$ admet c racines $u_i(z)$ (tendant vers 0 en 0) que l'on peut réinjecter dans l'équation pour expliciter alors toutes les inconnues du membre droit et l'on obtient $F(z,u) = \frac{1}{1-zP(u)} \prod_{i=1}^{c} 1 - u_i(z)/u$.

Nous avons ainsi montré que la série génératrice $F(z,u)$ des préfixes des mots de Dyck généralisés est algébrique ($f_{n,k}$ est le nombre de mots de longueur n, ayant pour somme k et qui sont préfixes de mots de Dyck généralisés). En particulier, la série génératrice $F(z,0)$ des mots de Dyck généralisés est algébrique et vaut $F(z,0) = \frac{(-1)^{c+1}}{z} \prod_{i=1}^{c} u_i(z)$, ce qui révèle la structure analytique des mots de Dyck généralisés : un intrigant polynôme symétrique des racines "convergentes". □

2 Nombre moyen de facteurs

Tout mot de Dyck w se factorise en un produit de mot de Dyck "primitifs" (*i.e.* qui n'ont pas de préfixe qui soit un mot de Dyck). Ainsi $abaabb = ab.aabb$ a deux facteurs, $ababab = ab.ab.ab$ a trois facteurs, $aababb$ a un seul facteur (c'est un mot de Dyck "primitif"). Il est naturel de se demander combien un mot de Dyck (généralisé) de longueur n a de facteurs en général. Le paradoxe des facteurs est le suivant : il est erroné de croire que plus un mot de Dyck (généralisé) est long, plus il a de facteurs.

Théorème 2 (Paradoxe des facteurs) *Un mot de Dyck généralisé a asymptotiquement un nombre constant de facteurs.*

Preuve Un chemin est une excursion suivie d'un "reste" (un bout de marche qui ne repasse pas par 0) et une excursion est une suite d'arches, on a donc les relations $\mathcal{C} = \mathcal{E} \times \mathcal{R}$ et $\mathcal{E} = \mathrm{Seq}\,\mathcal{A}$. On peut alors faire une traduction en séries génératrices, en marquant avec une variable t chaque passage par zéro (*i.e.* chaque nouveau facteur) : $C(z,t) = \frac{1}{1-tA(z)} \times \frac{C(z)}{E(z)}$. On obtient $F(z,u,t) = \frac{1}{1-t(1-1/F(z,0))} \frac{F(z,u)}{F(z,0)}$, ce qui montre au passage que le langage des mots de Dyck généralisés "primitifs"est algébrique.

Le nombre moyen μ_n de facteurs d'un mot de longueur n s'obtient en regardant $F(z,0,t)$:

$$\mu_n = \frac{[z^n]\partial_t F(z,0,1)}{[z^n]F(z,0,1)} = \frac{[z^n]E(z)^2}{[z^n]E(z)} - 1 = 2b_0 - 1 + O\left(\frac{1}{n}\right)$$

car une analyse de singularité [6] (omise ici et non triviale car il s'agit d'établir un comportement pour tout une classe de fonctions non explicites [1]) montre que l'on a $E(z) = b_0 - b_1\sqrt{\rho - z} + \ldots$ □

Pour les mots de Dyck, on a ainsi $\mu_n = 3 - \frac{12}{n} + \frac{48}{n^2} + O(\frac{1}{n^3})$: le facteur frappe toujours trois fois !

Références

[1] Cyril Banderier. *Combinatoire analytique des cartes et des chemins.* Thèse, 3001.

[2] Cyril Banderier, Mireille Bousquet-Mélou, Alain Denise, Philippe Flajolet, Danièle Gardy, and Dominique Gouyou-Beauchamps. Generating functions for generating trees. *To appear in Discrete Mathematics.*

[3] M. Bousquet-Mélou and M. Petkovšek. Linear recurrences with constant coefficients: the multivariate case. *Discrete Mathematics*, 225(1-3):51–75, 2000.

[4] M. Bousquet-Mélou and G. Schaeffer. Counting paths on the slit plane. In *Colloque Informatique et Mathématiques : Algorithmes, Arbres, Combinatoire. Versailles*, 2000.

[5] Philippe Duchon. On the enumeration and generation of generalized Dyck words. *Discrete Mathematics*, 225(1-3):121–135, 2000.

[6] Philippe Flajolet and Robert Sedgewick. *The average case of algorithms.* INRIA Research Reports, 1993.

[7] Jacques Labelle and Yeong-Nan Yeh. Generalized Dyck paths. *Discrete Mathematics*, 82:1–6, 1990.

[8] Robert A. Sulanke. Moments of generalized Motzkin paths. *Journal of Integer Sequences*, 3, 2000.