Récursivité au carrefour de la modélisation de séquences, des arbres aléatoires, des algorithmes stochastiques et des martingales

Peggy Cénac

UNIVERSITE DE BOURGOGNE

U.F.R. Sciences et Techniques

Mémoire

en vue de l'obtention de

l'Habilitation à Diriger des Recherces

Discipline : **Mathématiques Appliquées**

Présentée par

Peggy CÉNAC

le 15 Novembre 2013

Récursivité au carrefour de la modélisation de séquences, des arbres aléatoires, des algorithmes stochastiques et des martingales

---

Jury

Marc ARNAUDON,	Professeur à l'Université de Bordeaux 1,	Rapporteur,
Gérard BIAU,	Professeur à l'Université Pierre et Marie Curie,	Rapporteur,
Hervé CARDOT,	Professeur à l'Université de Bourgogne,	Examinateur,
Amaury LAMBERT,	Professeur à l'Université Pierre et Marie Curie,	Examinateur,
Eva LOCHERBACH,	Professeure à l'Université de Cergy-Pontoise,	Examinatrice,
Gilles PAGÈS,	Professeur à l'Université Pierre et Marie Curie,	Rapporteur,
Sylvie ROELLY,	Professeure à l'Université de Potsdam,	Présidente.

Remerciements

Je commence par un grand merci à Marc Arnaudon, Gérard Biau et Gilles Pagès pour leurs rapports et l'intérêt qu'ils ont porté à mon travail. C'est un honneur pour moi. Merci également pour le temps que vous consacrez à mon jury. Je tiens aussi à remercier chaleureusement l'ensemble des membres qui composent mon jury d'habilitation : Amaury Lambert, Eva Löcherbach et Sylvie Roelly. J'adresse également un merci particulier à mon « chef d'équipe » Hervé Cardot, pour sa disponibilité et pour m'avoir fait confiance dès mon arrivée à Dijon. J'ai beaucoup appris grâce à toi. Merci pour tes conseils et ta relecture minutieuse.

Bernard Bercu, Guy Fayolle et Jean-Marc Lasgouttes ont guidé mes premiers pas dans la recherche. Leur exigence mathématique, leur curiosité et leur enthousiasme, accompagnés d'une grande modestie m'ont permis de débuter dans les meilleures conditions. Ce manuscrit vous doit beaucoup. Je remercie aussi toutes les personnes que j'ai cotoyées toutes ces années, que ce soit à l'INRIA, au LMV, à Paris 5. Je ne citerai personne pour ne pas en oublier.

Je remercie tous les membres passés ou actuels de l'équipe SPAN, avec qui c'est un plaisir de travailler au quotidien. Merci également à tous les membres de l'IMB, avec une mention particulière pour le personnel administratif : Francis, Pierre et tout le secrétariat, sans oublier Véronique, Renée et Muriel ; il est parfois facile d'oublier que si tout fonctionne, c'est justement parce qu'ils sont là. Dans la série des « merci » du quotidien, j'ajoute aussi mes amis Jean-Charles, Dominique et Marc : votre soutien sans faille permet d'affronter n'importe quelle tempête.

Je remercie aussi tous mes co-auteurs, avec qui j'ai eu la chance et le plaisir de travailler, pour les nombreux échanges passionnants et enrichissants : Pierre-André Zitt, Véronique Maume-Deschamps, Clémentine Prieur, Samuel Herrmann, Jean-Marie Monnez, Pierre Vallois, Khalifa Es-Sebaiy, Stéphane Loisel et Stéphane Ginouillac. Parmi mes co-auteurs, les « mousquetaires » ont une place toute particulière. Merci pour toutes ces années de travail commun, votre enthousiasme, ces bulles de moments privilégiés à chercher ensemble. Votre manière d'être dans la vie, qu'elle soit scientifique ou non, ont beaucoup déteint sur moi. Merci Brigitte, Nicolas et Frédéric !

Arnaud, tu tiens pour moi le rôle de « grand frère » scientifique. Ton soutien et ton amitié n'ont pas de prix. Merci !

Enfin, mes pensées vont à mes proches pour leur patience, leur soutien sans faille, et tout ce qu'ils m'apprennent au quotidien.

Table des matières

Chapitre 1

Introduction

Ce document présente une synthèse de mes travaux de recherche. Ceux-ci se répartissent sur trois thèmes : structures aléatoires discrètes construites à partir de sources dynamiques, algorithmes de gradient stochastique et théorème de la limite centrale presque-sûr. Dans ce mémoire, je décris l'évolution de mes travaux depuis ma thèse tout en mettant en relief le fil conducteur commun. Mon travail de recherche se situe à l'intersection des systèmes dynamiques dans l'analyse statistique de séquences, de l'analyse d'algorithmes dans les arbres aléatoires et des processus stochastiques discrets. Les résultats établis ont des applications dans des domaines variés allant des séquences biologiques aux modèles de régression linéaire, processus de branchement, en passant par la statistique fonctionnelle et les estimations d'indicateurs de risque appliqués à l'assurance.

Mon travail de thèse (2003-2006) au sein de l'équipe PREVAL de l'INRIA Rocquencourt portait sur la modélisation de séquences biologiques ainsi que sur des propriétés de convergence de martingales vectorielles. L'objet de la première partie était une représentation de séquences en système dynamique par un jeu du chaos, la CGR (de l'anglais *Chaos Game Representation*). L'originalité tenait dans l'utilisation de cette représentation (appliquée pour la première fois aux séquences d'ADN par Jeffrey [1]) pour construire des distances permettant de quantifier des différences de structure de séquences biologiques. La masse d'information accumulée, suite aux différents séquençages d'espèces, a soulevé de nombreux problèmes de stockage : il était devenu nécessaire de développer des outils permettant de retrouver rapidement l'information *pertinente* dans cet amas de données. La première partie de ma thèse formalise cette représentation en jeu du Chaos et permet de répondre à la question « Comment utiliser la CGR et l'information qu'elle contient ? ». J'ai utilisé cette représentation pour caractériser la structure de dépendance dans une séquence et pour construire des arbres taxonomiques.

L'idée principale sur laquelle reposent les preuves des résultats énoncés est d'utiliser la *récursivité* sous-jacente à la représentation en jeu du Chaos pour exhiber des martingales et des séries génératrices que l'on sait contrôler. Le caractère récursif de la structure étudiée est de façon beaucoup plus générale à la base de la plupart des preuves des articles que je synthétise ici ; il peut se traduire en « *méthodes forward* » et « *méthodes backward* ». En décrivant l'état d'un processus à l'instant $n+1$ en fonction de son état à l'instant n, on peut faire apparaître des *invariants*, comme des martingales, en calculant

1. H.J. Jeffrey. « Chaos Game Representation of gene structure ». Dans : *Nucleic Acid. Res* 18 (1990), p. 2163–2170.

des espérances conditionnelles (méthodes *forward*). On décompose ensuite le processus étudié en un invariant sur lequel on dispose de résultats de convergence et on contrôle le reste. Dans un raisonnement *backward*, on décrit la situation à l'etape $n+1$ comme reproduisant la situation à l'étape n.

La seconde partie de ma thèse portait sur l'étude d'arbres aléatoires, les *arbres-CGR*, liés à des algorithmes de compression de données. Avec Brigitte Chauvin, Nicolas Pouyanne et Stéphane Ginouillac (LMV, Université de Versailles Saint-Quentin-en-Yvelines), nous avons établi le comportement asymptotique de certains paramètres (hauteur, niveau de saturation, niveau d'insertion) d'un arbre digital de recherche construit à partir des suffixes successifs d'un même mot infini.

Les résultats obtenus utilisent tant des outils purement probabilistes (martingales liées à la construction récursive de l'arbre, inégalités de concentration, théorème de Borel-Cantelli) que des outils d'analyse complexe (comportement du di-logarithme complexe, théorème de transfert basé sur l'analyse de singularités) en lien avec certaines séries génératrices. Les arbres digitaux poussent en insérant successivement de nouveaux suffixes. La vitesse de croissance possède des liens étroits avec les propriétés des temps d'attente d'occurrences de mots ainsi que leur auto-corrélation.

Ce premier travail sur les paramètres de cet arbre-CGR a été le début d'une fructueuse coopération avec Brigitte Chauvin et Nicolas Pouyanne qui se prolonge bien au-delà du travail de thèse.

L'objet *martingale* étant récurrent dans mes preuves, j'ai souhaité obtenir de nouvelles propriétés de convergence pour des martingales vectorielles. Dans la dernière partie de ma thèse, en collaboration avec mes deux directeurs de thèse, Bernard Bercu (IMB, Université de Bordeaux 1) et Guy Fayolle (INRIA Rocquencourt), nous avons prouvé la convergence des moments de martingales vectorielles dans le théorème de la limite centrale presque-sûr (TLCPS). Ce travail est la généralisation au cas vectoriel de celui qu'avait obtenu Bernard Bercu[2] dans le cas scalaire. Ces propriétés sont appliquées aux erreurs cumulées d'estimation et de prédiction dans des modèles stables de type processus auto-régressifs linéaires ou processus de branchement avec immigration.

Les tests de structure markovienne d'ordre fixe et déterministe effectués dans la première partie de ma thèse m'ont amenée à travailler sur un mécanisme plus général de production de mots permettant des corrélations de longue portée entre symboles et ainsi une modélisation plus réaliste : *la source dynamique*. En informatique, toute information se code sous forme de texte, que ce soit un texte initialement écrit en langue naturelle, une séquence biologique (génétique, protéique), musicale, le flux de données dans un réseau ou encore une base de données. L'algorithmique du texte est une branche de l'algorithmique qui vise à un traitement efficace de ces données. Il faut construire des

2. B. Bercu. « On the convergence of moments in the almost sure central limit theorem for martingales with statistical applications ». Dans : *Stochastic Processes and their applications* 111 (2004), p. 157–173.

structures qui permettent de rechercher et trier des « mots » et de déterminer le comportement des algorithmes correspondants. Le traitement doit être fiable, rapide et ne pas prendre trop de place en mémoire. Pour une source dynamique générale, l'analyse probabiliste des structures arborescentes construites sur ses mots s'avère beaucoup plus délicate et nécessite la confrontation entre des points de vue variés et complémentaires.

En collaboration avec Brigitte Chauvin, Nicolas Pouyanne (LMV, Université de Versailles Saint-Quentin-en-Yvelines) et Frédéric Paccaut (LAMFA, Université de Picardie Jules Verne) dans les articles [11, 12], nous nous sommes intéressés à un type particulier de sources, extensions « naturelles » de chaînes de Markov : les chaînes de Markov à mémoire variable, appelées aussi *VLMC* (de l'anglais *Variable Length Markov Chains*). Les VLMC étaient déjà abondamment utilisées dans des domaines d'applications variés, mais encore mal comprises du point de vue de la théorie « analytique » de l'information. Ce modèle est un bon compromis : suffisamment général pour pouvoir représenter des sources diverses et unifier le traitement de sources naturelles (sources sans mémoire et chaînes de Markov), suffisamment structuré pour pouvoir être précisément étudié, via en particulier l'opérateur de transfert du système dynamique. Dans l'article [11] nous explicitons une source dynamique permettant de produire des VLMC. Puis, nous nous intéressons à deux exemples particuliers pour lesquels nous donnons une condition nécessaire et suffisante d'existence et unicité de mesure invariante.

La manière dont les mots sont produits a une grande influence sur le comportement probabiliste des algorithmes ou des principales structures de données arborescentes de type dictionnaire associées (arbre binaire de recherche, arbre digital, arbre des suffixes, pour la compression). Une fois de plus la structure récursive de ces algorithmes est exploitée dans nos preuves. Ces arbres aléatoires dans lesquels on stocke des données sont connus dans la littérature pour « pousser » à vitesse logarithmique en fonction du nombre de données que l'on stocke. Dans l'article [12], il y a deux exemples d'arbres explicites, des *tries* des suffixes, dont la hauteur ou le niveau de saturation ne grandissent pas logarithmiquement. Ces deux exemples sont construits à partir d'un mot généré par une source VLMC.

Egalement à partir d'un mot généré par une VLMC, avec Brigitte Chauvin, Samuel Herrmann (IMB, Université de Bourgogne) et Pierre Vallois (Institut Elie Cartan, Université de Lorraine) nous avons étudié la marche aléatoire construite en sommant les codages des lettres du mot. Cette marche n'est en général pas markovienne. Elle est dite *persistante*. On montre que renormalisée, elle converge vers un processus continu de type processus zigzag.

Lors de mon séjour au LMV, des discussions avec Mariane Pelletier m'ont donné la curiosité de me pencher sur une autre structure récursive naturelle : les algorithmes de gradient. Pour calculer un minimum ou un zéro de fonction, lorsque la fonction a de nombreux minima locaux proches les uns des autres, les algorithmes déterministes risquent de rester piégés. De plus, très souvent en modélisation, la fonction dont on cherche un zéro ou un minimum n'est connue qu'à une perturbation près. Les algorithmes stochas-

tiques sont des estimateurs particulièrement bien adaptés en grande dimension. Mon premier article sur les algorithmes stochastiques utilise des outils développés pendant ma thèse pour établir la convergence des moments de tout ordre dans le TLCPS [8].

J'ai ensuite utilisé de tels algorithmes dans différents contextes : estimation de courbe médiane, de médiane conditionnelle, minimisation de risque sous contrainte en assurance, classification non-supervisée. Une fois encore les martingales sont bien adaptées à ces structures récursives et sont des outils précieux pour obtenir le comportement asymptotique des algorithmes.

Avec Hervé Cardot (IMB, Université de Bourgogne) et Pierre-André Zitt (LAMA, Université Paris-Est), nous avons proposé un algorithme séquentiel extrêmement rapide de type gradient stochastique pour estimer la médiane géométrique. La médiane géométrique est une extension naturelle de la médiane pour des vecteurs aléatoires à valeurs dans des espaces vectoriels normés. La médiane géométrique est, contrairement à la moyenne, robuste et donc peu sensible aux points atypiques. Il est généralement difficile de détecter de tels points lorsqu'on dispose de grands échantillons de variables à valeurs dans des espaces de grande dimension ou de dimension infinie.

Avec Hervé Cardot et Jean-Marie Monnez (Institut Elie Cartan, Université de Lorraine), nous avons ensuite développé une extension directe de cet algorithme pour la classification non supervisée en introduisant un critère de type k-médianes. Cet algorithme peut s'interpréter comme une variante des k-means (à la MacQueen).

Tous les articles autour de la médiane sont illustrés sur des courbes d'audience relevées seconde par seconde par Médiamétrie au cours d'une journée.

En collaboration avec Clémentine Prieur (LJK, Université Joseph Fourier) et Véronique Maume-Deschamps (ISFA, Université de Lyon 1), nous avons utilisé un algorithme stochastique pour estimer le minimum de certains indicateurs de risque, sous contrainte d'un capital initial fixé à partager entre les différentes filiales d'une assurance. Nous considérons des indicateurs de risque de processus vectoriels qui prennent en compte des dépendance entre les filiales ainsi que certaines dépendances temporelles. Cette minimisation correspond à une allocation de réserve optimale.

Enfin, en continuité avec le dernier chapitre de ma thèse sur le théorème de la limite centrale presque-sûr, nous avons démontré avec Khalifa Es-Sebaiy (ENSA de Marrakech, Université Cadi Ayyad) que la suite des estimateurs des moindres carrés du paramètre d'un processus d'Ornstein-Uhlenbeck fractionnaire satisfait le TLCPS. Nous avons considéré le cas d'une observation du processus à temps continu aussi bien qu'une observation discrétisée.

La première partie de ce mémoire porte sur la modélisation de séquences, les chaînes de Markov à mémoire variable ainsi que les arbres et marches aléatoires associés. Elle rassemble les travaux [9, 10, 11, 12, 13, 14]. La seconde partie est consacrée aux algorithmes stochastiques des articles [2, 3, 4, 5, 15, 16]. Enfin la dernière partie traite du TLCPS en rassemblant les papiers [1, 7, 8, 17].

Liste des travaux présentés

Articles dans des revues avec comité de lecture

[1] B. Bercu, P. C. et G. Fayolle. « On the Almost Sure Central Limit Theorem for Vector Martingales: Convergence of Moments and Statistical Applications ». Dans : *Journal of Applied Probability* 46 (2009), p. 151–169.

[3] H. Cardot, P. C. et J-M. Monnez. « A fast and recursive algorithm for clustering large datasets with k-medians ». Dans : *Computational Statistics & Data Analysis* 56.6 (2012), p. 1434–1449.

[4] H. Cardot, P. C. et P-A. Zitt. « Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. » Dans : *Bernoulli* 19.1 (2013), p. 18–43.

[5] H. Cardot, P C. et P.-A. Zitt. « Recursive estimation of the conditional geometric median in Hilbert spaces ». Dans : *Electron. J. Statist.* 6 (2012), p. 2535–2562. issn : 1935-7524. doi : 10.1214/12-EJS759.

[7] P. C. « Almost sure properties of Weighted Martingales Transforms with applications to Prediction for Linear Regression Models ». Dans : *Probability and Mathematical Statistics* 23.1 (2003), p. 61–76.

[8] P. C. « On the convergence of moments in the almost sure central limit theorem for stochastic approximation algorithms ». Dans : *ESAIM Probability and Statistics* 17 (2013), p. 179–194.

[9] P. C. « Test on the Structure of Biological Sequences via Chaos Game Representation ». Dans : *Statistical Applications in Genetics and Molecular Biology* 4.1 (2005). 34 pages.

[10] P. C., B. Chauvin, S. Herrmann et P. Vallois. « Persistent random walks, variable length Markov chains and piecewise deterministic Markov processes ». Dans : *Markov Processes and Related Fields* 19.1 (2013), p. 1–50.

[11] P. C., B. Chauvin, F. Paccaut et N. Pouyanne. « Context trees, variable length Markov chains and dynamical sources ». Dans : *Séminaire de Probabilités* XLIV (2012).

[12] P. C., B. Chauvin, F. Paccaut et N. Pouyanne. « Uncommon Suffix Tries ». Dans : *A paraître dans Random Structures and Algorithms* (2013).

[13] P. C., B. Chauvin, N. Pouyanne et S. Ginouillac. « Digital Search Trees and Chaos Game Representation ». Dans : *ESAIM Probability and Statistics* 13 (2009), p. 15–37.

[16] P. C., V. Maume-Deschamps et C. Prieur. « Some multivariate risk indicators; Minimization by using a Kiefer-Wolfowitz approach to the mirror stochastic algorithm ». Dans : *Statistics and Risk Modeling* 29 (2012), p. 47–71.

Acte de Congrès

[2] H. Cardot, P. C. et M. Chaouch. « Stochastic approximation to the multivariate and the functional median ». Dans : *COMPSTAT 2010*. Paris, France., 2010.

Articles Soumis

[15] P. C., S. Loisel, V. Maume-Deschamps et C. Prieur. « Risk Indicators with several lines of business: comparison, asymptotic behavior and applications to optimal reserve allocation ». 21 pages.

[17] P. C. et K. Es-Sebaiy. « Almost sure central limit theorems for random ratios and applications to LSE for fractional Ornstein-Uhlenbeck processes ». 15 pages.

Rapports de Recherche

[14] P. C., G. Fayolle et J.-M. Lasgouttes. « Dynamical Systems in the Analysis of Biological Sequences ». Dans : *Rapport de recherche INRIA* 5351 (oct. 2004). 47 pages.

Thèse

[6] P. Cénac. « Etude statistique de séquences biologiques et convergence de martingales ». Thèse de doct. Université Toulouse III, 2006.

Chapitre 2

Modélisation de séquences et structures aléatoires discrètes

Ce chapitre est une présentation des travaux [9, 10, 11, 12, 13, 14] listés au chapitre 1.

Sommaire

Avant toute chose, je commence par une petite mise en garde sur les notations. Les notations (X_n) et (U_n) ne sont pas homogènes sur le chapitre. J'ai choisi de garder des notations cohérentes avec les articles auxquels je fais référence, ce qui induit des variations entre sections. La suite (X_n) désignera tantôt une suite de lettres, tantôt une suite de points ou une suite de longueur de branches. La notation (U_n) pourra correspondre à une suite de lettres indexée par $\mathbb{N}$ dans certaines sections, ou indexée par $\mathbb{Z}$ dans d'autres. Ces notations sont toutes reprécisées en début de chaque section.

2.1 Introduction

En théorie de l'information, une source stochastique est un mécanisme aléatoire qui émet, à chaque instant, un symbole choisi dans un alphabet fini. Le plus souvent en in-

formatique, on se limite aux sources les plus simples : sources sans mémoire, où chaque symbole est produit indépendamment de l'histoire, ou chaînes de Markov, où la production d'un symbole ne dépend que d'une partie bornée de l'histoire. Une analyse plus réaliste des algorithmes (texte, recherche, tri) doit prendre en compte des modèles de sources plus généraux, où la possible corrélation entre symboles émis est plus forte, avec une portée plus étendue. Une chaîne de Markov à mémoire variable (VLMC, de l'anglais *Variable Length Markov Chain*) est une extension naturelle des chaînes de Markov. Ce modèle fournit un compromis intéressant : suffisamment général pour pouvoir représenter des sources diverses, suffisamment structuré pour pouvoir être précisément étudié.

La première partie de ce chapitre résume un travail que j'ai effectué en thèse dans les articles [12, 7]. Il s'agit d'utiliser un mécanisme de représentation de séquences par jeu du Chaos pour caractériser l'ordre de dépendance d'une chaîne de Markov. La bijection entre les points de la représentation et l'historique de toute la séquence parcourue permet d'éviter les méthodes basées sur le comptage de mots. En d'autres termes, plutôt que de s'intéresser au processus décrivant la suite des lettres $(u_n)_{n\geq 1}$ le long d'une séquence, il peut s'avérer plus utile de s'intéresser à la suite des historiques $(U_n)_{n\geq 1}$ avec $U_n \stackrel{\text{def}}{=} u_1 \dots u_n$. Quel que soit l'ordre de dépendance Markovienne sur la suite $(u_n)_{n\geq 1}$, le processus $(U_n)_{n\geq 1}$ reste une chaîne de Markov d'ordre un.

On effectue ce même changement d'échelle lorsque l'on étudie les chaînes de Markov d'ordre variable dans la section 2.3. Plutôt que d'étudier directement la chaîne avec ses ordres de dépendance variables, on s'intéresse au processus contenant l'historique. Comme la dépendance peut être d'ordre infini, l'historique est un mot infini à gauche. La VLMC est aussi une chaîne de Markov d'ordre un sur l'ensemble des mots infinis à gauche. La section 2.3 définit le modèle VLMC et montre que ce processus peut être produit par une source dynamique explicite. Le formalisme des sources dynamiques introduit par Clément, Flajolet et Vallee [5]) est rappelé dans la section 2.3.2.

Les deux dernières sections sont des études du comportement de structures aléatoires construites à partir de chaînes de Markov ou VLMC. Dans la section 2.4, l'objet d'étude est une marche aléatoire persistante, en dimension un, dont les incréments sont générés par une VLMC. La dernière partie contient des résultats asymptotiques sur le comportement d'arbres aléatoires de compression construits à partir de VLMC.

2.2 Représentation par jeu du Chaos

2.2.1 Définition

Le développement de la génétique et l'accélération des programmes de séquençage provoquent des besoins importants de représentation et de stockage de l'ADN. La *Chaos Game Representation* (CGR) est à la fois une méthode de représentation graphique et un outil de stockage. Cette méthode itérative fut appliquée pour la première fois aux séquences d'ADN par Jeffrey [29].

L'algorithme de représentation est défini de la façon suivante. Pour un alphabet fini $\mathcal{A}$ constitué de d lettres, pour un borélien borné $S \subset \mathbb{R}^q$, où q est un entier positif, on définit la collection de fonctions affines $\{T_u,\ u \in \mathcal{A}\}$, liées à un facteur de contraction réel ρ et à une famille $\{\ell_u,\ u \in \mathcal{A}\}$ d'éléments de $\mathbb{R}^q$ avec $0 < \rho < 1$, par

$$T_u(x) \stackrel{\text{def}}{=} \rho\big(x + \ell_u\big),\ u \in \mathcal{A},\ x \in S,$$

où, pour tout $u \in \mathcal{A}$, $T_u(S) \subset S$ et

$$T_u(S) \cap T_v(S) = \emptyset,\ \forall\,(u,v) \in \mathcal{A}^2,\ u \neq v.$$

Définition 2.2.1. *Soit $(U_n)_{1\leq n\leq N}$ une suite de mots construits à partir de l'alphabet $\mathcal{A}$: $U_n = u_1 \dots u_n$, avec pour tout $n \in \{1, \dots, N\}$, $u_n \in \mathcal{A}$. La CGR de la séquence U_N sur l'ensemble S est la suite de points $(X_n)_{1\leq n\leq N}$, définie par une position initiale arbitraire X_0 et par la relation récursive, pour $0 \leq n \leq N-1$,*

$$X_{n+1} \stackrel{\text{def}}{=} T_{u_{n+1}}(X_n) = \rho\big(X_n + \ell_{u_{n+1}}\big). \tag{2.1}$$

À partir d'une suite de symboles dans un alphabet fini, la CGR associe une trajectoire dans S. La définition de Jeffrey est le cas particulier de la CGR obtenue en choisissant l'alphabet $\mathcal{A} = \{A, C, G, T\}$ des nucléotides de l'ADN, $S = [0,1[^2$, $\rho = 1/2$. De plus, les 4 lettres sont situées aux quatre sommets du carré unité, avec

$$\ell_A = (0,0), \quad \ell_C = (0,1), \quad \ell_G = (1,1), \quad \ell_T = (1,0).$$

La relation (2.1) s'écrit alors, pour $0 \leq n \leq N-1$,

$$X_{n+1} = \frac{1}{2}(X_n + \ell_{u_{n+1}}),$$

avec $X_0 = (\frac{1}{2}, \frac{1}{2})$. La CGR peut bien entendu être définie sur d'autres ensembles que le carré.

La Figure 2.1 illustre la construction de la CGR pour le mot $ATGCGAGTGT$. On peut visualiser sur la Figure 2.2 deux exemples de CGR de séquences d'ADN de longueur 70 000.

On associe au mot $w = u_1 \dots u_n$ l'ensemble Sw défini par

$$Sw \stackrel{\text{def}}{=} \sum_{k=1}^{n} \rho^{n-k+1} \ell_{u_k} + \rho^n S, \tag{2.2}$$

comme l'illustre la Figure 2.3 pour $S = [0,1[^2$. Il est équivalent de compter le nombre de points dans le carré Sw ou de compter le nombre d'occurrences du mot w dans la séquence.

FIGURE 2.1: Représentation en jeu du Chaos des 10 premiers nucléotides du gène threonine « thrA » de *E. Coli* : ATGCGAGTGT. Les coordonnées de chaque nucléotide sont calculées récursivement à partir du point initial situé au centre du carré. Le point 3 correspond au premier mot de 3 lettres ATG. Il est situé dans le carré correspondant. Le second mot de 3 lettres TGC correspond au point 4, etc.

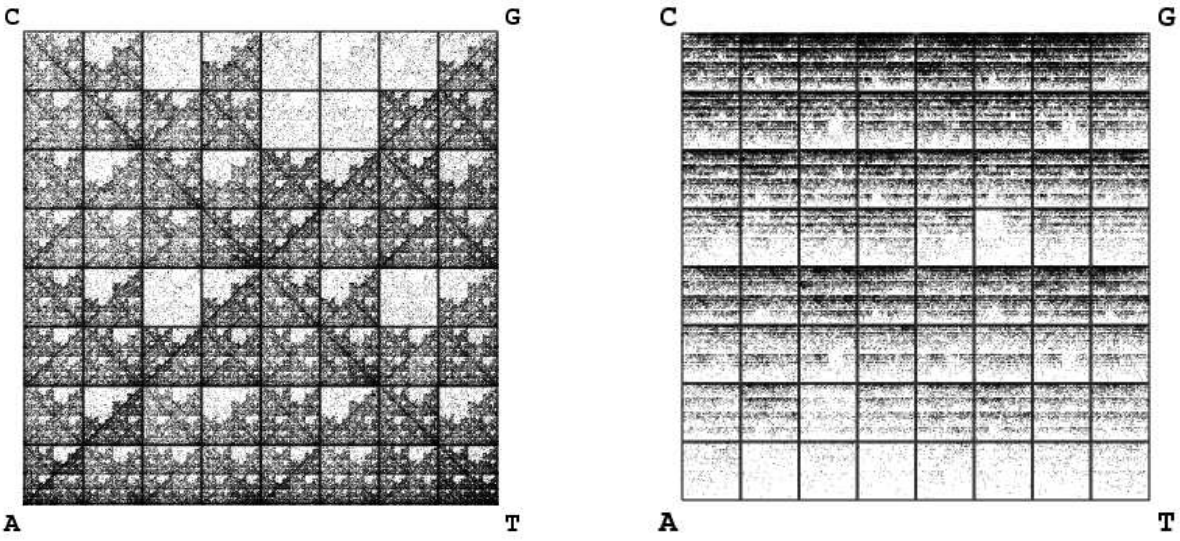


FIGURE 2.2: Chaos Game Representation des 70 000 premiers nucléotides du Chromosome 2 d'*Homo Sapiens* à gauche, et de *Streptomyces Coelicolor* sur la droite.

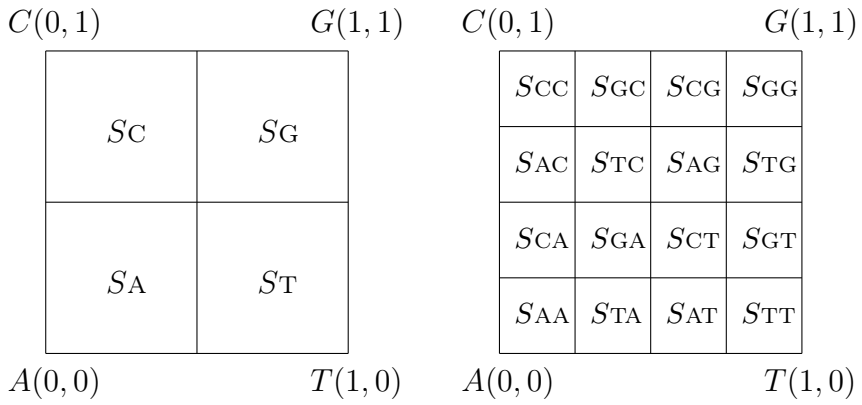


FIGURE 2.3: Définitions des carrés correspondant aux nucléotides (à gauche) et aux dinucléotides (à droite) pour la CGR sur $[0,1[^2$.

La suite (U_n) étant fixée, la CGR (X_n) est elle aussi déterministe. Lorsque l'on met un modèle aléatoire sur la séquence d'ADN, la CGR devient elle aussi aléatoire *de facto*. Il est clair que la suite (X_n) de points définissant la CGR forme une chaîne de Markov d'ordre 1, quel que soit le niveau de dépendance dans le mot aléatoire $U = u_1 u_2 \ldots$. En effet, tout le passé de la séquence à un instant donné n peut être retrouvé à partir de la valeur de X_n : l'équation (2.1) entraîne, par construction, que $X_n \in Su_n$. On identifie ainsi u_n. A partir de la relation de récurrence (2.1) : $X_{n-1} = \frac{X_n}{\rho} - \ell_{u_n}$. X_{n-1} est donc déterminé à partir de X_n, et on itère ainsi jusqu'au point initial X_0.

Ainsi, les tribus $\sigma\big(X_n\big)$ et $\sigma\big(X_0, (u_k)_{1\leq k\leq n}\big)$ sont égales et

$$\sigma\big(X_n\big) = \sigma\big(X_1, \ldots X_n\big).$$

Conditionner par rapport au point X_n équivaut à conditionner par rapport à tout le passé. Travailler avec la suite (X_n) revient donc à faire un changement d'échelle sur l'ordre de dépendance dans la suite observée. On utilise cette propriété pour caractériser la structure de dépendance.

2.2.2 Famille de tests asymptotiques

J'ai proposé dans [7] une famille de tests asymptotiques permettant de déterminer l'ordre de dépendance d'une chaîne de Markov en utilisant la représentation CGR.

La correspondance bijective entre l'ensemble des séquences possibles et l'ensemble des points de la CGR suggère de noter, pour un mot $w = v_1 \ldots v_m$ et pour un ensemble $B \subset S$,

$$Bw \stackrel{\text{def}}{=} T_{v_m} \circ \cdots \circ T_{v_1}(B).$$

Cette définition coïncide clairement avec (2.2) lorsque $B = S$. Autrement dit, un point X_n de la CGR se trouve dans Bw si $X_{n-m} \in B$ et $u_{n-m+1} \ldots u_n = v_1 \ldots v_m$.

Supposons maintenant que la suite de mots (U_n) soit stationnaire et ergodique, et notons π la mesure limite invariante de sa CGR sur S. La propriété de π énoncée ci-dessous permet de caractériser une suite sans mémoire.

Propriété 2.2.2. *La suite aléatoire stationnaire (u_n) est constituée de lettres indépendantes et identiquement distribuées si et seulement si*

$$\pi(Bu) = \pi(B)\pi(Su), \quad \forall u \in \mathcal{A}\ \forall B \subset S.$$

On étend cette caractérisation au cas Markovien :

Propriété 2.2.3. *Le processus des lettres aléatoires (u_n) est une chaîne de Markov d'ordre m si et seulement si,*

$$\pi(Sw)\pi(Bwu) = \pi(Swu)\pi(Bw), \quad \forall B \subset S, \quad \forall w \in \mathcal{A}^m, \quad \forall u \in \mathcal{A}.$$

En particulier le rapport $\pi(Bwu)/\pi(Bw)$ ne dépend pas de B.

On note respectivement H_0, H_m et H les hypothèses suivantes : « *$U_n = u_1 \dots u_n$ est une suite de lettres indépendantes et identiquement distribuées* », « *Le processus des lettres (u_n) est une chaîne de Markov d'ordre m* » et « *Le processus des lettres (u_n) est stationnaire ergodique* ».

A partir des caractérisations précédentes, on construit un test de H_0 contre $H \setminus H_0$, puis de H_m contre $H \setminus H_m$. La notation $\hat{\pi}_n$ désigne la mesure empirique associée à la CGR $(X_n)_{n\geq 0}$: pour tout ensemble B,

$$\hat{\pi}_n(B) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{X_j \in B\}}.$$

On définit la statistique R_n inspirée par la propriété 2.2.3 :

$$R_n(B,w,u) \stackrel{\text{def}}{=} \sqrt{n}\frac{\hat{\pi}_n(Sw)\hat{\pi}_n(Bwu) - \hat{\pi}_n(Swu)\hat{\pi}_n(Bw)}{\sqrt{\hat{\pi}_n(Sw)\hat{\pi}_n(Swu)\hat{\pi}_n(Bw)}}.$$

De plus, $q_\alpha(d)$ désigne le $(1-\alpha)$-quantile de la loi de chi-deux $\chi^2(d)$.

Théorème 2.2.4 (Test de caractérisation markovienne).
(i) Pour une partition $\mathcal{P}$ de S, avec $|\mathcal{P}| = K > 1$, l'ensemble

$$\left\{ \sum_{wu\in\mathcal{A}^m\times\mathcal{A},\ B\in\mathcal{P}} R_n^2(B,w,u) > q_\alpha\Big[d^m(d-1)(K-1)\Big] \right\}$$

est une région de rejet d'un test de niveau asymptotique α, de H_m contre $H \setminus H_m$.

(ii) Sous $H \setminus H_m$, et en supposant qu'il existe $B \in \mathcal{P}$ tel que

$$\pi(Sw)\pi(Bwu) \neq \pi(Bw)\pi(Swu),$$

le test construit à partir de la région de rejet (2.2.4) est asymptotiquement consistant.

Avec la même méthode, on peut construire des tests d'adéquation à une loi, basés sur des statistiques analogues à celles utilisées pour les tests de structure.

La preuve repose sur le théorème de la limite centrale pour martingales vectorielles et sur le théorème de Cochran. L'idée dans le cas de la caractérisation « sans mémoire » est de décomposer la différence

$$D_n(B,v) \stackrel{\text{def}}{=} \sqrt{n}\,(\hat{\pi}_n(Bv) - \hat{\pi}_n(B)\hat{\pi}_n(Sv))$$

sous la forme

$$D_n(B,v) = \frac{1}{\sqrt{n}} M_n(B,v)\,(1+\eta_n)\,,$$

avec

$$M_n(B,v) \stackrel{\text{def}}{=} \sum_{j=1}^{n} \varepsilon_j(v) V_{j-1}(B), \quad \varepsilon_j(v) \stackrel{\text{def}}{=} \mathbb{1}_{\{u_j=v\}} - \mathbb{P}(u_j = v), \quad V_j(B) \stackrel{\text{def}}{=} \mathbb{1}_{\{B\}}(X_j) - \pi(B)$$

et où $(\eta_n)_n$ est une suite de variables aléatoires tendant en probabilités vers 0. La suite $(M_n(B,v))$ est une martingale adaptée à la filtration naturelle $\sigma(X_n)$. Les propriétés bien connues pour les martingales permettent d'obtenir le théorème de test de caractérisation. L'idée est la même dans le cas de la caractérisation Markovienne, même si la décomposition demande un peu plus de minutie.

Pour éviter un choix rigide d'une partition unique, on peut proposer une généralisation du test précédent à une collection de partitions de S. L'idée est inspirée de la méthode de Bonferroni décrite dans Baraud, Huet et Laurent [2]. Pour une collection finie $\Pi = \{\mathcal{P}_1, \ldots \mathcal{P}_p\}$, avec $K_j \stackrel{\text{def}}{=} |\mathcal{P}_j|$, H_m est rejetée dès que l'une des partitions $\mathcal{P}_j$ la rejette, *i.e.*

$$Z_{(m)} \stackrel{\text{def}}{=} \sup_{1\leq j\leq p} \left\{ \sum_{B\in\mathcal{P}_j,\ v\in\mathcal{A}} R_n^2(B,w,v) - q_{\alpha_j}\Big[d^m(d-1)(K_j-1)\Big] \right\} > 0.$$

Il reste à choisir le niveau α_j de chaque partition $\mathcal{P}_j$ pour obtenir un niveau global égal à α pour la collection Π, à l'aide d'une procédure bien définie.

En pratique, le choix de la partition est crucial. Deux partitions de même taille n'ont pas nécessairement le même taux de rejet empirique. On constate à partir de simulations numériques que le taux de rejet de chaînes de Markov de longue dépendance augmente

lorsque le nombre d'ensembles constituant la partition augmente. Les partitions les plus grosses sont souvent les plus robustes mais leur niveau converge plus lentement vers le niveau asymptotique. La puissance du test construit à partir d'une collection de partitions est comparable à la puissance du meilleur des tests construits sur les partitions elles-mêmes. En mélangeant plusieurs partitions, on peut même parfois améliorer la performance du test. Considérer une collection peut être une solution au problème crucial du choix d'une partition unique.

Les statistiques basées sur les points de la CGR sont plus générales que les statistiques des tests classiques basés sur le comptage de mots : ces tests classiques ne sont pas consistants contre toutes les alternatives stationnaires ergodiques contrairement aux tests basés sur la CGR. Pour illustrer ce point, on peut considérer la famille suivante de chaînes de Markov *mixées* d'ordre $m > 1$: dans un premier temps, on génère m chaînes de Markov indépendantes $U^{(1)}, \ldots, U^{(m)}$ d'ordre 1 comme ci-dessus ; puis la séquence finale U est obtenue par l'agrégation

$$u_{km+i} = u_k^{(i)}, \text{ pour tout } k \geq 0 \text{ et } 1 \leq i \leq m.$$

U est une chaîne de Markov d'ordre m, où chaque nucléotide u_i ne dépend seulement que du nucléotide u_{i-m}, et est indépendant des u_{i-k} pour $1 \leq k < m$. Cependant, du point de vue de la statistique classique du test de Pearson, U se comporte comme une suite de lettres i.i.d. Au contraire, les tests basés sur la CGR avec des partitions ne correspondant pas à du comptage de mots détectent la régularité du modèle.

2.2.3 Signature génomique et arbres taxonomiques

Deschavanne et al. [15] utilisent la CGR pour caractériser et classifier les espèces. Ils utilisent les fréquences d'apparition de tous les mots, qui forment alors une « signature génomique ». En effet, une analyse de fréquences le long d'un gène permet de mettre en évidence des similarités et des différences entre les espèces. Dans leur étude, la CGR n'est qu'un outil permettant de représenter ces signatures sous forme d'images, dans lesquelles les zones les plus foncées correspondent aux mots les plus fréquents. De plus, ils affirment que cette spécificité de la signature génomique, qui permet de « caractériser le style d'écriture », est une conséquence de l'action de l'environnement d'une part, et des structures de contraintes d'autre part. La figure 2.6 permet de visualiser plusieurs CGR pour des espèces différentes.

Karlin et Burge [30], Karlin et Mràzek [31] utilisent un *profil de fréquences relatives de dinucléotides* comme signature génomique. Avec les notations de la proposition 2.2.2, le rapport d'abondance relative du dinucléotide uv peut s'écrire sous la forme

$$\rho_{uv} \stackrel{\text{def}}{=} \frac{\pi(Suv)}{\pi(Su)\pi(Sv)}.$$

Il semble remarquable que le profil d'abondance relative de dinucléotides ait un comportement stable dans le sens où, lorsqu'on le calcule sur des fenêtres de taille 50 000 nucléotides sur un génome donné, le profil est quasi identique à celui que l'on calculerait sur tout le génome de l'organisme (Karlin et Mràzek [32], Karlin, Mràzek et Campbell [33]).

Il m'a semblé naturel dans ce contexte d'étendre cette notion d'abondance relative de dinucléotides grâce à la CGR en faisant une partition de S et en définissant un *rapport d'abondance relative basé sur la CGR* par

$$\rho(B, v) \stackrel{\text{def}}{=} \frac{\pi(Bv)}{\pi(B)\pi(Sv)},$$

qui vérifie trivialement $\rho(Su, v) = \rho_{uv}$. On construit alors une matrice de distances entre espèces à partir de ces rapports d'abondance et on constate, en comparant ces matrices de distances, que l'utilisation de partitions qui ne correspondent pas à des mots, ni à des unions de mots, permet un gain de précision et d'information dans la comparaison des séquences biologiques ainsi que l'obtention de meilleurs arbres taxonomiques générés avec la méthode de *clustering* dite *Neighbor-Joining* (grâce à l'outil NJPLOT), comme l'illustrent les figures 2.4 et 2.5.

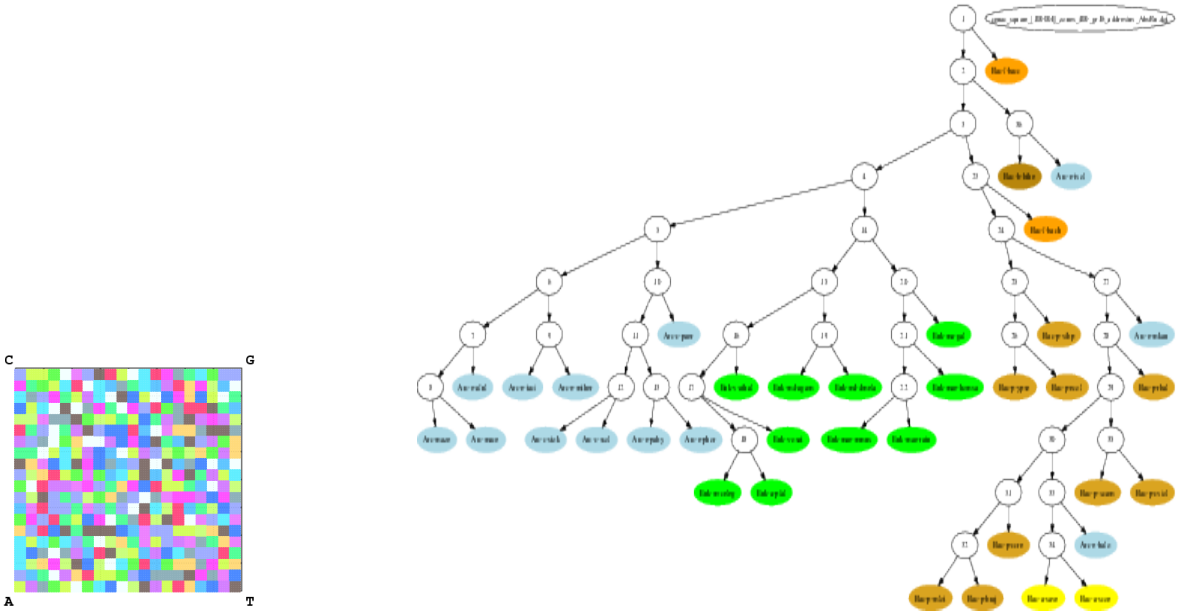


FIGURE 2.4: Arbre taxonomique (à droite) construit à partir des différences d'abondance relative basées sur la CGR avec une partitions de 400 zones régulières regroupées aléatoirement en 16 ensembles (représentée à gauche). Les différentes espèces d'archées sont en bleu, les eucaryotes en vert et les bactéries en jaune, orange ou marron.

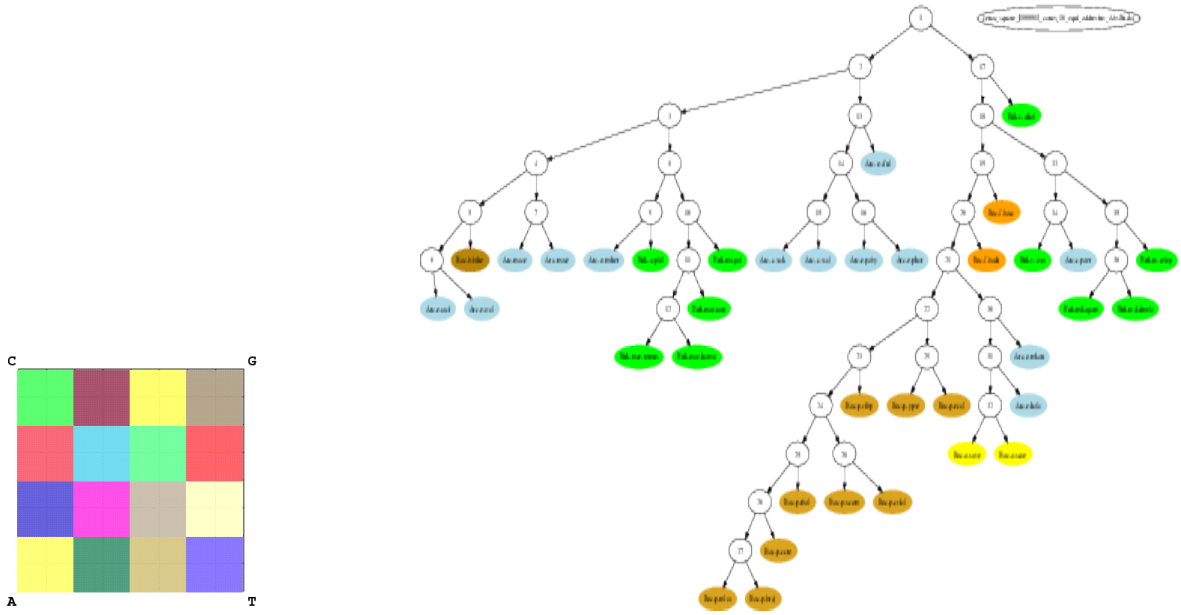
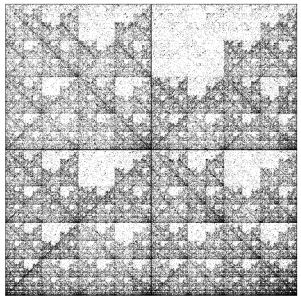
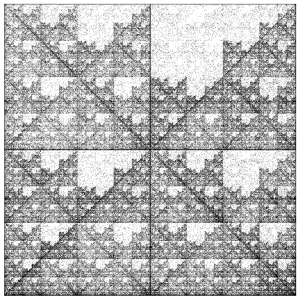


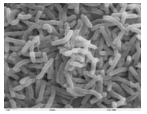
FIGURE 2.5: Arbre taxonomique (à droite) construit à partir des différences d'abondance relative basées sur la CGR avec la partition régulière de 16 zones (à gauche), correspondant au comptage des trinucléotides.

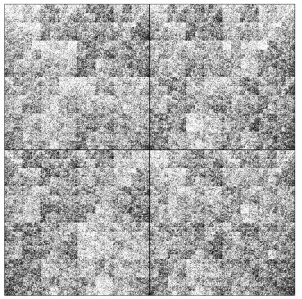
Homo sapiens (chr. 4)



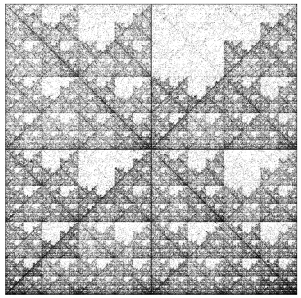
Souris (chr. 2) 



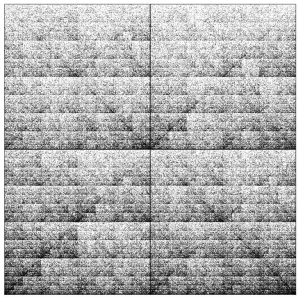
Choléra 



Canis familiaris (chr. 11) 



Saccharomyces cerevisiae (chr. 4) 



Arabidopsis thaliana (chr. 1) 

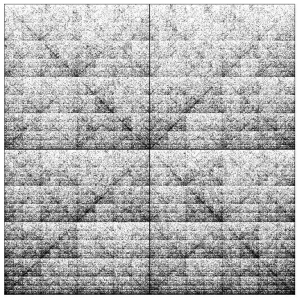


FIGURE 2.6: Exemples de CGR pour différentes espèces.

Le gain d'information sur la structure de la séquence est obtenu en utilisant le changement d'échelle qu'est la CGR, où chaque point garde en mémoire l'historique de toute la séquence. Cette idée est à la base du formalisme que nous avons introduit pour définir l'objet de la section suivante : les chaînes de Markov à mémoire variable. Plutôt que de regarder le processus de succession des lettres avec une longueur de dépendance fonction du motif observé, nous avons redéfini les chaînes de Markov à mémoire variable comme une chaîne de Markov d'ordre un sur les mots contenant tout le passé, *i.e* les mots infinis « à gauche ».

2.3 Chaînes de Markov à mémoire variable

Dans cette section, on modélise une suite, indexée par $\mathbb{Z}$, de lettres dans un alphabet fini $\mathcal{A}$. Comme pour la CGR où l'idée est de remplacer une suite de lettres de $\mathcal{A}$ par une suite de variables contenant l'historique de toute la séquence et formant de ce fait une chaîne de Markov d'ordre 1, il est naturel d'associer à une suite de lettres $(X_n)_{n\in\mathbb{Z}}$ à valeur dans $\mathcal{A}^{\mathbb{Z}}$, un processus $(U_n)_{n\in\mathbb{N}}$ à valeur dans l'ensemble des mots infinis à gauche $\mathcal{L} \stackrel{\text{def}}{=} \mathcal{A}^{-\mathbb{N}}$ ($\mathcal{L}$ pour *left*). A chaque étape on passe de $U_n = \dots X_{-1}X_0X_1 \dots X_n$ à U_{n+1} en ajoutant une lettre X_{n+1} à droite avec une évolution décrite par les probabilités de transition $\mathbb{P}(U_{n+1} = U_n\alpha|U_n)$, pour $\alpha \in \mathcal{A}$.

Dans le contexte des chaînes à proprement parler, jusqu'à présent le point de vue a surtout été d'ordre statistique depuis Harris [26] qui parle de « chaîne d'ordre infini » (*chains of infinite order*) pour exprimer le fait que la production d'une nouvelle lettre dépend d'un nombre fini mais non borné de lettres précédentes. Comets, Fernandez et Ferrari [6] s'intéressent à des chaînes à mémoire infinie. Rissanen [40] introduit une classe de modèles où la transition entre U_n et U_{n+1} dépend de U_n au travers d'un suffixe fini de U_n qu'il appelle *contexte*. Les contextes peuvent être stockés dans des feuilles d'un arbre appelé *arbre des contextes*. Ainsi le modèle est entièrement défini par une famille de distributions de probabilités indexées par les feuilles d'un arbre de contexte. Le nom *Variable Length Markov Chain* (VLMC) est dû à Bühlmann et Wyner [4]. On trouve dans Galves et Löcherbach [23] une bibliographie plus complète sur le sujet.

2.3.1 Définition

On commence par définir un *arbre des contextes probabilisé* ; on lui associe ensuite une chaîne de Markov à mémoire variable (VLMC).

Pour simplifier les notations, on se place dans le cas particulier d'un alphabet de deux lettres $\mathcal{A} = \{0, 1\}$ mais la définition s'étend à n'importe quel alphabet fini. On note $\mathcal{W}$ l'ensemble de tous les mots sur $\mathcal{A}$. On note également $\mathcal{R} \stackrel{\text{def}}{=} \mathcal{A}^{\mathbb{N}}$ l'ensemble des mots infinis à droite ($\mathcal{R}$ pour *right*). Pour un entier k et un mot $w = \alpha_{-k} \cdots \alpha_0$, $\overline{w}$ désigne le mot retourné

$$\overline{w} \stackrel{\text{def}}{=} \alpha_0 \cdots \alpha_{-k}.$$

Le *cylindre basé sur* w est l'ensemble des mots infinis à gauche ayant pour suffixe w :

$$\mathcal{L}w \stackrel{\text{def}}{=} \{s \in \mathcal{L},\ \forall j \in \{-k, \cdots, 0\},\ s_j = \alpha_j\}.$$

On considère un arbre planaire enraciné $\mathcal{T}$ et on définit l'ensemble de ses feuilles $\mathcal{C}(\mathcal{T})$ comme étant la réunion de l'ensemble des *feuilles finies* $\{u \in \mathcal{T}, \forall j \in \mathcal{A}, uj \notin \mathcal{T}\}$ et de l'ensemble des *feuilles infinies* $\{u \in \mathcal{R}, \forall v$ préfixe de $u, v \in \mathcal{T}\}$. Les *nœuds internes* sont les éléments du complémentaire, dans l'arbre $\mathcal{T}$, de toutes les feuilles.

Définition 2.3.1. *Un arbre est dit* saturé *lorsque tout nœud interne w a exactement $|\mathcal{A}|$ enfants.*

Définition 2.3.2 (Arbre des contextes probabilisé)**.** *Un* arbre des contextes *est un arbre saturé ayant un nombre fini ou dénombrable de feuilles. Les feuilles sont appelées* contextes*. Un* arbre des contextes probabilisé *est un couple*

$$\big(\mathcal{T}, (q_c)_{c \in \mathcal{C}(\mathcal{T})}\big)$$

où $\mathcal{T}$ est un arbre des contextes sur $\mathcal{A}$ et $(q_c)_{c \in \mathcal{C}(\mathcal{T})}$ une famille de mesures de probabilités sur $\mathcal{A}$, indexée par l'ensemble dénombrable $\mathcal{C}(\mathcal{T})$ de feuilles de $\mathcal{T}$.

Sur la figure 2.7 on trouve un exemple d'arbre des contextes probabilisé fini avec 5 contextes.

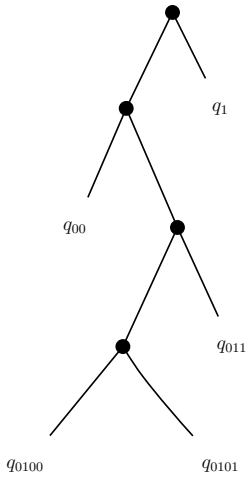


FIGURE 2.7: Exemple d'arbre des contextes probabilisés.

Définition 2.3.3. *On dit qu'un sous-ensemble $\mathcal{K}$ de $\mathcal{W} \cup \mathcal{R}$ est un* cutset *si les deux conditions suivantes sont satisfaites :*

i) aucun mot de $\mathcal{K}$ n'est un préfixe d'un autre mot de $\mathcal{K}$

ii) $\forall r \in \mathcal{R}$, $\exists u \in \mathcal{K}$, u est un préfixe de r.

L'ensemble $\mathcal{C}$ des contextes est un *cutset*. En effet, pour un arbre saturé, tout mot infini à droite $r \in \mathcal{R}$ est soit une feuille infinie, soit il existe une unique feuille finie qui soit préfixe de r. Autrement dit pour les mots infinis à gauche, on a

$$\mathcal{L} = \bigcup_{\overline{s} \text{ feuille infinie}} \{s\} \cup \bigcup_{\overline{w} \text{ feuille finie}} \mathcal{L}w.$$

Cette partition est abondamment utilisée dans [9], aussi bien sur les mots que sur les feuilles de l'arbre.

Définition 2.3.4 (Fonction Préfixe). *Soit $\mathcal{T}$ un arbre saturé et $\mathcal{C}$ l'ensemble de ses contextes. Pour tout $s \in \mathcal{L}$, $\overleftarrow{\text{pref}}(s)$ désigne l'unique contexte $\alpha_1 \dots \alpha_N$ tel que $s = \dots \alpha_N \dots \alpha_1$. La fonction*

$$\overleftarrow{\text{pref}} : \mathcal{L} \to \mathcal{C}$$

est appelée fonction préfixe. *Pour des raisons techniques, cette fonction est étendue à*

$$\overleftarrow{\text{pref}} : \mathcal{L} \cup \mathcal{W} \to \mathcal{T}$$

de la façon suivante :

- *si $\overline{w} \in \mathcal{T}$ alors $\overleftarrow{\text{pref}}(w) = \overline{w}$;*
- *si $\overline{w} \in \mathcal{W} \setminus \mathcal{T}$ alors $\overleftarrow{\text{pref}}(w)$ est l'unique contexte $\alpha_1 \dots \alpha_N$ tel que w ait $\alpha_N \dots \alpha_1$ comme suffixe.*

Définition 2.3.5 (VLMC). *On considère l'arbre des contextes probabilisé $(\mathcal{T}, (q_c)_{c \in \mathcal{C}})$. La chaîne de Markov à longueur variable (VLMC, de l'anglais* Variable Length Markov Chain*) est une chaîne de Markov d'ordre un $(U_n)_{n \geq 0}$ à espace d'états $\mathcal{L}$, définie par les probabilités de transition*

$$\forall n \geq 0,\ \forall \alpha \in \mathcal{A},\ \mathbb{P}\left(U_{n+1} = U_n \alpha | U_n\right) = q_{\overleftarrow{\text{pref}}(U_n)}(\alpha).$$

La lettre la plus à droite du mot infini $U_n \in \mathcal{L}$ est notée X_n. Le processus $(X_n)_{n \geq 0}$ n'est pas Markovien dès que l'arbre des contextes a au moins un contexte infini. Si l'arbre est fini, $(X_n)_{n \geq 0}$ est alors une chaîne de Markov d'ordre la hauteur de l'arbre.

On suppose maintenant qu'il existe une mesure stationnaire π sur $\mathcal{L}$ et on considère la VLMC en mode stationnaire (de loi π). Pour alléger les notations, pour un mot $w \in \mathcal{W}$, on écrit $\pi(w)$ plutôt que $\pi(\mathcal{L}w) : \pi(w) = \mathbb{P}(U_0 \in \mathcal{L}w) = \mathbb{P}(X_{-(|w|-1)} \dots X_0 = w)$.

Lorsque u est un nœud interne de l'arbre des contextes, on étend la notation q_u de la façon suivante : pour tout $\alpha \in \mathcal{A}$,

$$q_u(\alpha) = \begin{cases} \dfrac{\pi(\overline{u}\alpha)}{\pi(\overline{u})} \text{ si } \pi(\overline{u}) \neq 0 \\ \\ 0 \text{ si } \pi(\overline{u}) = 0. \end{cases}$$

Avec cette notation, la probabilité stationnaire de tout cylindre peut être décomposée par la simple formule suivante.

Lemme 2.3.6. *On considère une VLMC stationnaire de mesure invariante π sur $\mathcal{L}$. Alors on a :*

i) pour tout mot fini $w \in \mathcal{W}$ et toute lettre $\alpha \in \mathcal{A}$,

$$\pi(w\alpha) = \pi(w) q_{\overleftarrow{\text{pref}}(w)}(\alpha);$$

ii) pour tout mot fini $w = \alpha_1 \dots \alpha_N \in \mathcal{W}$,

$$\pi(w) = \prod_{k=0}^{N-1} q_{\overleftarrow{\text{pref}}(\alpha_1 \dots \alpha_k)}(\alpha_{k+1}) \tag{2.3}$$

(si $k = 0$, $\alpha_1 \dots \alpha_k$ désigne le mot vide $\emptyset$, $\overleftarrow{\text{pref}}(\emptyset) = \emptyset$, $q_\emptyset(\alpha) = \pi(\alpha)$ et $\pi(\emptyset) = \pi(\mathcal{L}) = 1$).

2.3.2 Source dynamique produisant une VLMC

Dans cette section, nous établissons le lien entre le point de vue « chaîne stochastique » de la chaîne de Markov à mémoire variable (U_n) et le point de vue source au sens de la théorie de l'information. Ici I désigne l'intervalle $[0, 1]$ et la mesure de Lebesgue d'un Borélien J est notée $|J|$.

Sources dynamiques probabilisées

On présente ici le formalisme classique des sources dynamiques probabilisées définies dans Clément, Flajolet et Vallee [5]). Une telle source est définie par quatre éléments :

- une partition topologique de I en intervalles $(I_\alpha)_{\alpha \in \mathcal{A}}$,
- une fonction de codage $\rho : I \to \mathcal{A}$, telle que, pour toute lettre α, la restriction de ρ à I_α est égale à α,
- une application $T : I \to I$,
- une mesure de probabilité μ sur I.

Cette source permet de définir un processus aléatoire $(Y_n)_{n \in \mathbb{N}}$ à valeurs dans $\mathcal{A}$ de la façon suivante. On choisit un réel aléatoirement selon la loi μ. L'application T permet de créer l'orbite $(x, T(x), T^2(x), \dots)$ de x. Grâce à la fonction de codage, on définit une suite infinie à droite $\rho(x)\rho(T(x))\rho(T^2(x)) \cdots$ dont les lettres sont $Y_n \stackrel{\text{def}}{=} \rho(T^n(x))$. Cette construction est illustrée sur la figure 2.8.

Pour un mot fini $w = \alpha_0 \dots \alpha_N \in \mathcal{W}$, on note B_w l'ensemble des réels x telle que la suite $(Y_n)_{n \in \mathbb{N}}$ ait w comme préfixe. La probabilité qu'une source émette un symbole commençant par w est égale à $\mu(B_w)$. Lorsque la mesure initiale μ sur I est T-invariante, la source dynamique génère un processus aléatoire stationnaire à valeurs dans $\mathcal{A}$.

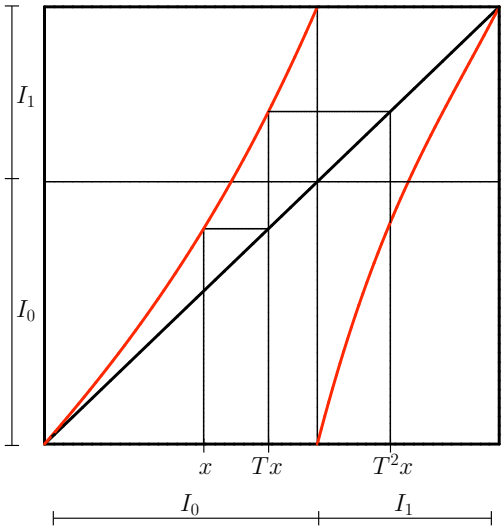


FIGURE 2.8: Un exemple de fonction T avec les intervalles I_0 et I_1 qui permettent de coder l'intervalle I par l'alphabet $\mathcal{A} = \{0, 1\}$ ainsi que les trois premiers points de l'orbite d'un point $x \in I$ du système dynamique correspondant. Le mot généré par cette source correspondant à l'orbite de x commence par $001\ldots$.

Définition de l'application T

Soit $(U_n)_{n\geq 0}$ une VLMC stationnaire définie par l'arbre des contextes probabilisé $(\mathcal{T}, (q_c)_{c\in\mathcal{C}})$ et de probabilité stationnaire π sur $\mathcal{L}$.

On commence par construire l'unique subdivision $\mathcal{A}$-adique $(I_w)_{w\in\mathcal{W}}$ de I associée à π définie par $\forall w \in \mathcal{W},\ |I_w| = \pi(\overline{w})$. On considère alors les trois partitions topologiques ordonnées suivantes :

- *la partition de codage* $I = I_0 + I_1$.
- *la partition verticale* résultant de la propriété de *cutset* de l'ensemble des contextes :

$$I = \sum_{c\in\mathcal{C}} \uparrow I_c.$$

- *la partition horizontale*

$$I = \sum_{\alpha\in\mathcal{A},\ c\in\mathcal{C}} \uparrow I_{\alpha c}.$$

Définition 2.3.7. *L'application $T : I \to I$ est la seule fonction continue à gauche telle que*

i) la restriction de T à tout intervalle $I_{\alpha c}$ est affine et croissante ;

ii) pour tout $(\alpha, c) \in \mathcal{A}\times\mathcal{C}$, $T(I_{\alpha c}) = I_c$.

Ainsi, lorsque $q_c(\alpha) \neq 0$, la pente de T sur l'intervalle $I_{\alpha c}$ est $1/q_c(\alpha)$.

Source dynamique associée à une VLMC stationnaire

On construit maintenant la source dynamique probabilisée $((I_\alpha)_{\alpha\in\mathcal{A}}, \rho, T, |.|)$ construite à partir d'une VLMC stationnaire. Elle fournit un processus aléatoire $(Y_n)_{n\in\mathbb{N}}$ à valeurs dans $\mathcal{A}$ défini par $Y_n \stackrel{\text{def}}{=} \rho(T^n\xi)$, où ξ est une variable aléatoire de loi uniforme sur I. Comme la mesure de Lebesgue est invariante par T par construction, toutes les variables aléatoires Y_n ont la même loi et $\mathbb{P}(Y_n = 0) = |I_0| = \pi(0)$.

Définition 2.3.8. *On dit que les deux processus aléatoires à valeurs dans $\mathcal{A}$ $(V_n)_{n\in\mathbb{N}}$ et $(W_n)_{n\in\mathbb{N}}$ sont* symétriquement distribués *si pour tout $N \in \mathbb{N}$, les mots $W_0W_1\dots W_N$ et $V_NV_{N-1}\dots V_0$ ont la même loi.*

Le lien entre l'approche source dynamique et l'approche processus stochastique est donné par le théorème suivant.

Théorème 2.3.9. *Soit $(U_n)_{n\in\mathbb{N}}$ une VLMC stationnaire de mesure invariante π sur $\mathcal{L}$. On note $(X_n)_{n\in\mathbb{N}}$ le processus des dernières lettres de $(U_n)_{n\in\mathbb{N}}$. L'application $T : I \to I$ désigne la fonction définie en Section 2.3.2. Alors,*

i) la mesure de Lebesgue est invariante par T.

ii) si ξ est une variable aléatoire de loi uniforme sur I, les processus $(X_n)_{n\in\mathbb{N}}$ et $(\rho(T^n\xi))_{n\in\mathbb{N}}$ sont symétriquement distribués.

La preuve est une conséquence du théorème de Thalès.

2.3.3 Le peigne

Le peigne est un cas particulier d'arbre des contextes possédant une unique branche infinie. De ce fait, le processus des lettres (X_n) de la VLMC n'est a priori pas Markovien. Plus précisément, on considère l'arbre des contextes dessiné sur la partie gauche de la figure 2.11. Il fournit un cas très concret de chaînes d'ordre infini où l'étude peut être traitée complètement « à la main ». Dans ce cas, nous sommes en mesure de donner une condition nécessaire et suffisante pour l'existence et l'unicité d'une mesure invariante et d'en donner son expression. Pour certaines valeurs de données appropriées, on obtient des exemples de sources dynamiques intermittentes. On précise également les propriétés de mélange de ce processus et on détermine l'expression de la fonction génératrice de la $r^{\text{ème}}$ occurrence d'un mot donné. Dans l'article [9], nous donnons également l'expression de la série de Dirichlet associée à cette source dynamique, que je ne présenterai pas dans ce mémoire.

Mesure de probabilité stationnaire

Pour le peigne, il y a une unique feuille infinie 0^∞ et toutes les feuilles finies sont de la forme 0^n1, pour $n \in \mathbb{N}$. Les données correspondant à cette VLMC sont donc les mesures de probabilité sur $\mathcal{A} = \{0, 1\}$: q_{0^∞} et $q_{0^n1},\ n \in \mathbb{N}$.

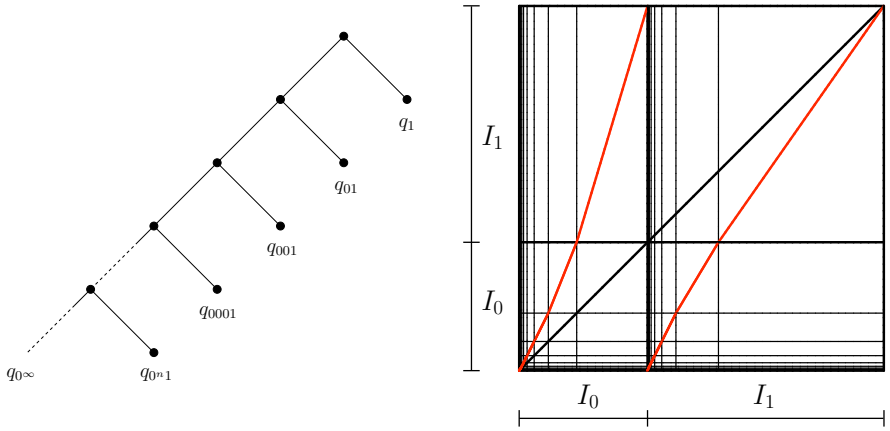


FIGURE 2.9: Arbre des contextes probabilisé du peigne infini (à gauche) et son système dynamique associé (à droite).

Proposition 2.3.10 (Mesures invariantes pour le peigne infini). *Soit $(U_n)_{n\geq 0}$ la VLMC définie par un peigne infini probabilisé.*

i) Cas irréductible : *on suppose que $q_{0^\infty}(0) \neq 1$.*

a) Existence

Le processus Markovien $(U_n)_{n\geq 0}$ admet une mesure de probabilité invariante sur $\mathcal{L}$ si et seulement si $\sum c_n$ converge, où c_n est définie par

$$c_n = \prod_{k=0}^{n-1} q_{0^k1}(0), \text{ pour } n \geq 1, \qquad \text{et} \quad c_0 = 1.$$

b) Unicité

On suppose que $\sum c_n$ converge et on note $S(1) \stackrel{\text{def}}{=} \sum_{n\geq 0} c_n$. Alors la mesure invariante π est unique et caractérisée par les formules (2.3) et $\pi(1) = \frac{1}{S(1)}$. On obtient les probabilités des contextes et des nœuds internes :

$$\pi(10^n) = \pi(1)c_n, \qquad \pi(0^n) = 1 - \pi(1)\sum_{k=0}^{n-1} c_k \quad \text{pour } n \geq 0. \tag{2.4}$$

De plus, π est triviale si et seulement si $q_1(0) = 0$, en quel cas $\pi(1^\infty) = 1$.

ii) Cas réductible : *on suppose que $q_{0^\infty}(0) = 1$.*

a) si $\sum c_n$ diverge, alors la mesure de probabilité triviale π sur $\mathcal{L}$ définie par $\pi(0^\infty) = 1$ est l'unique mesure de probabilité invariante.

b) Si $\sum c_n$ converge, alors il existe une famille de mesures invariantes à paramètre sur $\mathcal{L}$. Plus précisément, pour $a \in [0,1]$, il existe une unique mesure de

probabilités invariante π_a sur $\mathcal{L}$ telle que $\pi_a(0^\infty) = a$. La loi π_a est caractérisée par $\pi_a(1) = \frac{1-a}{S(1)}$ et les formules (2.3) et (2.4). De plus, π_a n'est pas triviale sauf sans les cas suivants :

- *$a = 1$ en quel cas π_1 est définie par $\pi_1(0^\infty) = 1$;*
- *$a = 0$ et $q_1(0) = 0$, en quel cas π_0 est définie par $\pi_0(1^\infty) = 1$.*

Le système dynamique associé

La partition verticale est constituée des intervalles I_{0^n1} pour $n \geq 0$, la partition horizontale des intervalles I_{00^n1} et I_{10^n1}, pour $n \geq 0$, ainsi que deux intervalles provenant des contextes infinis $I_{0\infty}$ et $I_{10\infty}$. Dans le cas irréducible, $\pi(0^\infty) = 0$ et les deux intervalles $I_{0\infty}$ et $I_{10\infty}$ viennent des points d'accumulation de la partition en 0 et $\pi(0)$.

Proposition 2.3.11. *Si la suite $(q_{0^n1}(0))_{n\in\mathbb{N}}$ converge, alors T est dérivable aux deux points d'accumulation 0 et $\pi(0)$ et*

$$T_r'(0) = \lim_{n\to+\infty} \frac{1}{q_{0^n1}(0)}, \quad T_r'(\pi(0)) = \lim_{n\to+\infty} \frac{1}{q_{0^n1}(1)}.$$

En particulier il découle de cette proposition que lorsque $(q_{0^n1}(0))_{n\in\mathbb{N}}$ converge vers 1, $T_r'(0) = 1$. Dans ce cas, 0 est un point fixe indifférent et $T_r'(\pi(0)) = +\infty$. La fonction T est alors une légère modification de la fonction de Wang (voir Wang [48]). La source dynamique correspondante est dite *intermittente*.

Fonction génératrice de la loi exacte des occurrences de mots dans un ordre généré par un peigne

Le comportement du temps d'entrée dans les cylindres est une question naturelle dans les systèmes dynamiques. Il existe une abondante littérature sur les propriétés asymptotiques de ces moments d'entrée pour différents types de systèmes (voir Abadi et Galves [1] pour une étude approfondie sur le sujet). La plupart des résultats utilise une approximation de la loi du temps d'entrée dans un petit cylindre par une loi exponentielle. Les résultats présentés dans ce paragraphe (issus de [9]) sont des résultats non asymptotiques.

Pour les suites indépendantes et identiquement distribuées, Blom et Thorburn [3] fournissent la fonction génératrice de la première occurrence d'un mot. Ce résultat est étendu aux chaînes de Markov par Robin et Daudin [41]. Leurs travaux sont basés sur une relation de récurrence sur les probabilités d'occurrence de motifs. D'autres approches sont envisagées dans la littérature : l'une des techniques plus générales est la méthode dite de plongement Markovien introduite par Fu [21] et développée ensuite dans Fu et Koutras [22], Koutras [34]. Il existe aussi une approche alternative avec des martingales (voir Gerber et Li [24], Li [35], Williams [52]). Ces deux approches sont comparées dans Pozdnyakov, Glaz, Kulldorff et Steele [38].

On considère le processus $X = (X_n)_{n\geq 0}$ des dernières lettres de $(U_n)_{n\geq 0}$, dans le cas particulier d'une VLMC stationnaire définie par un peigne infini. Soit $w = w_1 \dots w_k$ un mot de longueur $k \geq 1$. On dit qu'on a une occurrence de w en $n \geq k$ dans le mot infini X si le mot w se termine à la position n :

$$\{w \text{ en } n\} = \{X_{n-k+1} \dots X_n = w\} = \{U_n \in \mathcal{L}w\}.$$

Notons $T_w^{(r)}$ la position de la r ème occurrence de w dans X et $\Phi_w^{(r)}$ sa fonction génératrice :

$$\Phi_w^{(r)}(x) \stackrel{\text{def}}{=} \sum_{n\geq 0} \mathbb{P}\left(T_w^{(r)} = n\right) x^n.$$

Enfin, on utilise la notation suivante : pour tout mot fini u, pour tout contexte $c \in \mathcal{C}$ et $n \geq 0$,

$$q_c^{(n)}(u) \stackrel{\text{def}}{=} \mathbb{P}\left(X_{n-|u|+1} \dots X_n = u | X_{-(|c|-1)} \dots X_0 = \overline{c}\right).$$

Proposition 2.3.12. *Pour une VLMC stationnaire définie par un peigne infini, pour un mot w tel que $\overline{w}$ ne soit pas un nœud interne, la fonction génératrice de sa première occurrence est donnée, pour $|x| < 1$, par*

$$\Phi_w^{(1)}(x) = \frac{x^k \pi(w)}{(1-x) S_w(x)}$$

et la fonction génératrice de sa $r^{ème}$ occurrence est donnée pour $|x| < 1$, par

$$\Phi_w^{(r)}(x) = \Phi_w^{(1)}(x) \left(1 - \frac{1}{S_w(x)}\right)^{r-1},$$

où

$$\begin{aligned} S_w(x) &= C_w(x) + \sum_{j=k}^{\infty} q^{(j)}_{\overleftarrow{\text{pref}}\ (w)}(w) x^j, \\ C_w(x) &= 1 + \sum_{j=1}^{k-1} \mathbb{1}_{\{w_{j+1}\dots w_k = w_1 \dots w_{k-j}\}} q^{(j)}_{\overleftarrow{\text{pref}}\ (w)}\left(w_{k-j+1} \dots w_k\right) x^j. \end{aligned}$$

Propriétés de mélange

Pour une suite stationnaire $(U_n)_{n\geqslant 0}$ de mesure invariante π, on cherche à mesurer l'indépendance entre deux mots A et B séparés de n lettres. Pour $0 \leqslant m \leqslant +\infty$, on note $\mathcal{F}_0^m$ la tribu engendrée par les $\{U_k, 0 \leqslant k \leqslant m\}$ et on considère les mots $A \in \mathcal{F}_0^m$ et $B \in \mathcal{F}_0^\infty$. On s'intéresse en particulier au coefficient de mélange

$$\psi(n, A, B) \stackrel{\text{def}}{=} \frac{\sum_{|w|=n} \pi(AwB) - \pi(A)\pi(B)}{\pi(A)\pi(B)},$$

où la somme est calculée sur les mots finis w de longueur $|w| = n$. La suite $(U_n)_{n\geqslant 0}$ est dite *ψ-mélangeante* si

$$\lim_{n\to\infty} \sup_{m\geqslant 0, A\in\mathcal{F}_0^m, B\in\mathcal{F}_0^\infty} |\psi(n, A, B)| = 0.$$

Dans cette définition, la convergence vers zéro est uniforme sur tous les mots A et B. Dans l'article [10], nous établissons l'expression exacte de $\psi(n, A, B)$ pour un peigne général. Par souci de concision, je choisis de ne présenter ici le résultat de ce calcul que pour deux exemples : l'un fournissant une suite (U_n) ψ-mélangeante et l'autre une suite (U_n) qui n'est pas uniformément mélangeante.

Exemple 1 : le peigne logarithmique

Le peigne logarithmique est défini par $c_0 = 1$ et pour $n \geqslant 1$ par

$$c_n = \frac{1}{n(n+1)(n+2)(n+3)}.$$

Les probabilités conditionnelles correspondantes sur les feuilles de l'arbre sont donc $q_1(0) = \frac{1}{24}$ et pour $n \geqslant 1$, $q_{0^n1}(0) = 1 - \frac{4}{n+4}$.

Exemple 2 : le peigne factoriel

Le peigne factoriel est défini par les probabilités conditionnelles sur les feuilles $q_{0^n1}(0) = \frac{1}{n+2}$ pour $n \geqslant 0$. Par conséquent on a

$$c_n = \frac{1}{(n+1)!}.$$

Proposition 2.3.13. *La VLMC définie par le peigne logarithmique a un mélange polynomial non uniforme de la forme suivante : pour deux mots finis A et B, il existe une constante positive $C_{A,B}$ telle que pour tout $n \geqslant 1$,*

$$|\psi(n, A, B)| \leqslant \frac{C_{A,B}}{n^3}.$$

Remarque 2.3.14. Les constantes $C_{A,B}$ ne peuvent pas être bornées par une constante indépendante des mots A et B : par exemple on peut montrer que $\psi(n, 0, 0^n)$ tend vers une constante strictement positive $\frac{13}{6}$.

Proposition 2.3.15. *La VLMC définie par le peigne factoriel a un mélange exponentiel uniforme de la forme suivante : il existe une constante C positive telle que pour tout $n \geqslant 1$ et pour tous mots finis A et B,*

$$|\psi(n, A, B)| \leqslant \frac{C}{(2\pi)^n}.$$

Comme dans Isola [28], on utilise abondamment les propriétés de renouvellement du peigne infini. Le renouvellement nous permet de calculer explicitement les coefficients de mélange, les séries génératrices des instants d'occurrence, etc ... Les deux types de comportements extrêmes que fournissent le peigne factoriel et le peigne logarithmique en terme de mélange illustrent la richesse de ce modèle de renouvellement.

On a obtenu dans [9] des résultats analogues (existence et unicité de mesure stationnaire, fonctions génératrices des instants d'occurrences de mots, séries de Dirichlet, système dynamique) pour un autre exemple de VLMC : *le bambou*. Ce modèle est défini par l'arbre des contextes probabilisé donné par la partie gauche de la figure 2.10. Les

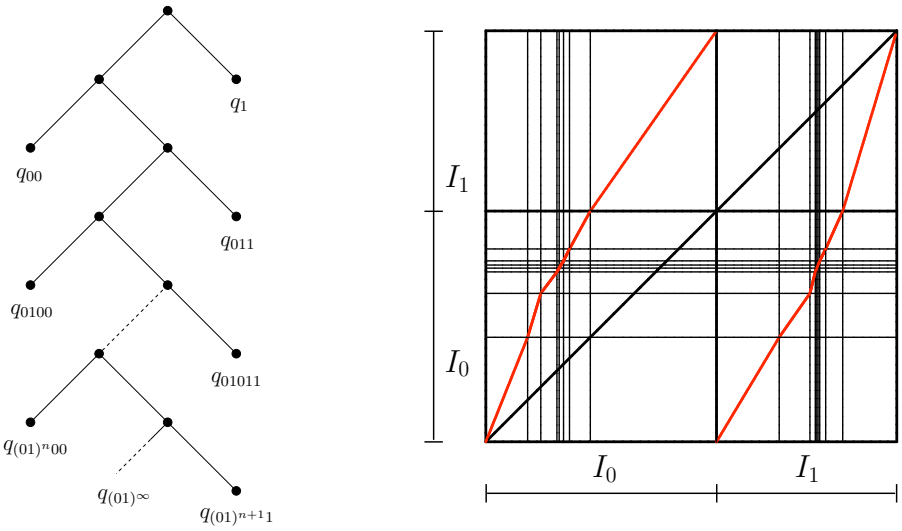


FIGURE 2.10: Arbre des contextes probabilisé du bambou (à gauche) et son système dynamique associé (à droite).

données associées à la VLMC sont les mesures de probabilités sur l'alphabet $\mathcal{A}$ indexées par les deux familles de contextes finis

$$(q_{(01)^n1})_{n\geq 0} \text{ et } (q_{(01)^n00})_{n\geq 0}$$

ainsi qu'une mesure de probabilités sur le contexte infini $q_{(01)^\infty}$.

J'ai choisi de ne pas présenter le détail de ces résultats et de me limiter au cas du peigne. Sur le fond, le traitement du bambou a des analogies avec celui du peigne, car il possède également une propriété de *renouvellement* induit par les motifs 11 ou 00.

2.4 Marches aléatoires persistantes

Les marches aléatoires classiques sont définies par

$$S_t \stackrel{\text{def}}{=} \sum_{n=0}^{t} X_n, \tag{2.5}$$

où $t \in \mathbb{N}$ et pour des incréments indépendants et identiquement distribués $(X_n)_{n\in\mathbb{N}}$. Prenons un exemple en finance en supposant que S_t représente le prix d'un actif à l'instant t. Dans le modèle de Cox, Ross et Rubinstein, la condition de non-arbitrage entraîne que les incréments relatifs $\left(\frac{S_t - S_{t-1}}{S_{t-1}};\ t \geq 1\right)$ sont indépendants.

Avec un bon changement d'échelle pour la marche et un passage à la limite au cas continu, il est possible d'obtenir un mouvement brownien standard. Lorsque les incréments forment une chaîne de Markov d'ordre 1, une courte mémoire est introduite dans la dynamique de la marche aléatoire : le processus est dit *persistant* ou est encore appelé marche aléatoire corrélée ou marche de Kac (voir Eckstein, Goldstein et Leggas [18], Renshaw et Henderson [39], Weiss [50, 51]). La marche n'est plus Markovienne dans ce cas et par un changement d'échelle adéquat, le processus renormalisé converge vers le processus du télégraphe intégré (Herrmann et Vallois [27], Vallois et Tapiero [47] et Vallois et Tapiero [46].

Dans l'article [8], nous généralisons ce résultat à une famille de marches aléatoires construites à partir d'incréments $(X_n)_{n\in\mathbb{N}}$ constituant une chaîne à mémoire variable. Plus précisément, pour une VLMC donnée $(U_n)_{n\geqslant 0}$, on définit X_n comme étant la dernière lettre de U_n pour $n \geq 0$. Lorsque $(X_n)_{n\in\mathbb{N}}$ est une chaîne de Markov d'ordre fini, il est naturel de penser que le processus limite sera très proche du télégraphe intégré. C'est pourquoi nous avons construit des chaînes à mémoire infinie afin d'identifier éventuellement un autre processus limite. La marche associée (S_t) définie par (2.5) n'est pas Markovienne, elle est en quelque sorte *très persistante*. Cette marche est-elle de même nature que dans le cas Markovien d'ordre un ? La marche renormalisée par le bon changement d'échelle conduit-elle à un processus analogue à celui du télégraphe intégré ?

2.4.1 Modèle

La chaîne de Markov à mémoire variable et non bornée que l'on considère dans cette section et que nous appellerons *double peigne* est définie de la façon suivante. On considère l'arbre des contextes donné sur la figure 2.11. Le double peigne possède deux feuilles infinies 0^∞ et 1^∞ ainsi qu'une famille dénombrable de feuilles finies 0^n1 et 1^n0, pour $n \in \mathbb{N}^*$, ainsi

$$\mathcal{C} = \{0^n1,\ n \geq 1\} \cup \{1^n0,\ n \geq 1\} \cup \{0^\infty\} \cup \{1^\infty\}.$$

Les données associées à la VLMC correspondantes sont donc les mesures de probabilité sur $\{0, 1\}$:

$$q_{0^\infty}, q_{1^\infty},\ \text{et}\ \ q_{0^n1}, q_{1^n0},\ n \in \mathbb{N}^*.$$

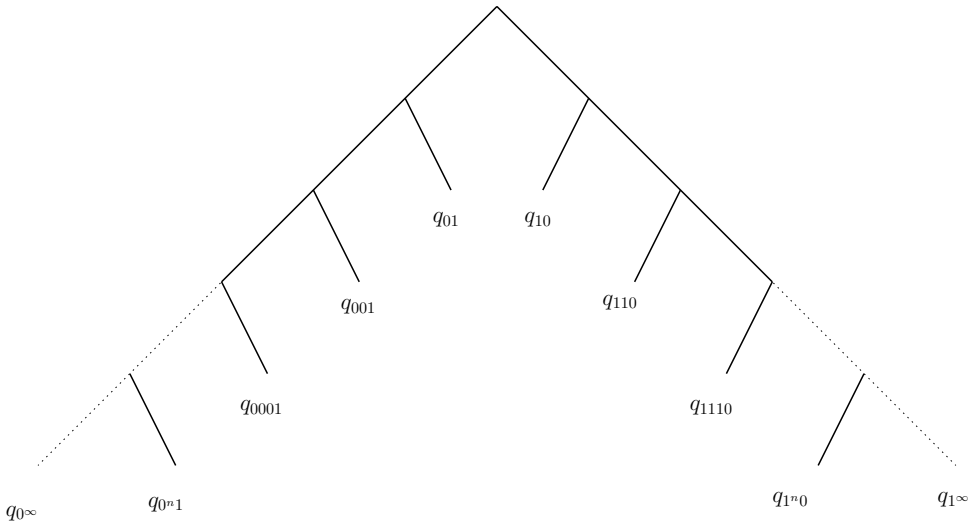


FIGURE 2.11: Arbre des contextes probabilisé du *double peigne*.

En généralisant les calculs pour le théorème d'existence et d'unicité de la mesure invariante pour le peigne simple, on montre facilement que si les deux séries

$$\Theta_1 \stackrel{\text{def}}{=} \sum_{n\geq 1}\prod_{k=1}^{n-1} q_{0^n1}(0) \quad \text{et} \quad \Theta_2 \stackrel{\text{def}}{=} \sum_{n\geq 1}\prod_{k=1}^{n-1} q_{1^n0}(1)$$

convergent, alors il existe une unique mesure invariante pour le double peigne que l'on calcule explicitement.

2.4.2 Comportement de la marche persistante

On effectue maintenant un changement de codage : '0' est codé en '-1'. La suite des lettres (X_n) de (U_n) est donc dans à valeur dans l'alphabet $\{-1, 1\}$.

Remarque 2.4.1. On définit la suite des instants de changements de cap :

$$T_0 = 0, \quad T_{k+1} = \inf\{n > T_k,\ X_n \neq X_{T_k}\}.$$

Il est facile de voir que $(X_n, T_n)_{n\geq 0}$ est un processus semi-Markovien.

La loi exacte de la marche aléatoire *très persistante* (S_n) est donnée dans [8]. Son expression est assez compliquée et peut être simplifiée en s'intéressant à la fonction génératrice de la variable $S_{\tau+1}$, où τ est une variable de loi géométrique indépendante de la source (U_n).

Une autre façon de comparer le processus (S_n) avec celui de la marche aléatoire « classique » est d'analyser les fluctuations à l'infini.

Proposition 2.4.2. *On suppose que les deux séries Θ_1 et Θ_2 convergent.*

(i) La suite $\dfrac{S_n}{n}$ converge p.s. et dans L^1 vers $\dfrac{\Theta_2 - \Theta_1}{\Theta_1 + \Theta_2}$ quand $n \to \infty$.

(ii) De plus, si

$$\sum_{n\geq 1} n \prod_{k=1}^{n-1} q_{0^n 1}(0) < \infty \quad et \quad \sum_{n\geq 1} n \prod_{k=1}^{n-1} q_{1^n 0}(1) < \infty, \tag{2.6}$$

alors on a le théorème de la limite centrale :

$$\frac{1}{\sqrt{n}\Upsilon}\left(S_n - n\frac{\Theta_2 - \Theta_1}{\Theta_1 + \Theta_2}\right) \xrightarrow[\mathrm{n}\to\infty]{\mathcal{L}} \mathcal{N}(0,1),$$

où la constante Υ est définie par

$$\Upsilon = \frac{4}{\Theta_1 + \Theta_2}\, \mathbb{E}\left[\left(T_1 - \frac{\Theta_2 T_2}{\Theta_1 + \Theta_2}\right)^2\right].$$

Sous la condition (2.6), on peut aussi prouver qu'il existe une constante $C \in \mathbb{R}$ telle que

$$\lim_{n\to\infty}\left\{\mathbb{E}(S_n) - n\,\frac{\Theta_2 - \Theta_1}{\Theta_1 + \Theta_2}\right\} = C.$$

Dans le cas particulier où $\Theta_1 = \Theta_2 < \infty$, la proposition entraîne que $\lim_{n\to\infty} \frac{\mathbb{E}(S_n)}{n} = 0$. De plus, sous la condition (2.6), on a

$$\frac{1}{\sqrt{n}} S_n \xrightarrow[\mathrm{n}\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

On observe donc sous ces conditions le même comportement asymptotique au premier et second ordre que celui de la marche « classique » construite avec des incréments de loi de Rademacher, la persistance n'a que peu d'influence dans ce cas. En revanche, cette condition (2.6) est assez forte et induit un mélange important pour la VLMC. Sur la chaîne, cette condition revient à supposer que l'espérance de la longueur de $\overleftarrow{\mathrm{pref}}\,(U_n)$ soit finie. Que se passe-t'il lorsque cette condition n'est pas satisfaite ? La question est encore ouverte.

2.4.3 Changement d'échelle

On introduit un paramètre d'échelle $\varepsilon > 0$ ainsi que deux fonctions $f_1, f_2 : [0,\infty[\to \mathbb{R}$ positives et continues à droite. On s'intéresse à la VLMC du double peigne (codée dans l'alphabet $\{-1,1\}$) définie par les probabilités de transition

$$\begin{aligned} q_{(-1)^k 1}(1) &= f_1(k\varepsilon)\varepsilon + o(\varepsilon), \\ q_{1^k(-1)}(-1) &= f_2(k\varepsilon)\varepsilon + o(\varepsilon). \end{aligned}$$

La chaîne a un comportement conservatif : lorsque la suite des lettres est dans l'état 1, la probabilité qu'elle change de signe pour la lettre suivante est faible et dépend de f_1, f_2 et ε.

On associe à cette VLMC la marche construite à partir de ces incréments, que l'on note (S_n^ε). Le processus $(S^\varepsilon(t),\ t \geq 0)$ est linéaire par morceaux et satisfait, pour tout $n \in \mathbb{N}$,

$$S^\varepsilon(n\varepsilon) = \varepsilon S_n^\varepsilon.$$

Théorème 2.4.3. *On considère une suite $(\xi_{n+1} - \xi_n,\ n \geq 0)$ de variables aléatoires indépendantes telles que $\xi_0 = 0$ et*

$$\mathbb{P}(\xi_{2n+1} - \xi_{2n} \geq t) = \exp\left(-\int_0^t f_2(u)\,du\right)$$
$$\mathbb{P}(\xi_{2n+2} - \xi_{2n+1} \geq t) = \exp\left(-\int_0^t f_1(u)\,du\right)$$

pour tous $t \geq 0$ et $n \geq 0$. On définit le processus de comptage associé aux instants de saut (ξ_n) :

$$N^0(t) \stackrel{\text{def}}{=} \sum_{n\geq 1} \mathbb{1}_{\{\xi_n \leq t\}}.$$

On définit enfin le processus du télégraphe intégré généralisé

$$S^0(t) \stackrel{\text{def}}{=} \int_0^t (-1)^{N^0(s)}\,ds, \quad t \geq 0.$$

On suppose que $U_0 \in \mathcal{L}w_0$ avec $w_0 = (-1)1$.

Le processus $(S^\varepsilon(t),\ t \geq 0)$ converge en loi (avec la topologie de Skorohod) lorsque $\varepsilon \to 0$ vers $(S^0(t),\ t \geq 0)$.

Remarque 2.4.4. Le processus $(S^0(t),\ t \geq 0)$ est à la fois un processus semi-Markov et un processus de Markov déterministe par morceaux (voir par exemple Davis [14, 13]). Ses trajectoires ressemblent à celles des processus zig-zag.

2.5 Arbres aléatoires

Dans la dernière partie de ce chapitre, on s'intéresse à une autre famille d'objets construits à partir de certaines sources dynamiques : des arbres digitaux de compression de données.

2.5.1 Arbre-CGR

Une propriété fondamentale de la CGR est que tous les mots possédant un même suffixe $w = w_1 \dots w_k$ sont regroupés dans une même zone Sw. À partir d'un mot $m =$

$a_1a_2\ldots$ à valeur dans $\mathcal{A}^{\mathbb{N}}$, on regroupe ainsi dans Sw tous les points X_n de la CGR tels que $a_{n-k+1}\ldots a_{n-1}a_n = w$. Une idée naturelle est de « ranger » ces sous-séquences dans les nœuds d'un arbre, tout en conservant la visualisation de répétitions de suffixes.

Algorithme de construction

On supposera les lettres classées arbitrairement dans l'ordre (A, C, G, T). On note $\mathcal{T}$ l'arbre quaternaire infini complet. À chaque étape de la construction, on insère un nœud dans $\mathcal{T}$. On construit ainsi une suite de sous-arbres finis $\mathcal{T}_n$ de $\mathcal{T}$, tous emboîtés $\mathcal{T}_1 \subset \mathcal{T}_2 \ldots \mathcal{T}_n \subset \ldots \subset \mathcal{T}$. Chaque sous-arbre $\mathcal{T}_n$ possède n nœuds étiquetés (sans compter la racine). Étant donné un mot aléatoire $m = a_1a_2\ldots$, l'*arbre-CGR* pousse en insérant successivement des mots $a_1\ldots a_i$ dans l'arbre infini complet. Chaque nœud de cet arbre possède 4 branches correspondant au quadruplet ordonné (A, C, G, T).

Tout d'abord, la lettre a_1 est insérée dans l'arbre infini complet au niveau 1, juste sous la racine, sous la branche correspondant à a_1. L'insertion du mot $a_1\ldots a_n$ est faite récursivement : on essaye tout d'abord de l'insérer au niveau 1 dans la branche correspondant à la dernière lettre rencontrée dans la lecture de ce mot, c'est-à-dire a_n ; si ce nœud est déjà occupé, on essaye de l'insérer au niveau 2 de l'arbre dans le sous-arbre correspondant à a_n, sous la branche correspondant à la lettre a_{n-1}. On répète l'opération jusqu'au premier nœud non occupé au niveau k dans la branche correspondant à la lettre a_{n-k+1} ; le mot $a_1\ldots a_n$ est alors inséré sur ce nœud. Si $k < n$, seul compte le suffixe $a_{n-k+1}\ldots a_n$ du mot $a_1\ldots a_n$ que l'on insère. On remarque que comme pour le processus de construction d'une VLMC, on analyse le préfixe retourné du mot parcouru.

La Figure 2.12 montre les premières étapes de la construction de l'arbre-CGR correspondant à tout mot m qui commence par $w = GAGCACAGTGGAAGGG$. Chaque nœud est étiqueté par son ordre d'insertion par souci de lisibilité. Cette construction est motivée par la volonté de mesurer des quantités statistiques *révélées* par la forme de l'arbre, afin de dégager de nouvelles caractéristiques pour une loi de génération fixée.

L'arbre-CGR d'un mot aléatoire $m = a_1a_2\ldots$ est un *arbre digital de recherche* (DST, de l'anglais *Digital Search Tree*) obtenu par l'insertion successive dans un arbre quaternaire des préfixes retournés du mot m :

$$\begin{aligned} m(1) &= a_1, \\ m(2) &= a_2a_1, \\ &\vdots \\ m(n) &= a_na_{n-1}\ldots a_1, \\ &\vdots \end{aligned}$$

Les mots insérés sont donc fortement dépendants, contrairement aux DST classiques où les mots insérés sont indépendants les uns des autres. Plusieurs résultats sont connus (voir chap. 6 dans Mahmoud [36]) sur la hauteur, la profondeur d'insertion et le profil,

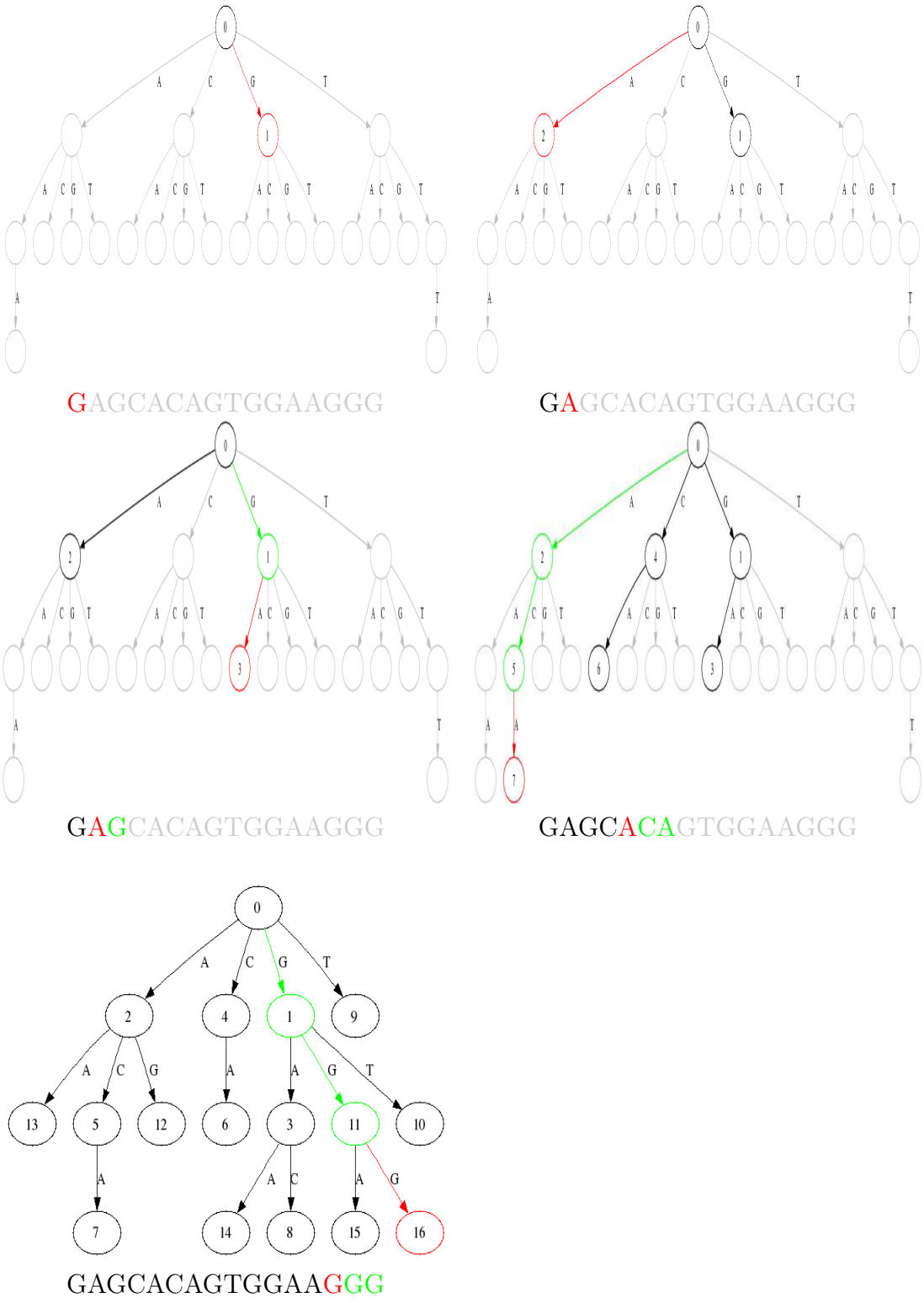


FIGURE 2.12: Étapes successives de construction de l'arbre-CGR représentant les premières étapes de l'insertion d'un mot GAGCACAGTGGAAGGG....

pour des DST construits sur des suites de mots *indépendants*, tous de même loi. Dans [11], nous nous intéressons aux propriétés asymptotiques de l'arbre-CGR, construit à partir de chaînes de Markov. En particulier, nous démontrons un résultat de convergence presque sûre pour les longueurs des branches, ainsi qu'une propriété de convergence en probabilité pour la profondeur d'insertion. Comme pour d'autres arbres de compression classiques (voir Fayolle [19] pour les tries et Devroye et Neininger [16] pour les arbres binaires de recherche), la différence de comportement asymptotique entre les DST classiques et les arbres-CGR, issus de séquences Markoviennes n'est pas visible au 1$^{\text{er}}$ ordre.

Dans le cas où les mots à insérer sont indépendants, mais à valeurs dans un alphabet fini et émises par des sources à *faible dépendance*, Pittel [37] obtient des résultats de convergence sur la profondeur d'insertion et sur la hauteur en supposant que la source est uniformément mélangeante.

La difficulté principale dans l'étude des propriétés asymptotiques des arbres-CGR provient de la dépendance des mots à insérer et des structures potentiellement auto-recouvrantes des mots. Les preuves utilisées dans ce travail font appel aux études donnant des résultats sur la loi des positions d'occurrences d'un mot le long d'une séquence. Blom et Thorburn [3] déterminent la fonction génératrice de ces lois dans le cas de suites i.i.d., à partir d'une relation de récurrence sur les probabilités. Ce résultat est généralisé par Robin et Daudin [41] dans le cas d'une séquence Markovienne d'ordre 1.

Convergence presque sûre de branches critiques

On suppose que le processus des lettres de $m = a_1 a_2 \dots a_n \dots$ forme une chaîne de Markov d'ordre 1 à espace d'états fini $\mathcal{A} = \{A, C, G, T\}$, irréductible, apériodique, stationnaire, de matrice de transition Q et de mesure invariante π.

Pour un mot infini fixée déterministe s, on note $s^{(n)}$ le mot constitué des n premières lettres de s, c'est-à-dire $s^{(n)} = s_1 \dots s_n$, où s_i désigne la $i^{\text{ème}}$ lettre de s. La mesure π est étendue aux mots $s^{(n)}$ retournés en posant $\pi(s^{(n)}) \stackrel{\text{def}}{=} \pi(a_1 = s_n, \dots, a_n = s_1)$. On inverse le mot du fait de la construction de l'arbre-CGR basée sur les préfixes inversés. On définit également les constantes

$$
\begin{aligned}
h_+ &\stackrel{\text{def}}{=} \lim_{n\to+\infty} \frac{1}{n} \max\Big\{ \log\Big(\frac{1}{\pi\big(s^{(n)}\big)}\Big) \;\Big|\; \pi\big(s^{(n)}\big) > 0 \Big\}, \\
h_- &\stackrel{\text{def}}{=} \lim_{n\to+\infty} \frac{1}{n} \min\Big\{ \log\Big(\frac{1}{\pi\big(s^{(n)}\big)}\Big) \;\Big|\; \pi\big(s^{(n)}\big) > 0 \Big\}, \\
h &\stackrel{\text{def}}{=} \lim_{n\longrightarrow+\infty} \frac{1}{n} \mathbb{E}\Big[\log\Big(\frac{1}{\pi\big(m^{(n)}\big)}\Big) \Big].
\end{aligned}
$$

Pittel [37] montre que ces limites sont bien définies dans son cadre plus général de faible dépendance, grâce à un argument de sous-additivité.

On introduit enfin les variables suivantes :

– ℓ_n, la longueur du plus court chemin de la racine à une feuille de l'arbre $\mathcal{T}_n$;
– $\mathcal{L}_n$, la longueur du chemin le plus long de la racine à une feuille de l'arbre $\mathcal{T}_n$;
– D_n, la profondeur d'insertion de $m(n)$ dans l'arbre $\mathcal{T}_{n-1}$ (pour créer $\mathcal{T}_n$) ;
– Δ_n, la longueur d'un chemin de l'arbre $\mathcal{T}_n$, choisi aléatoirement et uniformément parmi les n chemins possibles.

Notons que $\mathcal{L}_n$ représente la hauteur de l'arbre. Quant à ℓ_n, elle donne des renseignements sur le niveau de saturation. L'analyse de la hauteur et du niveau de saturation est généralement motivée par l'optimisation du coût de mémoire. La hauteur est clairement pertinente de ce point de vue et le niveau de saturation est algorithmiquement pertinent du fait que les nœuds internes en dessous du niveau de saturation sont souvent remplacés par un tableau moins coûteux. Ces paramètres dépendent essentiellement des caractéristiques de la source.

Les variables aléatoires définies ci-dessous jouent un rôle clé dans les preuves. Pour bien fixer le cadre, on rappelle que s est une donnée déterministe et que l'aléa n'est engendré que par le mot m, c'est-à-dire par la construction des arbres $\mathcal{T}_n$.

$$\begin{aligned} X_n(s) &\stackrel{\text{def}}{=} \begin{cases} 0 \text{ si } s_1 \text{ n'est pas dans } \mathcal{T}_n \\ \max\{k \geq 1 \mid \text{le mot } s^{(k)} \text{ est déjà inséré dans } \mathcal{T}_n\} \end{cases} \\ T_k(s) &\stackrel{\text{def}}{=} \min\{n \geq 1 \mid X_n(s) = k\}. \end{aligned}$$

$T_k(s)$ désigne ainsi la taille du premier arbre où $s^{(k)}$ est inséré. On peut noter que $T_0(s) = 0$. Ces deux variables sont en dualité au sens où

$$\{X_n(s) \geq k\} = \{T_k(s) \leq n\}.$$

Ici on obtient donc $\{T_k(s) = n\} \subset \{X_n(s) = k\}$. La variable $X_n(s)$ désigne la longueur de la branche correspondant à s dans l'arbre $\mathcal{T}_n$. Ainsi les longueurs ℓ_n et $\mathcal{L}_n$ peuvent-elles naturellement s'exprimer en fonction de X_n par les relations

$$\ell_n = \min_{s \in \mathcal{A}^{\mathbb{N}}} X_n(s) \quad \text{et} \quad \mathcal{L}_n = \max_{s \in \mathcal{A}^{\mathbb{N}}} X_n(s).$$

La variable $T_k(s)$ admet la décomposition

$$T_k(s) = \sum_{r=1}^{k} Z_r(s),$$

où $Z_r(s) \stackrel{\text{def}}{=} T_r(s) - T_{r-1}(s)$ est le nombre de symboles à lire pour que la longueur de la branche décrivant le mot s augmente de 1. Du point de vue de la séquence, c'est aussi le temps d'attente

$$Z_r(s) \stackrel{\text{def}}{=} \min\{n \geq 1 \big| a_{n+T_{r-1}(s)} \dots a_{n+T_{r-1}(s)-r+1} = s_1 \dots s_r\}.$$

Les variables aléatoires $Z_r(s)$ sont indépendantes. On utilise pour montrer le théorème suivant, l'expression de la fonction génératrice d'occurrence d'un motif, partant d'un

autre motif, dans une chaîne de Markov d'ordre un, donnée par Robin et Daudin [41]. A nouveau dans l'arbre des suffixes présenté dans la section 2.5.2 la preuve reposera sur l'expression d'une série génératrice d'occurrences de mot. Comme la source ne sera plus Markovienne mais une VLMC, on utilisera la proposition 2.3.12.

Théorème 2.5.1 (Convergence p.s. des branches critiques).

$$\frac{\ell_n}{\log n} \underset{n\to\infty}{\overset{\text{p.s.}}{\longrightarrow}} \frac{1}{h_+} \quad et \quad \frac{\mathcal{L}_n}{\log n} \underset{n\to\infty}{\overset{\text{p.s.}}{\longrightarrow}} \frac{1}{h_-}.$$

Quelques éléments de preuve :
La première partie de la preuve repose sur l'introduction d'une martingale résultant d'une somme de variables aléatoires centrées :

$$M_n(s) = \sum_{k=1}^{n} \left(Z_k(s) - \mathbb{E}\left[Z_k(s)\right]\right).$$

Avec un contrôle du crochet de cette martingale et en appliquant la loi des grands nombres pour les martingales, on obtient un premier résultat sur la vitesse de convergence presque sûre de $T_k(s)$. Par dualité, on en déduit des résultats pour les longueurs $X_n(s)$. Cette partie de la preuve permet de montrer les inégalités

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

Les deux autres inégalités complémentaires s'obtiennent avec un travail asymptotique plus fin, basé sur des inégalités de concentration de type inégalité de Chernoff en utilisant la série génératrice de Robin et Daudin [41]. On constitue deux familles de mots : l'une contenant les mots plutôt auto-recouvrant, pour lesquels la majoration obtenue est assez grossière mais compensée par le fait que le nombre de tels mots est faible, et l'autre famille pour laquelle la majoration est beaucoup plus fine et pour laquelle on utilise le comportement analytique de la fonction di-logarithme. On conclue la preuve avec le théorème de Borel-Cantelli.

Convergence en probabilité de la profondeur d'insertion

La profondeur d'insertion D_n est définie comme la longueur du chemin partant de la racine et conduisant au nœud où $m(n)$ est inséré. En d'autres termes, D_n est le nombre de lettres nécessaires à parcourir avant de trouver la position de $m(n)$. Le théorème 2.5.1 a des conséquences immédiates sur le comportement asymptotique de D_n. En effet, puisque $D_n = \ell_n$ lorsque $\ell_{n+1} > \ell_n$, ce qui se produit infiniment souvent presque sûrement puisque $\lim_{n\to\infty} \ell_n = \infty$ p.s., on en déduit que

$$\liminf_{n\to\infty} \frac{D_n}{\log n} = \liminf_{n\to\infty} \frac{\ell_n}{\log n} = \frac{1}{h_+}.$$

De même, puisque $D_n = \mathcal{L}_n$ lorsque $\mathcal{L}_{n+1} > \mathcal{L}_n$, on a

$$\limsup_{n\to\infty} \frac{D_n}{\log n} = \limsup_{n\to\infty} \frac{\mathcal{L}_n}{\log n} = \frac{1}{h_-}.$$

Théorème 2.5.2.

$$\frac{D_n}{\log n} \xrightarrow[n\to\infty]{P} \frac{1}{h} \quad et \quad \frac{\Delta_n}{\log n} \xrightarrow[n\to\infty]{P} \frac{1}{h}.$$

Remarque 2.5.3. Pour une suite i.i.d. $m = a_1 a_2 \ldots$, dans le cas où les variables aléatoires a_n ne sont pas uniformément distribuées sur l'alphabet $\{A, C, G, T\}$, le théorème 2.5.2 implique que $\frac{D_n}{\log n}$ ne converge pas presque sûrement. En effet, on a clairement

$$\limsup_{n\to\infty} \frac{D_n}{\log n} = \frac{1}{h-} > \frac{1}{h} > \frac{1}{h_+} = \liminf_{n\to\infty} \frac{D_n}{\log n}.$$

La visualisation des répétitions de sous-mots dans l'arbre-CGR suggère un rapprochement avec *l'arbre des suffixes*. Ces arbres ont été introduits par Weiner [49] pour accélérer les opérations de recherche de motifs. Ils se construisent sur le même schéma récursif que les *tries*, à partir de l'ensemble des suffixes d'un mot donné. Dans la section 2.5.2, on s'intéresse au comportement asymptotique de l'arbre des suffixes construit à partir d'une source VLMC.

2.5.2 Trie des Suffixes

Le *Trie* (abréviation de re*trie*val) est une structure de données introduite par Fredkin [20], efficace pour chercher des mots dans un ensemble donné et utilisée dans de nombreux algorithmes de compression de données comme les correcteurs d'orthographe ou les systèmes de recherche d'adresses IP. Le *trie* utilise une règle de construction récursive qui sépare les mots selon leurs préfixes. On stoppe la procédure dès que tous les mots sont distingués. Un *trie* est un arbre digital dans lequel les mots sont insérés sur les noeuds externes (les feuilles).

De même que pour les DST, dès qu'un ensemble de mots est donné, la façon dont ils sont insérés dans le *trie* est déterministe. Néanmoins, un *trie* devient aléatoire quand les mots sont tirés au hasard par exemple quand chaque mot est produit par une source probabiliste. Un *trie des suffixes* est un *trie* construit sur les suffixes de *un* mot infini. Le hasard vient alors de la source de production d'un tel mot infini et les mots successifs insérés dans l'arbre sont, comme pour l'arbre-CGR, loin d'être indépendants .

Les *tries des suffixes*, également appelés arbres des suffixes dans la littérature, ont été développés avec les outils de l'analyse d'algorithmes, de la théorie des probabilités et de la théorie ergodique. La construction de ces algorithmes date des travaux de Weiner [49] en 1973 et on trouve depuis de nombreuses applications en informatique et en biologie (voir par exemple le livre de Gusfield [25]). Comme application majeure des *tries des suffixes* on peut citer le célèbre algorithme de compression LZ77. Les premiers

résultats sur la hauteur des *tries des suffixes* sont dus à Szpankowski [45] et Devroye, Szpankowski et Rais [17]. La recherche d'un mot s'effectue efficacement dans un arbre des suffixes : il suffit de descendre dans la branche correspondant au mot. Si le nœud codant le mot est présent, alors ce mot a déjà été rencontré dans la séquence.

La vitesse à laquelle poussent les *tries des suffixes* (appelés aussi arbres des préfixes dans le livre de Shields [43]) est étroitement liée aux instants de deuxième occurrence des motifs. On trouve des résultats sur ce sujet dans les travaux traitant d'occurrences de mots, de temps d'attente ou d'instants de renouvellement. Citons à titre d'exemple les travaux de Shields [42] ou Wyner et Ziv [53].

On s'intéresse ici à la hauteur H_n et au niveau de saturation ℓ_n d'un *trie des suffixes* $\mathcal{T}_n$ contenant les n premiers suffixes d'un mot infini produit par une source générée par la VLMC du peigne.

Les arbres des suffixes que l'on rencontre dans la littérature ont une hauteur et un niveau de saturation à croissance logarithmique du nombre de mots insérés. Lorsque les mots insérés dans les *tries* sont indépendants (on parle de *tries* ordinaires par opposition aux *tries* des suffixes), les résultats de Pittel [37] reposent sur deux hypothèses concernant la source qui produit les mots : la source est supposée uniformément mélangeante et la probabilité d'un mot est supposée décroître exponentiellement avec sa longueur. Citons aussi l'analyse générale sur les *tries* de Clément, Flajolet et Vallee [5] construits à partir de sources dynamiques.

Pour les *tries des suffixes*, Szpankowski [44] obtient le même résultat, avec une hypothèse plus faible de mélange (toujours uniforme quand même) et avec la même hypothèse sur la mesure des mots. Néanmoins, Shields [42] indique un résultat sur les préfixes de processus ergodiques suggérant que le niveau de saturation des *tries des suffixes* ne va pas nécessairement croître de façon logarithmique avec le nombre de mots insérés.

Ces structures digitales ont habituellement dans la littérature une hauteur en $\log n$, n étant le nombre de données stockées. Dans l'article [10], nous présentons deux exemples explicites de tries des suffixes où soit la hauteur soit le niveau de saturation n'est pas logarithmique.

Algorithme de construction du *trie des suffixes*

Soit $(y_n)_{n\geqslant 1}$ une suite de mots infinis à droite sur $\{0,1\}$. On construit à partir de cette suite un **processus de tries** $(\mathcal{T}_n)_{n\geq 1}$ qui est une suite croissante d'arbres binaires construite de la façon suivante. L'arbre $\mathcal{T}_n$ contient les mots $y_1, \ldots, y_n$ sur ses feuilles. Il est obtenu par une construction séquentielle en insérant successivement les mots y_n.

Au début, $\mathcal{T}_1$ est l'arbre contenant la racine et la feuille $0\ldots$ (resp. la feuille $1\ldots$) si y_1 commence par 0 (resp. par 1). Pour $n \geq 2$, étant donné un arbre $\mathcal{T}_{n-1}$, le mot y_n est inséré de la façon suivante. On descend le long de la branche dont les nœuds codent pour les préfixes successifs de y_n ; quand la branche s'arrête, si un nœud interne est atteint, alors y_n est inséré dans la feuille libre, sinon on fait pousser la branche en comparant les

lettres suivantes jusqu'à ce qu'on puisse les insérer dans des feuilles différentes. Comme on le voit sur la figure 2.13, un *trie* n'est pas un arbre complet et l'insertion d'un mot peut faire grandir une branche de plus d'un niveau. Notez que si le *trie* contient un mot fini w comme nœud interne, il y a déjà au moins deux mots insérés qui ont pour préfixe w. Cela indique pourquoi la *seconde* occurrence d'un mot est importante dans la croissance d'un *trie*. Soit $m \stackrel{\text{def}}{=} a_1 a_2 a_3 \dots$ un mot infini sur $\{0, 1\}$. Le **trie des suffixes**

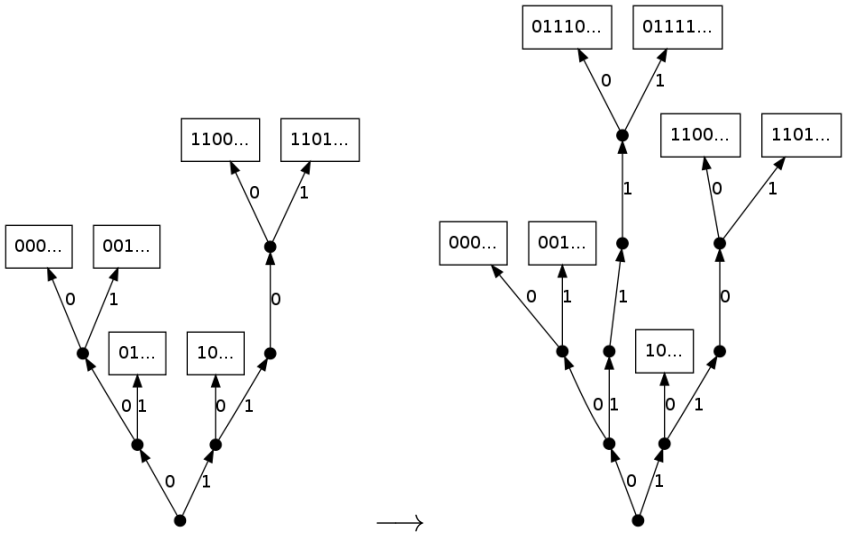


FIGURE 2.13: Dernières étapes de la construction du *trie* de l'ensemble $(000\dots, 10\dots, 1101\dots, 001, \dots, 01110\dots, 1100\dots, 01111\dots)$.

$\mathcal{T}_n$ (avec n feuilles) associé à m est le *trie* construit à partir des n premiers suffixes de m que l'on obtient en effaçant successivement la lettre la plus à gauche, c'est-à-dire

$$y_1 = m,\ y_2 = a_2 a_3 a_4 \dots,\ y_3 = a_3 a_4 \dots,\ \dots,\ y_n = a_n a_{n+1} \dots$$

On note $\mathcal{L}_n$ *la hauteur*. On note également l_n le *niveau de saturation* qui est le niveau maximal auquel tous les nœuds internes sont présents dans $\mathcal{T}_n$. Formellement, si $\partial\mathcal{T}_n$ désigne l'ensemble des feuilles de $\mathcal{T}_n$,

$$\begin{aligned} \mathcal{L}_n &= \max_{u \in \mathcal{T}_n \setminus \partial\mathcal{T}_n} \{|u|\} \\ l_n &= \max \{ j \in \mathbb{N} |\ \#\{u \in \mathcal{T}_n \setminus \partial\mathcal{T}_n, |u| = j\} = 2^j \}. \end{aligned}$$

On voit sur l'exemple de la figure 2.14 que le niveau de saturation n'est pas toujours la longueur du plus court chemin de la racine à une feuille. On rappelle que les constantes h_+ et h_- sont définies par (2.11) et (2.12). Dans leurs travaux, PITTEL [37] et SZPANKOWSKI [44] considèrent seulement les cas où $h_+ < +\infty$ et $h_- > 0$, ce qui revient à considérer que la probabilité de n'importe quel mot décroît exponentiellement avec sa longueur. Ici,

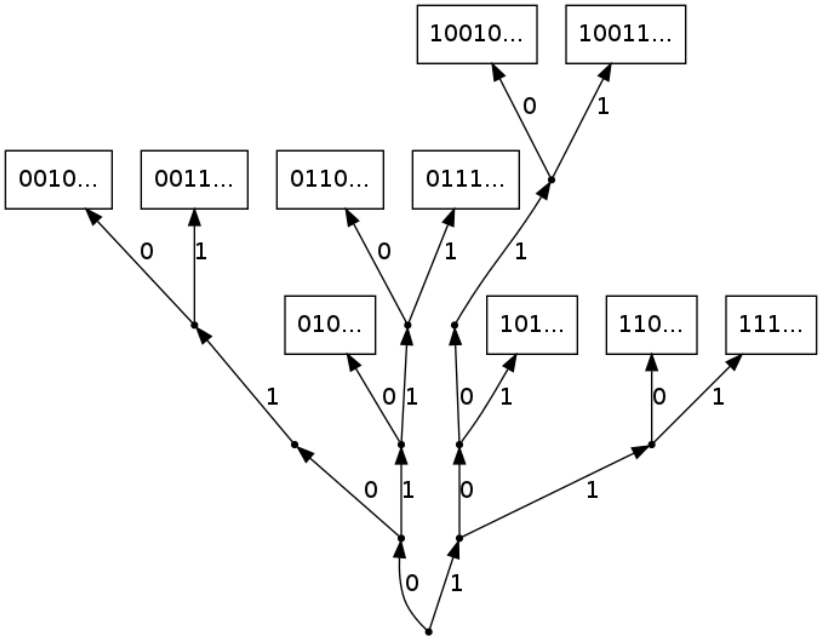


FIGURE 2.14: *Trie des suffixes* $\mathcal{T}_{10}$ associé au mot $1001011001110\ldots$. Ici, $H_{10} = 4$ et $\ell_{10} = 2$.

on s'intéresse à deux exemples où ces hypothèses ne sont pas satisfaites. Pour le peigne logarithmique (défini en section 2.3.3), on a

$$h_- \leqslant \lim_{n\to+\infty} \frac{1}{n} \ln\Big(\frac{1}{\pi\left(10^{n-1}\right)}\Big) = 4 \lim_{n\to+\infty} \frac{\ln n}{n} = 0.$$

Pour le peigne factoriel (voir section 2.3.3), $\pi(10^n)$ est d'ordre $\frac{1}{(n+1)!}$ donc

$$h_+ \geqslant \lim_{n\to+\infty} \frac{1}{n} \ln\Big(\frac{1}{\pi\left(10^{n-1}\right)}\Big) = \lim_{n\to+\infty} \frac{\ln n!}{n} = +\infty$$

Théorème 2.5.4 (Hauteur du peigne logarithmique). *On note $\mathcal{T}_n$ l'arbre des suffixes construit à partir des n premiers suffixes d'une suite générée par un peigne logarithmique. La hauteur $\mathcal{L}_n$ de $\mathcal{T}_n$ vérifie*

$$\forall \delta > 0, \qquad \frac{\mathcal{L}_n}{n^{\frac{1}{4}-\delta}} \xrightarrow[\mathrm{n}\to\infty]{\mathrm{P}} +\infty.$$

Théorème 2.5.5 (Niveau de saturation du peigne factoriel). *On note $\mathcal{T}_n$ l'arbre des suffixes construit à partir des n premiers suffixes d'une suite générée par un peigne factoriel. Pour tout $\delta > 1$, presque sûrement,on a*

$$l_n \in o\left(\frac{\ln n}{(\ln\ln n)^{\delta}}\right).$$

Pour prouver ces deux théorèmes, on utilise la dualité entre les variables $X_n(s)$ et $T_k(s)$, la série génératrice de la deuxième occurrence d'un motif pour une source générée par un peigne, ainsi que les propriétés de mélange de cette même source.

Le comportement dynamique asymptotique de la hauteur et du niveau de saturation des peignes logarithmique et factoriel est représenté sur la figure 2.15. Sur les deux graphiques, les courbes en pointillés longs représentent le comportement d'un troisième peigne, appelé *peigne en* $\ln n$, pour lequel $c_n = \frac{1}{3}\prod_{k=1}^{n-1}\left(\frac{1}{3} + \frac{1}{(1+k)^2}\right)$ pour $n \geqslant 1$. Ce peigne a un mélange exponentiel, la constante h_+ est finie et h_- est strictement positive. La hauteur et le niveau de saturation convergent tous les deux en $\ln n$ (ce résultat est une conséquence des travaux de Pittel [37] et Szpankowski [44]).

Ces comportements asymptotiques différents, provenant tous du même modèle, le peigne, illustrent sa richesse.

Références

[1] M. Abadi et A. Galves. « Inequalities for the occurrence times of rare events in mixing processes. The state of the art ». Dans : *Markov proc. Relat. Fields* 7(1) (2001), p. 97–112.

[2] Y. Baraud, S. Huet et B. Laurent. « Adaptive tests of linear hypotheses by model selection ». Dans : *Ann. Statist.* 31.1 (2003), p. 225–251. ISSN : 0090-5364.

[3] G. Blom et D. Thorburn. « How many random digits are required until given sequences are obtained ? » Dans : *Journal of Applied Probabilities* 19 (1982), p. 518–531.

[4] P. Bühlmann et A.J. Wyner. « Variable length Markov chains ». Dans : *Ann. Statist.* 27.2 (1999), p. 480–513.

[5] J. Clément, P. Flajolet et B. Vallee. « Dynamical sources in Information Theory: Analysis of general tries ». Dans : *Algorithmica* 29 (2001), p. 307–369.

[6] F. Comets, R. Fernandez et P. Ferrari. « Processes with long memory: Regenerative construction and perfect simulation ». Dans : *Ann. of Appl. Prob.* 12.3 (2002), p. 921–943.

[7] P. C. « Test on the Structure of Biological Sequences via Chaos Game Representation ». Dans : *Statistical Applications in Genetics and Molecular Biology* 4.1 (2005). 34 pages.

[8] P. C., B. Chauvin, S. Herrmann et P. Vallois. « Persistent random walks, variable length Markov chains and piecewise deterministic Markov processes ». Dans : *Markov Processes and Related Fields* 19.1 (2013), p. 1–50.

[9] P. C., B. Chauvin, F. Paccaut et N. Pouyanne. « Context trees, variable length Markov chains and dynamical sources ». Dans : *Séminaire de Probabilités* XLIV (2012).

[10] P. C., B. Chauvin, F. Paccaut et N. Pouyanne. « Uncommon Suffix Tries ». Dans : *A paraître dans Random Structures and Algorithms* (2013).

[11] P. C., B. Chauvin, N. Pouyanne et S. Ginouillac. « Digital Search Trees and Chaos Game Representation ». Dans : *ESAIM Probability and Statistics* 13 (2009), p. 15–37.

[12] P. C., G. Fayolle et J.-M. Lasgouttes. « Dynamical Systems in the Analysis of Biological Sequences ». Dans : *Rapport de recherche INRIA* 5351 (oct. 2004). 47 pages.

[13] M. H. A. Davis. *Markov models and optimization*. T. 49. Monographs on Statistics and Applied Probability. London : Chapman & Hall, 1993, p. xiv+295.

[14] M. H. A. Davis. « Piecewise-deterministic Markov processes: a general class of nondiffusion stochastic models ». Dans : *J. Roy. Statist. Soc. Ser. B* 46.3 (1984). With discussion, p. 353–388.

[15] P.J. Deschavanne et al. « Genomic Signature: Characterization and Classification of species Assessed by Chaos Game Representation of sequences ». Dans : *Mol. Bio. Evol.* 16 (1999), p. 1391–1399.

[16] L. Devroye et R. Neininger. « Random suffix search trees ». Dans : *Random Structures Algorithms* 23.4 (2003), p. 357–396. issn : 1042-9832.

[17] L. Devroye, W. Szpankowski et B. Rais. « A note on the height of suffix trees ». Dans : *SIAM J. Comput.* 21.1 (1992), p. 48–53.

[18] Eugene C. Eckstein, Jerome A. Goldstein et Mark Leggas. « The mathematics of suspensions: Kac walks and asymptotic analyticity ». Dans : *Proceedings of the Fourth Mississippi State Conference on Difference Equations and Computational Simulations (1999)*. T. 3. Electron. J. Differ. Equ. Conf. San Marcos, TX : Southwest Texas State Univ., 2000, p. 39–50.

[19] J. Fayolle. « Compression de données sans perte et combinatoire analytique ». Thèse de doct. Université Paris VI, 2006.

[20] Edward Fredkin. « Trie memory ». Dans : *Commun. ACM* 3.9 (sept. 1960), p. 490–499. issn : 0001-0782. doi : 10.1145/367390.367400. url : http://doi.acm.org/10.1145/367390.367400.

[21] J.C. Fu. « Bounds for reliability of large consecutive-K-out-of-N : F system ». Dans : *IEEE trans. Reliability* 35 (1986), p. 316–319.

[22] J.C. Fu et M.V. Koutras. « Distribution theory of runs: a Markov chain approach ». Dans : *J. Amer. Statist. Soc.* 89 (1994), p. 1050–1058.

[23] A. Galves et E. Löcherbach. « Stochastic chains with memory of variable length ». Dans : *TICSP Series* 38 (2008), p. 117–133.

[24] H. Gerber et S. Li. « The occurrence of sequence patterns in repeated experiments and hitting times in a Markov chain ». Dans : *Stochastic Processes and their Applications* 11 (1981), p. 101–108.

[25] D. Gusfield. *Algorithm on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge Press, 1997.

[26] T.E. Harris. « On chains of infinite order ». Dans : *Pacific J. Math.* 5 (1955), p. 707–724.

[27] S. Herrmann et P. Vallois. « From persistent random walk to the telegraph noise ». Dans : *Stoch. Dyn.* 10.2 (2010), p. 161–196.

[28] S. Isola. « Renewal sequences and intermittency ». Dans : *J. Statist. Phys.* 97.1-2 (1999), p. 263–280.

[29] H.J. Jeffrey. « Chaos Game Representation of gene structure ». Dans : *Nucleic Acid. Res* 18 (1990), p. 2163–2170.

[30] S. Karlin et C. Burge. « Dinucleotide relative abundance extremes: a genomic signature ». Dans : *Trends Genet.* 7 (1995), p. 283–290.

[31] S. Karlin et J. Mràzek. « Compositional differences within and between eukaryotic genomes ». Dans : *Proc. Natl. Acad. Sci. USA* 94 (1997), p. 10227–10232.

[32] S. Karlin et J. Mràzek. « Strand compositional asymmetry in bacterial and large viral genomes ». Dans : *Proc. Natl. Acad. Sci. USA* 95 (1998), p. 3720–3725.

[33] S. Karlin, J. Mràzek et A.M. Campbell. « Compositional biases of bacterial genomes and evolutionary implications ». Dans : *J. Bacteriol.* 179.12 (1997), p. 3899–3913.

[34] M. V. Koutras. « Waiting times and number of appearances of events in a sequence of discrete random variables ». Dans : *Advances in combinatorial methods and applications to probability and statistics*. Stat. Ind. Technol. Boston, MA : Birkhäuser Boston, 1997, p. 363–384.

[35] S.-Y. R. Li. « A martingale approach to the study of occurrence of sequence patterns in repeated experiments ». Dans : *Ann. Probab.* 8.6 (1980), p. 1171–1176.

[36] H. Mahmoud. « Evolution of Random Search Trees ». Dans : New York : John Wiley, 1992. Chap. 6.

[37] B. Pittel. « Asymptotic growth of a class of random trees ». Dans : *Annals Probab.* 13 (1985), p. 414–427.

[38] V. Pozdnyakov, J. Glaz, M. Kulldorff et J. M. Steele. « A martingale approach to scan statistics ». Dans : *Ann. Inst. Statist. Math.* 57.1 (2005), p. 21–37.

[39] Eric Renshaw et Robin Henderson. « The correlated random walk ». Dans : *J. Appl. Probab.* 18.2 (1981), p. 403–414.

[40] J. Rissanen. « A universal data compression system ». Dans : *IEEE Trans. Inform. Theory* 29.5 (1983), p. 656–664.

[41] S. Robin et J.J. Daudin. « Exact distribution of word occurences in a random sequence of letters ». Dans : *J. Appl. Prob.* 36 (1999), p. 179–193.

[42] P.C. Shields. « Entropy and prefixes. » Dans : *The Annals of Probability* 20, no 1 (1992), p. 403–409.

[43] P.C. Shields. *The Ergodic Theory of Discrete Sample Paths*. Graduate Studies in Mathematics. American Mathematical Society, 1996.

[44] W. Szpankowski. « Asymptotic properties of data compression and suffix trees ». Dans : *IEEE Trans. Information Theory* 39.5 (1993), p. 1647–1659.

[45] W. Szpankowski. « On the Height of Digital Trees and Related Problems ». Dans : *Algorithmica* 6 (1991), p. 256–277.

[46] Pierre Vallois et Charles S. Tapiero. « A claims persistence process and insurance ». Dans : *Insurance Math. Econom.* 44.3 (2009), p. 367–373.

[47] Pierre Vallois et Charles S. Tapiero. « Memory-based persistence in a counting random walk process ». Dans : *Phys. A.* 386.1 (2007), p. 303–307.

[48] X-J. Wang. « Statistical physics of temporal intermittency ». Dans : *Physical Review A* 40, no 11 (1989), p. 6647–6661.

[49] P. Weiner. « Linear pattern matching algorithm ». Dans : *14th Annual IEEE Symposium on Switching and Automata Theory*. Washington, DC, 1973, p. 1–11.

[50] George H. Weiss. *Aspects and applications of the random walk*. Random Materials and Processes. Amsterdam : North-Holland Publishing Co., 1994, p. xvi+361.

[51] George H. Weiss. « Some applications of persistent random walks and the telegrapher's equation ». Dans : *Phys. A* 311.3-4 (2002), p. 381–410.

[52] D. Williams. *Probability with martingales*. Cambridge Mathematical Textbooks. Cambridge : Cambridge University Press, 1991, p. xvi+251.

[53] A.D. Wyner et J Ziv. « Some Asymptotic Properties of the Entropy of a Stationary Ergodic Data Source with Applications to Data Compression ». Dans : *IEEE Transactions on Information Theory* 35, no 6 (1989), p. 1250–1258.

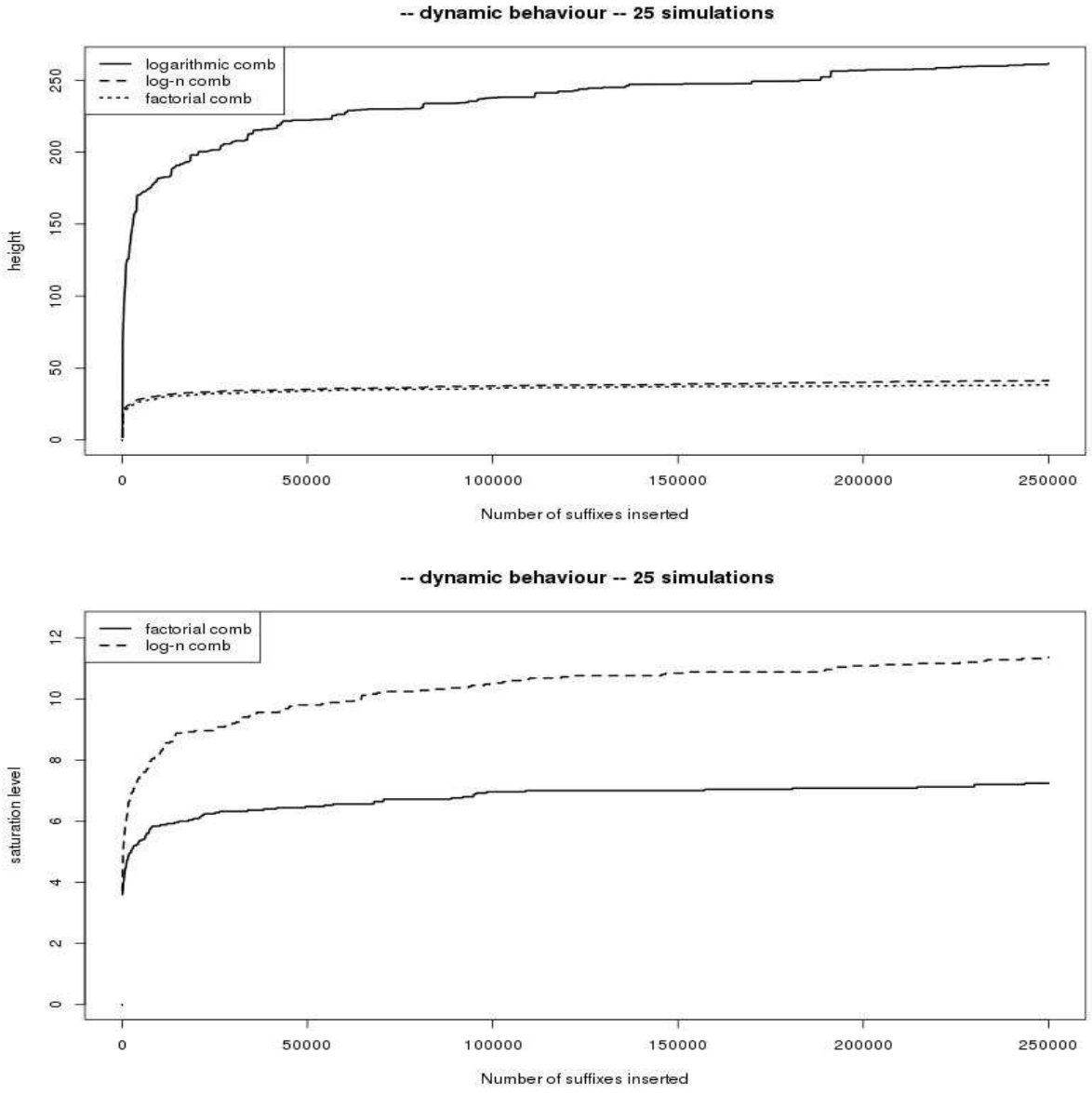


FIGURE 2.15: En haut : représentation de la hauteur d'un peigne logarithmique (trait plein) comparée à la hauteur d'un peigne en $\ln n$ (longs pointillés) et avec la hauteur d'un peigne factoriel (courts pointillés). En bas : représentation du niveau de saturation du peigne factoriel (trait plein) comparé au niveau de saturation du peigne en $\ln n$ (pointillés longs). Dans les deux graphiques, on représente la courbe moyenne sur 25 simulations.

Chapitre 3

Algorithmes Stochastiques

Ce chapitre est une présentation des travaux [2, 3, 4, 5, 15, 16] listés au chapitre 1.

Sommaire

3.1 Introduction

Dans ce chapitre, on utilise différents algorithmes de descente de gradient pour plusieurs problèmes d'optimisation. La première section est une présentation des algorithmes stochastiques utilisés dans les sections suivantes. On applique ensuite ces méthodes pour déterminer des courbes médianes, faire de la classification non hiérarchique et pour minimiser un indicateur de risque.

3.1.1 Algorithme du gradient

On recherche un zéro d'une fonction f à valeurs réelles. On suppose disposer d'une valeur très approximative x_0 de cette racine. L'idée naturelle de l'algorithme de Newton

est de remplacer la courbe représentative de la fonction f par sa tangente au point x_n. L'abscisse x_{n+1} du point d'intersection de cette tangente avec l'axe des abscisses est alors donnée pour $n \geq 1$ par

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

Sous des hypothèses sur la nature du point fixe de la fonction $x \mapsto x - \frac{f(x)}{f'(x)}$ on obtient la convergence et la vitesse de l'algorithme. Lorsque la dérivée de f n'est pas facilement calculable, on peut considérer une version déterministe de l'algorithme de Robbins-Monro qui consiste à remplacer le calcul de la dérivée par une suite positive décroissante tendant vers 0 de pas (γ_n) telle que la série $\sum \gamma_n$ diverge. A condition que f ait de bonnes propriétés de régularité et convexité (voir par exemple DUFLO [8]), la suite définie par

$$x_{n+1} = x_n - \gamma_n f(x_n)$$

converge vers le zéro de la fonction f, noté x^*, pour toute valeur initiale x_0.

FIGURE 3.1: Algorithme de Newton sur $f(x) = e^{x/2} - 2$ (à gauche) et de Robbins-Monro sur $f(x) = \ln(x+1)$ (à droite).

3.1.2 Algorithme de Robbins-Monro

Pour calculer un minimum ou un zéro de fonction, on emploie en général des algorithmes déterministes. Cependant, lorsque la fonction a de nombreux minima proches les uns des autres, ces algorithmes risquent de rester piégés dans des minima locaux. De plus, très souvent dans la modélisation stochastique, la fonction f n'est connue qu'à une perturbation près. On suppose donc que f est de la forme $f(x) = \mathbb{E}[F(x, \xi)]$, où ξ est une variable aléatoire. L'évaluation de la fonction f ne peut être faite que par des mesures approchées. L'algorithme de ROBBINS et MONRO [28] consiste à remplacer à chaque étape la valeur inconnue $f(X_n)$ par son observation bruitée $Y_{n+1} = F(X_n, \xi_{n+1})$ où les variables aléatoires $\xi_1, \ldots, \xi_N$ sont des différences de martingales. On a en particulier $\mathbb{E}[Y_{n+1}|\mathcal{F}_n] = f(X_n)$. L'algorithme de Robbins-Monro est donc défini par

$$X_{n+1} = X_n - \gamma_{n+1} Y_{n+1}, \quad \text{pour } n \geq 1$$

et X_0 est un point initial arbitrairement choisi. La littérature sur le sujet est très abondante, les hypothèses typiques associées à ce modèle étant généralement :

- un cadre réel ou multidimensionnel $\mathbb{R}^d$ pour les variables (Y_n),
- une suite de pas (γ_n) ni « trop grande » ni « trop petite »,
- un comportement « raisonnable » de la fonction f près du point,
- un comportement « raisonnable » à l'infini.

De telles conditions permettent d'obtenir la convergence de l'algorithme et la normalité asymptotique. Des références bibliographiques sur ce sujet sont données dans l'introduction de la partie 4.4.

En pratique, l'algorithme de Robbins-Monro est très sensible aux paramètres comme le point de départ X_0 ou la calibration de la suite de pas (γ_n). La vitesse optimale de l'algorithme est atteinte pour un pas de la forme $\gamma_n = c/n$ mais le choix de la constante c est extrêmement important et la déterminer en pratique s'avère très délicat (voir par exemple DUFLO [8, Th. 2.2.12]). Pour contourner ce choix critique, POLYAK et JUDITSKY [27] proposent une version «moyennisée» de cet algorithme en prenant la suite de pas (γ_n) à décroissance plus lente que la vitesse «optimale», par exemple $\gamma_n = n^{-3/4}$. Avec ce choix, les oscillations de l'algorithme (X_n) sont très fortes. Ils lissent alors cette suite d'estimateurs en considérant le «moyennisé» : $\frac{1}{n}\sum_{k=1}^{n} X_k$. Sous de bonnes conditions sur f, on peut montrer que la convergence est encore vraie et que la normalité asymptotique est également vérifiée, avec la variance optimale, au sens où l'on atteint la covariance de l'algorithme statique. En pratique, cet algorithme est beaucoup moins sensible au choix des paramètres.

3.1.3 Algorithme de Kiefer-Wolfowitz

L'algorithme de KIEFER et WOLFOWITZ [17] a été introduit en 1952 pour estimer le maximum d'une fonction $f : x \mapsto \mathbb{E}[F(x, \xi)]$, connue à une perturbation près. Lorsque le gradient de f n'est pas observable, cet algorithme permet d'approcher le maximum en utilisant une approximation du gradient de f par des différences finies. Il est défini pour $n \geq 1$, par l'itération

$$X_{n+1} = X_n - \gamma_{n+1}\left(\frac{F(X_n + c_n, \xi_n^1) - F(x_n - c_n, \xi_n^2)}{2c_n}\right),$$

où $(\xi_n^1)_n$ et $(\xi_n^2)_n$ sont deux suites de variables aléatoires indépendantes et identiquement distribuées, (c_n) et (γ_n) sont deux suites déterministes positives et décroissantes vers 0 vérifiant :

$$\sum_n \gamma_n = \infty, \quad \sum_n c_n\gamma_n < \infty, \quad \sum_n (\gamma_n/c_n)^2 < \infty.$$

La suite (c_n) désigne les largeurs des différences finies utilisées pour l'approximation du gradient, tandis que la suite (γ_n) représente les pas de la descente du gradient. Sous de bonnes hypothèses de régularité et convexité/concavité de la fonction f, cet algorithme converge vers le maximum/minimum visé (voir DUFLO [8]).

3.1.4 Descente en miroir

Il existe aussi des méthodes d'algorithmes stochastiques permettant de prendre en compte des espaces de contraintes. Parmi ces méthodes, l'algorithme de descente en

miroir a l'avantage de ne pas nécessiter de recalculer à chaque itération une projection sur l'ensemble des contraintes.

A partir de la version déterministe de l'algorithme de descente en miroir introduite par Nemirovsky et Yudin [24], la version stochastique a été proposée par Nesterov [25] et utilisée dans Yuditski, Nazin, Tsybakov et Vayatis [32], Juditsky, Rigollet et Tsybakov [14] pour résoudre des problèmes d'optimisation convexe non-lisses et stochastiques. Récemment, de nouveaux algorithmes stochastiques d'approximation avec des propriétés de convergence fortes, basées sur la méthode Nesterov ont été proposées (voir par exemple Lan [19], Nemirovsky, Juditsky, Lan et Shapiro [23]). Tous ces algorithmes reposent sur l'existence d'un *oracle stochastique*, qui est un mécanisme pour générer des versions bruitées de la pente.

On munit $\mathbb{R}^d$ de la norme L^1 $\|\|\|$ et son espace dual $(\mathbb{R}^d)^*$ de la norme duale $\|\|\|_*$:

$$\|x\| = \sum_{i=1}^{d} |x_i| \,, \quad \|\xi\|_* = \sup_{i=1,\dots,d} |\xi_i| \,.$$

Soit C un sous-ensemble convexe compact de $\mathbb{R}^d$. L'algorithme stochastique de descente en miroir requiert l'utilisation d'une *fonction auxiliaire* qui est utilisé pour pousser la trajectoire dans l'ensemble de contraintes. A partir d'une fonction fortement convexe et différentiable notée δ, on définit la *fonction auxiliaire* $V : \mathbb{R}^d \to \mathbb{R}$ par :

$$V(x) \stackrel{\text{def}}{=} \delta(x) - \delta(x_0) - \langle \Delta\delta(x_0), x - x_0 \rangle \,,$$

où x_0 est un point de C et Δ désigne l'opérateur de gradient. Une fonction continue δ est dite fortement convexe de paramètre α si elle vérifie pour tout $\lambda \in [0,1]$:

$$\delta(\lambda x + (1-\lambda)y) \leq \lambda\delta(x) + (1-\lambda)\delta(y) - \frac{\alpha}{2}\lambda(1-\lambda)\|x-y\|^2 \,.$$

Il est facile de voir que V est fortement convexe. Soit $\beta > 0$, on note W_β la transformée de Fenchel-Legendre de βV :

$$W_\beta(\xi) = \sup_{x \in C} \{\langle \xi, x \rangle > -\beta V(x)\} \,.$$

L'algorithme de descente en miroir permet de trouver un extremum d'une fonction pour laquelle on n'a accès qu'à une version bruitée du gradient que l'on note ψ. Cet algorithme utilise deux suite positives (β_n) et (γ_n) et une suite d'observations vectorielles aléatoires $(\mathcal{Y}^n)$ de la façon suivante :

Algorithme 1

Initialisation :
$$\begin{cases} \xi_0 = 0 \in (\mathbb{R}^d)^\star \\ \chi_0 \in C \end{cases}$$

Mise à jour :
$$\text{pour } n = 1, \ldots, N$$
$$\begin{cases} \xi_n = \xi_{n-1} - \gamma_n \Psi(\chi_{n-1}, \mathcal{Y}^n) \\ \chi_n = \nabla W_{\beta_n}(\xi_n) \end{cases}$$

Sortie :
$$S_N = \frac{\sum_{n=1}^N \gamma_n \chi_{n-1}}{\sum_{n=1}^N \gamma_n}$$

Le principe de cet algorithme est d'effectuer la descente de gradient dans l'espace dual. La transformée de Fenchel-Legendre de la fonction auxiliaire renormalisée par une suite de pas adéquats permet de « renvoyer » la trajectoire dans l'espace des contraintes. La figure 3.1.4 schématise cet algorithme, particulièrement bien adapté aux données en grande dimension.

Ces techniques d'optimisation stochastique sont appliquées dans la suite du chapitre à plusieurs problèmes relatifs à l'estimation de médianes ainsi qu'à l'estimation d'une allocation optimale sous contrainte.

3.2 Estimation de la médiane

Dans cette section, nous utilisons l'algorithme de Robbins-Monro pour l'estimation de médiane dans un espace de Hilbert ou pour la classification (non supervisée) à l'aide de k-médianes.

3.2.1 La médiane dans $\mathbb{R}$

Pour une variable aléatoire réelle Y à densité et de fonction de répartition strictement croissante, la médiane est définie comme étant l'unique valeur m telle que

$$\mathbb{E}\left(\text{sign}(Y-m)\right) = 0 = \mathbb{E}\left(\frac{Y-m}{|Y-m|}\right).$$

Il est ainsi naturel de définir la médiane comme étant

$$m \overset{\text{def}}{=} \arg\min_{z\in\mathbb{R}} \mathbb{E}\left(|Y-z|\right), \tag{3.1}$$

puisque si l'on note $G(z) \stackrel{\text{def}}{=} \mathbb{E}(|Y-z|)$ on a $G'(z) = -\mathbb{E}\left(\frac{Y-z}{|Y-z|}\right)$. La médiane est un *indicateur robuste* dans le sens où son point de rupture est non nul et sa fonction d'influence reste bornée, contrairement à la moyenne.

3.2.2 Médiane géométrique dans un Hilbert

Pour une généralisation dans un cadre multidimensionnel, étant donné que $\mathbb{R}^2$ n'a pas d'ordre naturel privilégié, la définition (3.1) suggère l'extension

$$m = \arg\min_{z\in\mathbb{R}} \mathbb{E}(\|Y-z\|). \tag{3.2}$$

Cette égalité définit la *médiane géométrique* qui n'est pas simple à calculer explicitement de façon générale.

Des algorithmes itératifs d'estimation ont été proposés dans le cadre multivarié par Gower [12] et Vardi et Zhang [30], pour des variétés riemanniennes par Arnaudon, Dombry, Phan et Yang [1], ainsi que par Gervini [11] pour des données fonctionnelles. Ce dernier algorithme nécessite d'inverser à chaque étape une matrice dont la dimension est égale à la dimension des données et nécessite donc d'importants efforts de calcul. L'algorithme proposé par Vardi et Zhang [30] est beaucoup plus rapide et ne nécessite que $\mathcal{O}(nd)$ opérations à chaque itération, où n est la taille de l'échantillon. Les propriétés de ces estimateurs figurent dans une étude détaillée récente de Möttönen, Nordhausen et Oja [22]. Ces procédures d'estimation ne sont pas adaptées lorsque les données arrivent séquentiellement, ils ont besoin de stocker toutes les données car la mise à jour nécessite toutes les observations antérieures.

La définition (3.2) est valable dans un espace vectoriel normé. Pour assurer l'unicité, la norme doit être strictement convexe (la boule unité doit l'être). Pour les espaces de Hilbert séparables et pour certains espaces de Banach comme L^p pour $1 < p < \infty$, Kemperman [16] démontre que la médiane géométrique est unique dès que le support de la loi n'est pas contenu dans une droite.

Dans les articles [5, 6, 4] nous considérons pour les applications l'espace de Hilbert séparable $H = L^2([0,T])$ de *dimension infinie*. On peut alors parler de «fonction médiane» d'une variable aléatoire à valeurs dans H. La médiane fonctionnelle a elle aussi de bonnes proriétés de robustesse. La détection automatique de courbes atypiques n'est pas facile en grande dimension. Estimer la médiane géométrique est par conséquent une alternative intéressante à la détection de courbes atypiques.

3.2.3 Algorithme de Robbins-Monro dans un espace de Hilbert pour l'estimation de la médiane

On suppose que les hypothèses suivantes sont vérifiées.

A1 Le support de Y n'est pas réduit à une droite.

A2 La loi de Y est un mélange : $\mu_Y = \lambda\mu_c + (1-\lambda)\mu_d$, avec

– μ_c vérifie, $\forall x \in H,\ \mu_c(\{x\}) = 0$ et

$$\alpha \mapsto \mathbb{E}\left(\|Y - \alpha\|^{-1}\right)$$

est bornée uniformément sur les boules.

– μ_d est une mesure discrète qui ne charge pas la médiane m.

L'hypothèse **A1** assure l'unicité de la médiane. Quant à **A2**, elle revient à considérer que la distribution n'est pas trop concentrée près de points isolés et elle peut se traduire en termes de petites boules :

$$\mathbb{E}\left(\|Y - m\|^{-1}\right) = \int_0^\infty \mathbb{P}\left(\|Y - m\| \leq t^{-1}\right) dt.$$

Si $\mathbb{P}\left(\|Y - m\| \leq \epsilon\right) \leq C\epsilon^d$, pour ϵ petit, alors on a pour $0 \leq \beta < d$,

$$\mathbb{E}\left(\|Y - m\|^{-\beta}\right) < \infty,$$

La fonction à minimiser est $G : z \mapsto \mathbb{E}\left(\|Y - z\|\right)$.

Proposition 3.2.1. *G est convexe. Sous les hypothèses **A1** et **A2**, excepté sur le support de la loi discrète μ_D, G est deux fois (Fréchet-)différentiable et*

$$D_x G = -\mathbb{E}\left(\frac{Y - x}{\|Y - x\|}\right)$$
$$D_x^2 G = \mathbb{E}\left(\frac{1}{\|Y - x\|}\left(\mathbf{I}_H - \frac{(Y - x) \otimes (Y - x)}{\|Y - x\|^2}\right)\right),$$

où $\mathbf{I}_H$ désigne l'opérateur identité sur H et $u \otimes v(h) = \langle u, h\rangle\, v$ pour u et v dans H.

On définit donc l'algorithme de Robbins-Monro d'estimation de la médiane de la façon suivante :

$$m_{n+1} \ = \ m_n + \gamma_n \frac{Y_{n+1} - m_n}{\|Y_{n+1} - m_n\|} \tag{3.3}$$

où la suite de pas $\gamma_n > 0$ vérifie

$$\sum_{n \geq 1} \gamma_n = \infty \quad \text{et} \quad \sum_{n \geq 1} \gamma_n^2 < \infty. \tag{3.4}$$

Cet algorithme est très rapide à exécuter et sa mise à jour est particulièrement simple : pour un échantillon de taille n de vecteurs de $\mathbb{R}^d$, il ne faut que $\mathcal{O}(nd)$ opérations. Il est de plus *séquentiel* et ne nécessite pas de stocker et sauvegarder les données antérieures.

Théorème 3.2.2. *Sous les hypothèses **A1**, **A2** et* (3.4)*, la suite (m_n) converge presque sûrement quand n tend vers l'infini,*

$$\|m_n - m\| \xrightarrow[\mathrm{n}\to\infty]{\mathrm{p.s.}} 0.$$

On lisse en prenant la moyenne des estimateurs : $\overline{m}_n = \frac{1}{n}\sum_{j=1}^{n} m_j$. La *mise à jour* est alors très simple :

$$\begin{cases} m_{n+1} = m_n + \gamma_n \dfrac{Y_{n+1} - m_n}{\|Y_{n+1} - m_n\|} \\ \\ \overline{m}_{n+1} = \dfrac{n}{n+1}\overline{m}_n + \dfrac{1}{n+1} m_{n+1}. \end{cases}$$

Théorème 3.2.3. *Sous les hypothèses* ***A1*** *et* ***A2*** *ainsi que sous l'hypothèse de moment :*

$$\sup_{h \in \mathcal{B}(m,\epsilon)} \mathbb{E}\left(\|Y - h\|^{-2}\right) < \infty,$$

avec une suite de pas (γ_n) *telle que* $\gamma_n = g/n^\alpha$ *avec* $0,5 < \alpha < 1$, $g \in \mathbb{R}^+$, *on a le théorème de la limite centrale*

$$\sqrt{n}\,(\overline{m}_n - m) \underset{\mathrm{n}\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}(0, \Gamma_m^{-1} V \Gamma_m^{-1}),$$

où Γ_m *désigne la Hessienne de* G *au point* m *et* V *est l'opérateur variance de* $\frac{X-\cdot}{\|X-\cdot\|}$.

La moyennisation permet d'atteindre asymptotiquement au premier ordre la même variance que dans le cas de l'estimateur statique. De plus, l'algorithme de Robbins-Monro est vraiment très rapide. A titre indicatif, pour des simulations en dimension deux où Y est un vecteur gaussien centré, en une seconde, on peut traiter des échantillons de taille :

- $n = 150$ avec l'algorithme de Vardi et Zhang [30] (en R),
- $n = 4500$ avec notre algorithme moyennisé (en R),
- $n = 90\,000$ avec notre algorithme moyennisé (en **C**).

Quelques éléments de preuve : La preuve de la convergence est une application du théorème de Robbins Siegmund. Pour démontrer le théorème de la limite centrale, on utilise un premier résultat sur une majoration de la norme L^2 de l'erreur d'estimation. Ce premier résultat s'appuie sur une décomposition de l'algorithme sous la forme :

$$Z_n - m = \beta_{n-1}(Z_1 - m) + \beta_{n-1} M_n - \beta_{n-1} R_{n-1},$$

où $\beta_n \overset{\text{def}}{=} (I_H - \gamma_n \Gamma_m)(I_H - \gamma_{n-1}\Gamma_m)\ldots(I_H - \gamma_1 \Gamma_m)$, M_n est la martingale définie par

$$M_n \overset{\text{def}}{=} \sum_{k=1}^{n-1} \gamma_k \beta_k^{-1} \left(\Phi(Z_k) - \frac{Y_{k+1} - Z_k}{\|Y_{k+1} - Z_k\|}\right), \quad \Phi(x) \overset{\text{def}}{=} D_x G,$$

et le terme de reste R_n est donné par

$$R_n \overset{\text{def}}{=} \sum_{k=1}^{n} \gamma_k \beta_k^{-1} \left(\Phi(Z_k) - \Gamma_m(Z_k - m)\right).$$

Avec une décomposition spectrale, on contrôle le terme déterministe β_{n-1}. On utilise ensuite la structure de martingale pour contrôler la vitesse de convergence de la martingale et on montre avec la décomposition spectrale que le terme de reste est bien un reste.

Pour prouver le théorème de la limite centrale à partir de la majoration de la vitesse on utilise une autre décomposition de l'algorithme moyennisé :

$$n\Gamma_m\left(\overline{Z}_n - m\right) = \sum_{k=1}^{n} \frac{1}{\gamma_k}(Z_k - Z_{k+1}) - \sum_{k=1}^{n} \left(\Phi(Z_k) - \Gamma_m(Z_k - m)\right) + \widetilde{M}_{n+1},$$

avec la martingale

$$\widetilde{M}_{n+1} \stackrel{\text{def}}{=} \sum_{k=1}^{n} \left(\Phi(Z_k) - \frac{Y_{k+1} - Z_k}{\|Y_{k+1} - Z_k\|} \right).$$

On montre dans un premier temps que le théorème de la limite centrale pour les martingales à valeurs dans un espace de Hilbert s'applique pour notre martingale, puis que les autres termes sont négligeables grâce à la majoration obtenue avec la première décomposition.

3.2.4 Applications

Courbe médiane d'audience télévisuelle

La société Médiamétrie nous a fourni leurs données sur les audiences télévisuelles pour la journée du 6 septembre 2010. Après avoir supprimé de l'échantillon les données des individus n'ayant pas allumé leur poste ce jour là, le jeu de données contient $n = 5423$ courbes d'audience. Pour chaque courbe, nous avons un vecteur $Y_i \in \{0,1\}^{86400}$ où 86400 est le nombre de secondes dans une journée. Chaque seconde Y_i vaut 1 à la seconde s si l'individu i regarde la télévision à cet instant et 0 sinon. La figure 3.2 représente un exemple de profil d'audience sur une journée. Grâce à l'algorithme de Robbins-Monro, on peut déterminer un profil médian d'audience du 6 septembre 2010. Ce profil médian est représenté sur la figure 3.3. Un indicateur d'audience classique est le profil moyen, représentant la proportion de gens qui regardent la télévision à chaque seconde au cours du temps. On a également ajouté ce profil moyen sur la même figure.

Courbe médiane de consommation électrique

Dans [3], nous analysons un échantillon de plus de 18902 courbes de consommation d'électricité mesurées chaque demi-heure sur une période d'une semaine. On représente sur la figure 3.4 la courbe moyenne de consommation ainsi que la courbe médiane. La différence marquée entre les deux courbes est dûe à la présence d'une petite fraction de clients possédant une très grande consommation électrique.

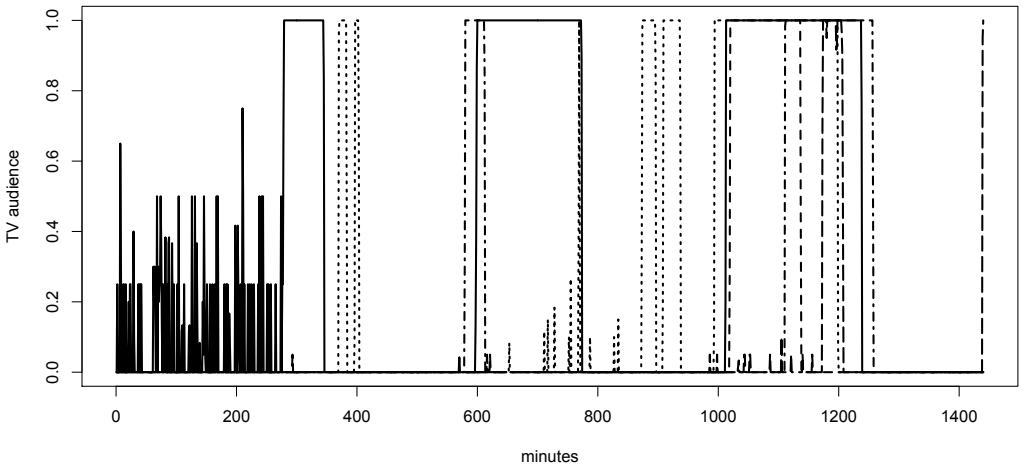


FIGURE 3.2: Exemple de 5 courbes d'audience sur une journée

3.2.5 Estimation de la médiane conditionnelle

On souhaite maintenant prendre en compte des variables corrélées avec la variable étudiée. L'estimation de la façon dont la forme de la courbe peut dépendre de variables réelles ou fonctionnelles donne lieu à une vaste littérature. Le principal inconvénient de tous ces estimateurs est qu'ils dépendent tous, explicitement ou non, d'une méthode basée sur les moindres carrés et sont donc sensibles aux valeurs atypiques aberrantes.

Soit (Y, X) une paire de variables aléatoires à valeurs dans $H \times \mathbb{R}$. On suppose que X est une variable continue. On veut estimer la *médiane géométrique* de Y conditionnellement à $X = x$, c'est-à-dire la valeur $m(x)$ définie par

$$m(x) = \mathbf{argmin}_{\alpha \in H} \mathbb{E}[\|Y - \alpha\| - \|Y\| \, | X = x].$$

De même que dans le cas non conditionné, il y a existence et unicité de $m(x)$ dès que le support de Y sachant $X = x$ n'est pas contenu dans une droite.

On introduit un *algorithme pondéré* par un noyau pour prendre en compte les covariables. L'algorithme de la médiane géométrique (3.3) devient pour la *médiane géométrique conditionnelle* :

$$Z_{n+1}(x) = Z_n(x) + \gamma_n \frac{Y_{n+1} - Z_n(x)}{\|Y_{n+1} - Z_n(x)\|} \frac{1}{h_n} K\left(\frac{X_{n+1} - x}{h_n}\right)$$

avec les deux suites déterministes (h_n) et (γ_n) et le noyau K en paramètres.

On suppose que les hypothèses suivantes sont satisfaites.

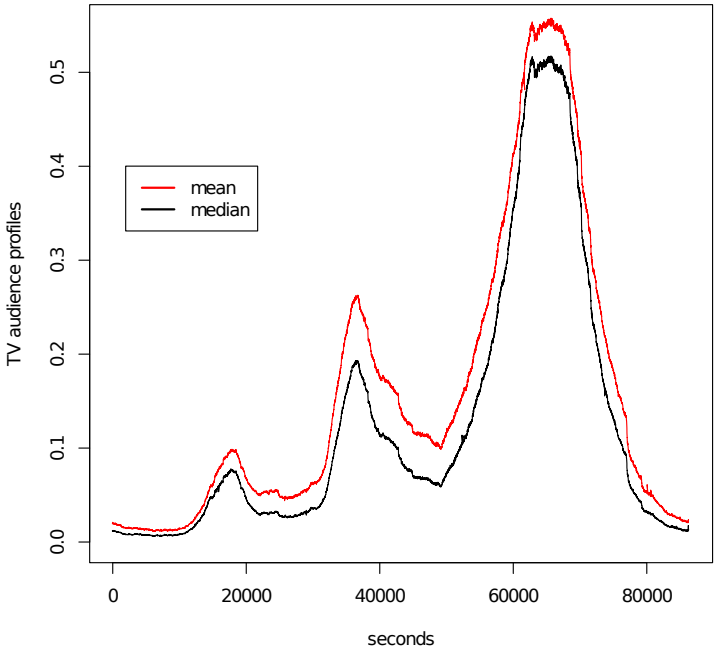


FIGURE 3.3: Courbes d'audience médiane et moyenne, à l'échelle de la seconde, le 6 septembre 2010. La moyenne est « au-dessus » de la médiane traduisant l'impact des personnes regardant beaucoup la télévision sur le profil moyen.

H0. Le noyau K est symétrique autour de 0, positif, à support compact,

$$\int_{\mathbb{R}} K(u)du = 1 \qquad\qquad \int_{\mathbb{R}} K(u)^2 du = \nu^2.$$

H1. Pour tout x dans le support de la densité de X, la variable Y sachant $X = x$ n'est pas concentrée sur une droite.

H2. La densité $p(x)$ de la loi de X est à support borné, continue, deux fois différentiable avec les dérivées bornées.

H3. Les lois conditionnelles $\mu_x = \mathcal{L}(Y|X = x)$ varient régulièrement en fonction de x : il existe deux constantes c et β telles que

$$\mathcal{W}_2\left(\mu_x, \mu_{x'}\right) \leq c\left|x - x'\right|^{\beta},$$

où $\mathcal{W}_2\left(\mu_x, \mu_{x'}\right)$ désigne la distance de Wasserstein entre μ_x et $\mu_{x'}$.

H4. Il existe une constante C telle que

$$\forall \alpha \in H, \forall x \in \mathbb{R}, \mathbb{E}[\|Y - \alpha\|^{-2} |X = x] \leq C.$$

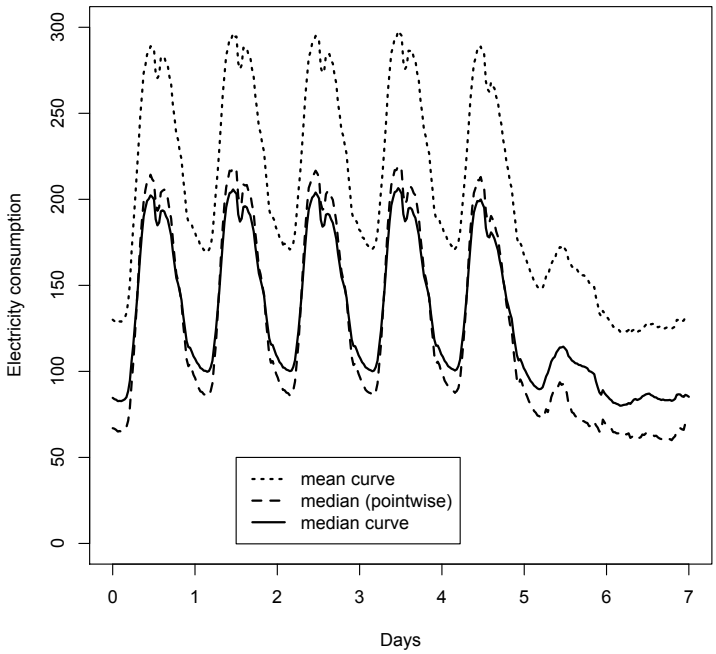


FIGURE 3.4: Comparaison du profil médian estimé avec la courbe de consommation d'électricité moyenne. La courbe en pointillés longs indique, pour chaque abscisse, le point médian des valeurs des courbes en cette abscisse. Ceci illustre que la courbe médiane n'est pas la courbe des points médians. On remarque également que la courbe médiane est moins sensible aux « gros consommateurs » d'électricité qui influencent fortement la moyenne.

L'hypothèse **H1** assure l'unicité (comme **A1** pour la médiane sans conditionnement par une covariable). Les hypothèses **H0** et **H2** sont classiques en estimation non paramétrique et pourraient être assouplies. Les hypothèses de régularité **H3** et **H4** permettent de contrôler l'erreur d'approximation et de prouver la convergence de l'algorithme. L'hypothèse **H4** pourrait être un peu allégée. Elle permet de forcer la loi à s'étendre et d'éviter des écueils pathologiques de comportement de l'algorithme. Enfin, nous avons supposé que la variable X est réelle mais notre algorithme peut être étendu sans difficulté aux covariables multidimensionnelles en considérant une fonction noyau multivariée (voir par exemple Wand et Jones [31]).

On définit l'algorithme moyennisé $\overline{Z}_{n+1}(x) = \frac{1}{n}\sum_{k=1}^{n} Z_k(x)$ et on choisit des suites de pas sous la forme

$$\gamma_n = \frac{c_\gamma}{n^\gamma}, \qquad\qquad h_n = \frac{c_h}{n^h}.$$

Théorème 3.2.4. *Soit x tel que $p(x) > 0$. On suppose que $\gamma < 1$, $2\gamma - h > 1$, $\gamma + \beta h > 1$ et $h > (2\beta+1)^{-1}$. Sous les hypothèses **H0**, **H1**,**H2**, **H3** et **H4**, on a*

$$\sqrt{nh_n}\left(\overline{Z}_n(x) - m(x)\right) \underset{\mathrm{n}\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}\left(0, \Gamma_{m(x)}^{-1}\Sigma_{m(x)}\Gamma_{m(x)}^{-1}\right), \tag{3.5}$$

avec

$$\Sigma_{m(x)} = \lim_{h\to 0}\frac{1}{h}\mathbb{E}\left[K^2\left(\frac{X-m(x)}{h}\right)\frac{(Y-m(x))}{\|Y-m(x)\|}\otimes\frac{(Y-m(x))}{\|Y-m(x)\|}\right],$$

et

$$\Gamma_{m(x)} = \mathbb{E}\left[\frac{1}{\|Y-m(x)\|}\left(\mathbf{I}_H - \frac{(Y-m(x))\otimes(Y-m(x))}{\|Y-m(x)\|^2}\right)\middle| X = x\right].$$

On montre que l'opérateur $\Sigma_{m(x)}$ a un inverse borné puisque, par hypothèse, le support de Y sachant $X = x$ n'est pas contenu dans une droite. Ainsi l'opérateur de variance limite est bien défini. Avec les pas (h_n) choisis, la convergence (3.5) se réécrit

$$\sqrt{nh_n}\left(\overline{Z}_n(x) - m(x)\right) \underset{\mathrm{n}\to\infty}{\overset{\mathcal{L}}{\rightsquigarrow}} \mathcal{N}\left(0, \frac{1}{1+h}\Gamma_{m(x)}^{-1}\Sigma_{m(x)}\Gamma_{m(x)}^{-1}\right).$$

Comme dans le cas de la régression réelle (voir Mokkadem, Pelletier et Slaoui [21]), il s'avère que l'estimateur moyennisé a une plus petite variance asymptotique que l'estimateur du noyau classique/statique, avec dans notre cas un facteur $(1+h)^{-1}$.

Comme pour la médiane, la preuve consiste en une décomposition spectrale minutieuse ainsi qu'une application du théorème de Robbins-Siegmund. On utilise de nouveau une approche de type *forward* permettant de dégager une martingale et d'utiliser le TLC pour martingales hilbertiennes. Enfin, ce théorème est une nouvelle fois une conséquence du théorème de Thalès.

On utilise cet algorithme sur les mêmes données de Médiamétrie pour estimer un profil d'audience télévisuel médian, conditionné par le temps total passé à regarder la

télévision. On cherche à estimer les comportements de consommation de télévision, sur une période de 24 heures. La covariable X représente ici la proportion du temps passé devant la télévision la journée du 6 Septembre 2010. On considère les valeurs quantiles de la covariable X qui sont, pour notre échantillon, $q_{25} = 0,0599$, $q_{50} = 0,128$, $q_{75} = 0,225$ et $q_{90} = 0,348$. Cela signifie par exemple que les dix pour cent des individus avec les niveaux d'audience les plus élevés passent plus de $34,8\%$ de leur temps à regarder la télévision alors que 25% des individus passent moins de 6% de leur temps à regarder la télévision.

Sur la figure 3.5, on représente les estimations des médianes conditionnelles avec une bande passante réglée sur $h_n = 0,05$ et un paramètre de descente $c_\gamma = 0,5$, pour $x \in \{q_{25}, q_{50}, q_{75}, q_{90}\}$. On peut noter que la forme des profils conditionnés dépend fortement de la valeur de la covariable : par exemple, les courbes médianes conditionnées à $x = q_{75}$ et $x = q_{90}$ sont vraiment distinctes en particulier aux instants 15 et 21. D'un point de vue rapidité de calcul, pour un point de départ donné, notre algorithme, qui prend moins de deux secondes, est environ 70 fois plus rapide que l'estimateur statique qui exige 140 secondes pour converger.

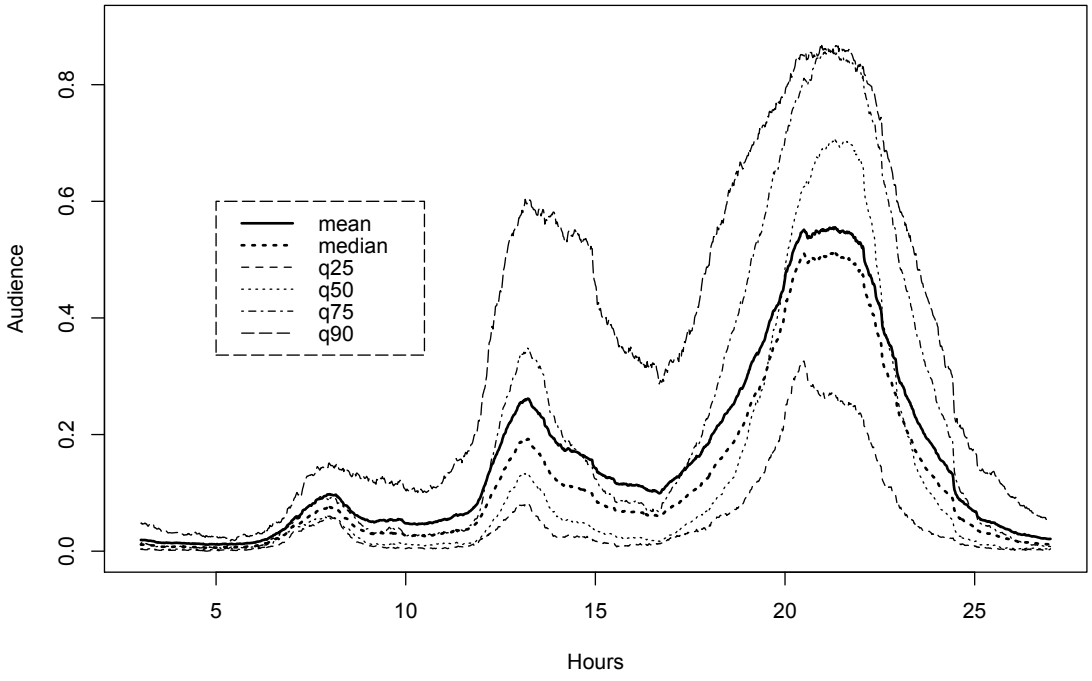


FIGURE 3.5: Estimation du profil median conditionné à différentes valeurs de temps passé devant la télévision le 6 Septembre 2010.

Pour résumer cette première partie, l'algorithme de gradient stochastique permettant d'estimer la médiane d'une distribution à valeurs dans un espace de Hilbert est un algorithme très simple à mettre en place, peu sensible au choix des paramètres lorsque l'on utilise la version *moyennisée*, très rapide en temps d'exécution pour une vitesse de convergence optimale, au sens où l'on atteint la covariance de l'algorithme statique.

3.2.6 Classification automatique non hiérarchique dans $\mathbb{R}^d$

Classifier rapidement de gros échantillons en grande dimension est un enjeu important en statistique computationnelle et en apprentissage, avec des domaines d'applications variés. La littérature sur les techniques de classification est riche comme en attestent les études référencées sur le sujet de Jain, Marty et Flynn [13] et Gan, Ma et Wu [10]. De plus, comme il est souligné dans Bottou [2], le développement d'algorithmes rapides est d'autant plus important que le temps de calcul est limité et l'échantillon important, étant donné que des procédures rapides seront en mesure de traiter un plus grand nombre d'observations et finiront par fournir de meilleures estimations que les plus lentes.

On souhaite trouver une partition de $\mathbb{R}^d$ en k ensembles (classes) homogènes. Le nombre de classes est fixé à l'avance. Chaque classe est caractérisée par son *centre* $\boldsymbol{\theta}^\ell \in \mathbb{R}^d$, $\ell = 1, \ldots, k$, en minimisant la fonction $g : \mathbb{R}^{dk} \mapsto \mathbb{R}$ définie par

$$g(\boldsymbol{\theta}) = \mathbb{E}\left(\min_{\ell=1,\ldots,k} \varphi(\|X - \boldsymbol{\theta}^\ell\|) \right), \tag{3.6}$$

où φ est une fonction croissante sur $\mathbb{R}^+$. Lorsque la fonction mesurant la distance $\varphi(u) = u^2$, trouver les centres de classe conduit à effectuer un *algorithme des k-means*. On trouve dans Forgy [9] une première version non-séquentielle de cet algorithme et dans MacQueen [20] une version séquentielle. Dans Pagès [26], il apparaît comme un cas particulier de l'algorithme de Kohonen (algorithme de *Self-Organizing Maps*). L'algorithme des *k-means* étant basé sur des estimations de moyennes, il est donc sensible aux valeurs atypiques.

Nous nous intéressons dans cette section au cas où l'on cherche à minimiser la norme L^1, c'est-à-dire $\varphi(u) = |u|$, dans (3.6). Nous effectuons un algorithme de type *k-médianes* ; les centres des classes sont les points médians de la classe. Cette approche est une première tentative pour obtenir des algorithmes de classification plus robustes, suggérée par MacQueen [20] et développée par Kaufman et Rousseeuw [15]. Il a été prouvé dans Laloë [18] que, sous des hypothèses générales, le minimum de la fonction objectif est unique. De nombreux algorithmes ont été proposés dans la littérature pour trouver ce minimum.

Dans [4] nous proposons une stratégie récursive de type k-médianes pour estimer les centres des classes. L'un des principaux avantages de notre estimateur est qu'il peut être calculé en seulement $\mathcal{O}(kn)$ opérations et permet donc de traiter de grandes bases de

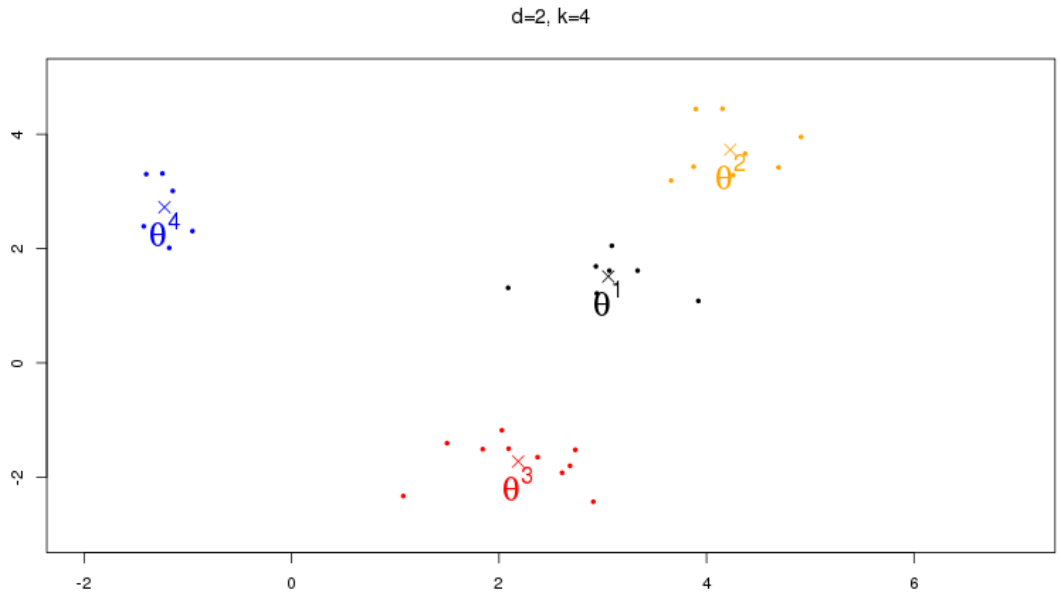


FIGURE 3.6: Exemple de classification en dimension $d = 2$ pour $k = 4$ classes. Les éléments de chaque classe sont représentés de la même couleur. Les centres des classes Θ_ℓ correspondent aux points moyens de la classe.

données. Par sa nature récursive, il permet une mise à jour automatique et n'a pas besoin de stocker toutes les données. Enfin, il est plus robuste que l'estimateur des *k-means*.

On dispose d'un échantillon $X_1, \ldots, X_N$ de vecteurs aléatoires indépendants et identiquement distribués de $\mathbb{R}^d$.

Algorithme récursif des k-means

On note I_ℓ l'indicatrice d'appartenance d'un point $\mathbf{x} \in \mathbb{R}^d$ à la classe ℓ,

$$I_\ell(\mathbf{x}, \boldsymbol{\theta}) = \prod_{j=1}^{k} \mathbb{1}_{\{\|\mathbf{x}-\boldsymbol{\theta}^\ell\| \leq \|\mathbf{x}-\boldsymbol{\theta}^j\|\}}.$$

L'algorithme des k-means est défini de la façon suivante. Pour la *phase d'initialisation*, on choisit arbitrairement k points initiaux de $\mathbb{R}^d$ notés $\boldsymbol{\theta}_1^\ell$. On effectue ensuite la *phase d'itération* de type algorithme de gradient stochastique

$$\boldsymbol{\theta}_{n+1}^\ell = \boldsymbol{\theta}_n^\ell - a_n^\ell I_\ell(\mathbf{X}_n, \boldsymbol{\theta}_n) \left(\boldsymbol{\theta}_n^\ell - \mathbf{X}_n\right),$$

où pour $n \geq 1$,

$$a_n^\ell \stackrel{\text{def}}{=} (1 + n_\ell)^{-1} \quad \text{et} \quad n_\ell \stackrel{\text{def}}{=} 1 + \sum_{r=1}^{n-1} I_\ell(\mathbf{X}_r, \boldsymbol{\theta}_r).$$

Avec cet algorithme, les centres de classes à la n-ième itération sont tout simplement la moyenne des points attribués à cette classe :

$$\boldsymbol{\theta}_{n+1}^\ell = \frac{1}{1 + n_\ell} \left(\boldsymbol{\theta}_1^\ell + \sum_{i=1}^{n} I_\ell(\mathbf{X}_i; \boldsymbol{\theta}_i) \mathbf{X}_i \right).$$

Algorithme récursif des k-médianes

Pour notre algorithme des k-médianes, les centres de classes sont les points médians (au sens de la médiane géométrique) de leur classe. On modifie donc l'algorithme précédent dans sa *phase d'itération* pour $n \geq 1$, et $\ell = 1, \ldots, k$, pour construire l'algorithme de k-médianes :

$$\boldsymbol{\theta}_{n+1}^\ell = \boldsymbol{\theta}_n^\ell - a_n^\ell I_\ell(\mathbf{X}_n, \boldsymbol{\theta}_n) \frac{\boldsymbol{\theta}_n^\ell - \mathbf{X}_n}{\left\|\boldsymbol{\theta}_n^\ell - \mathbf{X}_n\right\|}.$$

Pour des paramètres c_γ, c_α et $1/2 < \alpha \leq 1$ à régler, on choisit des pas de la forme

$$a_n^\ell = \begin{cases} a_{n-1}^\ell & \text{si } I_\ell(\mathbf{X}_n, \boldsymbol{\theta}_n) = 0, \\ \dfrac{c_\gamma}{(1 + c_\alpha n_\ell)^\alpha} & \text{sinon.} \end{cases}$$

Faire la moyenne de l'algorithme permet de diminuer sensiblement sa variabilité et d'améliorer considérablement ses performances. A nouveau on calcule l'*algorithme moyennisé* :

$$\bar{\boldsymbol{\theta}}_{n+1}^{\ell} = \begin{cases} \bar{\boldsymbol{\theta}}_{n}^{\ell} & \text{si } I_\ell(\mathbf{X}_n, \boldsymbol{\theta}_n) = 0 \\ \dfrac{n_\ell \bar{\boldsymbol{\theta}}_{n}^{\ell} + \boldsymbol{\theta}_{n+1}^{\ell}}{n_\ell + 1} & \text{sinon.} \end{cases}$$

On choisit comme points initiaux $\bar{\boldsymbol{\theta}}_1^{\ell} = \boldsymbol{\theta}_1^{\ell}$ pour $\ell = 1, \ldots, k$. On rappelle que l'on dispose d'un échantillon $X_1, \ldots, X_N$ de vecteurs indépendants et de même loi qu'un vecteur X de $\mathbb{R}^d$.

On définit les hypothèses :

(H1) a) Le vecteur aléatoire X a une densité absolument continue.

b) X est borné : $\exists K > 0 : \|X\| \leq K$ p.s.

c) $\exists C$ tel que : $\forall x \in \mathbb{R}^d$ satisfaisant $\|x\| \leq K + 1$, $\mathbb{E}\left[\frac{1}{\|X-x\|}\right] < C$.

(H2) a) $\forall n \geq 1$, $\min_\ell a_n^\ell > 0$.

b) $\max_\ell \sup_n a_n^\ell < \min(\frac{1}{2}, \frac{1}{8C})$ p.s.

c) $\sum_{n=1}^{\infty} \max_\ell a_n^\ell = \infty$ p.s.

d) $\sup_n \frac{\max_\ell a_n^\ell}{\min_\ell a_n^\ell} < \infty$ p.s.

(H3) $\sum_{\ell=1}^{k} \sum_{n=1}^{\infty} \left(a_n^\ell\right)^2 < \infty$ p.s.

(H3') $\sum_{\ell=1}^{k} \sum_{n=1}^{\infty} \mathbb{E}\left[\left(a_n^\ell\right)^2 I_\ell(X_n, \boldsymbol{\theta}_n)\right] < \infty$.

Les hypothèses (H2), (H3) et (H3') sont satisfaites lorsque les classes se remplissent avec une vitesse de même ordre.

Proposition 3.2.5. *On suppose que $\boldsymbol{\theta}_1$ est absolument continu, $\left\|\boldsymbol{\theta}_1^{\ell}\right\| \leq K$ et pour $\ell = 1, \ldots, k$, sous les hypothèses (H1) et (H2), (H3) ou (H3'),*

$$\nabla g(\boldsymbol{\theta}_n) \xrightarrow[\mathrm{n}\to\infty]{\mathrm{p.s.}} 0$$

et la distance entre $\boldsymbol{\theta}_n$ et l'ensemble des points stationnaires de g converge aussi presque sûrement vers 0.

Dans l'article [4], nous appliquons ce résultat aux données de Médiamétrie. Les techniques de classification « en ligne » sont intéressantes pour déterminer les principaux profils de téléspectateurs de façon automatique, puis pour relier ces profils aux variables

socio-économiques. Dans ces échantillons, Médiamétrie a noté la présence de comportements atypiques ; ainsi les techniques robustes s'avèrent utiles dans ce contexte.

Nous avons établi des profils d'audience du 6 Septembre 2010, à partir d'un échantillon de $n = 5422$ audiences individuelles, regroupées pour toutes les chaînes de télévision et mesurée toutes les minutes sur une période de 24 heures. Un échantillon de 5 profils temporels différents est représenté sur la figure 3.7.

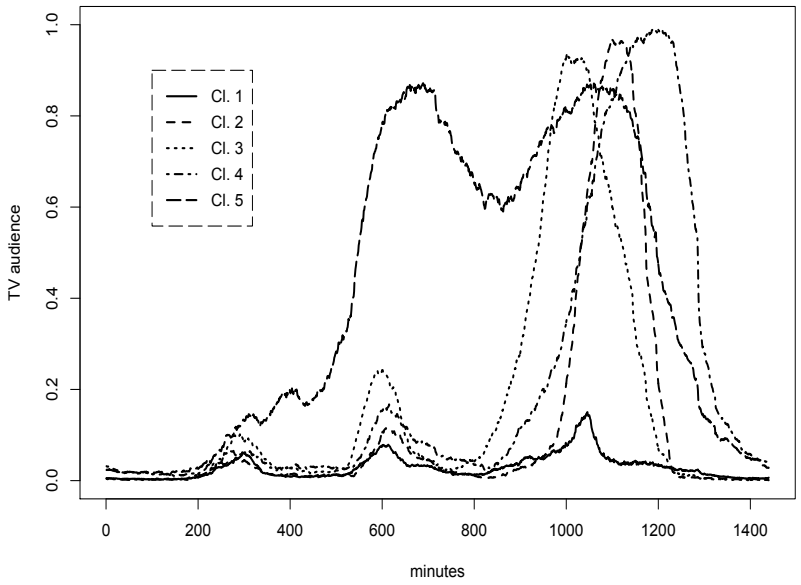


FIGURE 3.7: Le Groupe 1 (Cl.1) est le plus grand groupe et contient environ 35% des individus de l'échantillon. Il correspond à des personnes qui ne regardent pas beaucoup la télévision en journée. A l'opposé, le groupe 5 (représentant environ 12% des individus) est associé à des taux d'audience élevés tout au long de la journée. Les autres groupes correspondent à des niveaux de consommation intermédiaires et peuvent être distingués selon que l'audience télévisuelle se situe plutôt en soirée ou en journée.

3.3 Allocation optimale

Dans la dernière partie de ce chapitre, on utilise un autre algorithme de gradient stochastique, l'algorithme de descente en miroir, appliqué à un problème d'allocation optimale.

3.3.1 Indicateur de risque

Les nouvelles règles de régulation du secteur de l'assurance, appelées en Europe « *Solvency 2* », conduisent les entreprises à ajuster leurs marges de solvabilité aux risques sous-jacents. Le capital global u doit être réparti pour chaque secteur d'activité : u_k est le capital initial de la ligne k de l'entreprise et $u_1 + \ldots + u_d = u$, où d est le nombre de lignes de l'entreprise. Il s'agit donc de trouver, à partir d'un capital u fixé, l'allocation optimale minimisant certains indicateurs de risque.

On considère un processus vectoriel de risque $X_n = (X_n^1, \ldots, X_n^d)^t$, où X_n^k désigne les gains (c'est-à-dire les revenus auxquels on soustrait les pertes) de la ligne k pendant la période n. On note les gains cumulés

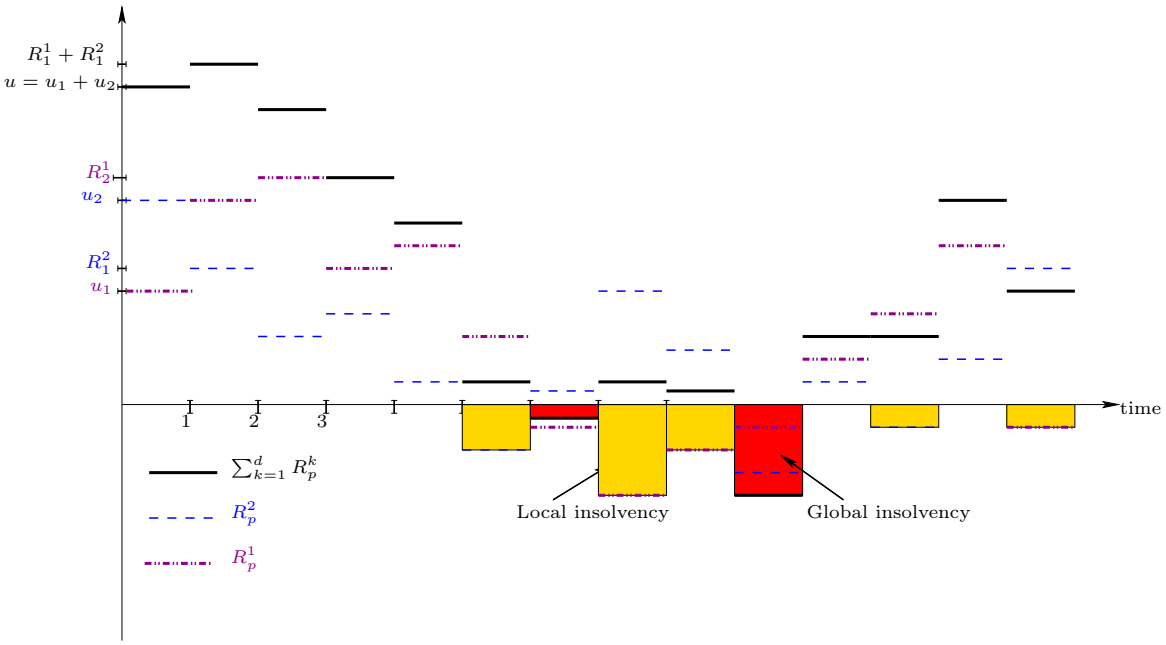
$$Y_n^k = \sum_{p=1}^{n} X_p^k.$$

On dit que la $k^{\text{ème}}$ ligne de l'entreprise fait défaut avant l'instant n s'il existe $i \in \{1, \ldots, n\}$ tel que $Y_i^k + u_k < 0$. On note $R_n^k = u_k + Y_n^k$ le capital de la ligne k de l'entreprise à l'instant n.

On considère un indicateur de risque permettant de tenir compte à la fois de la structure de dépendance entre les secteurs d'activité et de la gravité (ou du coût) de l'insolvabilité. A partir des fonctions de coût, convexes, $g_k : \mathbb{R} \to \mathbb{R}$, $k = 1, \ldots, d$ satisfaisant $g_k(x) \geq 0$ pour $x \leq 0$, et $\mathbb{E}(|g_k(R_p^k)|) < \infty$, on considère l'indicateur de risque

$$I(u_1, \ldots, u_d) = \sum_{k=1}^{d} \mathbb{E}\left(\sum_{p=1}^{n} g_k(R_p^k) \mathbb{1}_{\{R_p^k < 0\}} \mathbb{1}_{\{\sum_{j=1}^{d} R_p^j > 0\}} \right).$$

La fonction g_k représente le coût que la branche d'activité k doit payer quand elle devient insolvable. Minimiser I permet de contrôler l'insolvabilité partielle, c'est-à-dire l'insolvabilité mesurée pour certaines branches alors que l'ensemble de l'entreprise est solvable. Un choix naturel pour le coût pourrait être $g_k(x) = -x$. Le graphique ci-après représente l'indicateur de risque I (zone claire) pour $g_k(x) = -x$ et $d = 2$.



Il s'agit donc de déterminer un minimum sous contraintes : trouver $u^\star \in (\mathbb{R}_+)^d$ tel que

$$I(u^\star) = \inf_{\substack{v \in (\mathbb{R}_+)^d \\ v_1 + \ldots + v_d = u}} I(v).$$

Excepté pour des cas très particuliers, il n'y a pas de solution explicite à ce problème. Nous avons d'abord essayé une approche de type algorithme de descente de gradient, basée sur les multiplicateurs de Lagrange, mais l'algorithme ainsi construit ne convergeait pas dans nos simulations. Nous proposons dans [7] une méthode d'approximation stochastique de ce vecteur $u^\star$ basé sur un algorithme de descente stochastique en miroir. Dans notre problème, nous n'avons pas d'accès à une version bruitée du gradient, c'est la raison pour laquelle nous avons choisi de suivre une approche de type Kiefer-Wolfowitz. Notre preuve de la convergence de l'algorithme est inspirée par la thèse de Tauvel [29].

Alors que les algorithmes de Robbins-Monro et Kiefer-Wolfowitz sont très sensibles au choix du pas, il est remarquable que notre version de descente en miroir d'un algorithme de type Kiefer-Wolfowitz soit très stable et permette des simulations en grande dimension. Enfin, notre algorithme permet de prendre en compte une dépendance temporelle sur une période de longueur p. Néanmoins, pour faire tourner l'algorithme, il faut posséder N copies indépendantes de la distribution des gains sur une période de longueur p.

3.3.2 Descente en miroir

Pour une réalisation y de la matrice aléatoire

$$\mathcal{Y} = \begin{pmatrix} Y_1^1 & \cdots & Y_p^1 \\ \cdots & \cdots & \cdots \\ Y_1^d & \cdots & Y_p^d \end{pmatrix}$$

on pose

$$\mathcal{I}(u_1, \ldots, u_d, y) = \sum_{k=1}^{d} \sum_{n=1}^{p} g_k(y_n^k + u_k) \mathbb{1}_{\{y_n^k + u_k < 0\}} \mathbb{1}_{\{\sum_{k=1}^{d} y_n^k + u_k > 0\}},$$

de sorte que

$$I(u_1, \ldots, u_d) = \sum_{k=1}^{d} \mathbb{E}\left(\sum_{n=1}^{p} g_k(R_n^k) \mathbb{1}_{\{R_n^k < 0\}} \mathbb{1}_{\{\sum_{k=1}^{d} R_n^k > 0\}} \right).$$

On note également, pour une suite (c_n) convenablement choisie :

$$\begin{aligned} \mathcal{I}^k\left(c_n^+, \mathcal{Y}\right) &= \mathcal{I}(\chi_{n-1}^1, \ldots, \chi_{n-1}^{k-1}, \chi_{n-1}^k + c_n, \chi_{n-1}^{k+1}, \ldots, \chi_{n-1}^d, \mathcal{Y}), \\ \mathcal{I}^k\left(c_n^-, \mathcal{Y}\right) &= \mathcal{I}(\chi_{n-1}^1, \ldots, \chi_{n-1}^{k-1}, \chi_{n-1}^k - c_n, \chi_{n-1}^{k+1}, \ldots, \chi_{n-1}^d, \mathcal{Y}). \end{aligned}$$

On considère alors l'approximation du gradient $D_{c_n}\mathcal{I}$, vecteur aléatoire dont la $k^{\text{ème}}$ coordonnée $D_{c_n}^k \mathcal{I}(u_1, \ldots, u_d, \mathcal{Y})$ est définie par

$$\frac{\mathcal{I}^k\left(c_n^+, \mathcal{Y}\right) - \mathcal{I}^k\left(c_n^-, \mathcal{Y}\right)}{2c_n}.$$

On pose $\Psi_{c_n}(\chi_{n-1}, \mathcal{Y}_n) = D_{c_n}\mathcal{I}(\chi_{n-1}, \mathcal{Y}_n)$. L'algorithme est alors construit à partir de N copies $\mathcal{Y}_1, \ldots, \mathcal{Y}_N$, de la matrice aléatoire $\mathcal{Y}$ de la façon suivante.

Algorithme 2

Initialisation :
$$\begin{cases} \xi_0 = 0 \in (\mathbb{R}^m)^\star \\ \chi_0 \in C \end{cases}$$

Mise à jour : pour $n = 1, \ldots, N$

$$\begin{cases} \xi_n = \xi_{n-1} - \gamma_n \Psi_{c_n}(\chi_{n-1}, \mathcal{Y}_n) \\ \chi_n = \nabla W_{\beta_n}(\xi_n) \end{cases}$$

Sortie :
$$S_N = \frac{\sum_{n=1}^{N} \gamma_n \chi_{n-1}}{\sum_{n=1}^{N} \gamma_n}$$

Soit $C = \{(v_1, \ldots, v_d) \in (\mathbb{R}^+)^d | v_1 + \ldots + v_d = u\}$ convexe compact de $\mathbb{R}^d$. On se place dans le cadre des hypothèses suivantes :

(1) I est une fonction convexe de $\mathbb{R}^d$ dans $\mathbb{R}$.
(2) I est de classe C^2.
(3) I admet un unique minimum x^* sur C.
(4) Il existe un réel positif σ tel que pour tout vecteur $(v_1, \ldots, v_d) \in \mathbb{R}^d$,

$$\mathrm{var}\left(\mathcal{I}(v_1, \ldots, v_d, \mathcal{Y}_n)|\mathcal{F}_{n-1}\right) \leq \sigma^2,$$

où $\mathcal{F}_n$ désigne la tribu engendrée par $(\chi_0, \mathcal{Y}_1, \ldots, \mathcal{Y}_n)$.
(5) Il existe $a > 2$ tel que presque sûrement

$$\sup_{n>0} \mathbb{E}\left(|\mathcal{I}(v_1, \ldots, v_d, \mathcal{Y}_n)|^a \,|\mathcal{F}_{n-1}\right) < \infty.$$

Théorème 3.3.1. *Soient (β_n), (γ_n) et (c_n) des suites de $\left(\mathbb{R}_+^*\right)^{\mathbb{N}}$. On suppose que (β_n) est croissante et que les hypothèses suivantes sur les pas sont satisfaites :*

(i) $\lim_{N\to\infty} \dfrac{\beta_N}{\sum_{n=1}^N \gamma_n} = 0,$

(ii) $\lim_{N\to\infty} \dfrac{\sum_{n=1}^N \gamma_n c_n}{\sum_{n=1}^N \gamma_n} = 0,$

(iii) $\lim_{N\to\infty} \dfrac{\sum_{n=1}^N \frac{\gamma_n^2}{c_n^2 \beta_{n-1}}}{\sum_{n=1}^N \gamma_n} = 0,$

(iv) $\sum_{n=1}^{+\infty} \left(\dfrac{\gamma_n}{c_n}\right)^2 < \infty.$

Alors (S_N) converge presque sûrement et dans L^1.

Lorsque l'on veut déterminer un mimimum sur un simplexe, il est classique de choisir comme fonction strictement convexe

$$\delta_u(x) \stackrel{\text{def}}{=} \sum_{k=1}^d \frac{x_k}{u} \ln\left(\frac{x_k}{u}\right),$$

ce qui conduit à la fonction auxiliaire

$$V(x) = \ln d + \sum_{k=1}^d \frac{x_k}{u} \ln\left(\frac{x_k}{u}\right).$$

Dans ce cas, on calcule aisément

$$\nabla W_\beta(\xi) = \beta \ln\left(\frac{1}{d}\sum_{k=1}^d \exp\left(\xi_k \frac{u}{\beta}\right)\right),$$

et l'on peut mettre en place l'algorithme de descente en miroir, qui s'avère peu sensible aux paramètres et très rapide à mettre en œuvre.

Références

[1] Marc Arnaudon, Clément Dombry, Anthony Phan et Le Yang. « Stochastic algorithms for computing means of probability measures ». Dans : *Stochastic Process. Appl.* 122.4 (2012), p. 1437–1455.

[2] L. Bottou. « Large-scale machine learning with stochastic gradient descent. » Dans : sous la dir. de Springer. Lechevallier, Y., Saporta, G. (Eds.), Compstat 2010. Physica Verlag, 2010, p. 177–186.

[3] H. Cardot, P. C. et M. Chaouch. « Stochastic approximation to the multivariate and the functional median ». Dans : *COMPSTAT 2010*. Paris, France., 2010.

[4] H. Cardot, P. C. et J-M. Monnez. « A fast and recursive algorithm for clustering large datasets with k-medians ». Dans : *Computational Statistics & Data Analysis* 56.6 (2012), p. 1434–1449.

[5] H. Cardot, P. C. et P-A. Zitt. « Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. » Dans : *Bernoulli* 19.1 (2013), p. 18–43.

[6] H. Cardot, P C. et P.-A. Zitt. « Recursive estimation of the conditional geometric median in Hilbert spaces ». Dans : *Electron. J. Statist.* 6 (2012), p. 2535–2562. issn : 1935-7524. doi : 10.1214/12-EJS759.

[7] P. C., V. Maume-Deschamps et C. Prieur. « Some multivariate risk indicators; Minimization by using a Kiefer-Wolfowitz approach to the mirror stochastic algorithm ». Dans : *Statistics and Risk Modeling* 29 (2012), p. 47–71.

[8] M. Duflo. *Random Iterative Methods*. Springer-Verlag, 1997.

[9] E. Forgy. « Cluster analysis of multivariate data: efficiency vs. interpretability of classifications. » Dans : *Biometrics* 21 (1965), p. 768–769.

[10] Guojun Gan, Chaoqun Ma et Jianhong Wu. *Data clustering*. T. 20. ASA-SIAM Series on Statistics and Applied Probability. Theory, algorithms, and applications. Philadelphia, PA : Society for Industrial et Applied Mathematics (SIAM), 2007, p. xxii+466. isbn : 978-0-898716-23-8.

[11] Daniel Gervini. « Robust functional estimation using the median and spherical principal components ». Dans : *Biometrika* 95.3 (2008), p. 587–600.

[12] J. C. Gower. « Algorithm as 78: The mediancentre. » Dans : *ournal of the Royal Statistical Society. Series C (Applied Statistics)* 23.3 (1974), p. 466–470.

[13] A. Jain, M. Marty et P. Flynn. « Data clustering: a review. » Dans : *ACM Computing surveys* 31 (1999), p. 264–323.

[14] A. Juditsky, P. Rigollet et A. B. Tsybakov. « Learning by mirror averaging ». Dans : *Ann. Statist.* 36.5 (2008), p. 2183–2206.

[15] Leonard Kaufman et Peter J. Rousseeuw. *Finding groups in data*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. An introduction to cluster analysis, A Wiley-Interscience Publication. New York : John Wiley & Sons Inc., 1990, p. xvi+342. isbn : 0-471-87876-6.

[16] J. H. B. Kemperman. « The median of a finite measure on a Banach space ». Dans : *Statistical data analysis based on the L_1-norm and related methods (Neuchâtel, 1987)*. Amsterdam : North-Holland, 1987, p. 217–230.

[17] J. Kiefer et J. Wolfowitz. « Stochastic estimation of the maximum of a regression function ». Dans : *Ann. Math. Statistics* 23 (1952), p. 462–466.

[18] T. Laloë. « L_1-quantization and clustering in Banach spaces ». Dans : *Math. Methods Statist.* 19.2 (2010), p. 136–150. issn : 1066-5307.

[19] Guanghui Lan. « An optimal method for stochastic composite optimization ». Dans : *Math. Program.* 133.1-2, Ser. A (2012), p. 365–397.

[20] J. MacQueen. « Some methods for classification and analysis of multivariate observations ». Dans : *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)*. Berkeley, Calif. : Univ. California Press, 1967, Vol. I: Statistics, pp. 281–297.

[21] Abdelkader Mokkadem, Mariane Pelletier et Yousri Slaoui. « Revisiting Révész's stochastic approximation method for the estimation of a regression function ». Dans : *ALEA Lat. Am. J. Probab. Math. Stat.* 6 (2009), p. 63–114.

[22] Jyrki Möttönen, Klaus Nordhausen et Hannu Oja. « Asymptotic theory of the spatial median ». Dans : *Nonparametrics and robustness in modern statistical inference and time series analysis: a Festschrift in honor of Professor Jana Jurečková*. T. 7. Inst. Math. Stat. Collect. Beachwood, OH : Inst. Math. Statist., 2010, p. 182–193.

[23] A. Nemirovsky, A. Juditsky, G. Lan et A. Shapiro. « Robust stochastic approximation to stochastic programming ». Dans : *SIAM, Journal on Optimization* 19 (2009), p. 1574–1609.

[24] A. S. Nemirovsky et D. B. Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics. New York : John Wiley & Sons Inc., 1983.

[25] Y. Nesterov. « Primal-dual subgradient methods for convex problems ». Dans : *Mathematical Programming* 120 (2006), p. 221–259.

[26] Gilles Pagès. « A space quantization method for numerical integration ». Dans : *J. Comput. Appl. Math.* 89.1 (1998), p. 1–38. issn : 0377-0427.

[27] B.T. Polyak et A.B. Juditsky. « Acceleration of Stochastic Approximation. » Dans : *SIAM J. Control and Optimization* 30 (1992), p. 838–855.

[28] H. Robbins et S. Monro. « A stochastic approximation method ». Dans : *Ann. Math. Statistics* 22 (1951), p. 400–407.

[29] Claire Tauvel. « Optimisation stochastique à grande échelle ». Thèse de doct. Université Joseph Fourier, déc. 2008.

[30] Yehuda Vardi et Cun-Hui Zhang. « The multivariate L_1-median and associated data depth ». Dans : *Proc. Natl. Acad. Sci. USA* 97.4 (2000), 1423–1426 (electronic). issn : 1091-6490. doi : 10.1073/pnas.97.4.1423. url : http://dx.doi.org/10.1073/pnas.97.4.1423.

[31] M. P. Wand et M. C. Jones. *Kernel smoothing*. T. 60. Monographs on Statistics and Applied Probability. London : Chapman et Hall Ltd., 1995, p. xii+212.

[32] A. B. Yuditski, A. V. Nazin, A. B. Tsybakov et N. Vayatis. « Recursive aggregation of estimators by the mirror descent method with averaging ». Dans : *Problemy Peredachi Informatsii* 41.4 (2005), p. 78–96.

Chapitre 4

Théorème de la limite centrale presque-sûr

Ce chapitre est le résumé des travaux [1, 7, 8, 17] listés au chapitre 1.

Sommaire

4.1 Introduction

Soit (ξ_n) une suite de variables indépendantes et de même loi, centrées et de variance $\mathbb{E}[\xi_n^2] = \sigma^2$. Définissons la somme $Z_n \overset{\text{def}}{=} \xi_1 + \ldots + \xi_n$. D'après le célèbre théorème de la limite centrale, pour toute fonction h continue bornée,

$$\lim_{n\to\infty} \mathbb{E}\Big[h\Big(\frac{Z_n}{\sqrt{n}}\Big)\Big] = \int_{\mathbb{R}} h(x)dG(x),$$

où G est la mesure gaussienne $\mathcal{N}(0, \sigma^2)$. De plus, le théorème de la loi du logarithme itéré nous indique que

$$\limsup_{n\to\infty} \frac{Z_n}{\sqrt{n}} . \frac{1}{\sqrt{2\log\log n}} = 1 \quad \text{p.s.}$$

Le théorème de la limite centrale presque-sûr (TLCPS) fournit presque partout une convergence faible de la moyenne logarithmique de $(Z_n/\sqrt{n})$ avec des poids harmoniques, autrement dit la mesure de comptage dans la loi des grands nombres est remplacée par une mesure logarithmique $\mu(A) = \sum_{k\in A} \frac{1}{k}$, pour $A \subset \mathbb{N}$. Plus précisément on a : pour

toute fonction h continue bornée,

$$\lim_{n\to\infty} \frac{1}{\log n} \sum_{k=1}^{n} \frac{1}{k} h\Big(\frac{Z_k}{\sqrt{k}}\Big) = \int_{\mathbb{R}} h(x) dG(x) \quad \text{p.s.}$$

La première version du TLCPS a été énoncée sans preuve dans le livre de Lévy [31]. Ce théorème a été démontré par Brosamler [10], Schatte [39, 40], et dans sa forme présente par Lacey [28]. La version « universelle » de TLCPS présentée dans Berkes et Csáki [9] couvre une large classe de théorèmes limites pour les sommes partielles, les extrêmes, les fonctions de répartition empiriques, les temps locaux et pour les U-statistiques, construits à partir de variables indépendantes et identiquement distribuées. On trouve également des TLCPS pour les U-statistiques dans les travaux de Holzmann, Koch et Min [23] et de Min [34].

Le théorème de la limite centrale presque-sûr a aussi été établi dans un cadre martingales par Chaâbane [12, 11], Chaâbane et Maâouia [13] et Lifshits [33, 32] ou dans un cadre de variables mélangeantes par Gonchigdanzan [21]. On trouvera dans Berkes [8] et Atlagh et Weber [1] une étude détaillée des papiers sur le sujet.

Plus récemment, Thangavelu [42] a présenté des applications du TLCPS au contrôle qualité pour des estimations de quantiles, des tests d'adéquation et de comparaison, des statistiques de rang. L'avantage de ces méthodes basées sur le TLCPS est qu'elles permettent d'éviter l'estimation de la variance des observations.

La section 4.2 résume les résultats obtenus pendant ma thèse dans les articles [15, 5] qui établissent la convergence des moments dans le TLCPS pour des martingales vectorielles. Les deux autres sections sont consacrées à l'énoncé du TLCPS pour les algorithmes stochastiques et pour l'estimateur des moindres carrés du processus d'Ornstein Uhlenbeck.

4.2 TLCPS pour les martingales

On suppose que (M_n) est une martingale à valeurs dans $\mathbb{R}^d$, adaptée à une filtration $\mathbb{F}$. On note $(\langle M \rangle_n)$ son processus croissant. Une approche du TLCPS pour les martingales vectorielles discrètes a été développée dans Chaâbane, Maâouia et Touati [14]. L'une de leurs hypothèses porte sur le comportement asymptotique du processus croissant. Ils supposent qu'il existe une suite **déterministe** (U_n) de matrices réelles inversibles de la forme $U_n = \alpha_n I_d$ où (α_n) est une suite croissante vers l'infini avec $\lim_{n\to\infty} \frac{\alpha_n}{\alpha_{n-1}} = 1$, telle que

$$U_n^{-1} \langle M \rangle_n U_n^{-1} \xrightarrow[\mathrm{n}\to\infty]{\mathrm{p.s.}} C.$$

La matrice C peut être aléatoire ou déterministe.

On montre dans [5] que sous des hypothèses appropriées proches de celles de Chaâbane, Maâouia et Touati [14], mais en remplaçant les poids déterministes (U_n) par

la racine du processus croissant, il y a convergence des moments dans le TLCPS pour les martingales vectorielles. Plus précisément, l'article [5] est une généralisation du théorème de convergence des moments de Bercu [4] obtenue dans le cas scalaire, au cadre vectoriel.

Avant d'énoncer le théorème de convergence à proprement parler ainsi que ses applications, définissons plus précisément le cadre et les hypothèses. Soit (ε_n) une suite de différences de martingales adaptée à une filtration $\mathbb{F} \overset{\text{def}}{=} (\mathcal{F}_n)$. On suppose que (M_n) peut se décomposer sous la forme d'une transformée de martingales

$$M_n \overset{\text{def}}{=} M_0 + \sum_{k=1}^{n} \Phi_{k-1}\varepsilon_k,$$

avec M_0 arbitrairement choisie et où (Φ_n) est une suite de vecteurs aléatoires de $\mathbb{R}^d$, adaptée à la filtration $\mathbb{F}$. On note également

$$S_n \overset{\text{def}}{=} \sum_{k=0}^{n} \Phi_k \Phi_k^t + S,$$

où S est une matrice définie positive, symétrique et déterministe de sorte que S_n soit inversible pour tout $n \in \mathbb{N}$. On peut remarquer que, si le moment conditionnel d'ordre deux de la différence de martingales est constant, *i.e.* $\mathbb{E}[\varepsilon_{n+1}^2 \mid \mathcal{F}_n] = \sigma^2$ p.s., le processus croissant de (M_n) est la matrice

$$\langle M \rangle_n = \sigma^2 S_n.$$

Le coefficient d'explosion associé à (Φ_n) est défini par

$$f_n \overset{\text{def}}{=} \Phi_n^t S_n^{-1} \Phi_n = \frac{d_n - d_{n-1}}{d_n}, \quad \text{avec } d_n \overset{\text{def}}{=} \det(S_n).$$

Pour alléger les notations, on définit les hypothèses sur le bruit (ε_n) qui sont utilisées dans les résultats suivants. Pour $p \geq 1$, on note respectivement (H_{2p+}) et (C_{2p}) les assertions suivantes : la suite (ε_n) est une différence de martingales telle que

$$(H_{2p+}) \qquad \sup_{n \geq 0} \mathbb{E}\big[|\varepsilon_{n+1}|^a \big| \mathcal{F}_n\big] < \infty \quad \text{p.s.} \quad \text{pour un réel } a > 2p,$$

$$(C_{2p}) \qquad \forall n \geq 0, \quad \mathbb{E}\big[\varepsilon_{n+1}^{2p} \big| \mathcal{F}_n\big] \overset{\text{def}}{=} \sigma(2p) < \infty \quad \text{p.s.}$$

On note $\sigma(2) = \sigma^2$.

Théorème 4.2.1. *On suppose que (ε_n) est une différence de martingales satisfaisant (C_2) et (H_{2p+}) pour un entier $p \geq 1$. De plus, on suppose que le coefficient d'explosion f_n tend vers zéro p.s. et qu'il existe une suite aléatoire positive (α_n) croissante vers l'infini et une matrice inversible L telles que*

$$\lim_{n \to \infty} \alpha_n^{-1} S_n = L \quad \textit{p.s.} \tag{4.5}$$

Alors on a presque sûrement

$$\lim_{n\to\infty} \frac{1}{\log d_n} \sum_{k=1}^{n} f_k \big(M_k^t S_{k-1}^{-1} M_k\big)^p = \ell(p) \stackrel{\text{def}}{=} d\sigma^{2p} \prod_{j=1}^{p-1} \big(d+2j\big).$$

$$\lim_{n\to\infty} \frac{1}{\log d_n} \sum_{k=1}^{n} \big[\big(M_k^t S_{k-1}^{-1} M_k\big)^p - \big(M_k^t S_k^{-1} M_k\big)^p\big] = \lambda(p) \stackrel{\text{def}}{=} \frac{p}{d}\ell(p).$$

Remarque 4.2.2. La limite $\ell(p)$ est le moment d'ordre $2p$ de la norme d'un vecteur gaussien $\mathcal{N}(0, \sigma^2 I_d)$.

Plus généralement, sans faire l'hypothèse de moment d'ordre 2 constant, on peut aussi déterminer la vitesse de convergence de $\sum f_k ||S_{k-1}^{-1/2} M_k||^{2p}$.

Théorème 4.2.3. *On suppose que (ε_n) est une différence de martingales satisfaisant (H_{2p+}) pour un entier $p \geq 1$. De plus, on suppose que le coefficient d'explosion f_n tend vers zéro p.s. et qu'il existe une suite aléatoire (α_n) positive croissante vers l'infini et une matrice inversible L vérifiant* (4.5)*. Alors on a presque sûrement*

$$\lim_{n\to\infty} \frac{1}{\log d_n} \sum_{k=1}^{n} f_k \big(M_k^t S_{k-1}^{-1} M_k\big)^p = \mathcal{O}(1).$$

Dans les modèles étudiés ci-après, l'hypothèse de convergence (4.5) est vérifiée. Cette hypothèse revient à supposer que les valeurs propres de S_n tendent toutes vers l'infini à la même vitesse.

4.2.1 Applications

Ces propriétés asymptotiques pour les puissances de martingales vectorielles permettent d'établir des résultats de convergence sur les erreurs d'estimation et de prédiction associées aux modèles de régression linéaire. Ils sont définis, pour tout $n \geq 1$, par la relation

$$X_{n+1} = \theta^t \Phi_n + \varepsilon_{n+1}, \tag{4.8}$$

où $\theta \in \mathbb{R}^d$ est le paramètre inconnu du modèle. Les variables X_n, Φ_n, et ε_n sont respectivement l'observation scalaire, le vecteur de régression et le bruit scalaire du système. En particulier, on illustre les résultats sur les modèles autorégressifs linéaires et les processus de branchement avec immigration.

On considère l'estimateur des moindres carrés

$$\widehat{\theta}_n = S_{n-1}^{-1} \sum_{k=1}^{n} \Phi_{k-1} X_k.$$

En posant $M_0 = -S\theta$, on déduit de (4.8) et de (4.9) que

$$\widehat{\theta}_n - \theta = S_{n-1}^{-1} M_n.$$

On comprend alors comment le comportement asymptotique de (M_n) donne des informations sur la qualité de l'estimateur des moindres carrés.

Concentrons-nous sur l'erreur de prédiction $X_{n+1} - \widehat{\theta}_n^t \Phi_n$ et sur l'erreur d'estimation $\widehat{\theta}_n - \theta$. Il est plus approprié (voir par exemple GOODWIN et SIN [22]) de considérer les erreurs cumulées de prédiction et d'estimation, respectivement définies, pour tout $p \geq 1$, par

$$C_n(p) \stackrel{\text{def}}{=} \sum_{k=0}^{n-1} (X_{k+1} - \widehat{\theta}_k^t \Phi_k)^{2p} \quad \text{et} \quad G_n(p) \stackrel{\text{def}}{=} \sum_{k=1}^{n} k^{p-1} \|\widehat{\theta}_k - \theta\|^{2p}.$$

Estimation des moments, erreurs d'estimation et de prédiction

Pour $p \geq 1$,

$$\Gamma_n(2p) = \frac{1}{n} \sum_{k=0}^{n-1} (X_{k+1} - \widehat{\theta}_k^t \Phi_k)^{2p}$$

est un estimateur naturel et consistant du moment d'ordre $2p$ du bruit (ε_n) noté $\sigma(2p)$. On peut remarquer que $n\Gamma_n(2p) = C_n(p)$.

Corollaire 4.2.4. *On suppose qu'il existe $p \geq 2$ vérifiant (C_{2p}). De plus, on suppose qu'il existe une suite aléatoire (α_n) positive, croissante vers l'infini, ainsi qu'une matrice inversible L telles que l'hypothèse de convergence* (4.5) *soit vérifiée. On suppose également que (f_n) tend vers zéro p.s. Alors pour tout entier q vérifiant $1 \leq q \leq p$ et (C_{2q}), $\Gamma_n(2q)$ est un estimateur consistant de $\sigma(2q)$ et*

$$\Big(\Gamma_n(2q) - \frac{1}{n} \sum_{k=1}^{n} \varepsilon_k^{2q}\Big)^2 = \mathcal{O}\Big(\frac{\log d_n}{n}\Big) \quad p.s.$$

Sous les hypothèses du corollaire 4.2.4, la convergence (4.12) implique que $C_n(q)/n$ converge p.s. vers $\sigma(2q)$. De plus, si (ε_n) a un moment conditionnel fini d'ordre $a > 2q$, pour c vérifiant $2qa^{-1} < c < 1$, dès que $\log d_n = o\big(n^c\big)$, on a

$$\left|\frac{1}{n}\, C_n(q) - \sigma(2q)\right|^2 = o(n^{c-1}) \quad \text{p.s.}$$

Corollaire 4.2.5. *Sous les hypothèses du théorème 4.2.1, on a*

$$\lim_{n\to\infty} \frac{1}{\log d_n} \sum_{k=1}^{n} f_k \left((\widehat{\theta}_k - \theta)^t S_k (\widehat{\theta}_k - \theta)\right)^p = \ell(p) \quad p.s.$$

De plus, on suppose qu'il existe une matrice inversible L telle que

$$\lim_{n\to+\infty} \frac{1}{n}\, S_n = L \qquad p.s.$$

Alors on a aussi

$$\lim_{n\to\infty} \frac{1}{\log n} \sum_{k=1}^{n} k^{p-1} \left((\widehat{\theta}_k - \theta)^t L (\widehat{\theta}_k - \theta) \right)^p = \ell(p) \quad p.s.$$

Ainsi, puisque L est strictement définie positive, on déduit de (4.15) que

$$G_n(p) = \mathcal{O}\big(\log n\big) \quad \text{p.s.}$$

Modèles autorégressifs linéaires

Le modèle autorégressif linéaire est un cas particulier du modèle de régression (4.8). Il est défini pour tout $n \geq 1$, par

$$X_{n+1} = \sum_{k=1}^{d} \theta_k X_{n-k+1} + \varepsilon_{n+1}.$$

La matrice compagne C associée à ce modèle est donnée par

$$C = \begin{pmatrix} \theta_1 & \theta_2 & \dots & \theta_{d-1} & \theta_d \\ 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{pmatrix}.$$

Les résultats précédents s'appliquent dans le *cas stable*, c'est-à-dire quand $\rho(C) < 1$, où $\rho(C)$ désigne le rayon spectral de la matrice compagne C.

Processus de branchement avec immigration

On considère le processus de branchement à temps discret (X_n) sujet à une composante d'immigration indépendante à chaque génération : la population de référence peut donc s'enrichir d'apports extérieurs. On peut ainsi modéliser l'évolution d'un patrimoine génétique, de phénomènes en écologie, en physique des particules ou en épidémiologie. Le processus de branchement (X_n) est donné par la relation de récurrence

$$X_{n+1} = \sum_{k=1}^{X_n} Y_{n+1,k} + I_{n+1}, \tag{4.16}$$

avec $X_0 = 1$. La variable aléatoire (I_n) correspond à l'effectif de l'immigration à la génération n. Pour chaque individu k de la génération n, $Y_{n+1,k}$ désigne son nombre de descendants. On suppose que les familles de variables aléatoires (I_n) et $(Y_{n,k})$, indépendantes et identiquement distribuées, sont indépendantes entre elles. On pose alors

$$\mathbb{E}[Y_{n,k}] \stackrel{\text{def}}{=} m, \quad \mathbb{E}[I_n] \stackrel{\text{def}}{=} \lambda.$$

La relation de récurrence (4.16) peut s'écrire sous la forme

$$\widetilde{X}_{n+1} = \theta^t \widetilde{\Phi}_n + \widetilde{\varepsilon}_{n+1},$$

où les variables sont définies par

$$\widetilde{X}_{n+1} \stackrel{\text{def}}{=} c_n^{-1/2} X_{n+1}, \quad c_n \stackrel{\text{def}}{=} X_n + 1, \quad \theta^t \stackrel{\text{def}}{=} (m, \lambda), \quad \widetilde{\Phi}_n^t \stackrel{\text{def}}{=} c_n^{-1/2}(X_n, 1),$$

et $\widetilde{\varepsilon}_{n+1} \stackrel{\text{def}}{=} c_{n+1}^{-1/2} \left(X_{n+1} - mX_n - \lambda \right)$. Dans le cas stable $m < 1$, la convergence (4.5) a été établie par Wei et Winnicki [43]. On obtient ainsi le comportement asymptotique des erreurs d'estimation et de prédiction cumulées de l'estimateur des moindres carrés de la moyenne. Avec un raisonnement analogue et en décomposant le processus de bruit en un modèle linéaire, on obtient également les comportements asymptotiques de l'estimateur de la variance.

4.2.2 Sur l'estimateur des moindres carrés pondérés

Cette section résume le travail présenté dans [15]. Pour contourner certaines difficultés inhérentes au cas vectoriel et éviter une hypothèse de type (4.5), il est possible d'introduire des martingales pondérées. Dans cette section, (M_n) désigne la transformée de martingale pondérée

$$M_n = M_0 + \sum_{k=1}^{n} a_{k-1} \Phi_{k-1} \varepsilon_k,$$

où a_n est une suite décroissante adaptée à la filtration $\mathbb{F}$, avec $0 \leq a_n \leq 1$. Le coefficient d'explosion $f_n(a)$ est défini par

$$f_n(a) \stackrel{\text{def}}{=} a_n \Phi_n^t S_n^{-1}(a) \Phi_n, \quad \text{avec} \quad S_n(a) = \sum_{k=0}^{n} a_k \Phi_k \Phi_k^t + S.$$

On suppose que la suite (a_n) vérifie la convergence

$$\sum_{k=0}^{\infty} a_k f_k(a) < \infty \quad \text{p.s.},$$

L'équivalent du théorème 4.2.1 dans ce cadre de martingales pondérées est donné par le théorème suivant.

Théorème 4.2.6. *Si* $\sup_{n\geq 0} \mathbb{E}[\varepsilon_{n+1}^2 \mid \mathcal{F}_n] < \infty$ *p.s., alors, pour tout entier* $p \geq 1$, *on a*

$$\sum_{n=1}^{\infty} \Big(M_n^t S_{n-1}^{-1}(a) M_n - M_n^t S_n^{-1}(a) M_n \Big)^p < \infty \quad p.s.$$

Notons que grâce à la pondération, il suffit de supposer que le moment conditionnel est d'ordre 2.

Pour appliquer ce résultat aux modèles de régression linéaire présentés dans la Section 4.2.1, il est naturel de considérer l'estimateur des moindres carrés pondérés

$$\widehat{\theta}_n = S_{n-1}^{-1}(a) \sum_{k=1}^{n} a_{k-1} \Phi_{k-1} X_k.$$

Le théorème 4.2.6 s'applique aux modèles de régression, en choisissant convenablement la suite de pondération (a_n) et en considérant le cas stable :

$$\limsup_{n\to+\infty} f_n(a) < 1 \quad \text{p.s.}$$

Dans ce cas, on peut montrer, en supposant que les hypothèses (C_{2p}) et (H_{2p+}) sont vérifiées, qu'il existe un réel $0 < c < 1$ tel que

$$\left| \frac{1}{n}\, C_n(p) - \sigma(2p) \right| = o(n^{c-1}) \quad \text{p.s.}$$

De plus, s'il existe une matrice inversible L telle que la convergence (4.14) soit satisfaite, alors

$$G_n(p) = o\big((\log s_n)^{(p+1)(1+\gamma)} \big) \quad \text{p.s.}$$

4.3 TLCPS pour le processus d'Ornstein-Uhlenbeck

Dans un contexte de diffusion brownienne, Lamberton et Pagès [29, 30] établissent la convergence des mesures pondérées de type TLCPS pour obtenir une approximation de la mesure invariante de la diffusion. Le TLCPS est un corollaire de leur résultat.

Le TLCPS pour l'estimateur des paramètres d'un processus d'Ornstein-Uhlenbeck est établi dans Fathallah et Kebaier [19]. Dans l'article [17] nous démontrons le théorème de la limite centrale presque-sûr pour une suite d'estimateur du paramètre d'un processus d'Ornstein-Uhlenbeck fractionnaire, aussi bien pour une observation du processus à temps continu que pour une observation discrétisée. La preuve s'appuie sur un critère introduit par Ibragimov et Lifshits [26], Ibragimov et Lifshits [25] et basé sur la vitesse de convergence des fonctions caractéristiques. A partir de ce critère combiné avec du calcul de Malliavin, Bercu, Nourdin et Taqqu [6] obtiennent un critère de TLCPS pour des champs Gaussiens généraux.

Considérons le processus d'Ornstein-Uhlenbeck fractionnaire $X = \{X_t, t \geq 0\}$ défini par $X_0 = 0$ et

$$dX_t = -\theta X_t dt + dB_t, \quad t \geq 0, \tag{4.18}$$

où $B = \{B_t, t \geq 0\}$ est un mouvement Brownien fractionnaire de paramètre de Hurst $H \in (\frac{1}{2}, 1)$ et θ est un paramètre réel inconnu.

Pour estimer ce paramètre θ à partir de l'observation d'un *processus d'Ornstein-Uhlenbeck fractionnaire continu*, récemment Hu et Nualart [24] et Belfadli, Es-Sebaiy et Ouknine [2] ont étudié les propriétés de l'estimateur des moindres carrés $\widehat{\theta}_t$ de θ donné par

$$\widehat{\theta}_t \overset{\text{def}}{=} \frac{\int_0^t X_s dX_s}{\int_0^t X_s^2 ds}, \quad t \geq 0.$$

Hu et Nualart [24] prouvent la consistance forte et la normalité asymptotique de $\widehat{\theta}_t$ dans le cas ergodique, c'est-à-dire lorsque $\theta > 0$. Dans le cas non-ergodique $\theta < 0$, Belfadli, Es-Sebaiy et Ouknine [2] établissent la convergence presque sûre de $\widehat{\theta}_t$ vers θ ainsi que le comportement asymptotique de type loi de Cauchy.

Dans le cas discret, le processus X est observé en n points, à des instants réguliers de pas Δ_n, *i.e.* pour tout entier $i \in \{0, \ldots, n\}$, $t_i = i\Delta_n$. On considère alors l'estimateur des moindres carrés

$$\widetilde{\theta}_n \overset{\text{def}}{=} -\frac{\sum_{i=1}^n X_{t_{i-1}}(X_{t_i} - X_{t_{i-1}})}{\Delta_n \sum_{i=1}^n X_{t_{i-1}}^2}.$$

Lorsque $\theta > 0$, Es-Sebaiy [41] montre la convergence en probabilités de $(\widetilde{\theta}_n)$ et obtient également sa vitesse.

Dans la suite de la section, G désigne une variable aléatoire gaussienne centrée réduite $\mathcal{N}(0, 1)$. Les TLCPS établis pour $(\widehat{\theta}_t)$ et $(\widetilde{\theta}_n)$ dans [17] reposent sur les deux théorèmes suivants.

Théorème 4.3.1. *Soit (Z_n) une suite de variables aléatoires réelles satisfaisant un théorème de la limite centrale presque-sûr. On suppose que (R_n) est une suite positive de variables convergeant presque sûrement vers 1. Alors la suite (Z_n/R_n) vérifie le théorème de la limite centrale presque-sûr. En d'autres termes, presque sûrement pour tout $z \in \mathbb{R}$, on a*

$$\lim_{n\to\infty} \frac{1}{\log n} \sum_{k=1}^n \frac{1}{k} \mathbb{1}_{\{Z_k \leq z R_k\}} = \mathbb{P}\left(G \leq z\right).$$

Théorème 4.3.2. *Soit (Z_n) une suite de variables aléatoires réelles satisfaisant le théorème de la limite centrale presque-sûr. On suppose que (R_n) est une suite positive de*

variables convergeant presque sûrement vers 0. *Alors la suite* $(Z_n + R_n)$ *satisfait le théorème de la limite presque-sûr, et on a pour tout* $z \in \mathbb{R}$,

$$\frac{1}{\log n}\sum_{k=1}^{n}\frac{1}{k}\mathbb{1}_{\{Z_k+R_k\leq z\}} \xrightarrow[\mathrm{n}\to\infty]{\text{p.s.}} \mathbb{P}\left(G \leq z\right).$$

Remarque 4.3.3. Un résultat similaire au théorème 4.3.2 établissant un TLCPS de $(G_n + R_n)$ où (R_n) converge dans L^2 vers 0 et vérifie

$$\sum_{n\geq 2}\frac{1}{n\log^2 n}\sum_{k=1}^{n}\frac{1}{k}\mathbb{E}|R_k|^2 < \infty,$$

est démontré par Nourdin et Peccati [36].

4.3.1 Observation du processus à temps continu

On considère le processus d'Ornstein-Uhlenbeck $X = \{X_t, t \geq 0\}$ défini par l'équation différentielle stochastique linéaire (4.18). Dans le cas ergodique $\theta > 0$, lorsque l'exposant de Hurst $H \in (1/2, 3/4)$, la suite $(\sqrt{n}(\theta - \widehat{\theta}_n))$ satisfait le TLCPS :

Théorème 4.3.4. *On se place dans le cas ergodique* $\theta > 0$ *et on suppose que* $H \in (1/2, 3/4)$. *Alors presque sûrement pour toute fonction continue bornée* φ

$$\frac{1}{\log n}\sum_{k=1}^{n}\frac{1}{k}\varphi\left(\frac{\sqrt{k}}{\sigma_k}(\theta - \widehat{\theta}_k)\right) \underset{\mathrm{n}\to\infty}{\longrightarrow} \mathbb{E}(\varphi(G)),$$

où (σ_t) *désigne la normalisation positive*

$$\sigma_t \overset{\text{def}}{=} \theta^{-2H}H\Gamma(2H)\sqrt{\mathbb{E}\left(\frac{1}{t}\int_0^t dB_s e^{\theta s}\int_0^t dB_r e^{-\theta r}\right)}. \tag{4.19}$$

4.3.2 Observation à temps discret

On se place dans le cas ergodique $\theta > 0$ et on suppose que le processus est observé en n instants régulièrement espacés d'un pas $\Delta_n = n^{-\alpha}$, pour $\alpha \in (\frac{1}{2H+1}, 1)$. On note $T_n = n\Delta_n$ la longueur de la fenêtre d'observation.

Théorème 4.3.5. *On suppose que* $H \in (1/2, 1)$ *ainsi* $\lim_{n\to\infty}\Delta_n = 0$ *et* $\lim_{n\to\infty} n\Delta_n = \infty$. *On a*

$$\lim_{n\to\infty}\widetilde{\theta}_n = \theta \quad p.s.$$

Théorème 4.3.6. *On suppose que* $H \in (1/2, 3/4)$. *Alors presque sûrement pour toute fonction continue bornée* φ, *on a*

$$\lim_{n\to\infty}\frac{1}{\log n}\sum_{k=1}^{n}\frac{1}{k}\varphi\left(\frac{\sqrt{T_k}}{\sigma_{T_k}}(\theta - \widetilde{\theta}_k)\right) = \mathbb{E}(\varphi(N)),$$

avec (σ_t) *définie par* (4.19).

4.4 TLCPS pour les algorithmes stochastiques

Dans le cas scalaire, la convergence des moments dans le TLCPS a été étudiée par Bercu [4], Bercu et Fort [7] dans un cadre martingales. En reprenant les notations de l'introduction du chapitre, lorsque les variables (ξ_n) sont indépendantes et identiquement distribuées, la convergence des moments s'écrit :

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \frac{1}{k}\Big(\frac{Z_k}{\sqrt{k}}\Big)^{2p} = \frac{\sigma^{2p}(2p)!}{2^p p!} \quad \text{p.s.},$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \frac{1}{k}\Big(\frac{Z_k}{\sqrt{k}}\Big)^{2p-1} = 0 \quad \text{p.s.}$$

A partir de la convergence des moments de tout ordre, Bercu et Fort [7] démontrent en utilisant *le Théorème de Carleman* que les transformées de martingales réelles vérifient le TLCPS. L'article [16] reprend cette idée en montrant que les algorithmes stochastiques d'approximation vérifient aussi la convergence de moments. Le TLCPS est alors une conséquence de la convergence des moments, ce qui fournit une autre preuve au TLCPS établi par Pelletier [37].

On considère l'algorithme stochastique de la forme

$$Z_{n+1} = Z_n + \gamma_n\big[h(Z_n) + R_{n+1}\big] + \sigma_n\varepsilon_{n+1}, \tag{4.22}$$

où la fonction h est définie sur $\mathbb{R}$ et à valeur dans $\mathbb{R}$. Les deux suites aléatoires (R_n) et (ε_n) sont deux perturbations adaptées à la filtration $\mathbb{F}$. Les pas (γ_n) et (σ_n) sont deux suites déterministes positives qui tendent vers zéro. Ce modèle est une généralisation des algorithmes de Robbins-Monro, Kiefer-Wolfowitz et des algorithmes avec perturbations Markoviennes (voir Duflo [18]). L'algorithme de Robbins-Monro correspond au cas $R_n = 0$ et $\sigma_n = \gamma_n$.

Soit z^* le zéro de h. De très nombreux résultats basés sur différents critères garantissent la convergence presque sûre de (Z_n) vers z^*. Dans la vaste littérature sur le sujet, on citera Benveniste, Métivier et Priouret [3], Duflo [18], Kushner et Clark [27]. Si (Z_n) converge presque sûrement vers z^*, la vitesse de convergence est donnée par

$$\sqrt{\frac{\gamma_n}{\sigma_n^2}}\big(Z_n - z^*\big) \Longrightarrow \mathcal{N}(0, \Sigma^2), \tag{4.23}$$

où Σ^2 est un réel positif lié au moment d'ordre 2 du bruit (ε_n) et à la dérivée de la fonction h au point cible z^*. On définit la vitesse $v_n \stackrel{\text{def}}{=} \gamma_n\sigma_n^{-2}$.

4.4.1 Hypothèses et résultat principal

On commence par définir une classe de suites positives introduite par Mokkadem et Pelletier [35].

Définition 4.4.1. *Soient $\alpha \in \mathbb{R}$ et (v_n) une suite positive déterministe. On dit que (v_n) est dans l'ensemble $\mathcal{GS}(\alpha)$ si*

$$\lim_{n\to\infty} n\Big(1 - \frac{v_{n-1}}{v_n}\Big) = \alpha.$$

Par exemple, les suites $n^\alpha(\log n)^\beta$ ou $n^\alpha(\log\log n)^\beta$, pour $\alpha, \beta \in \mathbb{R}$ sont dans $\mathcal{GS}(\alpha)$. On définit les hypothèses :

$(H1)$ Z_n converge presque sûrement vers z^*.

$(H2)$ La fonction h est définie sur $\mathbb{R}$ et z^* est un zéro de h tel que, sur un voisinage de z^*,

$$h(z) = H(z - z^*) + \mathcal{O}(|z - z^*|^2),$$

avec $H < 0$.

$(H3)$ Le bruit (ε_n) est une différence de martingales telle que

$$\lim_{n\to\infty} \mathbb{E}\big[\varepsilon_{n+1}^2 | \mathcal{F}_n\big] = \sigma^2 \quad \text{p.s.}$$

$(H4)$ La perturbation (R_n) peut se décomposer en deux termes

$$R_n = o\big(v_n^{-1/2}(\log s_n)^{-q}\big) + \mathcal{O}\big(|Z_{n-1} - z^*|^2\big) \quad \text{p.s.} \quad \forall q \geq 0,$$

avec $s_n \overset{\text{def}}{=} \sum_{k=1}^n \gamma_k$.

$(H5)$ Les pas (γ_n) et (σ_n) vérifient

$$\begin{aligned} (\gamma_n) \in \mathcal{GS}(-\alpha) \qquad & \text{avec} \qquad \alpha \in \Big]\max\Big\{\frac{1}{2}, \frac{2}{a}\Big\}, 1\Big], \\ (\sigma_n) \in \mathcal{GS}(-\beta) \qquad & \text{avec} \qquad \beta \in \Big]\frac{\alpha}{2}, \alpha\Big], \\ \lim_{n\to\infty} n\gamma_n \quad & > \quad -\frac{2\beta - \alpha}{2H}. \end{aligned}$$

On note $\xi \overset{\text{def}}{=} \lim_{n\to\infty}\big(n\gamma_n\big)^{-1}$ et

$$\Sigma^2 \overset{\text{def}}{=} \frac{-\sigma^2}{2H + \xi(2\beta - \alpha)}.$$

Cette constante Σ^2 est bien celle qui apparaît comme étant la variance asymptotique de l'équation (4.23). Grâce à $(H5)$, Σ^2 est strictement positive.

Commentaires sur les hypothèses

Les gains usuels

$$\gamma_n = \frac{\gamma_0}{n^\alpha} \quad \text{et} \quad \sigma_n = \frac{\sigma_0}{\sqrt{n^{\alpha+\beta}}}, \quad \text{avec} \quad \gamma_0 > 0,\ \sigma_0 > 0, \quad \text{et} \quad 0 < \beta \leq \alpha,$$

pour $\alpha \in]\max\{1/2, 2/a\}, 1[$ ou ($\alpha = 1$ et $\beta < -2H\gamma_0$) vérifient l'hypothèse $(H5)$.

Théorème 4.4.2. *Soit $p \geq 1$ un entier. On suppose que le bruit (ε_n) vérifie la condition de moment (H_{2p+}) définie en (4.3). Sous les hypothèses $(H1)$ à $(H5)$, on a*

$$\lim_{n\to\infty} \frac{1}{s_n} \sum_{k=1}^{n} \gamma_k \Big[\sqrt{v_k}(Z_k - z^*)\Big]^{2p} = \frac{\Sigma^{2p}(2p)!}{2^p p!} \quad p.s.$$

$$\lim_{n\to\infty} \frac{1}{s_n} \sum_{k=1}^{n} \gamma_k \Big[\sqrt{v_k}(Z_k - z^*)\Big]^{2p-1} = 0 \quad p.s.$$

Dans le cas particulier $p = 1$, la convergence (4.24) est la loi forte quadratique établie par Pelletier [38]. Les constantes limites dans (4.24) et (4.25) correspondent aux moments de la loi gaussienne $\mathcal{N}(0, \Sigma^2)$.

Le *théorème de Carleman* (voir par exemple Feller [20]) fournit une condition sur les moments garantissant que la connaissance de tous ces moments caractérise la loi.

Théorème 4.4.3 (Carleman). *Une loi de probabilités est entièrement déterminée par ses moments (m_n) si*

$$\sum_{n=1}^{\infty} m_{2n}^{-1/2n} = \infty.$$

Puisque les moments de la Gaussienne satisfont la condition de Carleman, on en déduit les deux corollaires suivants.

Corollaire 4.4.4 (TLCPS). *On suppose que (ε_n) est une différence de martingales telle que, pour tout entier $p \geq 1$,*

$$\sup_{n\geq 0} \mathbb{E}\big[|\varepsilon_{n+1}|^p | \mathcal{F}_n\big] < \infty \quad p.s.$$

Alors sous les hypothèses $(H1)$ à $(H5)$, on a

$$\frac{1}{s_n} \sum_{k=1}^{n} \gamma_k \delta_{\sqrt{v_k}(Z_k - z^*)} \Longrightarrow \mathcal{N}(0, \Sigma^2) \quad p.s.$$

Corollaire 4.4.5. *On suppose que (ε_n) est une différence de martingales telle que, pour tout entier $p \geq 1$,*

$$\sup_{n\geq 0} \mathbb{E}\big[|\varepsilon_{n+1}|^p | \mathcal{F}_n\big] < \infty \quad p.s.$$

De plus, on suppose que f est une fonction presque partout continue, à croissance polynômiale au voisinage de l'infini. Sous les hypothèses $(H1)$ à $(H5)$, on a alors

$$\lim_{n\to\infty} \frac{1}{s_n} \sum_{k=1}^{n} \gamma_k f\left(\sqrt{v_k}(Z_k - z^*)\right) = \int f dG_\Sigma \quad p.s.,$$

où G_Σ désigne la loi $\mathcal{N}(0, \Sigma)$.

Ce corollaire permet d'approcher une intégrale gaussienne d'une fonction presque partout continue et à croissance polynomiale à l'infini.

4.4.2 Exemples d'applications

Les trois exemples d'applications sont des procédures de Robbins-Monro.

Paramètre de Translation

Soit (Y_n) une suite de variables aléatoires indépendantes de même loi de densité f par rapport à la mesure de Lebesgue. Cette suite (Y_n) de variables centrées n'est pas observable. On n'a accès qu'à un échantillon translaté (X_n) avec $X_n = Y_n + \theta$. Le paramètre de translation $\theta = \mathbb{E}[X_n]$ est inconnu. Sans connaître f, on suppose que cette fonction est paire, strictement positive et de classe C^1. L'estimateur récursif défini par

$$\widehat{\theta}_{n+1} = \widehat{\theta}_n - \gamma_n\Big(\mathbb{1}_{\{X_{n+1}\leq\widehat{\theta}_n\}} - \frac{1}{2}\Big)$$

est un cas particulier du modèle (4.22) avec $\sigma_n = \gamma_n$, $R_{n+1} = 0$ et

$$h(z) = \frac{1}{2} - \mathbb{E}\big[\mathbb{1}_{\{X_{n+1}\leq z\}}|\mathcal{F}_n\big],$$

où $\mathcal{F}_n$ est la tribu des événements antérieurs à l'instant n et $\mathbb{F} \stackrel{\text{def}}{=} (\mathcal{F}_n)$ est la filtration naturelle. Le théorème 4.4.2 s'applique et fournit une asymptotique de l'erreur cumulée d'estimation de la médiane. De plus, la suite des estimateurs du paramètre de translation $(\widehat{\theta}_n)$ satisfait le TLCPS :

$$\frac{1}{s_n}\sum_{k=1}^{n}\gamma_k\delta_{\frac{1}{\sqrt{\gamma_k}}(\widehat{\theta}_k-\theta)} \Longrightarrow \mathcal{N}(0,\Sigma^2) \quad \text{p.s.}$$

avec $\Sigma^2 = \big[4\big(2f(0) - \alpha\xi\big)\big]^{-1}$.

Estimation récursive de quantiles

Dans l'exemple précédent, puisque l'on a supposé la densité paire, le paramètre de translation est la médiane. Avec une procédure analogue, on obtient des propriétés asymptotiques sur l'erreur d'estimation de quantiles. A partir d'un échantillon (Y_n) de fonction de répartition F strictement croissante, sans connaître F on peut estimer le quantile q d'ordre δ, *i.e.* $\delta \stackrel{\text{def}}{=} F(q)$ avec la procédure récursive :

$$\widehat{q}_{n+1} = \widehat{q}_n - \gamma_n\Big(\mathbb{1}_{\{Y_{n+1}\leq\widehat{q}_n\}} - \delta\Big).$$

Cet algorithme est de nouveau un cas particulier de (4.22) avec $\sigma_n = \gamma_n$, $R_{n+1} = 0$,

$$h(z) = \delta - \mathbb{E}\big[\mathbb{1}_{\{Y_{n+1}\leq z\}}|\mathcal{F}_n\big] = \delta - F(z),$$

et

$$\varepsilon_{n+1} = \mathbb{E}\big[\mathbb{1}_{\{Y_{n+1}\leq\widehat{q}_n\}}|\mathcal{F}_n\big] - \mathbb{1}_{\{Y_{n+1}\leq\widehat{q}_n\}}.$$

Si l'on suppose que la densité $f = F'$ est continuement différentiable, on peut alors appliquer le théorème 4.4.2.

Estimation récursive de la moyenne

Soit (Y_n) une suite de variables aléatoires indépendantes et de même loi, de moyenne μ et de variance σ^2. L'estimateur récursif de la moyenne s'écrit également comme un cas particulier du modèle (4.22) sous la forme :

$$\widehat{\mu}_{n+1} = \widehat{\mu}_n + \gamma_n\big(Y_{n+1} - \widehat{\mu}_n\big),$$

avec $h(z) = \mu - z$ et $\varepsilon_{n+1} = Y_{n+1} - \mu$. En faisant les hypothèses de moments appropriées sur la loi de Y_1, on peut également appliquer le théorème 4.4.2.

Références

[1] M. Atlagh et M. Weber. « Le théorème central limite presque sûr ». Dans : *Expo. Math.* 18.2 (2000), p. 97–126.

[2] R. Belfadli, K. Es-Sebaiy et Y. Ouknine. « Parameter Estimation for Fractional Ornstein-Uhlenbeck Processes: Non-Ergodic Case. » Dans : *Frontiers in Science and Engineering* 1.1 (2011), p. 1–16.

[3] A. Benveniste, M. Métivier et P. Priouret. *Adaptive Algorithms and Stochastic Approximations*. T. 22. Applications of Mathematics. New York : Springer-Verlag, 1990.

[4] B. Bercu. « On the convergence of moments in the almost sure central limit theorem for martingales with statistical applications ». Dans : *Stochastic Processes and their applications* 111 (2004), p. 157–173.

[5] B. Bercu, P. C. et G. Fayolle. « Convergence des moments dans le théorème de la limite centrale presque sûr pour les martingales vectorielles ». Dans : *Rapport de recherche INRIA* 6056 (déc. 2006). 24 pages.

[6] Bernard Bercu, Ivan Nourdin et Murad S. Taqqu. « Almost sure central limit theorems on the Wiener space ». Dans : *Stochastic Process. Appl.* 120.9 (2010), p. 1607–1628.

[7] B. Bercu et J.-C. Fort. « A moment approach for the almost sure central limit theorem for martingales ». Dans : *Studia Scientiarum Mathematicarum Hungarica* (2006).

[8] I. Berkes. « Results and problems related to the pointwise central limit theorem ». Dans : *Asymptotic methods in probability and statistics (Ottawa, ON, 1997)*. Amsterdam : North-Holland, 1998, p. 59–96.

[9] I. Berkes et E. Csáki. « A universal result in almost sure central limit theory ». Dans : *Stochastic Process. Appl.* 94.1 (2001), p. 105–134.

[10] G. A. Brosamler. « An almost everywhere central limit theorem ». Dans : *Math. Proc. Cambridge Philos. Soc.* 104.3 (1988), p. 561–574. issn : 0305-0041.

[11] F. Chaâbane. « Invariance principles with logarithmic averaging for martingales ». Dans : *Studia Sci. Math. Hungar.* 37.1-2 (2001), p. 21–52. issn : 0081-6906.

[12] F. Chaâbane. « Version forte du théorème de la limite centrale fonctionnel pour les martingales ». Dans : *C. R. Acad. Sci. Paris Sér. I Math.* 323.2 (1996), p. 195–198. issn : 0764-4442.

[13] F. Chaâbane et F. Maâouia. « Théorèmes limites avec poids pour les martingales vectorielles ». Dans : *ESAIM Probab. Statist.* 4 (2000), 137–189 (electronic). issn : 1292-8100.

[14] F. Chaâbane, F. Maâouia et A. Touati. « Généralisation du théorème de la limite centrale presque-sûr pour les martingales vectorielles ». Dans : *C. R. Acad. Sci. Paris Sér. I Math.* 326.2 (1998), p. 229–232.

[15] P. C. « Almost sure properties of Weighted Martingales Transforms with applications to Prediction for Linear Regression Models ». Dans : *Probability and Mathematical Statistics* 23.1 (2003), p. 61–76.

[16] P. C. « On the convergence of moments in the almost sure central limit theorem for stochastic approximation algorithms ». Dans : *ESAIM Probability and Statistics* 17 (2013), p. 179–194.

[17] P. C. et K. Es-Sebaiy. « Almost sure central limit theorems for random ratios and applications to LSE for fractional Ornstein-Uhlenbeck processes ». 15 pages.

[18] M. Duflo. *Random Iterative Methods*. Springer-Verlag, 1997.

[19] Hamdi Fathallah et Ahmed Kebaier. « Weighted limit theorems for continuous-time vector martingales with explosive and mixed growth ». Dans : *Stoch. Anal. Appl.* 30.2 (2012), p. 238–257. issn : 0736-2994.

[20] W. Feller. *An introduction to probability theory and its applications*. T. II. New York : John Wiley, 1966.

[21] K. Gonchigdanzan. « Almost Sure Central Limit Theorems ». Thèse de doct. University of Cincinnati, 2001.

[22] G.C. Goodwin et K.S. Sin. *Adaptative Filtering Prediction and Control*. Prentice-Hall, Englewood Cliffs, N.J., 1984.

[23] H. Holzmann, S. Koch et A. Min. « Almost sure limit theorems for U-statistics ». Dans : *Statist. Probab. Lett.* 69.3 (2004), p. 261–269.

[24] Yaozhong Hu et David Nualart. « Parameter estimation for fractional Ornstein-Uhlenbeck processes ». Dans : *Statist. Probab. Lett.* 80.11-12 (2010), p. 1030–1038.

[25] I. A. Ibragimov et M. A. Lifshits. « On limit theorems of "almost sure" type ». Dans : *Teor. Veroyatnost. i Primenen.* 44.2 (1999), p. 328–350.

[26] Ildar Ibragimov et Mikhail Lifshits. « On the convergence of generalized moments in almost sure central limit theorem ». Dans : *Statist. Probab. Lett.* 40.4 (1998), p. 343–351.

[27] H. J. Kushner et D. S. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Berlin : Springer-Verlag, 1978.

[28] M. Lacey. « Laws of the iterated logarithm for partial sum processes indexed by functions ». Dans : *J. Theoret. Probab.* 2.3 (1989), p. 377–398. issn : 0894-9840.

[29] Damien Lamberton et Gilles Pagès. « Recursive computation of the invariant distribution of a diffusion ». Dans : *Bernoulli* 8.3 (2002), p. 367–405.

[30] Damien Lamberton et Gilles Pagès. « Recursive computation of the invariant distribution of a diffusion: the case of a weakly mean reverting drift ». Dans : *Stoch. Dyn.* 3.4 (2003), p. 435–451.

[31] P. Lévy. *Théorie de l'addition des variables aléatoires*. Gauthiers-Villars., 1937.

[32] M. A. Lifshits. « Almost sure limit theorem for martingales ». Dans : *Limit theorems in probability and statistics, Vol. II (Balatonlelle, 1999)*. Budapest : János Bolyai Math. Soc., 2002, p. 367–390.

[33] M. A. Lifshits. « Lecture Notes on Almost Sure Limit Theorems ». Dans : *Publications IRMA* 54 (2001), p. 1–25.

[34] A. Min. « Limit theorems for statistical functionals with applications to dimension estimation ». Thèse de doct. Institute of Mathematical Stochastics, University of Göttingen, Germany, 2004.

[35] A. Mokkadem et M. Pelletier. « A companion for the Kiefer-Wolfowitz-Blum stochastic approximation algorithm ». Dans : *The Annals of Statistics* (2007).

[36] I. Nourdin et G. Peccati. « Universal Gaussian fluctuations of non-Hermitian matrix ensembles: from weak convergence to almost sure CLTs ». Dans : *ALEA Lat. Am. J. Probab. Math. Stat.* 7 (2010), p. 341–375.

[37] M. Pelletier. « An almost sure central limit theorem for stochastic approximation algorithms ». Dans : *J. Multivariate Anal.* 71 (1999), p. 76–93.

[38] M. Pelletier. « On the almost sure asymptotic behaviour of stochastic algorithms ». Dans : *Stochastic processes and their applications* 78 (1998), p. 217–244.

[39] P. Schatte. « On strong versions of the central limit theorem ». Dans : *Math. Nachr.* 137 (1988), p. 249–256. issn : 0025-584X.

[40] P. Schatte. « On the central limit theorem with almost sure convergence ». Dans : *Probab. Math. Statist.* 11.2 (1990), 237–246 (1991). issn : 0208-4147.

[41] K Es-Sebaiy. « Berry-Esseen bounds for the least squares estimator for discretely observed fractional Ornstein-Uhlenbeck processes. » Dans : *Submitted for publication* (http://arxiv.org/abs/1202.5061).

[42] K. Thangavelu. « Quantile estimation based on the almost sure central limit theorem ». Thèse de doct. Georg–August University of Göttingen, Göttingen, Germany., 2005.

[43] C. Z. Wei et J. Winnicki. « Estimation of the means in the branching process with immigration ». Dans : *Ann. Statist.* 18.4 (1990), p. 1757–1773.

Chapitre 5

Perspectives

Je présente dans ce dernier chapitre un bref panorama de mes perspectives de recherche. Mes projets s'orientent autour de deux thèmes principaux mais non exhaustifs : les chaînes de Markov à mémoire variable et les algorithmes stochastiques.

5.1 Chaînes de Markov à mémoire variable

5.1.1 Etude du modèle probabiliste

Le modèle de source VLMC a le mérite d'être suffisamment général pour modéliser de la dépendance non bornée, inclure des processus ne possédant pas de propriétés de renouvellement, tout en permettant de faire des calculs explicites *« à la main »*. Je souhaite préciser le lien entre l'arbre des contextes probabilisé et la complexité de la source. Qu'est-ce qui résulte de la forme de l'arbre et des lois de probabilités associées aux feuilles ? Quel phénomène sur l'arbre est responsable des propriétés de mélange et du degré de corrélation ? Je souhaite également établir des conditions nécessaires et suffisantes d'existence et d'unicité de mesure invariante pour des arbres plus généraux que le peigne ou le bambou, en évitant les hypothèses classiques et sans doute trop restrictives de non-nullité et continuité :

- $\inf_{c\in\mathcal{C},\alpha\in/CA} q_c(\alpha) > 0$,
- l'application $c \mapsto q_c$ est continue.

Si une mesure invariante existe, y a-t-il convergence vers cette mesure et éventuellement à quelle vitesse ? Une première direction de travail est de s'intéresser à un arbre des contextes relativement simple, par exemple un arbre avec un nombre fini de branches infinies quelconques. Sans utiliser le renouvellement comme nous l'avons fait dans les cas du peigne et du bambou, il s'agit de comprendre quels paramètres sur l'arbre des contextes probabilisé influent sur le calcul ou l'existence de la mesure invariante.

Ensuite, on pourra considérer des arbres plus généraux, comme l'arbre possédant toutes les branches codant en base 2 pour des éléments de $\mathbb{Q}$, avec par conséquent un nombre dénombrable de branches infinies.

Dans les questions relatives à la vitesse de convergence vers l'équilibre, je souhaite pouvoir identifier des conditions garantissant des propriétés de mélange. Dans quels cas a-t-on un mélange uniforme ? Quelle notion de mélange ?

5.1.2 Applications à la neurobiologie

Dans le cadre d'applications à la neurobiologie avec Bruno Cessac (INRIA Sophia-Antipolis), nous aurions également besoin d'avoir des résultats de type grandes déviations ou des inégalités de concentration non-asymptotique pour adapter les résultats au jeu de données. Généralement, les instants d'émission d'un neurone dépendent de l'activité des neurones sur un voisinage. Ce voisinage est fonction de la configuration des potentiels cumulés. Ce type d'interaction est une « interaction à portée variable ». Ces processus à mémoire (spatiale ou non) variable sont des modèles pertinents dans ce contexte, comme en attestent les travaux de Bruno Cessac [1] et Antonio Galvès et Eva Löcherbach [2]. L'un de nos objectifs est d'utiliser ces modèles pour résoudre des problèmes provenant de la neurobiologie et déterminer les vitesses de convergence pour obtenir un ordre de grandeur sur la taille d'échantillon garantissant une estimation fiable.

5.1.3 Marches aléatoires persistantes

Nous avons vu dans le chapitre 2 que pour une VLMC $(U_n)_{n\geq 0}$, avec $U_n = \dots X_{n-1}X_n$, dont l'arbre de contextes possède une branche infinie, le processus des lettres (X_n) n'est en général pas markovien. La marche aléatoire associée aux incréments (X_n)

$$S_t := \sum_{n=0}^{t} X_n,$$

pour $t \in \mathbb{N}$, est dite *persistante* et appartient à la famille des processus à mémoire longue. Cette procédure permet d'agrandir la collection des modèles stochastiques utilisés en mathématiques appliquées. Néanmoins, dans de nombreux cas, des modèles en temps continu sont préférés, par exemple en économie et finance. En général, les modèles mathématiques considérés dans ces cadres sont markoviens (mouvement brownien géométrique, processus Cox-Ingersoll-Ross, ...). Par conséquent, ils ne sont pas particulièrement adaptés à la mémoire des agents du marché. Le défi consiste maintenant à utiliser la marche aléatoire associée à des augmentations VLMC afin d'améliorer le cadre continu.

La clé des preuves dans [9] réside dans l'utilisation de propriétés de renouvellement inhérentes aux modèles de type peignes. Il serait intéressant d'étudier des modèles plus généraux : des marches construites à partir d'autres arbres des contextes. Quel est l'impact de la «persistance» dans ce cas plus général ?

Avec Arnaud Le Ny (LAMA, Université Paris-Est), nous avons commencé à étudier un modèle de marche à mémoire variable en dimension deux, sur le réseau carré $\mathbb{Z}^2$. Cette

1. B. Cessac. « A discrete time neural network model with spiking neurons: II: Dynamics with noise ». Dans : *J. Math. Biol.* 62.6 (2011), p. 863–900.

2. A. Galvès et E. Löcherbach. « Infinite systems of interacting chains with memory of variable length - a stochastic model for biological neural nets ». Dans : *Journal of Statistical Physics* 151.5 (June 2013), p. 896–921.

marche est construite à partir des incréments à valeurs dans les quatre orientations cardinales $\{N, S, E, W\}$ avec un arbre des contextes dont la forme est donnée sur la figure 5.1. Y a-t'il un changement de phase, en fonction des lois de probabilités associées

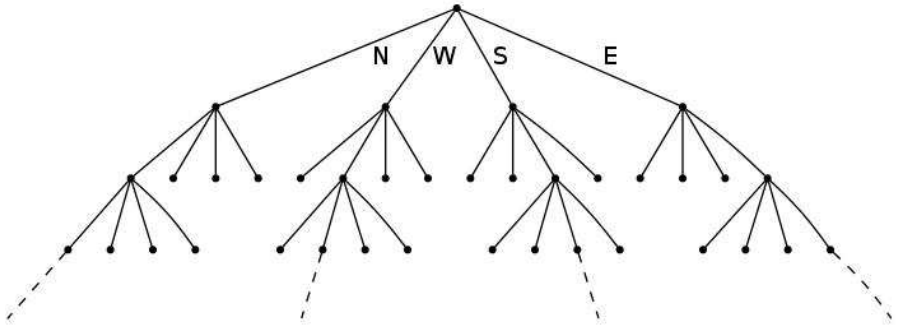


FIGURE 5.1: Arbre de contextes quaternaire associé à la construction d'une marche aléatoire dans le réseau $\mathbb{Z}^2$.

à cet arbre de contexte, qui permette de passer d'une chaîne récurrente à une chaîne transiente ?

5.2 Algorithmes Stochastiques

Les algorithmes stochastiques, en raison de leur caractère récursif, sont particulièrement utiles lorsqu'on observe les données « en ligne ». Ils permettent de faire des estimations avec des procédures, généralement simples, de mise à jour qui ne nécessitent pas la ré-estimation complète du modèle statistique considéré, ré-estimation parfois très coûteuse en temps de calcul. Par ailleurs un grand intérêt de ces approches itératives est qu'elles ne nécessitent pas de stocker en mémoire tous les objets à analyser.

Dans les articles [4, 5], nous avons proposé un algorithme récursif d'estimation de la médiane géométrique, efficace au sens où il se comporte asymptotiquement en loi comme l'estimateur « statique » qui prendrait en compte l'ensemble des observations simultanément. Les résultats théoriques obtenus sur ces algorithmes stochastiques à valeur dans un espace de Hilbert sont tous asymptotiques. Il serait intéressant, en vue d'applications pratiques à des courbes de consommation électrique, d'obtenir des bornes non asymptotiques, à partir d'inégalités de concentration pour des martingales à valeurs dans un espace de Hilbert. L'obtention de bornes qui prennent en compte les différents paramètres du problème permettra de mieux comprendre les performances de l'algorithme à distance finie.

Une autre piste de recherche concerne l'estimation de la matrice, ou l'opérateur pour les espaces fonctionnels, de covariance limite de l'estimateur de la médiane ou de la médiane conditionnelle, à l'aide d'un estimateur récursif correctement pondéré. Il est pour

cela nécessaire d'obtenir un résultat de type loi forte quadratique pour des algorithmes stochastiques à valeurs dans un espace de Hilbert. Ceci permettra de calculer des bandes de confiance pour la courbe médiane.

Dans la partie classification non supervisée via l'algorithme séquentiel de type k-médianes, il a été prouvé que cet algorithme converge presque sûrement vers un point stationnaire de la fonctionnelle à minimiser. Cependant, l'algorithme peut se retrouver « piégé » par des minima locaux. Il serait intéressant d'obtenir des informations sur la vitesse de convergence et de pouvoir ainsi calibrer les pas de l'algorithme pour contourner ces pièges. Les hypothèses que nous avons utilisées pour établir la convergence ne sont pas facilement vérifiables en pratique et assez techniques. Une méthode analogue à celle utilisée par Pagès [3] pour établir la convergence de l'algorithme de k-means à l'aide d'un algorithme de Kohonen pourrait sans doute être une piste intéressante qui permettrait d'obtenir des hypothèses plus élégantes, comme le sont celles de Pagès. De plus, la preuve de la convergence a été établie uniquement en dimension finie et il serait intéressant de la généraliser pour des espaces fonctionnels afin de pouvoir effectuer de la classification de courbes.

5.3 Vers la biologie ?

Ces pistes de recherche et ces projets sont loin d'être exhaustifs. En particulier, les travaux sur lesquels je me suis penchée depuis ma thèse m'ont éloignée des applications vers la biologie. J'ai toujours été fascinée par la complexité et la beauté du monde vivant et j'ai envie de revenir vers ces problématiques. Ma participation à la Commission Interdisciplinaire 51 du CNRS intitulée « Modélisation et analyse des données et des systèmes biologiques : approches informatiques, mathématiques et physiques » m'a permis de créer des liens avec plusieurs biologistes. J'aimerais réussir à interagir avec eux et m'ouvrir vers de nouvelles thématiques qui me permettront de dégager de nouvelles structures récursives.

3. Gilles Pagès. « A space quantization method for numerical integration ». Dans : *J. Comput. Appl. Math.* 89.1 (1998), p. 1–38. ISSN : 0377-0427.

Récursivité au carrefour de la modélisation de séquences, des arbres aléatoires, des algorithmes stochastiques et des martingales

Résumé : Ce mémoire est une synthèse de plusieurs études à l'intersection des systèmes dynamiques dans l'analyse statistique de séquences, de l'analyse d'algorithmes dans des arbres aléatoires et des processus stochastiques discrets. Les résultats établis ont des applications dans des domaines variés allant des séquences biologiques aux modèles de régression linéaire, processus de branchement, en passant par la statistique fonctionnelle et les estimations d'indicateurs de risque appliqués à l'assurance. Tous les résultats établis utilisent d'une façon ou d'une autre le caractère *récursif* de la structure étudiée, en faisant apparaître des *invariants* comme des martingales. Elles sont au cœur de ce mémoire, utilisées comme outils dans les preuves ou comme objets d'étude.

Mots-clés : Arbres digitaux de recherche, chaîne de Markov à mémoire variable, temps d'occurrences de motifs, système dynamique, trie des suffixes ; lois fortes de martingales discrètes, modèles auto-régressifs, erreur d'estimation et de prédiction, algorithmes de gradient stochastiques, optimisation stochastique.

Recursion at the crossroads of sequence modeling, random trees, stochastic algorithms and martingales

Abstract : This monograph synthesizes several studies spanning from dynamical systems in the statistical analysis of sequences, to analysis of algorithms in random trees and discrete stochastic processes. These works find applications in various fields ranging from biological sequences to linear regression models, branching processes, through functional statistics and estimates of risk indicators for insurances. All the established results use, in one way or another, the *recursive property* of the structure under study, by highlighting *invariants* such as martingales, which are at the heart of this monograph, as tools as well as objects of study.

Key-words : Digital search trees, variable length Markov chain, occurrences time, dynamical system, suffix trie ; strong laws for discrete martingales, auto-regressive models, estimation and prediction error, stochastic gradient algorithms, stochastic optimization.