Quelques contributions au carrefour de la géométrie, de la combinatoire et des probabilités.

Nicolas Pouyanne

Université de Versailles - Saint-Quentin-en-Yvelines
Ecole doctorale SoFt

Quelques contributions au carrefour de la géométrie, de la combinatoire et des probabilités

Nicolas Pouyanne

Laboratoire de mathématiques de Versailles
CNRS, UMR 8100
Université de Versailles - Saint-Quentin
45, avenue des Etats-Unis
78035 Versailles cedex
pouyanne@math.uvsq.fr

Rapport de synthèse en vue de l'obtention de l'habilitation à diriger des recherches en mathématiques, 28 novembre 2006

Jury :
M. Michael Drmota
M. Philippe Flajolet
M. Laurent Gruson
M. Svante Janson, rapporteur
M. Gérard Letac, rapporteur
M. Michel Merle
M. Didier Piau
M. Alain Rouault

Remerciements

A l'adage britannique *time is money*[1], l'universitaire d'aujourd'hui substitue plus volontiers *la liberté, c'est le temps*. Cela donne la mesure de ma reconnaissance à Philippe Flajolet qui a bien voulu se porter "tuteur" de cette habilitation, à Svante Janson et Gérard Letac, rapporteurs et à Michael Drmota, Laurent Gruson, Michel Merle, Didier Piau et Alain Rouault, examinateurs, pour avoir accepté de consacrer cette journée à mon travail après l'avoir lu et commenté.

A l'origine des temps, enseigner les mathématiques s'imposa comme une vocation, dont les fondements sont sans doute à rechercher entre les murs du lycée Louis Barthou à Pau. Vint plus tard l'heure des premiers pas vers la recherche qui, après tout, ne pouvait pas nuire au futur enseignant. Lentement se forgea la conviction que chercher et enseigner forment deux piliers complémentaires sans lesquels la science elle-même vacillerait. Ce furent alors la thèse de géométrie algébrique à Grenoble puis les trois années lyonnaises où j'eus la chance d'embrasser goulûment d'autres mathématiques sans que ma soif ne s'étanchât jamais.

Puis le port : Versailles. Il y eut d'emblée le contact avec toute la réalité universitaire, diverse et complexe. Ce fut Mokka qui, quelle audace pour un statisticien, me proposa cette question combinatoire sur les permutations, moi le géomètre de service. Il y eut la timide visite dans le bureau de Philippe Flajolet, à l'INRIA, puis la découverte de l'analyse d'algorithmes et de l'accueillante communauté *Algo* sous la houlette de ce dernier. Naturellement, il fallut bien se (re)mettre aux probabilités, alors précautionneusement rangées dans un tiroir paraphiné depuis l'Agrégation.

Pour apprendre à se servir d'un outil en informatique, la meilleure solution consiste – c'est une idée répandue – à dégoter quelque acolyte qui sache mieux et vous invite à le solliciter à toute heure du jour ou de la nuit. Pour le bonheur de cette ré-initiation aux probabilités, il en fut ainsi. Sans la complicité quotidienne de Brigitte Chauvin, il en eût été bien autrement.

C'est ainsi, sous le regard bienveillant de ces deux personnalités et dans la bonne humeur de la collaboration avec mes co-auteurs que put naître ce travail. Bien entendu, il y eut l'équipe de probabilités et statistiques du laboratoire. On ne cherche jamais vraiment seul, c'est tout le sens d'un laboratoire de mathématiques. Il y eut aussi les géomètres avec le séminaire et les groupes de travail. Peut-être se demandent-ils encore quelle science nous partageons. Mais voyons, nous autres mathématiciens en notre Académie, ne sommes-nous pas tous géomètres ?

Il serait très incomplet de se limiter aux seuls cadres du laboratoire et de la communauté Algo pour rendre compte des conditions de ce travail. Chemin faisant, il y eut les collègues enseignants qui permettent quand il le faut de se rappeler qu'il fut un temps où nous-mêmes n'avions pas la connaissance que nous tâchons de faire acquérir. Depuis mon arrivée à Versailles, la collaboration avec Aline Robert fut permanente, rigoureux

[1] B. Franklin, *Advice to a Young Tradesman*, ca.1748

étai de mes questionnements didactiques, épistémologiques et mêmes scientifiques tout court à propos des mathématiques et de leur environnement. Il y eut les jeunes étudiants dont l'attitude parfois déroutante nous montre qu'on est loin d'embrasser la complexité cognitive, psychologique et sociale de l'enseignement. Il y eut les jeunes chercheurs dont l'enthousiasme parfois prudent ne cesse d'être moteur. Il faut aussi veiller à ne pas sous-estimer le travail de l'ombre des secrétaires du département et du laboratoire. Enfin, le dernier aspect de la vie universitaire est le domaine politique et syndical. Il fut toujours intense à Versailles Saint-Quentin et les actions collectives qui en font l'essence furent aussi l'occasion de toucher du doigt la diversité des pratiques entre les domaines disciplinaires.

La recherche scientifique, univers dont le présent travail est une particule, prend son sens dans l'équilibre si complexe de l'édifice universitaire ; à rebours, il contribue à lui donner du sens.

Par ailleurs, le gougeage des anches, les salles de concert, les cailloux du Billare et la fenaison estivale à Lescun ne sont pas non plus pour rien dans cette affaire, vous imaginez.

Vous tous qui, peu ou prou, reconnaissez votre rôle à la lecture de ce récit, soyez-en chaleureusement remerciés.

Enfin, supporter un chercheur en mathématiques en guise d'époux et de père n'est certainement pas une sinécure. Pour plus de détail, allez donc demander à Hendrike, Marthe, Matthias et Thomas. Si ces quelques pages devaient être dédiées, c'est à eux, assurément, qu'elles le seraient, ainsi qu'à ma mère et feu mon père qui, pour son plus grand plaisir, s'étonna de porter la seule cravate de l'assistance lors de la soutenance de ma thèse grenobloise.

Lescun, août 2006.

Table des matières

1 Introduction

Ce texte entreprend de faire la synthèse de travaux de recherche dont les thèmes principaux sont empruntés à la géométrie algébrique, la combinatoire analytique et les probabilités. Proclamer le rattachement de ces articles aux domaines cités présente l'avantage évident d'une classification interne aux mathématiques. Mais s'en tenir à cette seule taxinomie occulterait un aspect essentiel du discours scientifique : l'apport des points de vue des différentes disciplines aux problématiques des autres. C'est pour cette raison, et parce que les publications en jeu s'y prêtent convenablement que le choix a été fait de porter autant que possible l'accent sur le carrefour des domaines abordés.

La première partie traite de variétés algébriques complexes de dimension trois. Briques de la classification encore en cours de ces dernières, les singularités-quotient par des groupes finis sont au centre du texte.

La conjecture de McKay conduit naturellement, dans le cas des groupes abéliens, à la cohomologie entière des variétés toriques lisses non complètes. Ce sont des considérations combinatoires élémentaires autour de la longue suite exacte de Mayer-Vietoris qui amènent aux résultats de topologie algébrique présentés.

Par ailleurs, pour toute singularité-quotient, on fournit un modèle birationnel dont les singularités sont des quotients par des groupes abéliens. Une fois la liste des sous-groupes finis de $\mathrm{PGL}(3,\mathbb{C})$ établie à conjugaison près, le calcul de ces modèles nécessite celui des configurations de ces groupes dans leur action naturelle sur le plan projectif complexe. Ici, c'est de combinatoire énumérative des actions de groupes finis qu'il s'agit.

Dans la deuxième partie, on part de la question suivante : un entier naturel m étant donné, quelle est la proportion asymptotique des permutations du groupe symétrique $\mathfrak{S}_n$ qui sont des puissances $m^{\text{ièmes}}$, lorsque n tend vers l'infini ? La méthode symbolique en combinatoire permet sans effort de calculer la fonction génératrice exponentielle de ces permutations comme un produit infini. Surgit alors la difficulté suivante : cette fonction est holomorphe dans le disque unité mais admet le cercle comme frontière naturelle, ce qui interdit l'usage de l'analyse des singularités. Par ailleurs, ses singularités ne sont pas assez violentes pour qu'opère la méthode du col comme dans l'exemple célèbre de l'asymptotique du nombre de partitions de G. H. Hardy et S. A. Ramanujan.

Une analyse directe de cette fonction génératrice permet de trouver un équivalent de la proportion cherchée par des méthodes de l'analyse élémentaire. Mais ce produit

infini, après réécriture, fait apparaître un hiérarchie naturelle dans son infinité de points singuliers qui relèvent tous isolément de l'analyse des singularités. Une telle forme avait aussi été rencontrée par ceux qui devinrent co-auteurs de [8]. Fruit de cette collaboration, une hybridation de la méthode de Darboux et de l'analyse des singularités permet de développer à un ordre arbitraire, lorsque n tend vers l'infini, le $n^{\text{ième}}$ coefficient d'une telle série entière. Ce résultat trouve des applications dans de nombreux autres problèmes combinatoires ou probabilistes.

Les arbres m-aires de recherche sont une des structures fondamentales de l'algorithmique des ensembles de données informatiques ; ils sont le point d'entrée de la troisième partie. Modélisés comme processus aléatoires, ils font l'objet d'une transition de phase qui resta inexpliquée par la combinatoire analytique. C'est en conservant le point de vue vectoriel du processus que ce phénomène fut élucidé par des techniques probabilistes de martingales, en collaboration avec B. Chauvin.

Ce processus s'exprime dans le cadre des urnes de Pólya-Eggenberger, qui se généralisent naturellement aux *processus de Pólya*. Ces derniers se séparent en deux catégories qui traduisent le changement de phase des arbres m-aires de recherche : les *petits*, qui relèvent la plupart du temps d'un théorème de la limite centrale essentiellement normal et les *grands*, qui admettent toujours une asymptotique presque sûre au second ordre. Le cas des grands processus, traité par des méthodes de martingales, d'algèbre linéaire et de géométrie dans les réseaux, fait apparaître de nouvelles lois pour lesquelles beaucoup de questions restent ouvertes. Ces processus modélisent de nombreuses situations des mathématiques, de l'informatique et de la physique théorique.

La dernière partie, qui retrace un travail en commun avec les co-auteurs de [9] est encore un exemple au carrefour des disciplines. La question vient de la biologie du génome en quête de modèle et de représentation des énormes bases de données que constituent les séquences d'ADN. Leur traitement informatique suggère un algorithme par structure arborescente, appelé *arbre-CGR* dont l'analyse trouve un cadre probabiliste naturel. C'est pour l'essentiel la hauteur et la profondeur d'insertion de ces arbres aléatoires qui fait l'objet de cette étude.

Les publications qui font l'objet du présent rapport sont numérotées de [1] à [9] dans la liste des références bibliographiques. Cette liste est reportée à la fin du texte.

2 Combinatoire de singularités algébriques

Dans ce chapitre, il est question de variétés algébriques complexes de dimension trois. Sous une hypothèse rendue naturelle par la nature combinatoire de l'approche, on calcule les groupes de cohomologie entière de variétés toriques lisses non complètes. Par ailleurs, les quotients de variétés algébriques quasi-projectives lisses par des groupes finis d'automorphismes ont des germes analytiques de singularités qui sont ceux des quotients de $\mathbb{C}^3$ par l'action naturelle des sous-groupes finis de $\mathrm{GL}(3,\mathbb{C})$. La compréhension de la géométrie de ces singularités-quotient passe par les propriétés combinatoires de l'action de ces groupes sur le plan projectif, qui sont accessibles par diverses formes d'équations aux classes.

2.1 Cohomologie entière de variétés toriques lisses non complètes

2.1.1 Soient M un réseau de rang 3. Si σ est un cône polyédral rationnel de dimension 3 de $M_{\mathbb{R}} = M \otimes \mathbb{R}$, on note X_σ la variété algébrique affine normale de dimension 3

$$X_\sigma = \operatorname{Spec} \mathbb{C}[\sigma \cap M],$$

variété torique dont l'anneau des fonctions régulières $\mathbb{C}[\sigma \cap M]$ est l'algèbre complexe du monoïde $\sigma \cap M$. Si Σ est un éventail du réseau dual $N = \mathrm{Hom}_{\mathbb{Z}}(M, \mathbb{Z})$, on note X_Σ la variété torique complexe canoniquement définie par Σ, dont un système de cartes affines équivariantes est donné par la famille des cônes duaux de Σ.

La géométrie torique fait l'objet d'une littérature devenue abondante. Les propriétés géométriques d'une variété torique se traduisent par des propriétés combinatoires de son éventail. Ce dictionnaire est un aspect fondamental de cette géométrie. Par exemple, un cône de $M_{\mathbb{R}}$ est dit *régulier* lorsqu'il est engendré par une base du réseau M ; une variété torique X_Σ est lisse si, et seulement si les cônes de dimension maximale de son éventail Σ sont tous réguliers (voir Danilov [20]). Ces variétés constituent une source d'exemples et un terrain d'expérimentation pour les conjectures de la géométrie algébrique. En outre, certains problèmes généraux se réduisent à des problèmes toriques qui s'expriment alors en termes combinatoires dans un réseau. On trouvera dans l'article de V. I. Danilov [20] et dans les livres de G. Kempf, F. Knudsen, D. Mumford et B. Saint-Donat [42] ou de W. Fulton [31] un développement de la géométrie torique élémentaire.

Il est démontré dans Danilov [20] qu'une variété torique est complète si, et seulement si le support de son éventail est le réseau tout entier. Dans le même article, les groupes de cohomologie rationnelle d'une telle variété sont exprimés en fonction du nombre de cônes de différentes dimensions de l'éventail – la structure de l'anneau de Chow, dans ce cas isomorphe à l'anneau de cohomologie, y est calculée. En revanche, le cas de la cohomologie entière et des variétés non complètes n'est pas traité par la théorie générale. C'est l'objet de l'article [3] dont la motivation initiale fut guidée par une formulation de M. Reid de la conjecture de McKay (voir ci-dessous). Un article très récent de M. Franz ([29]) approfondit le sujet.

2.1.2 Une variété torique lisse X_Σ étant donnée, le système de cartes affines isomorphes à $\mathbb{C}^3$ fourni par les cônes de dimension maximale de l'éventail régulier Σ permet d'envisager le calcul de la cohomologie entière *via* la longue suite exacte de Mayer-Vietoris.

Si Σ est un éventail de N, on définit dans [3] la *section sphérique* de Σ l'intersection du support de Σ avec la sphère euclidienne unité $\mathbf{S}^2$ de $N_\mathbb{R}$. Pour être précis, c'est de la classe d'homéomorphisme de cette intersection qu'il s'agit. En effet, ainsi donnée, cette définition dépend du choix d'une base de N, mais les seules considérations que nous aurons sur les sections sphériques seront relatives à leur topologie.

2.1.3 On donne un premier exemple : soit X une variété torique lisse dont la section sphérique est la réunion des deux triangles du papillon dessiné à gauche de la figure 1. Elle est la réunion de deux copies de $\mathbb{C}^3$ (les ailes du papillon) dont l'intersection, produit d'un tore complexe de dimension 2 et d'une droite affine complexe, se rétracte par déformation sur le produit de deux cercles $\mathbf{S}^1 \times \mathbf{S}^1$. On déduit alors de la longue suite exacte de

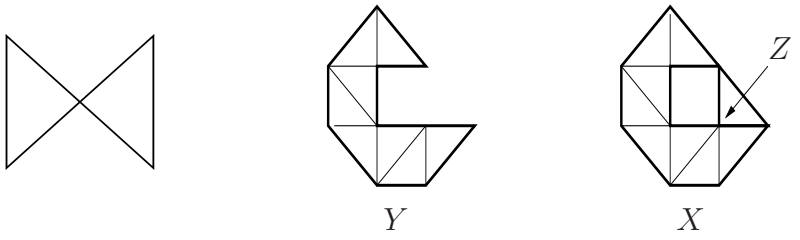

Figure 1: Exemples de sections sphériques d'éventails qui définissent des variétés toriques lisses ayant des groupes de cohomologie impaire non nuls.

Mayer-Vietoris que les groupes de cohomologie entière sont tous libres, respectivement isomorphes à

$$\mathrm{H}^0(X,\mathbb{Z}) \simeq \mathbb{Z}, \quad \mathrm{H}^2(X,\mathbb{Z}) \simeq \mathbb{Z}^2, \quad \mathrm{H}^3(X,\mathbb{Z}) \simeq \mathbb{Z} \ \text{ et } \ \mathrm{H}^q(X,\mathbb{Z}) = (0) \ \text{ si } \ q \neq 0,2,3.$$

2.1.4 Le fait qu'un des groupes de cohomologie impaire soit non nul dans l'exemple précédent provient de la forme singulière du papillon. Le théorème suivant est démontré dans [3].

Théorème 2.1 *Soit X une variété torique complexe lisse de dimension trois. Si la section sphérique de l'éventail Σ de X est homéomorphe au disque fermé $\mathbf{D}^2$, alors :*

1- les groupes de cohomologie entière de X sont libres, nuls en dimensions 1, 3 et ≥ 5 ;

2- si A désigne le nombre d'arêtes du 1-squelette de Σ et I le nombre d'arêtes du 1-squelette de Σ intérieures au support de Σ, alors

$$\operatorname{rg} \mathrm{H}^0(X,\mathbb{Z}) = 1, \quad \operatorname{rg} \mathrm{H}^2(X,\mathbb{Z}) = A - 3 \ \text{ et } \ \operatorname{rg} \mathrm{H}^4(X,\mathbb{Z}) = I.$$

On a vu plus haut que la cohomologie des variétés *complètes* et la structure de leur anneau de Chow sont bien connues, exposées dans Danilov [20]. Les variétés qui vérifient les hypothèses du théorème 2.1 ne sont jamais complètes, puisque un éventail Σ définit une variété complète si, et seulement si sa section sphérique est $\mathbf{S}^2$.

La nullité des groupes de cohomologie impaire est assurée par l'hypothèse. En particulier, le théorème 2.1 s'applique à toute désingularisation crépante et équivariante d'une singularité-quotient $\mathbb{C}^3/G$ où G est n'importe quel sous-groupe abélien de $\mathrm{SL}(3,\mathbb{C})$. Cela apporte une réponse, développée dans [3], à la conjecture de McKay sur la cohomologie entière dans le cas des groupes abéliens, telle que M. Reid l'avait formulée dans [60].

2.1.5 On donne un dernier exemple de calcul qui utilise le théorème 2.1 : soit X une variété torique dont la section sphérique est la réunion des triangles dessinés à droite sur la figure 1. Elle est la réunion d'une variété torique Y dont la section sphérique – dessinée au centre de la figure 1 – vérifie les hypothèses du théorème 2.1 et d'une copie Z de $\mathbb{C}^3$. Par ailleurs, l'intersection de Y et de Z est un espace affine complexe de dimension trois privé de deux droites complexes sécantes : elle se rétracte par déformation sur la somme connexe de $\mathbf{S}^4$ et de deux copies de $\mathbf{S}^3$ ($\mathbf{S}^m$ désigne la sphère euclidienne de dimension m). Là encore, les groupes de cohomologie entière sont libres, tous nuls à l'exception de

$$\mathrm{H}^0(X,\mathbb{Z}) \simeq \mathbb{Z}, \quad \mathrm{H}^2(X,\mathbb{Z}) \simeq \mathbb{Z}^7, \quad \mathrm{H}^4(X,\mathbb{Z}) \simeq \mathbb{Z}^2, \quad \mathrm{H}^5(X,\mathbb{Z}) \simeq \mathbb{Z}.$$

A l'image de ce qui précède, le théorème 2.1 est source de nombreux exemples de calculs des groupes de cohomologie entière de variétés algébriques non complètes. Le calcul de la cohomologie d'une variété torique non complète arbitraire en fonction de la topologie de sa section sphérique et de la combinatoire de son 1-squelette reste une perspective. Enfin, une question : les groupes de cohomologie entière d'une variété torique complexe lisse sont-ils toujours sans torsion ?

2.2 Singularités-quotient de dimension trois

2.2.1 Le programme de classification birationnelle des variétés algébriques complexes de dimension trois amène naturellement à l'étude des singularités-quotient Gorenstein, germes (analytiques) à l'origine des quotients de $\mathbb{C}^3$ par l'action naturelle des sous-groupes finis de $\mathrm{SL}(3,\mathbb{C})$. Le cas de la dimension deux est bien connu : ces quotients sont les points doubles rationnels, qui font l'objet d'une très vaste littérature. La thèse de doctorat [1] fut l'une des premières études sur la dimension trois.

Soit G un sous-groupe fini de $\mathrm{SL}(3,\mathbb{C})$. On décrit ci-dessous comment est construite dans [1] une variété algébrique quasi-projective X_G qui soit un modèle à singularités toriques de $\mathbb{C}^3/G$.

• Si G est abélien et non trivial, la variété algébrique singulière $\mathbb{C}^3/G$ est une variété torique affine simpliciale, c'est-à-dire une variété torique affine définie par un cône engendré par trois vecteurs linéairement indépendants. On pose dans ce cas $X_G = \mathbb{C}^3/G$. Ces variétés sont bien comprises, notamment par le dictionnaire reliant les propriétés géométriques d'une variété torique aux propriétés combinatoires de son éventail.

• Si G est non abélien et stabilise une (nécessairement unique) droite de $\mathbb{C}^3$, on éclate cette droite stable. Cet éclatement est équivariant pour les actions de G ; la variété X_G est le quotient géométrique de cet éclaté.

• Si G ne stabilise aucun sous-espace propre de $\mathbb{C}^3$, on éclate l'origine de $\mathbb{C}^3$, puis les fibres de cet éclatement dont les groupes d'isotropie ne sont pas abéliens. Ces éclatements sont équivariants pour les actions de G ; la variété X_G est le quotient géométrique de cette suite d'éclatements.

Théorème 2.2 *Soit G un sous-groupe fini de $\mathrm{SL}(3,\mathbb{C})$. Dans chacun des cas décrits ci-dessus, le morphisme canonique $X_G \to \mathbb{C}^3/G$ est propre et birationnel ; c'est un isomorphisme au-dessus du lieu régulier de $\mathbb{C}^3/G$. En outre, les singularités de X_G sont toutes toriques simpliciales.*

Ce théorème est démontré dans [2] pour toutes les singularités-quotient de $\mathbb{C}^3/G$ par les sous-groupes finis de $\mathrm{GL}(3,\mathbb{C})$. La construction est analogue lorsque le groupe n'est pas dans $\mathrm{SL}(3,\mathbb{C})$. Il a depuis été généralisé en dimensions supérieures par F. Pan dans [55].

2.2.2 Dans Miller, Blichfeldt et Dickson [52] est établie une classification des sous-groupes finis de $\mathrm{SL}(3,\mathbb{C})$ à conjugaison près, héritée de travaux de géomètres du dix-neuvième siècle (F. Klein et H. Maschke, notamment). On en trouvera une synthèse – et la rectification d'une petite erreur – dans [1]. La liste des classes de conjugaison des sous-groupes finis de $\mathrm{SL}(3,\mathbb{C})$ est la suivante (l'action de $\mathrm{SL}(3,\mathbb{C})$ sur $\mathbb{P}^2_{\mathbb{C}}$ que l'on considère est l'action naturelle).

• Les groupes abéliens, qui fixent trois points non alignés de $\mathbb{P}^2_{\mathbb{C}}$, conjugués aux groupes finis de matrices diagonales spéciales unitaires.

• Les groupes non abéliens *réductibles*, qui fixent un point de $\mathbb{P}^2_{\mathbb{C}}$; en écriture matricielle par blocs, ce sont les conjugués des sous-groupes de $\mathrm{SL}(3,\mathbb{C})$ de la forme

$$\left\{ \begin{pmatrix} \det\gamma^{-1} & \\ & \gamma \end{pmatrix},\ \gamma \in \Gamma \right\}$$

où Γ est n'importe quel sous-groupe fini non abélien de $\mathrm{U}(2,\mathbb{C})$.

• Les groupes non réductibles *imprimitifs*, qui admettent une orbite formée de trois points dans $\mathbb{P}^2_{\mathbb{C}}$ (les groupes réels de rotations d'un tétraèdre régulier ou d'un cube appartiennent à cette classe) ; ce sont les sous-groupes de $\mathrm{SL}(3,\mathbb{C})$ conjugués à un produit semi-direct interne de la forme $D \rtimes T$ où D est un groupe fini de matrices diagonales

spéciales unitaires et où T est soit le groupe cyclique engendré par $t = \begin{pmatrix} & & 1 \\ 1 & & \\ & 1 & \end{pmatrix}$ soit le groupe diédral engendré par t et une matrice de la forme $\begin{pmatrix} -1 & & \\ & & a^{-1} \\ & a & \end{pmatrix}$ avec $|a| = 1$.

• Les groupes *primitifs*, dont toutes les orbites dans $\mathbb{P}^2_{\mathbb{C}}$ ont au moins quatre points. Ce sont les groupes conjugués à une liste de huit sous-groupes finis de $\mathrm{SU}(3,\mathbb{C})$ dont on trouvera une description dans [1]. Les notations ci-dessous sont empruntées à Miller, Blichfeldt et Dickson [52]. Brièvement, on a une chaîne $\mathcal{E} \lhd \mathcal{F} \lhd \mathcal{G}$ de groupes d'ordres respectifs 108, 216 et 648, le groupe $\mathcal{G}$ étant le groupe dit *Hessien* (voir Jordan [41], page 209). Suivent le groupe simple $\mathcal{H}$ d'ordre 60, groupe réel des rotations d'un icosaèdre régulier et son extension centrale $\mathcal{H}'$, d'ordre 180. On trouve ensuite une extension centrale $\mathcal{I}$ d'ordre 1080 du groupe des permutations alternées $\mathfrak{A}_6$ (voir Wiman [66]). Pour finir, viennent le groupe simple $\mathcal{J}$ d'ordre 168, groupe des automorphismes de la quartique de Klein (Klein [43]), et son extension centrale $\mathcal{J}'$ d'ordre 504.

La démonstration que cette liste est complète est due à G. A. Miller, H. F. Blichfeldt et L. E. Dickson ([52]), à coups de considérations arithmétiques sur les ordres et les traces des éléments de ces groupes, ainsi que sur leurs sous-groupes de Sylow. Une perspective : à l'image de la classification des sous-groupes finis de $\mathrm{SL}(2,\mathbb{C})$ issue du revêtement double $\mathrm{SU}(2,\mathbb{C}) \to \mathrm{SO}(3,\mathbb{R})$ et des polyèdres platoniciens, quelle géométrie doit-on considérer pour établir de manière unifiée la classification des sous-groupes finis de $\mathrm{SL}(3,\mathbb{C})$ (ou plus généralement, celle des sous-groupes finis de $\mathrm{SL}(n,\mathbb{C})$) ?

2.2.3 Les algèbres d'invariants des sous-groupes imprimitifs de $\mathrm{SL}(3,\mathbb{C})$ sont pour l'essentiel calculées par F. Klein ([44], [43]), H. Maschke ([51]), et A. Wiman ([66]) ; la liste exhaustive est complétée dans [1]. Ces calculs montrent en particulier que les quotients de $\mathbb{C}^3$ par ces groupes sont des intersections complètes. On utilise ce résultat ainsi qu'un théorème de M. Reid ([61]) pour établir dans [2] l'énoncé suivant, liste de toutes les singularités-quotient terminales de dimension trois.

Théorème 2.3 *Un sous-groupe fini de* $\mathrm{GL}(3,\mathbb{C})$ *sans pseudo-réflexion définit une singularité-quotient terminale si, et seulement s'il est conjugué à un groupe cyclique engendré par une matrice diagonale de la forme* $\mathrm{Diag}(\zeta_r, \zeta_r^{-1}, \zeta_r^a)$ *où* r *est un entier* ≥ 1, $\zeta_r = \exp(2i\pi/r)$ *et* $\mathrm{pgcd}(a,r) = 1$.

On trouvera dans l'article [46] de J. Kollar une présentation (peut-être devenue un peu ancienne) de la théorie de Mori. Le rôle joué par les singularités terminales dans ce programme de classification birationnelle des variétés algébriques de dimension 3 y est notamment exposé. On se référera naturellement aussi au travail de M. Reid sur le sujet ([60] et [61] par exemple).

2.2.4 Le calcul des singularités du modèle à singularités toriques simpliciales du théorème 2.2 passe, pour chaque sous-groupe fini G de $\mathrm{GL}(3,\mathbb{C})$, par la détermination des

points de $\mathbb{P}^2_{\mathbb{C}}$ à isotropie non triviale pour l'action naturelle de G. On appellera *configuration* de G la répartition de ces points en orbites sous l'action de G – ou de son image $\mathbb{P}G$ dans $\mathrm{PSL}(3,\mathbb{C})$ – ainsi que la donnée des classes de conjugaison des sous-groupes d'isotropie correspondants. Une *orbite exceptionnelle* est l'orbite d'un point fixe isolé d'un élément du groupe. Une *orbite semi-exceptionnelle* est une orbite de droite de points fixes d'un élément du groupe qui ne soit pas une homothétie – l'action de G sur les droites que l'on considère est l'action naturelle.

Les configurations des groupes primitifs est donnée dans [2]. Cependant les preuves n'y sont pas exposées ; ce parti avait été pris au nom de la concision de l'article. A défaut d'une meilleure interprétation géométrique de ces groupes, ces calculs ont été menés pour l'essentiel à l'aide des théorèmes de Sylow et de propriétés combinatoires relatives à l'action de sous-groupes. A titre d'exemple, on trouvera au paragraphe suivant une preuve du calcul de la configuration du groupe $\mathcal{I}$ de Wiman.

2.3 Un exemple : configuration du sous-groupe $\mathbb{P}\mathcal{I}$ de Wiman

On développe dans ce paragraphe le calcul de la configuration du groupe $\mathbb{P}\mathcal{I}$, qui contient le groupe $\mathbb{P}\mathcal{H}$ de l'icosaèdre. A. Wiman dit de ce groupe que *sa véritable condition est de contenir deux systèmes "égaux en droits"*[2] *de six groupes icosaédraux*. En termes plus contemporains, $\mathbb{P}\mathcal{I}$ est isomorphe à $\mathfrak{A}_6$ et contient de ce fait douze sous-groupes isomorphes à $\mathfrak{A}_5$ répartis en deux classes de conjugaison qui sont les deux systèmes dont parle Wiman. Ces deux classes sont échangées par les automorphismes non intérieurs du groupe.

2.3.1 On note respectivement μ et μ' les racines (réelles) du polynôme X^2+X-1, $e_6=\exp(2i\pi/6)$, et

$$t=\begin{pmatrix} & & 1\\ 1 & & \\ & 1 & \end{pmatrix},\ s=\begin{pmatrix} 1 & & \\ & -1 & \\ & & -1\end{pmatrix},\ m=\frac{1}{2}\begin{pmatrix} -1 & \mu' & \mu\\ \mu' & \mu & -1\\ \mu & -1 & \mu'\end{pmatrix},\ n=\begin{pmatrix} -1 & & \\ & & e_6\\ & 1/e_6 & \end{pmatrix}$$

les classes dans $\mathrm{PSL}(3,\mathbb{C})$ d'un système de générateurs de $\mathbb{P}\mathcal{I}$. Le calcul de ces matrices est essentiellement du à A. Wiman ([66]). Le groupe $\mathbb{P}\mathcal{I}$ est simple, isomorphe au groupe alterné $\mathfrak{A}_6$. Un tel isomorphisme est fourni par l'action de $\mathbb{P}\mathcal{I}$ par conjugaison sur les six conjugués de son sous-groupe $\mathbb{P}\mathcal{H}=\langle s,t,m\rangle\simeq\mathfrak{A}_5$ (voir 2.2.2), qui constituent l'un de ses deux systèmes de six sous-groupes icosaédraux ; on considérera, parmi ces isomorphismes, celui qui est défini par

$$\begin{aligned} t &\mapsto (123)\\ s &\mapsto (12)(34)\\ m &\mapsto (12)(45)\\ n &\mapsto (12)(56). \end{aligned} \tag{1}$$

2.3.2 Cet isomorphisme permet immédiatement de déterminer les classes de conjugaison dans $\mathbb{P}\mathcal{I}$: outre la classe de l'unité, on trouve la classe de s contenant les 45 éléments

[2]Dans le texte, (...) *zwei Systeme von je sechs gleichberechtigten Ikosaedergruppen* (...).

d'ordre 2, les classes de t et de tmn contenant chacune 40 éléments d'ordre 3, la classe de stn formée des 90 éléments d'ordre 4 et enfin les classes de tsm et de son carré, formées chacune de 72 éléments d'ordre 5.

2.3.3 On déduit de cette liste qu'il n'y a qu'une seule orbite semi-exceptionnelle, celle de la droite des points fixes de s. En effet, seuls les éléments d'ordre 2 sont des homologies – une *homologie* est, selon le vocabulaire de Miller, Blichfeldt et Dickson [52], une homographie de $\mathbb{P}^2_{\mathbb{C}}$ admettant une droite de points fixes.

• L'orbite sous $\mathbb{P}\mathcal{H}$ du point fixe $p_1 = (1 : 0 : 0)$ de s est réelle et contient 15 éléments ; c'est l'orbite dans $\mathbb{P}^2_{\mathbb{C}}$ des droites des milieux des arêtes de l'icosaèdre. Par ailleurs, le groupe d'isotropie de p_1 dans $\mathbb{P}\mathcal{I}$ contient le centralisateur de s qui est d'ordre 8, isomorphe à l'unique extension non centrale $(\mathbb{Z}/2\mathbb{Z})^2 \rtimes \mathbb{Z}/2\mathbb{Z}$ (le voir dans $\mathfrak{A}_6$). Puisqu'elle n'est pas formée que de points à coordonnées réelles, l'orbite de p_1 contient donc 45 points, qui sont les points fixes isolés des éléments d'ordre 2. Enfin, le groupe d'isotropie de p_1 est le centralisateur de s, sous-groupe engendré par tst^{-1} et n dont la structure est décrite ci-dessus.

• L'orbite du point fixe $p_2 = (1 : 1 : 1)$ de t contient au plus 60 éléments, puisque le sous-groupe $\langle t, m\rangle$, isomorphe à $\mathfrak{S}_3$, stabilise p_2. Comme les points fixes de t sont tous dans la même orbite (ce sont p_2, np_2 et $tmnp_2$), les points fixes des conjugués de t sont tous dans l'orbite de p_2. Chacun des dix 3-sous-groupes de Sylow – en abrégé, on dira 3-Sylow – de $\mathbb{P}\mathcal{I}$ contient quatre conjugués de t, qui fixent six points de $\mathbb{P}^2_{\mathbb{C}}$. Ces 3-Sylow sont conjugués dans $\mathrm{PSL}(3, \mathbb{C})$ au groupe non cyclique d'ordre 9 engendré par t et l'homographie $\mathrm{Diag}(1, e_6^2, e_6^4)$.

Par ailleurs, tout 3-Sylow du groupe d'isotropie $\mathbb{P}\mathcal{I}_{p_2}$ de p_2 est d'ordre 3. En effet, l'intersection de deux 3-Sylow de $\mathbb{P}\mathcal{I}$ est triviale (le voir dans $\mathfrak{A}_6$) et le 3-Sylow de $\mathbb{P}\mathcal{I}$ contenant t – à savoir $\langle t, tmn\rangle$ – n'est pas inclus dans $\mathbb{P}\mathcal{I}_{p_2}$. Ainsi, si le groupe d'isotropie de p_2 contenait deux 3-Sylow distincts, il contiendrait m, t et un élément d'ordre 3 qui ne commute pas avec t ; on montre aisément du côté de $\mathfrak{A}_6$ que cela imposerait qu'il contienne une copie de $\mathfrak{A}_5$ ce qui n'est pas puisque l'orbite de p_2 contient au moins les dix points provenant des centres des faces de l'icosaèdre, orbite de p_2 sous l'action de $\mathbb{P}\mathcal{H}$.

Ainsi, on vient de montrer que le stabilisateur de p_2 contient un unique 3-Sylow, engendré par t. Cela implique que si deux éléments d'ordre 3 conjugués à t dans $\mathbb{P}\mathcal{I}$ fixent un même point de $\mathbb{P}^2_{\mathbb{C}}$, ils engendrent le même sous-groupe. En particulier, ils sont dans le même 3-Sylow de $\mathbb{P}\mathcal{I}$. Avec la conclusion de l'avant-dernier paragraphe, on en déduit que l'orbite de p_2 a 60 points, qui sont les points fixes des conjugués de t. En outre, le groupe d'isotropie de t est engendré par t et m (voir ci-dessus).

• La situation est analogue pour l'orbite des points fixes de tmn : elle est constituée des 60 points fixes des conjugués de tmn, dont les groupes d'isotropie sont des copies de $\mathfrak{S}_3$, tous conjugués à $\langle tmn, n\rangle$.

• Les orbites de points fixes des éléments d'ordre 5 sont celles des points fixes de tsm. Soit p un tel point ; on note $\mathbb{P}\mathcal{I}_p$ (respectivement $\mathbb{P}\mathcal{H}_p$) son groupe d'isotropie sous l'action de $\mathbb{P}\mathcal{I}$ (resp. $\mathbb{P}\mathcal{H}$). On va montrer que $\mathbb{P}\mathcal{I}_p = \mathbb{P}\mathcal{H}_p$.

Supposons que $g \in \mathbb{P}\mathcal{I}_p \setminus \mathbb{P}\mathcal{H}_p$. Vu du côté de $\mathfrak{A}_6$, quitte à renuméroter, cela signifie que $\mathbb{P}\mathcal{I}_p$ contient le 5-cycle (12345) et une permutation g qui ne fixe pas 6. Alors, $\mathbb{P}\mathcal{I}_p$ contient les deux 5-Sylow engendrés respectivement par (12345) et $g(12345)g^{-1}$; donc il contient au moins six 5-Sylow. Cela entraîne que son ordre est ≥ 25. Ainsi, cet ordre est-il un diviseur ≥ 30 de 360. Le nombre 40 est impossible car un groupe d'ordre 40 n'a qu'un seul 5-Sylow. Enfin, aucun élément d'ordre 3 de $\mathbb{P}\mathcal{I}$ ne fixe p, puisqu'on a montré que les groupes d'isotropie des points fixes des éléments d'ordre 3 ne contiennent pas d'élément d'ordre 5 (ils ont 6 éléments) ; les diviseurs de 360 qui sont ≥ 30 et $\neq 40$ sont donc également interdits, puisqu'ils sont tous multiples de 3. L'hypothèse $\mathbb{P}\mathcal{I}_p \neq \mathbb{P}\mathcal{H}_p$ ne tient pas.

Ainsi, les points fixes des éléments d'ordre 5 sont-ils répartis en deux orbites. La première provient des sommets de l'icosaèdre (c'est celle de $p_3 = (\mu : 0 : 1)$) ; elle a 36 éléments et son groupe d'isotropie est conjugué au groupe diédral d'ordre 10. La seconde est celle des points fixes non réels de tsm ; elle contient 72 éléments et son groupe d'isotropie est d'ordre 5.

• On note r le nombre d'homologies de $\mathbb{P}\mathcal{I}$, et s le nombre d'orbites exceptionnelles. Si ω est une orbite exceptionnelle, on note aussi c_ω l'ordre de son groupe d'isotropie et r_ω le nombre d'homologies de ce groupe (identité comprise). En calculant le cardinal de

$$\big\{(g,p) \in \mathbb{P}\mathcal{I} \times \mathbb{P}^2_{\mathbb{C}},\ g.p = p \ \text{ et } g \ \text{ n'est pas une homologie}\big\}$$

de deux manières, selon la première ou la seconde composante du produit, on obtient la formule

$$3\left(1 - \frac{r}{|\mathbb{P}\mathcal{I}|}\right) = s - \sum_{\substack{\omega \text{ orbite} \\ \text{exceptionnelle}}} \frac{r_\omega}{c_\omega}.$$

Cela impose que $s \geq 6$. Par ailleurs, les points fixes de $tst^{-1}n$ – qui est d'ordre 4 – sont p_1, $p_4 = (0 : i : e_6)$ et $tst^{-1}p_4$, d'où $s \leq 6$. On a donc une dernière orbite exceptionnelle de points fixes des éléments d'ordre 4, celle de p_4. La formule ci-dessus impose que son quotient r_ω/c_ω vaille 1/2. Comme aucun élément d'ordre 3 ou 5 ne fixe p_4 (les groupes d'isotropie des points fixes de ces éléments sont calculés), $c_\omega \in \{4, 8\}$. Mais si $c_\omega = 8$, le groupe d'isotropie correspondant est un 2-Sylow de $\mathfrak{A}_6$, qui contient 5 éléments d'ordre 2 (c'est un groupe diédral), imposant $r_\omega = 6$. Donc $c_\omega = 4$: l'orbite de p_4 contient 90 points, et son groupe d'isotropie est cyclique d'ordre 4, engendré par $tst^{-1}n$.

• Pour conclure, la configuration dans $\mathbb{P}^2_{\mathbb{C}}$ du groupe $\mathbb{P}\mathcal{I}$ est la suivante : une orbite semi-exceptionnelle dont le groupe d'isotropie est d'ordre 2, et six orbites exceptionnelles dont les groupes d'isotropie sont respectivement d'ordres 8 (diédral), 6 (diédral aussi), encore 6 (*idem*), 4 (cyclique), 5 et 10 (diédral).

2.3.4 Les positions ensemblistes relatives des composantes irréductibles du lieu singulier du modèle à singularités toriques $X_{\mathcal{I}}$ (notations du théorème 2.2) sont exposées dans [1]. On y trouvera également le calcul explicite de ces singularités toriques, qui résulte du calcul présenté ci-dessus. Ce travail y a été fait pour tous les sous-groupes primitifs de $\mathrm{SL}(3, \mathbb{C})$.

3 Une méthode hybride en combinatoire analytique

La démarche de la combinatoire analytique, présentée de manière unifiée dans Flajolet et Sedgewick [28] consiste *grosso modo* à calculer la fonction génératrice d'une structure combinatoire *via* la méthode symbolique puis à déduire de ses singularités dominantes une asymptotique précise de ses coefficients de Taylor à l'origine. Les méthodes usuelles d'extraction des coefficients sont pour l'essentiel l'analyse des singularités, les techniques de points cols, la méthode de Darboux et des procédés tauberiens.

Il arrive parfois que toutes ces méthodes échouent. C'est le cas par exemple des permutations qui admettent des racines $m^{\text{ièmes}}$ dans le groupe symétrique $\mathfrak{S}_n$, dont la série génératrice à m donné présente une frontière naturelle sans que ses singularités ne relèvent de la méthode du col. On peut trouver un équivalent de ses coefficients par analyse détaillée de la fonction. Mais cette dernière a une forme particulière : c'est un produit infini qui présente une hiérarchie de singularités qui toutes, prises séparément, relèvent de l'analyse des singularités. On montre comment une hybridation de la méthode de Darboux et de l'analyse des singularités permet de fournir à toute une classe d'objets combinatoires contenant ces permutations un développement asymptotique à un ordre arbitraire.

3.1 Combien de permutations admettent-elles une racine $m^{\text{ième}}$?

La question du nombre de permutations qui admettent une racine carrée dans le groupe symétrique $\mathfrak{S}_n$ est un classique de l'analyse combinatoire, comme en atteste par exemple le livre *Generatingfunctionology* [65] de H. S. Wilf. Ce problème admet une généralisation immédiate aux puissances $m^{\text{ièmes}}$, $m \geq 2$.

3.1.1 Le sous-ensemble de $\mathfrak{S}_n$ formé des puissances $m^{\text{ièmes}}$ est stable par conjugaison. Par conséquent, une permutation de $\mathfrak{S}_n$ étant donnée, la propriété d'appartenir à ce sous-ensemble se lit sur la partition de n déterminée par les longueurs des cycles de sa décomposition canonique en produit de cycles à supports disjoints. Pour tout entier naturel k, on note

$$k^{\infty} \wedge m = \lim_{t \to +\infty} \operatorname{pgcd}(k^t, m).$$

L'arithmétique élémentaire montre alors qu'une permutation w est une puissance $m^{\text{ième}}$ si, et seulement si pour tout entier naturel k, le nombre de cycles de longueur k de w est un multiple de $k^\infty \wedge m$. Cette assertion dont on trouvera une preuve dans [4], était auparavant établie par P. Turán ([64]) lorsque m est un nombre premier et par H. S. Wilf ([65]) dans le cas général.

3.1.2 Au moyen de cette caractérisation, la méthode symbolique fournit immédiatement la fonction génératrice exponentielle P_m du nombre de permutations qui admettent une racine $m^{\text{ième}}$. Dans leur best-seller impatiemment attendu [28], P. Flajolet et R. Sedgewick offrent un développement approfondi de la méthode symbolique en analyse combinatoire, auquel on pourra se référer. Pour tout entier naturel non nul d, soit e_d la serie formelle (ou la fonction entière)

$$e_d(z) = \sum_{n\geq 0} \frac{z^{dn}}{(dn)!} = \frac{1}{d}\sum_{k=0}^{d-1} \exp\left(e^{2ik\pi/d} z\right).$$

On note $p_n(m)$ la probabilité qu'une permutation uniforme de $\mathfrak{S}_n$ soit une puissance $m^{\text{ième}}$. Alors,

$$P_m(z) = \sum_{n\geq 0} p_n(m) z^n = \prod_{k=1}^{+\infty} e_{k^\infty \wedge m}\left(\frac{z^k}{k}\right).$$

Ce produit se scinde naturellement en deux termes $P_m = C_m R_m$. Le premier facteur C_m est la fonction génératrice exponentielle des permutations qui sont produits de cycles dont les orbites ont toutes un cardinal premier avec m. C'est la série algébrique

$$C_m(z) = \prod_{\substack{k\geq 1 \\ \text{pgcd}(k,m)=1}} \exp\left(\frac{z^k}{k}\right) = \prod_{k|m} \left(1-z^k\right)^{-\mu(k)/k},$$

où μ désigne la fonction de Möbius. Le second facteur est la fonction génératrice exponentielle des autres puissances $m^{\text{ièmes}}$, celles dont les longueurs des cycles ont toutes un facteur commun non trivial avec m. C'est la série transcendante

$$R_m(z) = \prod_{\substack{k\geq 1 \\ \text{pgcd}(k,m)\neq 1}} e_{k^\infty \wedge m}\left(\frac{z^k}{k}\right). \tag{2}$$

On trouvera davantage de développements sur ces différentes expressions dans [4] et dans Wilf [65].

3.1.3 Le cercle trigonométrique est une frontière naturelle pour la fonction holomorphe R_m, *i.e.* les points singuliers de R_m constituent un sous-ensemble dense de son cercle de convergence[3]. En effet, dans le cas des racines carrées ($m = 2$), en réorganisant le produit

[3]Cette propriété démontre la transcendance de R_m. Voir le livre de P. Flajolet et R. Sedgewick [28] pour un exposé sur l'analyse des singularités des fonctions algébriques.

infini après intégration de la dérivée logarithmique du cosinus hyperbolique, on démontre aisément l'égalité suivante, valide dans le disque ouvert :

$$R_2(z) = \prod_{k\geq 1} \cosh\left(\frac{z^{2k}}{2k}\right) = \exp\left(\sum_{n\geq 1} \frac{(-1)^{n-1}\tau_{n-1}}{n2^{2n+1}} \mathrm{Li}_{2n}(z^{4n})\right), \tag{3}$$

où $\mathrm{Li}_\nu(z) = \sum_{k\geq 1} z^k/k^\nu$ désigne le $\nu^{\text{ième}}$ polylogarithme usuel et où les τ_k sont les coefficients (strictement positifs) de la série tangente $\tan(z) = \sum_{k\geq 0} \tau_k z^{2k+1}$. Comme tout polylogarithme est singulier en 1, cette expression montre immédiatement que toutes les racines $4n^{\text{ièmes}}$ de l'unité sont des points singuliers de R_2.

Dans le cas général, une telle formule peut également être établie, qui permet de déterminer les singularités de R_m. On donne très rapidement ce résultat non publié. On note $\tau_{d,k}$ les coefficients de Taylor de la dérivée logarithmique de e_d, définis par $e_d'/e_d = \sum_{k\geq 1} \tau_{d,k-1} z^{dk-1}$. On note également $\mathcal{D}S(m) = \{k^\infty \wedge m,\ k \geq 1\} \setminus \{1\}$ l'ensemble des *diviseurs saturés* de m et $q(t)$ le radical sans carré du nombre entier t, produit des facteurs premiers distincts de t. Alors, après réorganisation du produit infini (2), il vient

$$R_m(z) = \exp\left\{ \sum_{n\geq 1} \frac{1}{n} \sum_{\substack{(t,d) \\ t\in \mathcal{D}S(m),\ d\mid \frac{m}{t}}} \mu(d) \frac{\tau_{t,n-1}}{tq(t)^{nt} d^{nt}} \mathrm{Li}_{nt}\big(z^{ntdq(t)}\big) \right\}. \tag{4}$$

3.1.4 Puisque la fonction holomorphe P_m admet une frontière naturelle, l'analyse des singularités échoue dans la détermination de l'asymptotique de ses coefficients. Une approche tauberienne permet de trouver un équivalent sans terme d'erreur pour le cas des racines carrées ($m = 2$, Bender [14] et Blum [16]), mais la preuve s'appuie sur la décroissance des $p_n(2)$ qui ne s'étend pas au cas général. Enfin, la croissance des fonctions R_m est trop modérée au voisinage du cercle de convergence pour que l'asymptotique de ses coefficients puisse être atteinte par la méthode du col.

Par une analyse détaillée des suites des coefficients de C_m et R_m, on calcule dans [4] un équivalent de la suite $p_n(m)$ lorsque n tend vers $+\infty$.

Théorème 3.1 *Si m est un entier naturel non nul, la probabilité $p_n(m)$ qu'une permutation de $\mathfrak{S}_n$ admette une racine carrée vérifie*

$$p_n(m) \underset{n\to+\infty}{\sim} \frac{\pi_m}{n^{1-\frac{\varphi(m)}{m}}}$$

où φ est la fonction d'Euler et π_m la constante

$$\pi_m = \frac{R_m(1)}{\Gamma\left(\frac{\varphi(m)}{m}\right)} \prod_{k\mid m} k^{-\frac{\mu(k)}{k}}.$$

La constante $R_m(1)$ peut se calculer *via* le produit (2) qui converge au point $z = 1$, mais la convergence de la série (4) en 1 est beaucoup plus rapide. La preuve de [4] approche les coefficients de C_m par sa partie principale au voisinage de 1, qui est proportionnelle à $(1-z)^{-\varphi(m)/m}$. Une étude détaillée de la monotonie et de la comparaison des suites en présence permet de conclure au moyen de méthodes classiques de l'analyse.

3.2 Darboux *et* analyse des singularités : une méthode hybride

La formule (3), on l'a vu, montre que les racines $4n^{\text{ièmes}}$ de l'unité sont des singularités de la fonction génératrice R_2. Par ailleurs, 1 est l'unique point singulier du polylogarithme Li_ν. Ce dernier admet un développement asymptotique complet au voisinage de 1 dans $\mathbb{C} \setminus [1, +\infty[$ dans l'échelle $(1-z)^a \log^b \frac{1}{1-z}$. La partie singulière principale de ce développement est $\frac{(-1)^\nu}{(\nu-1)!}(1-z)^{\nu-1} \log \frac{1}{1-z}$ lorsque ν est entier naturel non nul. Cette propriété fait apparaître dans la formule (3) une hiérarchie des singularités de R_2 par ordre décroissant.

C'est à la fois l'existence d'une telle hiérarchie de singularités relevant chacune de l'analyse des singularités et l'ordre de dérivabilité des polylogarithmes en 1 qui sont à l'origine de la méthode hybride développée dans [8].

La méthode symbolique amène fréquemment à exprimer les fonctions génératrices d'objets combinatoires sous la forme d'un produit infini. Le traitement de l'asymptotique de ses coefficients de Taylor diffère selon la nature de son terme général, notamment dans les cas de frontière naturelle. Un tableau de cas typiques est présenté à la fin de [8].

3.2.1 Si a est un réel négatif ou nul, une fonction f, holomorphe dans le disque trigonométrique ouvert D, est dite d'*ordre global* a lorsque $f(z)/(1-|z|)^a$ est bornée sur D. Par exemple, la fonction génératrice exponentielle des permutations admettant une racine carrée dans le groupe symétrique

$$\sqrt{\frac{1+z}{1-z}} \prod_{k \geq 1} \cosh\left(\frac{z^{2k}}{2k}\right)$$

est d'ordre global $-1/2$.

Si s est un entier positif ou nul, une fonction f, holomorphe dans D est dite *de classe $\mathcal{C}^s$ sur le disque fermé $\overline{D}$* lorsqu'elle est s fois continûment différentiable sur $\overline{D}$ pour la topologie induite de $\mathbb{R}^2$, c'est-à-dire lorsque ses dérivées complexes $f^{(k)}$, $0 \leq k \leq s$ sont prolongeables par continuité sur $\overline{D}$. Le polylogarithme Li_ν est de classe $\mathcal{C}^{\nu-2}$ dès que $\nu \geq 2$; ainsi, pour tout entier $N \geq 1$, le reste

$$\exp\left(\sum_{n \geq N} \frac{(-1)^{n-1}\tau_{n-1}}{n 2^{2n+1}} \mathrm{Li}_{2n}(z^{4n})\right)$$

du troisième terme de la formule (3) est-il de classe $\mathcal{C}^{2N-2}$.

3.2.2 Si Z est une partie finie du cercle unité, on dira qu'une fonction f, holomorphe dans D, admet un *développement log-puissance de classe $\mathcal{C}^s$ en Z* lorsqu'il existe une combinaison linéaire Σ de fonctions $(1-z/\zeta)^a \log^b \frac{1}{1-z/\zeta}$, $a \in \mathbb{R}$, $b \in \mathbb{N}$, $\zeta \in Z$ telle que $f - \Sigma$ soit de classe $\mathcal{C}^s$ sur $\overline{D}$.

La *méthode de Darboux* pour le calcul de l'asymptotique des coefficients de Taylor d'une fonction holomorphe s'appuie sur le résultat suivant de l'analyse élémentaire : si une fonction f, holomorphe dans D, est de classe $\mathcal{C}^s$ sur $\overline{D}$, alors son $n^{\text{ième}}$ coefficient vérifie $[z^n]f = o(n^{-s})$. Ainsi, si f admet un développement log-puissance Σ de classe $\mathcal{C}^s$ en une partie finie Z du cercle, $[z^n]f = [z^n]\Sigma + o(n^{-s})$. Par ailleurs, on dispose d'un développement complet dans l'échelle $n^a \log^b n$ des coefficients des fonctions log-puissances, dont on trouvera le détail dans Flajolet et Sedgewick [28]. Ainsi, cette proposition permet-elle, sous les hypothèses requises, de développer $[z^n]f$ dans l'échelle $\zeta^{-n} n^a \log^b n$, $\zeta \in Z$.

3.2.3 Une autre approche pour le calcul de telles asymptotiques est l'*analyse des singularités*. S'appuyant elle aussi sur l'analyse élémentaire des fonctions holomorphes, elle a été développée à l'origine par P. Flajolet et A. M. Odlyzko ([27]) et s'impose comme un outil essentiel de l'analyse d'algorithmes. Son principe de base est un théorème de transfert qui assure que si une fonction f, holomorphe dans D, singulière en 1, se prolonge analytiquement sur un voisinage du disque fermé édenté en 1[4] et vérifie $f(z) = O(1-z)^a$ lorsque z tend vers 1 dans ce voisinage édenté, alors $[z^n]f(z) = O(n^{-1-a})$ lorsque n tend vers $+\infty$. Un tel théorème de transfert admet une version "petit o" et s'étend à la comparaison avec les fonctions $(1-z)^a \log^b \frac{1}{1-z}$.

Soit f une fonction holomorphe sur D, dont les singularités sur le cercle constituent un ensemble fini Z. Si t est un nombre réel, on dit que f admet un *développement log-puissance de type $\mathcal{O}^t$ en Z* lorsque sont satisfaites les deux conditions : (i) f est analytique dans un voisinage édenté en Z du disque, c'est-à-dire sur un ouvert de la forme $\mathcal{D} = \bigcap_{\zeta \in Z} \zeta.\Delta$ où Δ est un voisinage du disque édenté en 1 ; (ii) il existe une fonction log-puissance $\Sigma(z) = \sum_{\zeta \in Z} \sigma_\zeta(z/\zeta)$ où chaque σ_ζ est une combinaison linéaire de fonctions $(1-z)^a \log^b \frac{1}{1-z}$, $a \in \mathbb{R}$, $b \in \mathbb{N}$, telle que, pour tout $\zeta \in Z$, on ait $f(z) - \sigma_\zeta(z/\zeta) = O(1-z/\zeta)^t$ lorsque z tend vers ζ dans $\mathcal{D}$.

Par la vertu du théorème de transfert évoqué plus haut, si f est holomorphe sur D, n'a que des singularités isolées et admet un développement log-puissance Σ de type $\mathcal{O}^t$ en la partie finie Z du cercle, son coefficient de Taylor vérifie $[z^n]f(z) = [z^n]\Sigma(z) + O(n^{-1-t})$ lorsque n tend vers l'infini. Là encore, le développement asymptotique des coefficients des fonctions log-puissances dans l'échelle $n^a \log^b n$ permet, sous ces hypothèses, de développer $[z^n]f(z)$ dans l'échelle $\zeta^{-n} n^a \log^b n$, $\zeta \in Z$.

3.2.4 On trouvera dans Flajolet et Sedgewick [28] une brève étude comparative de la méthode de Darboux et de l'analyse des singularités. L'analyse combinatoire trouve mieux son compte dans la seconde car y interviennent souvent des fonctions non bornées

[4] *i.e.* sur un ouvert de la forme $\Delta = \{z, |z| < 1+\delta,\ z \neq 1,\ \theta < \arg(z-1) < 2\pi - \theta\}$ où $\delta > 0$ et $0 < \theta < 2\pi$.

au voisinage de leurs singularités ; cependant, certaines fonctions ayant une frontière naturelle comme $\sum_{n\geq 0} z^{2^n}/2^{nr}$, $r \geq 1$, sont du ressort de la méthode de Darboux et pas de l'analyse des singularités (même si, dans le cas présent, la question de l'asymptotique des coefficients n'est pas du premier intérêt...).

A l'instar de la fonction génératrice des permutations admettant une racine $m^{\text{ième}}$, certaines fonctions de l'analyse combinatoire apparaissent comme des produits infinis admettant une frontière naturelle, qui ne relèvent d'aucune des deux méthodes prise isolément. On en trouvera quelques exemples ci-dessous, tous tirés de [8]. Il apparaît que, lorsque les singularités d'une telle fonction ne sont pas trop violentes comme celles, par exemple, de la fonction génératrice des partitions qui relève de la méthode du col, la méthode de Darboux et l'analyse des singularités peuvent se combiner et fournir le développement asymptotique cherché. C'est l'objet de l'article [8] qui développe une *méthode* dite *hybride*. Elle fournit à la fois un théorème d'existence et un algorithme de calcul du développement asymptotique complet des coefficients de Taylor de toute fonction à laquelle elle s'applique.

Le principe théorique sur lequel la méthode hybride s'appuie est le suivant.

Théorème 3.2 *Soit f une fonction holomorphe dans le disque unité ouvert D. On suppose que f se factorise en un produit $f = PQ$ où P et Q sont des fonctions holomorphes dans D et vérifient :*

(i) Q est de classe $\mathcal{C}^s$ sur le disque fermé $\overline{D}$, $s \in \mathbb{N}$;

(ii) P est d'ordre $a \leq 0$ et admet un développement log-puissance Σ de classe $\mathcal{C}^t$, $t \in \mathbb{N}$ en une partie finie Z du cercle ∂D ;

(iii) $t \geq u$ où $u = \left\lfloor \frac{s+\lfloor a \rfloor}{2} \right\rfloor$ est supposé ≥ 0.

Alors, si $c = \left\lfloor \frac{s-\lfloor a \rfloor}{2} \right\rfloor$ et si H désigne le polynôme d'interpolation de Hermite dont les dérivées d'ordre $\leq c-1$ coïncident avec celles de Q en tous les point de Z,

$$[z^n]f = [z^n]\Sigma H + o(n^{-u})$$

lorsque n tend vers l'infini.

L'intérêt de cet énoncé réside notamment dans le fait que le produit ΣH est une fonction log-puissance dont les coefficients de Taylor sont développés à tout ordre. L'hybridation apparaît dans le théorème 3.3, qui constitue la méthode proprement dite. Son énoncé, cependant, nécessite une dernière définition.

Soient f une fonction holomorphe dans D, t un nombre réel et ζ un point du cercle. Lorsqu'il existe, le *développement radial à l'ordre t de f en ζ* est la plus petite (en terme de nombre de monômes) combinaison linéaire σ de fonctions $(1-z/\zeta)^a \log^b \frac{1}{1-z/\zeta}$, $a \in \mathbb{R}$, $b \in \mathbb{N}$ telle que $f(z) = \sigma(z) + O(1-z/\zeta)^t$ lorsque z tend radialement vers ζ dans D (*i.e.* pour $z = (1-x)\zeta$ et x tend vers 0 dans $]0,1]$). Dans ces conditions, on note $\sigma = \text{asympt}(f, \zeta, t)$.

S'il est en général difficile d'évaluer le comportement d'une fonction holomorphe le long de son cercle de convergence, l'analyticité dans le disque ouvert permet en revanche de calculer, souvent sans peine, de tels développements radiaux.

Théorème 3.3 *Soit f une fonction holomorphe dans le disque unité ouvert D. On suppose que f se factorise en un produit $f = PQ$ où P et Q sont des fonctions holomorphes dans D et vérifient :*

(i) Q est de classe $\mathcal{C}^s$ sur le disque fermé $\overline{D}$, $s \in \mathbb{N}$;

(ii) P est d'ordre global $a \leq 0$ et admet un développement log-puissance de type $\mathcal{O}^t$ en une partie finie Z du cercle, où $t \in \mathbb{R}_+$;

(iii) $t > u$ où $u = \left\lfloor \frac{s+|a|}{2} \right\rfloor$ est supposé ≥ 0.

Alors, f admet un développement radial à l'ordre u en chaque point de Z et

$$[z^n]f = [z^n]A + o(n^{-u}) \text{ où } \quad A = \sum_{\zeta \in Z} \operatorname{asympt}(f, \zeta, u).$$

En d'autres termes, dans la situation d'un produit $f = PQ$ où le facteur de Darboux Q est assez régulier sur le disque fermé et où le facteur singulier P satisfait les hypothèses de l'analyse des singularités à un ordre suffisant, le développement asymptotique des coefficients de Taylor de f peut se faire, jusqu'à un certain ordre, comme si Q était analytique au voisinage du disque fermé, c'est-à-dire comme si f elle-même vérifiait les hypothèses de l'analyse des singularités. En particulier, l'asymptotique se déduit d'une analyse séparée d'un nombre fini de singularités.

3.2.5 A titre d'exemple, la situation de la fonction génératrice exponentielle des carrés dans les groupes symétriques est pleinement du ressort du théorème 3.3. Selon l'ordre de développement que l'on cherche à atteindre, sachant que Li_{2n} est de classe $\mathcal{C}^{2n-2}$ sur le disque fermé, on découpera le produit

$$P_2(z) = \sqrt{\frac{1+z}{1-z}} \prod_{n \geq 1} \exp\left(\frac{(-1)^{n-1}\tau_{n-1}}{n 2^{2n+1}} \mathrm{Li}_{2n}(z^{4n}) \right)$$

en un facteur singulier, produit de la racine carrée et d'un nombre fini de premiers termes du produit infini, et un facteur de Darboux constitué du reste correspondant du produit infini. Le théorème 3.3 fournit alors l'existence d'un développement asymptotique complet de la probabilité $p_n(2)$ qu'une permutation de $\mathfrak{S}_n$ soit un carré ; le début de ce développement est, après calcul,

$$p_n(2) = \sqrt{\frac{2}{\pi n}} e^G \left[1 - \frac{\log n}{n} + \frac{c_3 + (-1)^n}{4n} \right] - 2e^G \frac{(-1)^{\lfloor n/2 \rfloor}}{n^2} + O\left(\frac{\log n}{n^{5/2}} \right) \tag{5}$$

où les constantes G et c_3 sont données par les formules

$$e^G = \prod_{k \geq 1} \cosh\left(\frac{1}{2k} \right) = \exp \sum_{n \geq 1} \frac{(-1)^{n-1}\tau_{n-1}}{n 2^{2n+1}} \zeta(2n)$$

– la convergence de la série est rapide et permet une approximation numérique efficace – et

$$c_3 = -12 + 16 \log 2 + 4\gamma + 2c_2 \text{ avec } \quad c_2 = \sum_{k \geq 1} \left(\frac{1}{2k} - \tanh \frac{1}{2k} \right) \ ;$$

dans ces formules, ζ est la fonction de Riemann et γ la constante d'Euler. On notera, dans la formule (5), les termes périodiques en n de périodes 2 et 4, provenant de la contribution des points singuliers -1 et $\sqrt{-1}$ de P_2. Un développement à un ordre plus élevé ferait apparaître des périodes 8, 12, 16... issues des racines $4m^{\text{ièmes}}$ de l'unité qui sont les points singuliers de R_2.

3.2.6 Dans l'article [8], d'autres exemples d'applications de la méthode hybride sont présentés. Dans tous ces cas, les problèmes combinatoires aboutissent à des fonctions génératrices qui ont une frontière naturelle tout en admettant des singularités modérées du ressort de la méthode. Les résultats sont les suivants.

• La probabilité pour que les cycles disjoints d'une permutation uniforme de $\mathfrak{S}_n$ aient des longueurs toutes distinctes vaut asymptotiquement $e^{-\gamma}(1+1/n)+O(\log n/n^2)$.

• La probabilité pour que deux permutations uniformes soient conjuguées dans $\mathfrak{S}_n$ vaut asymptotiquement $W(1)/n^2+O(\log n/n^3)$ où $W(1)=\prod_{k\geq 1} I(1/k^2)$, la fonction I étant définie par $I(z)=\sum_{n\geq 0} z^n/(n!)^2$.

• La probabilité pour qu'un polynôme de degré n sur le corps fini $\mathbb{F}_q$ ait des facteurs irréductibles de degrés tous distincts vaut asymptotiquement $\delta(q)+O(1/n)$ où $\delta(q)$ est la constante $\delta(q)=\prod_{k\geq 1}(1+I_k/q^k)(1-q^{-k})^{I_k}$, l'entier I_k désignant le nombre de polynômes irréductibles unitaires de degré k sur $\mathbb{F}_q$.

• La probabilité pour qu'une forêt d'arbres planaires enracinés de taille n ne contienne que des arbres de tailles différentes est équivalente, lorsque n tend vers l'infini, à $\frac{1}{K}e^{L+1/2}$, où

$$K=\exp\sum_{k\geq 1}\frac{1}{2k}\left(1-\sqrt{1-4^{1-k}}\right) \text{ et } L=\sum_{m\geq 2}\frac{(-1)^{m-1}}{m}\sum_{n\geq 1}\left(\frac{1}{n}\binom{2n-2}{n-1}4^{-n}\right)^m.$$

D'autres cas d'applications de la méthode hybride sont évoqués dans [8], sans qu'un traitement complet ne leur soit appliqué.

4 Arbres m-aires de recherche, processus de Pólya

Où l'on montre comment une conjecture de l'analyse d'algorithmes, formulée dans un cadre probabiliste, put être résolue en adoptant une démarche vectorielle aussi intrinsèque – *i.e.* sans coordonnées – que possible ([5]). Où l'on voit également comment le vecteur des complexités en mémoire des arbres m-aires de recherche peut se réécrire comme une urne de Pólya-Eggenberger et comment la démarche intrinsèque pour ces urnes amène naturellement aux processus de Pólya, objets d'étude de [6] et [7].

4.1 Complexité en mémoire des arbres m-aires de recherche

4.1.1 "Les arbres m-aires de recherche sont des structures fondamentales utilisées pour le tri et la recherche de données en informatique" (citation de J. A. Fill et N. Kapur ([25])). Ils généralisent l'arbre binaire de recherche, pierre d'angle du tri et de la recherche de données en informatique par le biais de l'incontournable algorithme Quicksort. En stockant les données dans un arbre dont le facteur de branchement est supérieur à deux, on cherche à réduire la longueur des chemins et, conséquemment, le coût de la recherche.

Soit m un entier, $m \geq 2$. On procède au tirage aléatoire d'une suite de nombres réels appelés *clefs* dans $[0, 1]$, chaque clef étant tirée indépendamment, selon la loi uniforme. Ces clefs sont insérées les unes après les autres dans un arbre m-aire complet infini, chaque nœud de l'arbre ayant la capacité de contenir au plus $m - 1$ clefs. La première clef est placée à la racine. Récursivement, une clef nommée k est insérée dans l'arbre comme suit :

• si la racine n'est pas *saturée*, c'est-à-dire si elle contient strictement moins de $m - 1$ clefs, on place k à la racine. On ordonne usuellement les clefs d'un nœud donné par ordre croissant.

• Si la racine est saturée, soient $k_1 < \cdots < k_{m-1}$ les clefs qui y sont déjà insérées. A chaque intervalle $I_1 =]0, k_1[$, $I_2 =]k_1, k_2[$, $\ldots$, $I_{m-1} =]k_{m-2}, k_{m-1}[$, $I_m =]k_{m-1}, 1[$ on fait correspondre un sous-arbre de l'arbre m-aire. Dans les représentations graphiques, on dessine usuellement les branches de l'arbre de gauche à droite dans l'ordre de ces intervalles. Récursivement, on insère la clef k dans le sous-arbre correspondant à l'intervalle

I_j qui contient le nombre k ; chaque nœud de l'arbre est la racine d'un sous-arbre qui est lui-même un arbre m-aire de recherche.

On donne dans la figure 2 un exemple, insertions successives des clefs $0,2$; $0,7$; $0,1$ $0,8$; $0,28$; $0,18$; $0,284$; $0,5$; $0,9$; $0,04$; $0,52$; $0,3$; $0,53$; $0,6$; $0,02$; $0,87$; $0,4$ $0,71$; $0,35$; $0,26$; $0,62$ dans un arbre quaternaire de recherche.

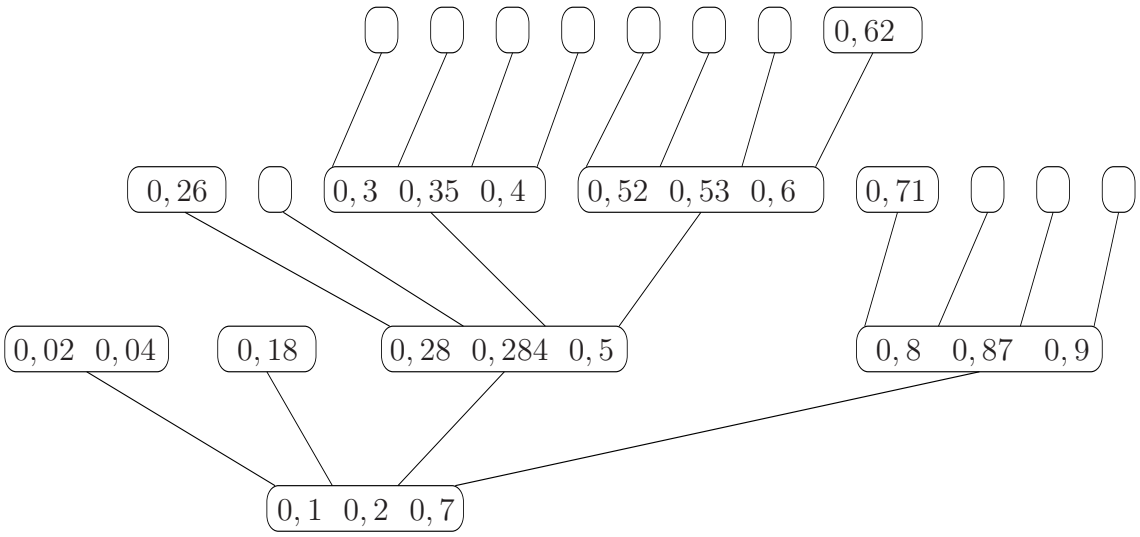

Figure 2: Insertions des clefs 0,2 ; 0,7 ; 0,1 ; 0,8 ; 0,28 ; 0,18 ; 0,284 ; 0,5 ; 0,9 ; 0,04 ; 0,52 ; 0,3 ; 0,53 ; 0,6 ; 0,02 ; 0,87 ; 0,4 ; 0,71 ; 0,35 ; 0,26 ; 0,62 dans un arbre quaternaire de recherche.

Pour un formalisme sur les arbres, on se référera au papier [54] de J. Neveu. Le modèle probabiliste des arbres m-aires de recherche est celui des permutations aléatoires (uniformes) ; dans ce modèle qui équivaut à l'hypothèse de jets indépendants et équidistribués des clefs, à chaque instant n, les ordres relatifs des clefs dans $[0,1]$ induisent des permutations équiprobables de $\mathfrak{S}_n$.

4.1.2 Pour tout $k \in \{1,\dots,m-1\}$, on appelle *nœud de type k* un nœud contenant $k-1$ clefs. Pour tout $n \geq 2$, on note $Y_n^{(k)}$ le nombre de nœuds de type k dans l'arbre obtenu après insertion de la $(n-1)^{\text{ième}}$ clef. On convient pour cela que la saturation d'un nœud après l'insertion d'une clef provoque l'apparition de m nœuds vides ayant le nœud nouvellement saturé pour père. La somme des $Y_n^{(k)}$, $1 \leq k \leq m-1$, est la *complexité en mémoire*[5] de l'arbre m-aire (en anglais, *space requirement*). Plus synthétiquement, on

[5]Pour être plus correct, le vocable de complexité en mémoire devrait être réservé au nombre total de nœuds *non vides* de l'arbre, c'est-à-dire à $\sum_{k=2}^{m} Y_n^{(k)}$, le symbole $Y_n^{(m)}$ désignant le nombre de nœuds saturés au temps n. La relation déterministe $n = \sum_{k=1}^{m}(k-1)Y_n^{(k)}$ qui traduit la répartition des clefs dans les nœuds de différents types permet cet écart de langage.

désigne par Y_n le vecteur aléatoire de $\mathbb{R}^{m-1}$ défini par

$$Y_1 = {}^t(1,0,\dots,0) \text{ et } Y_n = {}^t(Y_n^{(1)},\dots,Y_n^{(m-1)}) \text{ si } n \geq 2.$$

C'est le *vecteur des complexités en mémoire (space requirements)* de l'arbre. Le début de cette suite est déterministe :

$$\begin{cases} Y_1 = {}^t(1,0,\dots,0) \\ Y_2 = {}^t(0,1,\dots,0) \\ \quad \vdots \\ Y_{m-1} = {}^t(0,\dots,0,1) \\ Y_m = {}^t(m,0,\dots,0) \\ Y_{m+1} = {}^t(m-1,1,0,\dots). \end{cases}$$

Les vecteurs suivants sont aléatoires. Lorsque $n-1$ clefs ont été insérées, l'arbre contient n places libres, toutes équiprobables ; en effet, selon le modèle aléatoire des permutations, lors du jet de n clefs uniformes et indépendantes dans $[0,1]$, la probabilité que la $n^{\text{ième}}$ d'entre elles appartienne à l'un quelconque des n intervalles définis par les $n-1$ premières égale $1/n$. Par ailleurs, chaque nœud de type k contient k places libres pour l'insertion d'une nouvelle clef. Ainsi le processus markovien $(Y_n)_n$ est-il régi par les probabilités de transition

$$\text{Prob } (Y_{n+1} = Y_n + \Delta_k | Y_n) = \frac{kY_n^{(k)}}{n}, \tag{6}$$

où les Δ_k, vecteurs d'incréments déterministes, sont définis par

$$\begin{cases} \Delta_1 = {}^t(-1,1,0,0,\dots) \\ \Delta_2 = {}^t(0,-1,1,0,\dots) \\ \quad \vdots \\ \Delta_{m-2} = {}^t(0,\dots,0,-1,1) \\ \Delta_{m-1} = {}^t(m,0,\dots,0,-1). \end{cases}$$

4.1.3 Il résulte d'un calcul élémentaire que l'espérance de Y_{n+1} conditionnée à l'état du processus au temps n s'écrit

$$E^{\mathcal{F}_n}(Y_{n+1}) = \left(\text{Id} + \frac{A}{n}\right) Y_n \ ,$$

où Id désigne la matrice de l'identité et A la matrice diagonalisable (sur $\mathbb{C}$)

$$A = \begin{pmatrix} -1 & & & & & m(m-1) \\ 1 & -2 & & & & \\ & 2 & -3 & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & -(m-2) & \\ & & & & m-2 & -(m-1) \end{pmatrix}.$$

Le vecteur Y_n est la somme de trois de ses projections

$$Y_n = \pi_1 Y_n + \pi_{>1/2} Y_n + \pi_{\leq 1/2} Y_n. \tag{7}$$

Dans cette décomposition, π_1 désigne la projection sur la droite des points fixes de A et $\pi_{>1/2}$ (respectivement $\pi_{\leq 1/2}$) la projection sur la somme des espaces propres de A associés aux valeurs propres différentes de 1 dont la partie réelle est $> 1/2$ (respectivement $\leq 1/2$). Les directions de ces projections sont toutes relatives à la décomposition de $\mathbb{C}^{m-1}$ en somme de droites propres pour A.

Dans l'article [5], on évalue l'asymptotique des moments des coordonnées de Y_n dans une base de vecteurs propres pour A. Il résulte de cette étude les trois points suivants :

• le vecteur $\pi_1 Y_n$ est déterministe, égal à $n\pi_1 Y_1$;

• le vecteur $\pi_{>1/2} Y_n$ se normalise en une martingale qui converge dans tout espace L^p, $p \geq 1$;

• le vecteur $\pi_{\leq 1/2} Y_n$ est négligeable devant toute puissance $n^{\frac{1}{2}+\varepsilon}$, $\varepsilon > 0$, presque sûrement et dans tout espace L^p, $p \geq 1$.

4.1.4 Lorsque $m \leq 26$ et seulement dans ce cas, les valeurs propres de A ont toutes une partie réelle $< 1/2$ à l'exception de 1 qui est toujours valeur propre simple. Il en résulte que le vecteur $(Y_n - EY_n)/\sqrt{n}$ converge en distribution vers un vecteur gaussien. Cela est établi par plusieurs résultats antérieurs à la publication de [5] ; parmi les différentes approches utilisées, citons des méthodes basées sur des calculs de moments et de fonctions génératrices (Mahmoud et Pittel [49], Lew et Mahmoud [47], Smythe [63], Mahmoud et Smythe [50]), un plongement du processus en temps continu interprété en termes de branchement multitype (Athreya et Karlin [11], Janson [36]) ou des méthodes de contraction (Neininger et Rüschendorf [53]).

Certains auteurs n'hésitent pas à parler de transition de phase entre les valeurs $m = 26$ et $m = 27$. Par exemple, H. H. Chern et H. K. Hwang, dans [18], montrent que lorsque $m \geq 27$, aucune normalisation non triviale ne fournit de convergence en loi de Y_n, mais que les moments de ce processus fluctuent, régis asymptotiquement par des fonctions périodiques de $\log n$.

Dans [5], l'assertion *2-* du théorème suivant est démontrée en suivant la démarche basée sur la décomposition (7) décrite plus haut. Elle apporte une réponse à l'asymptotique des "grands" arbres m-aires de recherche.

Théorème 4.1 (Asymptotique des arbres m-aires de recherche) *Soit $m \geq 2$. Soit $(Y_n)_n$ le processus des complexités en mémoire d'un arbre m-aire de recherche. Avec les notations ci-dessus, soit λ_2 la valeur propre de la matrice A qui soit différente de 1, qui ait la plus grande partie réelle ($\Re(\lambda_2) < 1$) et dont la partie imaginaire soit strictement positive. Soit $v_1 = \pi_1 Y_1$, vecteur déterministe de $\mathbb{R}^{m-1}$.*

1- Si $m \leq 26$, alors $\Re(\lambda_2) < 1/2$; le vecteur $(Y_n - nv_1)/\sqrt{n}$ converge en distribution vers un vecteur gaussien.

2- Si $m \geq 27$, $\Re(\lambda_2) > 1/2$; il existe une variable aléatoire complexe W et un vecteur déterministe v_2 dans $\mathbb{C}^{m-1}$ tels que

$$Y_n = nv_1 + 2\Re\left(n^{\lambda_2} W v_2\right) + o\left(n^{\Re(\lambda_2)}\right),$$

le reste o étant presque sûr et dans tous les L^p, $1 \leq p \leq 2$.

La variable aléatoire W apparaît comme la limite d'une martingale. Cet énoncé est complété dans l'article [7] : l'asymptotique est valide dans tous les L^p, $p \geq 1$ et les moments EW, $E(W^2)$ et $E|W^2|$ sont explicitement calculés par des formules closes, et pas seulement par des relations de récurrence. La distribution de W a depuis été caractérisée comme unique solution d'une équation fonctionnelle en loi par J. A. Fill et N. Kapur ([25]), en utilisant une méthode de contraction.

Cela dit, la question de la détermination de cette loi reste entière. Par exemple, peut-elle s'exprimer à l'aide d'opérations usuelles sur des distributions usuelles ?

4.1.5 En termes géométriques, le théorème 4.1 *2-* s'interprète de la façon suivante. Si $\rho \geq 0$ et $\varphi \in [-\pi, \pi]$ sont respectivement l'amplitude et la phase de la variable aléatoire complexe $2W = \rho \exp(i\varphi)$ et si σ_2 et τ_2 désignent la partie réelle et la partie imaginaire de λ_2, alors

$$\frac{1}{n^{\sigma_2}} \left(Y_n - n v_1\right) \underset{n \to +\infty}{\sim} \rho \left(\cos(\tau_2 \log n + \varphi) \Re(v_2) - \sin(\tau_2 \log n + \varphi) \Im(v_2)\right),$$

presque sûrement. Ainsi l'arc paramétré par $x = t$, $y + iz = \rho t^{\sigma_2} \exp[-i(\tau_2 \log t + \varphi)]$, tracé sur la surface d'équation $\rho^2 x^{2\sigma_2} = y^2 + z^2$ dans le sous-espace réel de dimension trois repéré par les vecteurs $(v_1, \Re(v_2), \Im(v_2))$ est-il presque sûrement arc asymptote de Y_n dans $\mathbb{R}^{m-1}$. Une illustration de cette courbe asymptote est donnée dans la figure 3.

Là encore, la question des lois de la variable aléatoire positive ρ et de la variable aléatoire φ sur le tore reste ouverte.

Mentionnons pour finir que le théorème 4.1 *2-* fournit une asymptotique de n'importe quel paramètre sur les grands arbres m-aires de recherche qui soit fonction continue du vecteur Y_n des complexités en mémoire.

4.2 Des arbres m-aires de recherche aux urnes de Pólya-Eggenberger

Les probabilités de transition (6) du processus des complexités en mémoire des arbres m-aires de recherche suggèrent le changement de variables

$$(y_1, \ldots, y_{m-1}) \mapsto (y_1, 2y_2, \ldots (m-1)y_{m-1}).$$

Le nouveau processus $(X_n)_n$ obtenu est encore markovien, garde le même vecteur initial $X_1 = Y_1$ et est régi par les nouvelles probabilités de transition

$$\text{Prob}\left(X_{n+1} = X_n + \Delta'_k | X_n\right) = \frac{X_n^{(k)}}{n}$$

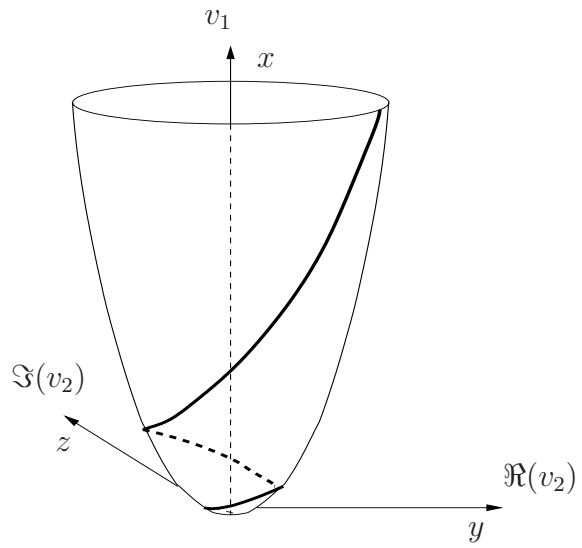

Figure 3: Une spirale logarithmique $y + iz = \rho t^{\sigma_2} \exp[-i(\tau_2 \log t + \varphi)]$ enroulée sur une surface d'équation $\rho^2 x^{2\sigma_2} = y^2 + z^2$ dans le sous-espace de $\mathbb{R}^{m-1}$ engendré par les vecteurs v_1, $\Re(v_2)$ et $\Im(v_2)$. Une telle courbe est presque sûrement asymptote à la trajectoire du vecteur aléatoire Y_n des complexités en mémoire d'un "grand" arbre m-aire de recherche (*i.e.* $m \geq 27$). La phase φ et l'amplitude ρ sont aléatoires.

où les nouveaux vecteurs d'incréments Δ'_k sont donnés par

$$\begin{cases} \Delta'_1 = {}^t(-1, 2, 0, 0, \dots) \\ \Delta'_2 = {}^t(0, -2, 3, 0, \dots) \\ \quad \vdots \\ \Delta'_{m-2} = {}^t(0, \dots, 0, -m+2, m-1) \\ \Delta'_{m-1} = {}^t(m, 0, \dots, 0, -m+1). \end{cases}$$

On reconnaît là le processus d'une urne de Pólya-Eggenberger équilibrée (la balance est 1) à $m-1$ couleurs. Dans le modèle d'urne de Pólya-Eggenberger de balance $S \geq 1$ à d couleurs, on dispose d'une urne de capacité infinie et de boules d'un nombre fini de couleurs notées $1, \dots, d$. On part d'une configuration initiale et on procède à des tirages successifs d'une boule prise au hasard, chaque boule de l'urne ayant la même probabilité d'être tirée. A chaque instant, on note la couleur de la boule tirée, on la replace dans l'urne et on ajoute d'autres boules selon une règle qui reste la même pendant tout le processus. Cette règle est décrite par la matrice de remplacement de l'urne $R = (r_{j,k})_{1 \leq j,k \leq d} \in \mathcal{M}_d(\mathbb{Z})$, où $r_{j,k}$ désigne le nombre de boules de couleur k que l'on ajoute (algébriquement) lorsqu'une boule de couleur j vient d'être tirée. A un coefficient négatif $r_{j,k}$ correspond le retrait de $-r_{j,k}$ boules de couleur k. L'urne est dite *équilibrée de balance* S lorsque le nombre total de boules ajoutées est invariablement S à chaque opération. Cela signifie que la somme des coefficients de chaque ligne de R est S. On associe à une telle urne le processus aléatoire de $\mathbb{R}^d$, dont la $k^{\text{ième}}$ coordonnée est le nombre de boules de couleur k.

Le processus vectoriel de l'urne est la suite $(X_n)_n$ de vecteurs aléatoires dans $\mathbb{R}^d$, où la $k^{\text{ième}}$ coordonnée de X_n est le nombre de boules de couleur k que contient l'urne à l'issue du $(n-1)^{\text{ième}}$ tirage.

Les urnes de Pólya-Eggenberger font l'objet d'une vaste littérature. On mentionne ici l'article fondateur de G. Pólya [58], le livre référence de N. L. Johnson et S. Kotz [40], les articles récents de S. Janson [36] et de P. Flajolet, J. Gabarró et H. Pekari [26], ainsi que la thèse de V. Puyhaubert [59] qui contiennent une base bibliographique et illustrent la variété des méthodes de traitement.

Le changement de variables dans le processus des complexités en mémoire d'un arbre m-aire de recherche fournit l'urne de Pólya-Eggenberger à $m-1$ couleurs et de balance 1 dont le vecteur initial est $X_1 = {}^t(1, 0, \dots, 0)$ et dont la matrice de remplacement est

$$R_m = \begin{pmatrix} -1 & 2 & & & & & \\ & -2 & 3 & & & & \\ & & -3 & 4 & & & \\ & & & \ddots & \ddots & & \\ & & & & -(m-2) & m-1 & \\ m & & & & & -(m-1) & \end{pmatrix}.$$

4.3 Processus de Pólya

4.3.1 Introduction, définition

Comme on vient d'en voir un exemple à la section précédente, les processus d'urnes de Pólya-Eggenberger équilibrées ne sont pas invariants par changement de coordonnées. Or, dans l'étude asymptotique de ces urnes, interviennent naturellement les propriétés spectrales des matrices de remplacement, qui imposent précisément un système privilégié de coordonnées. Par ailleurs, les coordonnées des processus d'urnes normalisés après division par la balance S sont toujours rationnelles. La classe minimale des processus vectoriels aléatoires de dimension finie qui contienne les urnes de Pólya-Eggenberger équilibrées et qui soit stable par changements linéaires de coordonnées réelles (ou complexes) est celle des *processus de Pólya* tels qu'ils sont définis dans [6]. Cette définition est reportée ci-dessous.

Définition 4.2 *Soit V un espace vectoriel réel de dimension finie $d \geq 1$. Soient X_1, $w_1, \ldots, w_d$ des vecteurs de V et $(l_1, \ldots, l_d)$ une base de formes linéaires sur V satisfaisant les hypothèses suivantes:*
i- (hypothèse d'initialisation)

$$X_1 \neq 0 \quad \text{et} \quad \forall k \in \{1, \ldots, d\},\ l_k(X_1) \geq 0\ ; \tag{8}$$

ii- (hypothèse d'équilibre) pour tout $k \in \{1, \ldots, d\}$,

$$\sum_{j=1}^{d} l_j(w_k) = 1\ ; \tag{9}$$

iii- (condition suffisante de viabilité) pour tous $k, k' \in \{1, \ldots, d\}$,

$$\begin{cases} k \neq k' \Longrightarrow l_k(w_{k'}) \geq 0, & (10.a) \\ l_k(w_k) \geq 0 \quad \text{ou} \quad l_k(X_1)\mathbb{Z} + \displaystyle\sum_{j=1}^{d} l_k(w_j)\mathbb{Z} = l_k(w_k)\mathbb{Z}. & (10.b) \end{cases} \tag{10}$$

Le **processus de Pólya** *associé à ces données est la marche aléatoire $(X_n)_{n\geq 1}$ à valeurs dans V et à incréments dans $\{w_1, \ldots, w_d\}$, définie par X_1 et par la récurrence suivante : pour tout $n \geq 1$ et pour tout $k \in \{1, \ldots, d\}$,*

$$\text{Prob}\ (X_{n+1} = X_n + w_k | X_n) = \frac{l_k(X_n)}{n + \tau_1 - 1} \tag{11}$$

où τ_1 est le nombre réel strictement positif défini par

$$\tau_1 = \sum_{k=1}^{d} l_k(X_1).$$

Le processus est défini sur l'espace des trajectoires de $X_1 + \sum_{1\leq k\leq d} \mathbb{Z}_{\geq 0} w_k$ muni de la filtration naturelle $(\mathcal{F}_n)_{n\geq 0}$, où $\mathcal{F}_n$ est la tribu engendrée par $X_1, \ldots, X_n$.

Les conditions (8) et (9) sont nécessaires et suffisantes pour que X_2 soit défini par les relations (11). Elles entraînent la relation déterministe

$$\forall n \geq 1, \quad \sum_{k=1}^{s} l_k(X_n) = n + \tau_1 - 1. \tag{12}$$

par une récurrence immédiate. Une autre récurrence sans difficultés montre que les conditions (10) suffisent à ce que le processus ne s'éteigne pas, c'est-à-dire à ce que les nombres $l_k(X_n)$ restent positifs ou nuls pour tous k et n. Cette condition arithmétique est la traduction naturelle de l'hypothèse devenue classique que l'on trouve déjà dans l'article de A. Bagchi et A. K. Pal [12] pour les processus d'urnes. En fait, les résultats sur l'asymptotique des processus de Pólya qui suivent s'affranchissent sans modification de la condition suffisante de viabilité si l'on conditionne le processus à sa non-extinction.

4.3.2 Opérateur de transition d'un processus de Pólya

Soit $(X_n)_n$ un processus de Pólya de dimension $d \geq 1$ défini par sa condition initiale X_1, ses vecteurs d'incréments $w_1, \ldots, w_d$ et sa base de formes linéaires $(l_1, \ldots, l_d)$. La recherche de ses lois fini-dimensionnelles ou de sa distribution asymptotique conduit à calculer les espérances $Ef(X_n)$ pour une classe de fonctions f suffisamment large. Si f est une fonction définie sur V et à valeurs dans n'importe quel espace vectoriel réel ou complexe, l'espérance de $f(X_{n+1})$, conditionnellement à la tribu $\mathcal{F}_n$, s'écrit

$$E^{\mathcal{F}_n} f(X_{n+1}) = \left(\mathrm{Id} + \frac{1}{n+\tau_1-1}\Phi\right)(f)(X_n) \tag{13}$$

où Id désigne l'application identique et Φ l'*opérateur de transition*, opérateur aux différences finies défini par

$$\Phi(f)(v) = \sum_{1\leq k\leq d} l_k(v)\Big[f(v+w_k) - f(v)\Big]. \tag{14}$$

Par une récurrence immédiate, il s'ensuit que

$$Ef(X_n) = \gamma_{\tau_1,n}(\Phi)(f)(X_1) \tag{15}$$

où $\gamma_{\tau_1,n}$ est le polynôme à coefficients réels défini par

$$\gamma_{\tau_1,n}(t) = \prod_{k=1}^{n-1}\left(1 + \frac{t}{k+\tau_1-1}\right).$$

Le traitement des processus de Pólya développé dans [6] est basé sur la constatation suivante : pour tout entier naturel e, l'opérateur de transition stabilise l'espace des fonctions polynomiales de degré total inférieur ou égal à e. Cette propriété suggère une approche des lois fini-dimensionnelles ou asymptotique de X_n *via* ses moments polynomiaux, par

le biais d'une décomposition spectrale de Φ sur des espaces de polynômes de dimensions finies.

On peut remarquer que les formules (14) et (15) restent valables pour des processus de Pólya étendus à la dimension infinie dénombrable, sous réserve de convergence de la somme (14). Cela concerne par exemple le polynôme de niveau des arbres binaires de recherche ou des arbres récursifs (voir Jabbour-Hattab [35] pour une définition) qui peuvent être vus comme des urnes de Pólya à une infinité dénombrable de couleurs – une couleur est un niveau de l'arbre. On sait que la hauteur d'un tel arbre de taille n est presque sûrement $c \log n$ où c est une constante explicite ; cela impose, dans le modèle d'urnes, que le nombre de couleurs qui interviennent au temps n soit de l'ordre de $\log n$ à une constante multiplicative près.

4.3.3 Bases de Jordan, polynômes réduits

On reprend les notations du paragraphe précédent. L'action naturelle de Φ sur les formes linéaires de V induit un endomorphisme du dual V^*. La réduction de cet endomorphisme sur le corps des nombres complexes amène à la définition suivante. On notera $u_1 = \sum_{1 \le k \le d} l_k$, forme linéaire fixée par Φ.

Définition 4.3 *On appelle* **base de Jordan** *d'un processus de Pólya toute base de formes linéaires $(u_1, \dots, u_d)$ sur V dans laquelle la matrice de la restriction de Φ à V^* admet une forme diagonale par blocs $J = \mathrm{Diag}(1, J_{p_1}(\lambda_{k_1}), \dots, J_{p_t}(\lambda_{k_t}))$, où $J_p(z)$ désigne la matrice de Jordan de dimension p*

$$J_p(z) = \begin{pmatrix} z & 1 & & \\ & z & \ddots & \\ & & \ddots & 1 \\ & & & z \end{pmatrix}.$$

On notera σ_2 le nombre réel ≤ 1 défini par

$$\sigma_2 = \begin{cases} 1 \text{ si } 1 \text{ est valeur propre multiple de } \Phi_{|V^*} \ ; \\ \\ \max\{\Re(\lambda),\ \lambda \text{ valeur propre de } \Phi_{|V^*},\ \lambda \neq 1\} \text{ sinon.} \end{cases}$$

Par ailleurs, une base de Jordan étant choisie, un bloc diagonal $J_p(\lambda)$ de J sera dit **bloc principal** *si $\Re(\lambda) = \sigma_2$ et s'il est de taille maximale parmi les blocs diagonaux $J_q(\mu)$ de J tels que $\Re(\mu) = \sigma_2$.*

On fixe pour la suite du paragraphe une base de Jordan $(u_1, \dots, u_d)$ et on note $(v_1, \dots, v_d)$ sa base duale de vecteurs de V. On note λ_k la valeur propre associée à v_k et $\lambda = (\lambda_1, \dots, \lambda_d) \in \mathbb{C}^d$. Pour tout $\alpha = (\alpha_k)_{1 \le k \le d} \in \mathbb{Z}^d$, on note aussi

$$|\alpha| = \sum_{1 \le k \le d} \alpha_k \quad \text{et} \quad \langle \alpha, \lambda \rangle = \sum_{1 \le k \le d} \alpha_k \lambda_k$$

et, lorsque tous les α_k sont des entiers naturels,

$$\mathbf{u}^\alpha = \prod_{1\leq k\leq d} u_k^{\alpha_k}.$$

Par ailleurs, on considère sur $\mathbb{Z}_+^d$ l'*ordre degré-lexicographique inverse* défini, si l'on note $\alpha = (\alpha_1, \ldots, \alpha_d)$ et $\beta = (\beta_1, \ldots, \beta_d)$, par $\alpha < \beta$ lorsque $\Big(|\alpha| < |\beta|\Big)$ ou bien

$$\Big(|\alpha| = |\beta| \text{ et } \exists r \in \{1, \ldots, d\} \text{ tel que } \alpha_r < \beta_r \text{ et } \alpha_t = \beta_t \text{ pour tout } t > r\Big).$$

Avec ces notations, la famille $(\mathbf{u}^\alpha)_{\alpha\in\mathbb{Z}_+^d}$ est une base de l'espace des polynômes à d indéterminées sur $\mathbb{C}$. On démontre dans [6] que pour tout $\alpha \in \mathbb{Z}_+^d$, le sous-espace de dimension finie

$$S_\alpha = \mathrm{Vect}\{\mathbf{u}^\beta,\ \beta \leq \alpha\}$$

est stable par Φ et que les valeurs propres de $\Phi_{|S_\alpha}$ sont les $\langle\beta, \lambda\rangle$, $\beta \leq \alpha$. Pour chaque nombre complexe z, on note $\ker(\Phi - z)^\infty = \bigcup_{n\geq 0} \ker(\Phi - z)^n$ l'espace caractéristique de la restriction de Φ aux fonctions polynomiales sur $\mathbb{C}^d$.

Définition 4.4 *Une base de Jordan $(u_1, \ldots, u_d)$ d'un processus de Pólya étant fixée, pour tout $\alpha \in \mathbb{Z}_+^d$, on appelle $\alpha^{\text{ième}}$* **polynôme réduit** *le projeté de $\mathbf{u}^\alpha$ sur $\ker(\Phi - \langle\alpha, \lambda\rangle)^\infty$ parallèlement à $\bigoplus_{z\neq\langle\alpha,\lambda\rangle} \ker(\Phi - z)^\infty$. On le notera Q_α. Le nombre entier naturel ν_α désigne l'indice de nilpotence de Q_α dans l'espace caractéristique $\ker(\Phi - \langle\alpha, \lambda\rangle)^\infty$:*

$$\nu_\alpha = \max\{p \geq 0,\ (\Phi - \langle\alpha, \lambda\rangle)^p(Q_\alpha) \neq 0\}.$$

Les polynômes réduits réalisent la décomposition spectrale de Φ sur l'espace des fonctions polynomiales sur V annoncée plus haut. Notamment, chaque Q_α est un vecteur caractéristique de Φ, associé à la valeur propre $\langle\alpha, \lambda\rangle$. On verra ci-dessous comment ils interviennent naturellement dans l'asymptotique des processus de Pólya. Ces résultats sont établis dans [6] ; on trouvera dans cet article-là, ainsi que dans [7] une présentation d'un calcul récursif de ces polynômes propice à une implémentation en calcul formel et certaines formules closes dans des cas particuliers.

4.3.4 Géométrie dans l'espace des exposants, asymptotique des moments

On fixe une base de Jordan $(u_1, \ldots, u_d)$ d'un processus de Pólya $(X_n)_n$, avec les notations des sous-sections précédentes. La clef de l'approche développée dans [6] réside dans l'évaluation asymptotique des $\mathbf{u}$-moments $E\,\mathbf{u}^\alpha(X_n)$ lorsque n tend vers l'infini. Pour cela, on développe les $\mathbf{u}^\alpha$ dans la base des polynômes réduits : soient $q_{\alpha,\beta}$ les nombres complexes définis par

$$\mathbf{u}^\alpha = Q_\alpha + \sum_{\substack{\beta<\alpha \\ \langle\beta,\lambda\rangle\neq\langle\alpha,\lambda\rangle}} q_{\alpha,\beta} Q_\beta. \tag{16}$$

La formule (15) combinée avec les propriétés des polynômes réduits assure que $EQ_\alpha(X_n)$ a pour ordre de magnitude $n^{\langle\alpha,\lambda\rangle}\log^{\nu_\alpha} n$ lorsque n tend vers $+\infty$. Cela pose, en reportant ce résultat dans le développement (16), la double question :

i) quels nombres $q_{\alpha,\beta}$ s'annulent-ils ?

ii) Pour un α donné, parmi les $\beta < \alpha$ tels que $q_{\alpha,\beta} \neq 0$, pour lesquels $\Re\langle\beta,\lambda\rangle$ est-elle maximale ?

La réponse optimale qui soit valable pour tous les processus de Pólya fait intervenir, dans l'espace des exposants $\mathbb{Z}^d \otimes \mathbb{R}$, un cône rationnel polyédral Σ et un polytope rationnel A_α pour chaque exposant α. On note $(\delta_1, \ldots, \delta_d)$ la base canonique de $\mathbb{R}^d$.

Définition 4.5 *Pour chaque $(i,j) \in \{1,\ldots,d\}^2$, $i \neq j$, soit $\delta_{(i,j)} = 2\delta_i - \delta_j$. Le cône Σ est, par définition, le cône engendré par ces vecteurs :*

$$\Sigma = \sum_{\substack{(i,j)\in\{1,\ldots,d\}^2 \\ i\neq j}} \mathbb{R}_+\delta_{(i,j)}.$$

Soient $\varepsilon_2, \ldots, \varepsilon_d$ les nombres de $\{0,1\}$ tels que $u_k \circ A = \lambda_k u_k + \varepsilon_k u_{k-1}$ pour tout $k \geq 2$ (*cf.* la définition 4.3, les u_k forment une base de Jordan). Si A et B sont deux parties de $\mathbb{R}^d$, la notation $A - B$ désigne la différence $A - B = \{a - b,\ a \in A,\ b \in B\}$.

Définition 4.6 *Pour tout $\alpha \in \mathbb{Z}_+^d$, soit A_α le polytope rationnel défini par*

$$A_\alpha = (\alpha - D_\alpha) \cap (\mathbb{R}_+)^d,$$

où D_α est le cône engendré par les vecteurs $\delta_k - \delta_{k-1}$ tels que $\alpha_k \geq 1$ et $\varepsilon_k = 1$.

On démontre dans [6] que la formule (16) se raffine, quel que soit le processus de Pólya, quel que soit le choix de la base de Jordan $(u_k)_{1\leq k\leq d}$ en

$$\mathbf{u}^\alpha = Q_\alpha + \sum_{\substack{\beta\in A_\alpha - \Sigma \\ \langle\beta,\lambda\rangle\neq\langle\alpha,\lambda\rangle}} q_{\alpha,\beta} Q_\beta. \tag{17}$$

Grâce à cette nouvelle formule, on peut établir le théorème 4.8 sur l'asymptotique des $\mathbf{u}$-moments de $(X_n)_n$. Sa preuve repose sur le lemme suivant.

Lemme 4.7 *Soient $\alpha = (\alpha_1, \ldots, \alpha_d)$ et $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{Z}_+^d$.*

1. Si α est tel que $\alpha_k = 0$ pour tous les indices k vérifiant $\Re(\lambda_k) > 1/2$ et si $\beta \in \alpha - \Sigma$, alors $\Re\langle\beta,\lambda\rangle < \frac{1}{2}|\alpha|$.

2. Si α est tel que $\alpha_k = 0$ pour tous les indices k vérifiant $\Re(\lambda_k) \leq 1/2$ et si $\beta \in \alpha - \Sigma$, alors $\Re\langle\beta,\lambda\rangle < \Re\langle\alpha,\lambda\rangle$.

3. Si $\beta \in A_\alpha$, alors $\Re\langle\beta,\lambda\rangle = \Re\langle\alpha,\lambda\rangle$.

Une preuve de ce lemme est exposée dans [6]. Elle consiste en une étude de la géométrie de Σ et des A_α dans l'espace des exposants. Notamment, les équations des faces de Σ y sont établies à l'aide d'une division barycentrique de son cône dual sur laquelle agit naturellement le groupe symétrique $\mathfrak{S}_d$.

Théorème 4.8 (Asymptotique de moments joints d'un processus de Pólya) *Soit $(u_k)_{1\leq k\leq d}$ une base de Jordan d'un processus de Pólya $(X_n)_n$, avec les notations des paragraphes précédents. Soit $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{Z}_+^d$.*

1. *Si α est tel que $\alpha_k = 0$ pour tous les indices k vérifiant $\Re(\lambda_k) > 1/2$, alors il existe un entier naturel ν tel que*

$$E\,\mathbf{u}^\alpha(X_n) \in O\left(n^{|\alpha|/2}\log^\nu n\right)$$

   lorsque n tend vers l'infini.

2. *Si α est tel que $\alpha_k = 0$ pour tous les indices k vérifiant $\Re(\lambda_k) \leq 1/2$, alors*

$$E\,\mathbf{u}^\alpha\,(X_n) = cn^{\langle\alpha,\lambda\rangle}\log^{\nu_\alpha} n + o\left(n^{\Re\langle\alpha,\lambda\rangle}\log^{\nu_\alpha} n\right)$$

   lorsque n tend vers l'infini, où

$$c = \frac{1}{\nu_\alpha!}\frac{\Gamma(\tau_1)}{\Gamma(\tau_1 + \langle\alpha,\lambda\rangle)}(\Phi - \langle\alpha,\lambda\rangle)^{\nu_\alpha}(Q_\alpha)(X_1)$$

   (Γ est la fonction d'Euler).

4.4 Asymptotique des processus de Pólya

Dans cette sous-section, il est question de l'asymptotique "en trajectoire" des processus de Pólya. La "transition de phase" établie dans la sous-section 4.1 pour les arbres m-aires de recherche trouve une généralisation naturelle dans le cadre des processus de Pólya. On pose pour cela la définition suivante.

Définition 4.9 *On dira qu'un processus de Pólya est* **grand** *lorsque $\sigma_2 > 1/2$ (notations de la définition 4.3). Dans le cas contraire, on dira qu'il est* **petit**.

Théorème 4.10 (Petits processus de Pólya irréductibles) *Soient $(X_n)_n$ un petit processus de Pólya de dimension d et v_1 le vecteur de $\mathbb{R}^d$, premier vecteur de n'importe quelle base de Jordan duale. On suppose le processus* irréductible *au sens de Janson [36].*

1. *Si $\sigma_2 < 1/2$, alors*

$$(X_n - nv_1)/\sqrt{n}$$

   converge en loi vers un vecteur gaussien centré.

2. *Si $\sigma_2 = 1/2$ et si $\nu + 1$ est la taille des blocs de Jordan principaux, alors*

$$(X_n - nv_1)/\sqrt{n\log^{2\nu+1} n}$$

   converge en loi vers un vecteur gaussien centré.

On trouvera une preuve de ce théorème, une bibliographie sur le sujet ainsi que des extensions dans l'article [36] de S. Janson. Les matrices de covariance des vecteurs limites y sont calculées. La méthode utilisée consiste à plonger le processus en temps continu, obtenir la convergence par des techniques de martingales et revenir au cadre discret *via* un temps d'arrêt. L'hypothèse d'irréductibilité invoquée s'exprime dans le langage des urnes en termes de couleurs dominantes ; elle est un peu plus faible que l'irréductibilité au sens des chaînes de Markov ou des processus de branchements multitypes.

La convergence après normalisation vers des vecteurs gaussiens centrés disparaît si l'on s'affranchit de l'hypothèse d'irréductibilité. On trouvera par exemples des études de processus d'urnes dont la matrice de remplacement est triangulaire dans Janson [37], Puyhaubert [59], Flajolet, Gourdon et Panario [26] ou encore [6]. Dans tous ces cas de petits processus de Pólya, la convergence en loi subsiste après renormalisation, mais les vecteurs limites sont en général non gaussiens.

Théorème 4.11 (Grands processus de Pólya) *Soient $(X_n)_n$ un grand processus de Pólya, $(u_k)_{1\leq k\leq d}$ une base de Jordan associée et $(v_k)_{1\leq v\leq d}$ sa base duale. On suppose que les blocs principaux sont tous de taille* 1*, au nombre de $r-1$ et respectivement associés aux valeurs propres $\lambda_2, \dots \lambda_r \in \mathbb{C}\setminus\{1\}$; en particulier, la partie réelle commune aux λ_k, $2\leq k\leq r$ est $\sigma_2 \in]1/2, 1]$.*

Alors, il existe des variables aléatoires complexes $W_2, \dots, W_r$, uniques, telles que

$$X_n = nv_1 + \sum_{2\leq k\leq r} n^{\lambda_k} W_k v_k + o\left(n^{\sigma_2}\right),$$

le o étant presque sûr et dans tous les L^p, $p\geq 1$. En outre, si $(Q_\alpha)_\alpha$ désigne la famille des polynômes réduits associés à la base de Jordan $(u_k)_k$, tous les moments joints des W_k existent et pour tous $\alpha_2, \dots, \alpha_r \in \mathbb{Z}_+$,

$$E\left(\prod_{2\leq k\leq r} W_k^{\alpha_k}\right) = \frac{\Gamma(\tau_1)}{\Gamma(\tau_1 + \langle \alpha, \lambda\rangle)} Q_\alpha(X_1)$$

où $\alpha = \sum_{2\leq k\leq r} \alpha_k \delta_k = (0, \alpha_2, \dots, \alpha_r, 0, \dots)$ (comme plus haut, $(\delta_1, \dots, \delta_d)$ désigne la base canonique de $\mathbb{R}^d$).

On peut également énoncer un théorème semblable en retirant l'hypothèse sur la taille des blocs de Jordan ; la présence de nilpotents dans les blocs principaux entraîne l'apparition d'un facteur en $\log^\nu n$ dans le terme du deuxième ordre, si l'entier $\nu+1$ est la taille des blocs principaux. Le théorème 4.11 sous sa forme la plus générale est démontré dans [6] ; la preuve s'appuie sur tous les éléments développés plus haut, notamment sur le théorème 4.8, et sur des arguments de martingales.

Afin de caractériser la forme des termes du deuxième ordre dans leur asymptotique et de fixer un vocabulaire, une classification des grands processus de Pólya est proposée dans [7]. La classification permet de distinguer, dans cette asymptotique, d'une part

la possibilité de normaliser $X_n - nv_1$ pour obtenir une convergence ("processus principalement réel") ou non ("principalement imaginaire"), et d'autre part la présence d'un facteur logarithmique ("principalement non semi-simple") ou non ("principalement semi-simple"). Le vocabulaire proposé est relatif à la somme directe des blocs principaux du processus.

4.5 Exemples et questions ouvertes

4.5.1 Dans [7], le calcul des trois premiers moments de la variable aléatoire limite W du grand processus de Pólya de dimension deux *général* est fait explicitement. On déduit de ce calcul que la loi de W est *génériquement non normale*, dans le sens où cela ne peut arriver que si le quadruplet des paramètres réels du processus appartient à une hypersurface analytique de $\mathbb{R}^4$. Savoir si certaines valeurs de ces paramètres peuvent conférer à la variable limite une distribution normale est un travail qui reste à faire.

4.5.2 Le cas particulier des processus de Pólya $(X_n)_n$ pour lesquels $\sigma_2 = 1$ est développé dans [6]. Sous cette hypothèse, X_n/n converge presque sûrement et dans tous les L^p, $p \geq 1$, vers un vecteur aléatoire à distribution de Dirichlet dans le sous-espace de $\mathbb{R}^d$ fixé par ${}^t\Phi$. Les paramètres de cette distribution dépendent de la condition initiale X_1 du processus. On rappelle qu'une distribution de Dirichlet de paramètres $t_1, \ldots, t_r$ sur le simplexe $\{\sum_{k=1}^r x_k = 1\}$ de $\mathbb{R}^r$ a pour densité

$$(x_1, \ldots, x_r) \mapsto \Gamma\left(\sum_{k=1}^r t_k\right) \prod_{k=1}^r \frac{x_k^{t_k}}{\Gamma(t_k)}.$$

4.5.3 On trouvera dans [6] et dans [7] de nombreux exemples de processus de Pólya, certains étant empruntés à l'algorithmique des ensembles de données informatiques.

4.5.4 Citons un exemple de processus aléatoire équivalent à un processus de Pólya. Cet exemple est davantage développé, sous un angle différent, dans l'article [67] d'Abraham, Dhersin et Ycart. On considère un nombre réel $p \in [0, 1]$ et une urne contenant au départ b boules blanches, v boules vertes et une boule rouge. Comme dans le cas des urnes de Pólya-Eggenberger, on procède à des tirages successifs, avec les règles de remplacement suivantes. Si on tire une boule blanche, on la remet avec une autre boule blanche ; si on tire une boule verte, on la remet avec une autre boule verte ; si on tire une boule rouge, on la remet et l'on ajoute une autre boule dont la couleur est blanche avec probabilité p ou verte avec probabilité $1 - p$.

Ainsi décrit, ce processus n'est pas Pólya. Cependant, il est équivalent au processus de Pólya $(X_n)_n$ de $\mathbb{R}^4$ pour lequel les formes linéaires l_k sont les formes coordonnées, les vecteurs d'incréments w_k sont les lignes de la matrice

$$R = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

et la condition initiale est $X_1 = {}^t(b, v, p, 1-p)$. On peut voir ce processus comme une urne de Pólya-Eggenberger à quatre couleurs blanche, verte, rouge foncé et rouge clair dont la matrice de remplacement est R, à ceci près que le vecteur initial (et les suivants également) n'est pas à coordonnées entières. La matrice R admet 1 comme valeur propre double, ce qui entraîne que X_n/n converge presque sûrement vers un vecteur aléatoire dont la distribution est une loi de Dirichlet sur le plan des deux premières coordonnées de $\mathbb{R}^4$.

4.5.5 Comme dans l'exemple des grands arbres m-aires de recherche, des questions naturelles se posent. Par exemple, les distributions des variables aléatoires limites W_k dans le théorème 4.11 (ou leur loi jointe) sont-elles caractérisées par leurs moments ? Peut-on exprimer ces lois en termes de lois usuelles en les combinant par des opérations usuelles ?

On prend deux grands processus de Pólya X et X', et on note respectivement Φ et Φ' leurs opérateurs de transition. On suppose que les restrictions de Φ et de Φ' aux formes linéaires sont conjuguées. Quand ces processus sont des urnes de Pólya-Eggenberger, cela signifie que les matrices de remplacement des deux processus sont semblables. Le théorème 4.11 montre alors que les asymptotiques au deuxième ordre de X et X' ont la même forme ; elles font notamment intervenir le même nombre de variables limites, $W_2, \ldots, W_r$ et $W'_2, \ldots, W'_r$ respectivement. Cependant, les lois des W_k n'égalent pas celles des W'_k, puisque les polynômes réduits respectifs de X et de X' n'admettent pas les mêmes relations algébriques (on trouvera un exemple précis dans [6]). Cela dit, existe-t-il une relation fonctionnelle entre les lois des variables limites de deux tels processus ?

5 Arbres digitaux de recherche et représentation des séquences d'ADN

Les récents progrès dans le séquençage de l'acide désoxyribonucléique (ADN) ouvrent, en biologie, des perspectives dans le décryptage du génome. En termes modélisés, une séquence d'ADN est une suite (de longueur arbitraire, virtuellement infinie) de lettres de l'alphabet $\mathcal{A} = \{A, C, G, T\}$, chaque lettre correspondant à l'une des quatre bases azotées – l'adénine, la cytosine, la guanine et la thymine – qui composent les brins d'ADN. A ce jour, les recherches ont principalement porté sur l'étude de séquences particulières, leur structure, la répétition de mots, la fréquence de lettres ou de combinaisons, *etc* (Roy, Raychaudhury, Nandy [62]).

Ces développements demandent des méthodes de stockage et de représentation d'un volume immense de données dont le traitement n'est envisageable que par l'outil informatique. Immanquablement, les algorithmes en jeu doivent être analysés. C'est sur ce sujet qu'interviennent les mathématiques présentées ici.

5.1 Arbre-CGR d'une séquence

5.1.1 *Chaos game representation* (CGR)

La CGR consiste à représenter une suite $(u_n)_n$ de $\mathcal{A}^{\mathbb{N}^*}$ par une suite de points $(X_n)_{n\in\mathbb{N}}$ d'un carré de $\mathbb{R}^2$ de la manière suivante. A chaque sommet du carré est affectée une lettre de $\mathcal{A}$ – il importe à la biologie de placer A et G sur une même diagonale – : on note ℓ_A le sommet du carré correspondant à A, *idem* pour les trois autres lettres. Le point X_0 est le centre du carré. Par récurrence, X_{n+1} est le milieu du segment $[X_n, \ell_{u_{n+1}}]$. Cette construction permet de la même façon de représenter une suite finie.

Cette méthode de représentation des séquences d'ADN fut appliquée pour la première fois par H. J. Jeffrey [39]. Elle fournit des images comme celle de la figure 4 – c'est d'*Homo Sapiens* qu'il s'agit. Une simple observation visuelle de ces images permet de discerner grossièrement les espèces. Roy, Raychaudhury, Nandy [62] mentionne par exemple que les CGR des vertébrés font apparaître des régions contenant très peu de points, celles des invertébrés présentent moins de motifs que celles des vertébrés, celles des moisissures ou des plantes contiennent des stries parallèles, celles des bactéries montrent des carrés remplis de points uniformément répartis, *etc*.

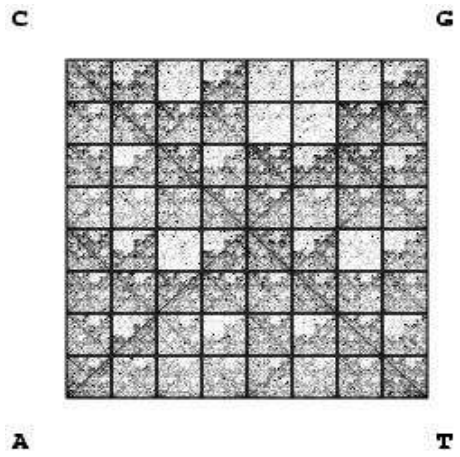

Figure 4: La CGR des 70 000 premiers nucléotides du chromosome 2 d'*Homo Sapiens*.

Si le carré de base est $[0,1]^2$, le $n^{\text{ième}}$ point X_n est au centre d'un unique carré de la forme $[\frac{j}{2^n}, \frac{j+1}{2^n}] \times [\frac{k}{2^n}, \frac{k+1}{2^n}]$. Par conséquent, la position de X_n contient la donnée de toute la suite des préfixes u_1, u_1u_2, ..., $u_1u_2 \cdots u_n$.

D'autres modes de représentation sont envisagés, par exemple en remplaçant le carré par d'autres parties de la droite, du plan ou de l'espace. La thèse de P. Cénac [17] contient une synthèse sur ces représentations et montre que "la CGR fournit plus d'information sur la distribution d'une séquence que les méthodes classiques liées au comptage de mots".

5.1.2 Arbre-CGR

On propose dans [9] d'associer à une suite de lettres de $\mathcal{A}$ un arbre quaternaire, son *arbre-CGR*, qui rende compte de la répétition des suffixes des préfixes successifs, à l'instar de la CGR elle-même. Sa construction est la suivante.

Une suite $u = (u_n)_n$ de $\mathcal{A}^{\mathbb{N}^*}$ étant donnée, on insère dans un arbre quaternaire digital de recherche la suite des *préfixes retournés* de u, c'est-à-dire la suite de mots (les *clefs*)

$$u_1,\ u_2u_1,\ u_3u_2u_1,\ u_4u_3u_2u_1,\ \cdots.$$

La croissance de la longueur de ces mots rend la construction possible, même si la suite u est finie. Pour une définition des arbres digitaux de recherche, on se référera aux livres de D. E. Knuth [45] ou de H. Mahmoud [48].

Dans la figure 5, on donne l'exemple de la construction de l'arbre-CGR de la suite finie $GAGCACAGTGGAAGGG$ (issue de *Mus Musculus*, nous dit-on). La racine reste vide. A chaque branche au-dessus de la racine correspond une lettre de $\mathcal{A}$; on ordonne ces branches de gauche à droite par ordre alphabétique et on procède de même pour chacun des nœuds de l'arbre quaternaire. On insère G au niveau 1, dans le nœud au bout de la branche correspondant à la lettre G – la place est libre. On insère AG au niveau 1, dans

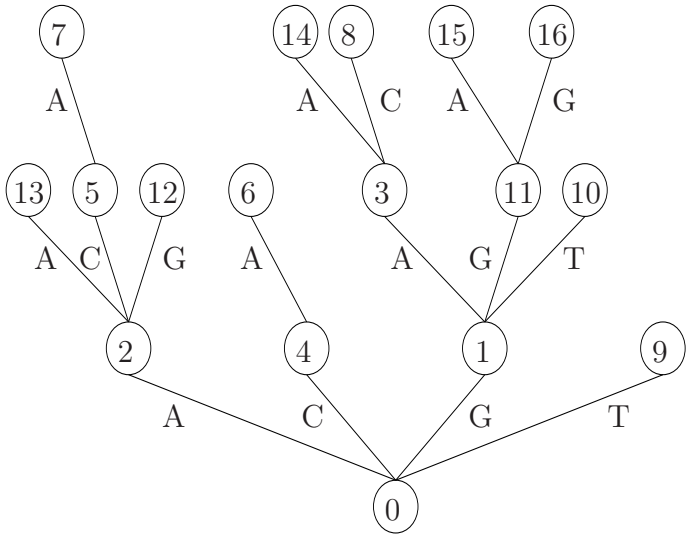

Figure 5: L'arbre-CGR de la suite finie $GAGCACAGTGGAAGGG$. Les étiquettes sont les ordres d'insertion des clefs.

le nœud au bout de la branche correspondant à la lettre A – de nouveau, la place est libre. On essaye d'insérer GAG au niveau 1 dans le nœud au bout de la branche correspondant à la lettre G, mais la place est prise ; on l'insère alors au niveau 2, dans le sous-arbre issu du nœud où l'on a déjà inséré G, au bout de la branche issue de A – la place est libre. On insère ainsi successivement G, AG, GAG, $CGAG$, $ACGAG$, $CACGAG$ *etc.*

La donnée de l'arbre-CGR d'une suite, étiqueté par les ordres d'insertion des clefs équivaut à celle de la suite elle-même. En revanche, deux suites finies distinctes peuvent avoir le même arbre non-étiqueté – on s'en convaincra aisément sur un exemple à deux lettres distinctes.

On pourrait penser à représenter une séquence par l'arbre digital de recherche de ses suffixes. Il ne serait complètement défini que lorsque la suite est infinie, ce qui n'arrive jamais dans les applications biologiques. Par ailleurs, l'arbre digital de recherche des préfixes n'a guère d'intérêt puisqu'il n'a qu'une seule branche. C'est la conjonction des propriétés des arbres digitaux de recherche et l'exigence de représenter les préfixes successifs comme dans la CGR qui a conduit à cette définition de l'arbre-CGR par insertion de la suite des préfixes renversés.

5.2 Propriétés asymptotiques des arbres-CGR

5.2.1 Modèle probabiliste

On s'intéresse à l'asymptotique des arbres-CGR des séquences dont la taille tend vers l'infini. Le modèle probabiliste que l'on considère est le suivant : la suite des lettres est une suite de variables aléatoires $(U_n)_{n\geq 1}$, chaîne de Markov d'ordre 1 irréductible, apériodique et stationnaire dont l'espace des états est $\mathcal{A} = \{A, C, G, T\}$. On notera p sa

mesure invariante et Q sa matrice de transition. Ces hypothèses contiennent naturellement le cas où les lettres sont indépendantes et identiquement distribuées, qu'elles soient équiprobables ou non.

On doit noter que les clefs que l'on insère dans l'arbre digital de recherche (l'arbre-CGR), c'est-à-dire les préfixes retournés successifs, forment une suite de variables aléatoires *fortement dépendantes*. Cette situation semble n'avoir pas été envisagée jusque-là dans la littérature. Mahmoud [48], par exemple, étudie des arbres digitaux de recherche à clefs indépendantes et ayant toutes la même loi. Une synthèse des résultats connus sur les arbres digitaux de recherche et des méthodes utilisées est faite dans Cénac [17]. Citons pêle-mêle et de façon lacunaire les travaux de D. Aldous et P. Shields ([10]), M. T. Barlow, R. Pemantle et E. A. Perkins ([13]), M. Drmota ([22]), B. Pittel ([57]). Les hypothèses présentes ne relèvent d'aucune de ces publications.

5.2.2 Plus longue et plus courte branches d'un arbre CGR

Si $s = (s_n)_{n\geq 1}$ est une séquence de $\mathcal{A}^{\mathbb{N}^*}$, on note $s^{(n)} = s_1 s_2 \dots s_n$ son $n^{\text{ième}}$ préfixe. On prolonge la mesure invariante p aux mots finis retournés par la formule

$$p(s^{(n)}) = \text{Prob}\,(U_1 = s_n, U_2 = s_{n-1}, \cdots, U_n = s_1)$$

et on définit les constantes h, h_+ et h_- par

$$\begin{aligned} h_+ &= \lim_{n\to+\infty} \frac{1}{n} \max_s \Big\{ \log\Big(\frac{1}{p\big(s^{(n)}\big)}\Big),\ p\big(s^{(n)}\big) > 0 \Big\}, \\ h_- &= \lim_{n\to+\infty} \frac{1}{n} \min_s \Big\{ \log\Big(\frac{1}{p\big(s^{(n)}\big)}\Big),\ p\big(s^{(n)}\big) > 0 \Big\}, \\ h &= \lim_{n\to+\infty} \frac{1}{n} E\Big[\log\Big(\frac{1}{p\big(U^{(n)}\big)}\Big)\Big]. \end{aligned}$$

La lettre E désigne l'espérance. L'existence de ces limites est assurée par un argument de sous-additivité que donne B. Pittel ([57]). Ces nombres ne dépendent que des paramètres de la chaîne de Markov des lettres de la séquence.

On note ℓ_n (respectivement $\mathcal{L}_n$) la plus petite (resp. la plus grande) longueur d'une branche d'un arbre-CGR d'une séquence aléatoire de n lettres. Le théorème suivant est prouvé dans [9].

Théorème 5.1 *Presque sûrement, lorsque n tend vers l'infini,*

$$\ell_n \sim \frac{\log n}{h_+} \quad \text{et} \quad \mathcal{L}_n \sim \frac{\log n}{h_-}.$$

La preuve présentée dans [9] fait intervenir les variables suivantes : si $s \in \mathcal{A}^{\mathbb{N}^*}$ est une suite déterministe, pour chaque $n \geq 1$ et chaque $k \geq 1$, on note $X_n(s)$ la longueur du plus long préfixe de s inséré dans l'arbre-CGR au temps n, et $T_k(s)$ le premier instant où l'arbre contient le préfixe $s^{(k)}$. Ces deux variables sont liées par la relation de dualité

$$X_n(s) \geq k \Longleftrightarrow T_k(s) \leq n$$

qui est un point-clef du raisonnement. A cause de l'hypothèse Markov du modèle, s étant donnée, les différences $Z_r(s) = T_r(s) - T_{r-1}(s)$ sont des variables aléatoires indépendantes, avec la convention $T_0 = 0$. Cela permet de voir les $T_k(s)$ comme sommes de variables indépendantes, dont les fonctions génératrices sont calculées par J. J. Daudin et S. Robin ([21]). Ces dernières, qui rendent compte de la structure des chevauchements des mots, sont des séries entières dont le rayon est > 1. Ces ingrédients permettent de démontrer le lemme suivant ([9]), *via* la loi des grands nombres pour les martingales – une preuve alternative est donnée dans Cénac [17].

Lemme 5.2 *Soit $s \in \mathcal{A}^{\mathbb{N}^*}$. On suppose que la limite $\lim_{n\to+\infty} \frac{1}{n} \log 1/p(s^{(n)}) = h(s)$ existe et est strictement positive. Alors, presque sûrement, lorsque n tend vers l'infini,*

$$X_n(s) \sim \frac{\log n}{h(s)}.$$

En passant, ce lemme redémontre et étend le résultat suivant de Erdös et Révész [24] et Petrov [56] : *soit $(V_n)_n$ une suite de variables aléatoires i.i.d. dans un alphabet fini ; si la probabilité de la lettre L est p, la longueur de la plus longue sous-suite de L consécutifs dans le mot $V_1 \cdots V_n$ est presque sûrement équivalente à $\log n / \log \frac{1}{p}$.*

Une fois ce lemme établi, comme les nombres h_+ et h_- sont atteints pour des suites s particulières (Pittel [57]), on obtient sans difficulté les inégalités presque sûres

$$\limsup_{n\to+\infty} \frac{\ell_n}{\log n} \leq \frac{1}{h_+} \quad \text{et} \quad \liminf_{n\to+\infty} \frac{\mathcal{L}_n}{\log n} \geq \frac{1}{h_-}.$$

Les inégalités presque sûres complémentaires

$$\liminf_{n\to+\infty} \frac{\ell_n}{\log n} \geq \frac{1}{h_+} \quad \text{et} \quad \limsup_{n\to+\infty}, \frac{\mathcal{L}_n}{\log n} \leq \frac{1}{h_-}$$

sont obtenues dans [9] grâce au lemme de Borel-Cantelli, en considérant avec précaution la structure de chevauchement des mots, notamment par le biais de la fonction génératrice de Daudin et Robin [21].

5.2.3 Profondeur d'insertion dans un arbre CGR

La *profondeur d'insertion* au temps n d'un arbre-CGR est la longueur de la branche au bout de laquelle on insère la $n^{\text{ième}}$ clef. On la note D_n. Il résulte du théorème 5.1 que

$$\liminf_{n\to+\infty} \frac{D_n}{\log n} = \frac{1}{h_+} \quad \text{et} \quad \limsup_{n\to+\infty} \frac{D_n}{\log n} = \frac{1}{h_-}$$

presque sûrement. On démontre par ailleurs dans [9] que $D_n / \log n$ *converge en probablité vers* $1/h$. A cause de ce qui précède, cette convergence n'est pas presque sûre lorsque les lettres de $\mathcal{A}$ ne sont pas équidistribuées.

5.2.4 Simulations

Pour illustrer les convergences des théorèmes ci-dessus, on a simulé des séquences de 100 000 lettres de $\mathcal{A}$ indépendantes et identiquement distribuées. Dans la figure 6, les probabilités respectives de A, C, G et T sont $0,4$, $0,3$, $0,2$ et $0,1$. Dans la figure 7, les lettres sont équiprobables. Dans les deux cas, on trace le graphe (interpolé) de la profondeur d'insertion normalisée $D_n/\log n$ en fonction de n ; c'est la courbe oscillante, dont on voit une fenêtre dans les dessins du dessous (100 valeurs de n). Sur le même graphe, on trace la longueur normalisée de la plus courte branche $\ell_n/\log n$ et la longueur normalisée de la plus longue branche $\mathcal{L}_n/\log n$, qui apparaissent comme les sortes d'"enveloppes" inférieure et supérieure de la courbe de $D_n/\log n$. On trace également les limites $1/h$, $1/h_+$ et $1/h_-$ de ces trois variables aléatoires. Dans la figure 7, naturellement, ces trois derniers nombres sont égaux.

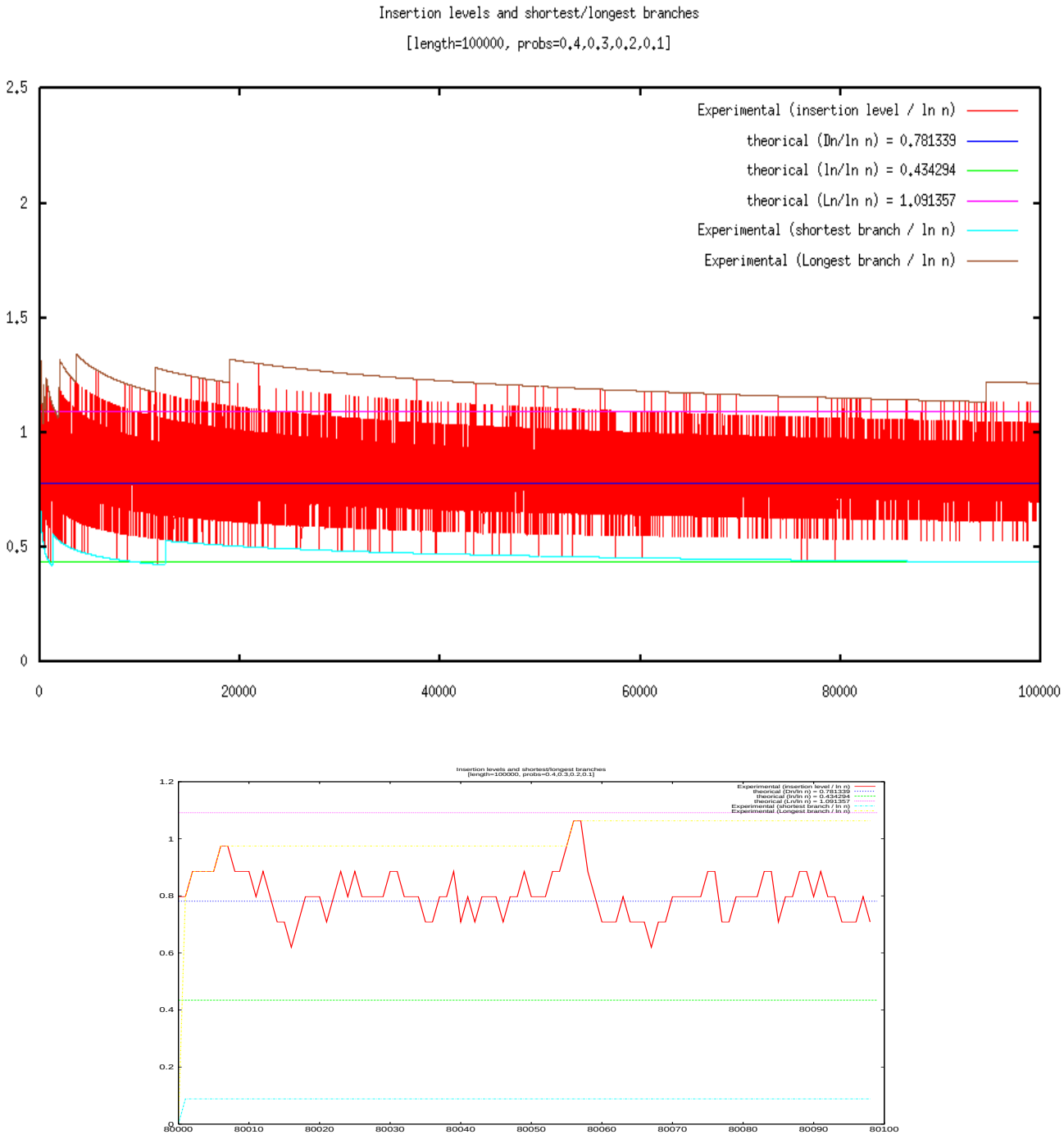

Figure 6: Simulation d'une séquence de 100 000 lettres i.i.d. non équiprobables. Sont représentés, en fonction de la longueur n des préfixes, les rapports $D_n/\log n$ (courbe oscillante), les longueurs des plus petite et plus longue branches ("enveloppes" inférieure et supérieure). Les droites horizontales sont les limites théoriques $1/h$, $1/h_+$ et $1/h_-$. Le second dessin représente un fenêtre du graphe précédent, entre les instants $n = 80\,000$ et $n = 80\,100$.

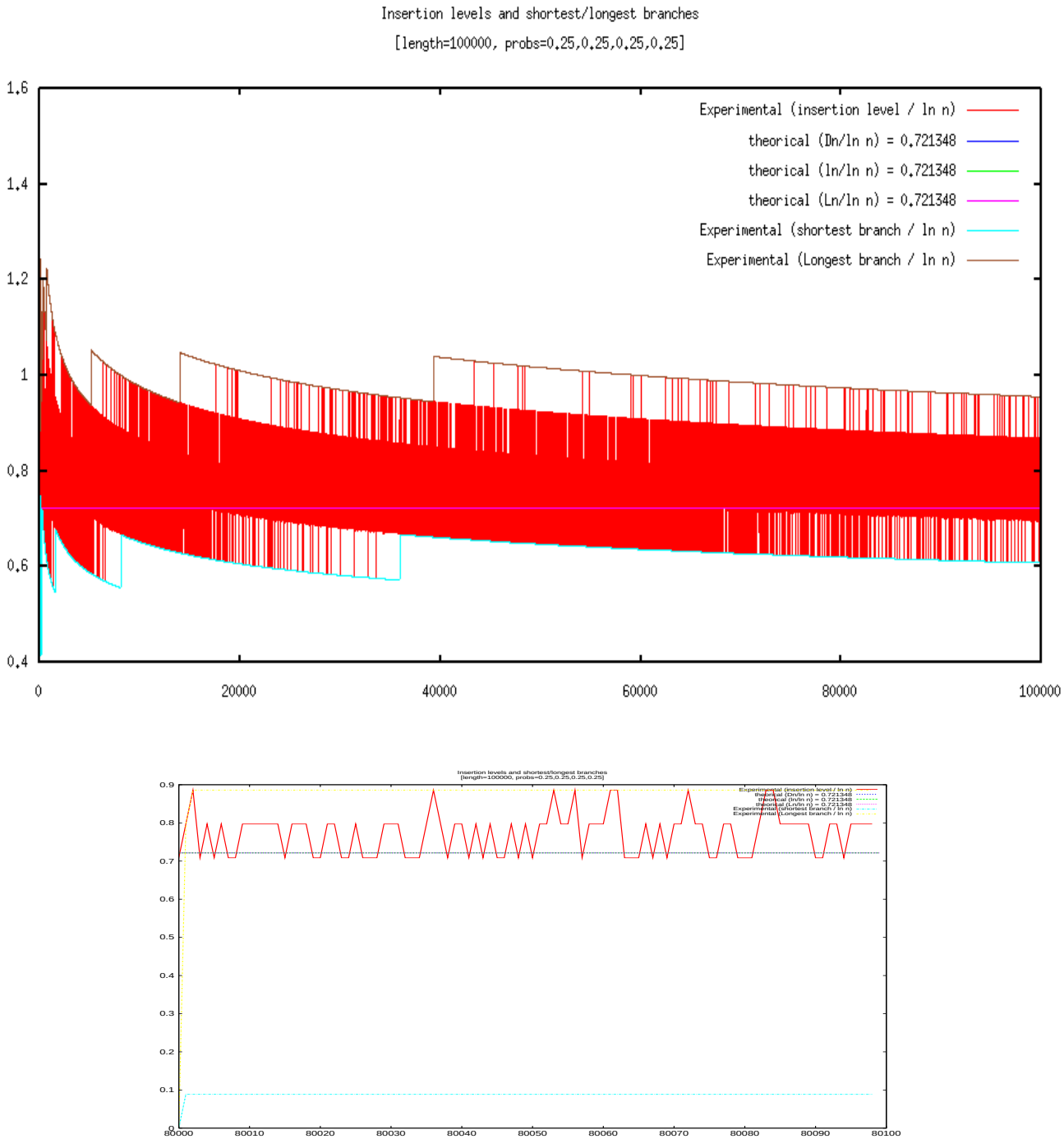

Figure 7: Simulation d'une séquence de 100 000 lettres i.i.d. équiprobables. Sont représentés, en fonction de la longueur n des préfixes, les rapports $D_n/\log n$ (courbe oscillante), les longueurs des plus petite et plus longue branches (enveloppes inférieure et supérieure). La droites horizontales est la limite théorique commune $1/h = 1/h_+ = 1/h_-$.

Références bibliographiques

[1] N. Pouyanne *Singularités-quotient par des groupes finis de dimension trois : un modèle à singularités toriques.* Thèse de l'université Joseph Fourier, Grenoble (1992).

[2] N. Pouyanne *Une résolution en singularités toriques simpliciales des singularités-quotient de dimension trois.* Annales de la Faculté des Sciences de Toulouse, **I**, no. 3 (1992), 363–398.

[3] N. Pouyanne *Cohomologie entière de certaines variétés toriques complexes lisses de dimension trois.* Comptes Rendus de l'Académie des Sciences de Paris, **t. 324**, Série I (1997), p. 61–66.

[4] N. Pouyanne *On the number of permutations admitting an m-th root.* Electronic Journal of Combinatorics, **9** (2002), no. 1, R3:1–12.

[5] B. Chauvin, N. Pouyanne *m-ary search trees when $m \geq 27$: a strong asymptotics for the space requirements.* Random Structures and Algorithms **24** (2004), 133–154.

[6] N. Pouyanne *An algebraic approach to Pólya processes.* Annales de l'Institut Henri Poincaré, à paraître (2006), 44 pages

[7] N. Pouyanne *Classification of large Pólya-Eggenberger urns with regard to their asymptotics.* Discrete Mathematics and Theoretical Computer Science, **AD** (2005), 275–286.

[8] P. Flajolet, E. Fusy, X. Gourdon, D. Panario, N. Pouyanne *A hybrid of Darboux's method and singularity analysis in combinatorial asymptotics.* Electronic Journal of Combinatorics, **13** (2006), no. 1, R103:1–35.

[9] P. Cénac, B. Chauvin, S. Ginouillac, N. Pouyanne *Digital search trees and chaos game representation.* ESAIM: Probability and Statistics, à paraître (2006), 26 pages.

A l'exception de [1], ces textes sont disponibles à l'adresse
http://www.math.uvsq.fr/∼pouyanne/

---

[10] D. Aldous, P. Shields *A diffusion limit for a class of randomly-growing binary search trees.* Probability Theory and Related Fields **79** (1988), 429–437.

[11] K. B. Athreya, S. Karlin *Embedding of urn schemes into continuous time Markov branching processes and related limit theorems.* Ann. Math. Statist. **39** (1968), 1801–1817.

[12] A. Bagchi, A. K. Pal *Asymptotic normality in the generalized Pólya-Eggenberger urn model, with an application to computer data structures.* SIAM Journal on Alegebraic and Discrete Methods, 6 **3** (1985), 394–405.

[13] M. T. Barlow, R. Pemantle, E. A. Perkins *Diffusion-limited aggregation on a tree.* Probability Theory and Related Fields **107**(1) (1997), 1–60.

[14] E. A. Bender *Asymptotic methods in enumeration.* SIAM Review **16** (1974), no. 4, 485–515.

[15] P. Billingsley *Probability and measure.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, second edition, John Wiley & Sons, 1986.

[16] J. Blum *Enumeration of the square permutations in* $\mathfrak{S}_n$. Journal of Combinatorial Theory, series A **17** (1974), 156–161.

[17] P. Cénac *Etude statistique de séquences biologiques et convergence de martingales.* Thèse de l'Université Paul Sabatier Toulouse III (2006).

Disponible à l'adresse http://www-rocq.inria.fr/∼cenac/these.html

[18] H.-H. Chern, H.-K. Hwang *Phase changes in random m-ary search trees and generalized quicksort.* Random Structures and Algorithms **19** (2001), 316–358.

[19] H.-H. Chern, M. Fuchs, H.-K. Hwang *Phase changes in random point quadtrees.* Submitted, 50 pages. Disponible à l'adresse http://algo.stat.sinica.edu.tw/HK/

[20] V. I. Danilov *The geometry of toric varieties.* Russian Mathematical Surveys **33**, II (1978), 97–154.

[21] J. J. Daudin, S. Robin *Exact distribution of word occurences in a random sequence of letters.* Journal of Applied Probability **36** (1999), 179–193.

[22] M. Drmota *The variance of the height of digital search trees.* Acta informatica **38** (2002), 261–276.

[23] F. Eggenberger, G. Pólya *Ueber die Statistik verketter Vorgänge.* Zeitschrift für reine und angewandte Mathematik und Mechanik **1** (1923), 279–289.

[24] P. Erdös, P. Révész *On the length of the longest head-run.* Topics in Information Theory, Keszthely 1975, Colloq. Math. Soc. Janos Bolyai **16** (1977) 219–228.

[25] J.A. Fill, N. Kapur *The space requirements of m-ary search trees: distributional asymptotics for* $m \geq 27$. Prépublication, 10 pages, disponible à l'adresse http://www.mts.jhu.edu/∼fill/.

[26] P. Flajolet, J. Gabarró, H. Pekari *Analytic urns.* The Annals of Probability, **33**(3) (2005), 1200–1233.

[27] P. Flajolet, A. M. Odlyzko *Singularity analysis of generating functions.* SIAM Journal on Algebraic and Discrete Methods, **3** (1990), no. 2, 216–240.

[28] P. Flajolet et R. Sedgewick *Analytic combinatorics,* Cambridge University Press (avril 2006), 717 pages, à paraître, disponible sur le site internet de P. Flajolet.

[29] M. Franz *The integral cohomology of toric manifolds.* Proc. Steklov Inst. Math., **252** (2006), 10 pages.

[30] B. Friedman *A simple urn model.* Comm. Pure Appl. Math., **2** (1949), 59–70.

[31] W. Fulton *Introduction to toric varieties,* Princeton University Press, Princeton, 1993.

[32] R. Gouet *Strong Convergence of Proportions in a Multicolor Pólya urn.* Journal of Applied Probability **34** (1997), 426–435.

[33] R.L. Graham, D.E. Knuth, O. Patashnik *Concrete Mathematics,* second edition, Addison-Wesley, 1995.

[34] P. Hall, C.C. Heyde *Martingale Limit Theory and Its Applications,* Academic Press, 1980.

[35] J. Jabbour-Hattab *Martingales and large deviations for binary search trees.* Random Structures and Algorithms, **19**(2) (2001), 112–127.

[36] S. Janson *Functional limit theorem for multitype branching processes and generalized Pólya urns.* Stochastic Processes and Applications, **110** (2004), 177–245.

[37] S. Janson *Limit theorems for triangular urn schemes.* Probability Theory and Related Fields, **134** (2005), 417–452.

[38] S. Janson *Congruence properties of depths in some random trees (2005).* Alea Lat. Am. J. Probab. Math. Stat., A paraître

Disponible à l'adresse arXiv:math.PR/0509471

[39] H. J. Jeffrey *Chaos game representation of gene structure.* Nucleic Acids Research, **18** (1990), 2163–2170.

[40] N. L. Johnson, S. Kotz *Urn models and their application,* John Wiley, 1977.

[41] C. Jordan *Mémoire sur les équations différentielles linéaires à intégrales algébriques.* Journal für die reine und angewandte Mathematik, **84** (1878), 89–215.

[42] G. Kempf, F. Knudsen, D. Mumford, B. Saint-Donat *Toroidal embeddings.* Lecture notes in Mathematics **339**, Springer, 1973.

[43] F. Klein *Ueber die Transformation siebenter Ordnung der elliptischen Funktionen.* Mathematische Annalen, **14** (1879), 428–471 (Gesammelte Mathematische Abhandlungen, t. III, 90–135).

[44] F. Klein *Vorlesung über das Ikosaeder und die Auflösung der Gleichungen vom fünften Grade.* B. G. Teubner éd., Leipzig, (1884), 1–260 (monographie numérisée par la Bibliothèque Nationale de France).

Disponible à l'adresse http://gallica.bnf.fr/ark:/12148/bpt6k996986.item

[45] D.E. Knuth *The art of computer programming,* vol. 3, Addison-Wesley, 1973.

[46] J. Kollar *The structure of algebraic threefolds: an introduction to Mori's program.* Bulletin of the American Mathematical Society, **17**, no. 2 (1987), 211–273.

[47] W. Lew et H. M. Mahmoud *The joint distribution of elastic buckets in multiway search trees.* SIAM Journal on Computing **23**(5) (1994), 1050–1074.

[48] H.M. Mahmoud *Evolution of random search trees.* Wiley, New-York, 1992.

[49] H. M. Mahmoud et B. Pittel *Analysis of the space of search trees under the random insertion algorithm.* Journal of Algorithms **10** (1989), 52–75.

[50] H. M. Mahmoud et R. T. Smythe *Probabilistic analysis of bucket recursive trees.* Theoretical Computer Science **144** (1995), 180–205.

[51] H. Maschke *Aufstellung des vollen Formensystems einer quaternären Gruppe von* 51810 *Substitutionen.* Mathematische Annalen **33** (1889), 317–344.

[52] G. A. Miller, H. F. Blichfeldt, L. E. Dickson *Theory and applications of finite groups.* Dover Publications inc., New-York, 1916.

[53] R. Neininger et L. Rueschendorf *A general limit theorem for recursive algorithms and combinatorial structures.* The Annals of Applied Probability **14** (2004) no. 1, 378–418.

[54] J. Neveu *Arbres et processus de Galton-Watson.* Annales de l'Institut Henri Poincaré **22**(2) (1986), 199–207.

[55] F. Pan *Singularités quotient et produits symétriques.* Thèse de l'université Joseph Fourier, Grenoble (1996).

[56] V. Petrov *On the probabilities of large deviations for sums of independent random variables.* Theory of Probability and Its Applications **10** (1965), 287–298.

[57] B. Pittel *Asymptotic growth of a class of random trees.* The Annals of Probabiility **13** (1985), 414–427.

[58] G. Pólya *Sur quelques points de la théorie des probabilités.* Annales de l'Institut Poincaré **1** (1930), 117–161.

[59] V. Puyhaubert *Modèles d'urnes et phénomènes de seuils en combinatoire analytique.* Thèse de l'Ecole Polytechnique (2005).

Disponible à l'adresse http://algo.inria.fr/puyhaubert/

[60] M. Reid *Canonical* 3*-folds.* Journées de géométrie algébrique d'Angers, A. Beauville Ed., Sijthoff en Noordhoff, Aalphen aan den Rijn (1980), 273–310.

[61] M. Reid *Minimal models of canonical* 3*-folds.* Advanced Studies in Pure Mathematics, **1** (1983), Algebraic Varieties and Analytic Varieties, 131–180.

[62] A. Roy, C. Raychaudhury, A. Nandy *Novel techniques of graphical representation and analysis of DNA sequences – A review.* Journal of Biosciences, **23**, no. 1 (1998), 55–71.

[63] R. T. Smythe *Central limit theorems for urn models.* Stochastic Processes and their Applications **65** (1996), 115–137.

[64] P. Turán *On some connections between combinatorics and group theory.* Colloquia Math. Soc. Janos Bolyai, P. Erdös, A. Renyi et V. T. Sos, Ed., Vol. **4**, North holland, Amsterdam (1970), 1055–1082.

[65] H. S. Wilf *Generatingfunctionology,* Academic Press, 1994, second edition. Disponible depuis la page internet de H. S. Wilf.

[66] A. Wiman *Ueber eine einfache Gruppe von* 360 *ebenen Collineationen.* Mathematische Annalen **47** (1896), 531–556.

[67] B. Ycart, R. Abraham, J. S. Dhersin *Strong convergence for urn models with reducible replacement policy.* Prépublication soumise (2005).

Résumé

Quelques contributions au carrefour de la géométrie, de la combinatoire et des probabilités

Ce travail est la synthèse de travaux de recherches en mathématiques, dont les thèmes sont empruntés à la géométrie algébrique, la combinatoire analytique et les probabilités.

La première partie concerne les variétés algébriques complexes de dimension trois. On y présente un calcul de la cohomologie singulière de variétés toriques lisses non complètes, ainsi que la construction d'un modèle toroïdal des singularités-quotient, dont le calcul nécessite l'étude combinatoire fine de l'action des groupes finis de matrices unitaires sur le plan projectif.

La deuxième partie développe une adaptation "hybride" de la méthode de Darboux et de l'analyse des singularités pour le développement asymptotique des coefficients d'une série entière dans certains cas de frontière naturelle d'analyticité. De nombreux exemples issus de l'analyse combinatoire sont ainsi traités, dont celui de l'analyse d'algorithmes de factorisation de polynômes sur les corps finis qui sont utilisés en calcul formel et pour les codes correcteurs d'erreurs.

La troisième partie résout une conjecture sur les arbres m-aires de recherche qui sont une structure fondamentale de l'algorithmiques des ensembles de données. Le modèle considéré est un modèle d'urnes qui se généralise en la notion de processus aléatoires de Pólya dont le comportement asymptotique général est étudié.

Dans la quatrième partie, on construit un arbre aléatoire associé à la *Chaos Game Representation* utilisée en bio-mathématique et en bio-informatique du génôme. Les asymptotiques de la hauteur et de la profondeur d'insertion de ces arbres y sont établies.

Abstract

Where geometry, combinatorics and probabilities meet: some contributions

The present text is a synthesis of research papers in mathematics, dealing with algebraic geometry, analytic combinatorics and probabilities.

The first part is about three-dimensional complex algebraic varieties. It begins with the computation of the singular cohomology of non complete smooth toric varieties under some topological assumption on their fans. Afterwards, we construct a toroidal model for any quotient-singularity, whose computation requires a precise combinatorial study of the action of all finite unitary groups on the projectif plane.

The second part develops a "hybrid" adaptation of Darboux's method and of singularity analysis for the coefficients' asymptotic expansion of power series that admit a natural boundary. Numerous applications in analytic combinatorics are given, including the analysis of factorization algorithms for polynomials on finite fields that are used in symbolic computation and for error-correcting codes.

The third part gives an answer to a conjecture on m-ary search trees that are fundamental data structures in computer science used in searching and sorting. To this end, we consider them as urn processes that can be generalized to so called Pólya processes, whose general asymptotics is studied.

In the last part, we give the construction of a random tree associated with the *Chaos Game Representation* of DNA sequences used in bioinformatics and biomathematics. Results on the height's and insertion depth's asymptotics are established.