l'algorithmique: la fouille de données et l'arithmétique

Loïck Lhote

Université de Caen / Basse-Normandie

U.F.R. Sciences **École doctorale SIMEM**
Département d'Informatique

Algorithmes du PGCD et Fouille de Données : le point de vue de l'analyse dynamique

THÈSE

présentée et soutenue publiquement le 6 Septembre 2006

pour l'obtention du

Doctorat de l'Université de Caen

Spécialité Informatique

(Arrêté du 30 mars 1992)

par

Loïck LHOTE

Composition du jury

Rapporteurs :	Richard Brent	Professeur	Australian National University Canberra, Australie
	Danièle Gardy	Professeur	Laboratoire PRISM, Université de
	Jean-Claude Bajard	Professeur	Laboratoire LIRMM, Université de
Examinateurs :	Michèle Sebag	Directrice de recherche au CNRS	Laboratoire LRI, Université d'Orsa
	Jérôme Buzzi	Chargé de recherche au CNRS	Laboratoire CMLS, Ecole Polytechn Palaiseau
	Philippe Flajolet	Directeur de recherche à l'INRIA	INRIA Rocquencourt, Le Chesnay
	Brigitte Vallée	Directrice de recherche au CNRS	GREYC, Université de Caen (Direc

Mis en page avec la classe thloria.

Remerciements

Le travail contenu dans ce manuscrit n'existerait pas sans l'aide considérable apportée par ma directrice de thèse Brigitte Vallée. Toutes ces années, elle n'a cessé de me motiver, m'a poussé à aller de l'avant, m'a fait mûrir (un peu) en pointant certains de mes défauts. Je lui serai toujours reconnaissant d'avoir déployé tant d'énergie à faire de moi un encore très jeune chercheur. Mais outre son énergie, sa connaissance, son intelligence ou ses capacités d'analyse, je garderai aussi en mémoire son rire communicatif qui, lorsqu'il survient, procure systématiquement un moment de plaisir aux personnes proches :-)

Jean-Claude Bajard, Richard Brent et Danièle Gardy m'ont fait l'immense honneur d'être les rapporteurs de ma thèse. Je les remercie de leurs commentaires qui ont contribué à améliorer le manuscrit. Je suis également très honoré de la présence dans mon jury de Philippe Flajolet, Jérôme Buzzi et Michèle Sébag. J'ai déjà eu l'occasion de rencontrer plusieurs fois Philippe Flajolet, mais je n'ai fait que croiser Jérôme Buzzi et Michèle Sébag dans des conférences. J'espère qu'il en sera autrement dans l'avenir.

Au cours de cette thèse, j'ai également eu le plaisir de collaborer avec Véronique Maume-Deschamps, Benoît Daireaux, François Rioult et Arnaud Soulet. Véronique et Benoît (et Brigitte) m'ont initié aux algorithmes d'Euclide rapides et à leur analyse. Ils ont de fait beaucoup contribué à cette thèse et je les en remercie. François Rioult et Arnaud Soulet m'ont ouvert au monde de la Fouille de Données. Je n'y connaissais rien et une grande dose de patience leur a été nécessaire pour répondre à mes incessantes questions. Merci.

La réussite d'une thèse tient aussi au cadre de travail. Les membres du GREYC m'ont très gentiment accueilli et je leur en suis très reconnaissant. Je tiens tout particulièrement à remercier mes compères de bureau ou mes voisins de bureau qui m'ont offert leur amitié : Arnaud, François, Céline, Pédro, Nicolas, Nathalie. L'équipe d'Analyse Dynamique est composée de personnes que j'ai souvent le plaisir de croiser et avec qui j'ai également beaucoup de plaisir à discuter.

Au moment où j'écris ces mots, ma femme Céline s'occupe de notre petit Lilian qui pleure. Céline m'a toujours soutenu dans les moments difficiles même si elle trouvait parfois que la rédaction prenait trop de temps. J'ai vu s'épaissir le manuscrit en même temps que son ventre grossissait. Aujourd'hui, j'ai deux joyaux à la maison qui contribuent à mon bien être personnel et professionnel. Merci Céline pour tout ce que tu m'as apporté avant et pendant cette thèse. Je sais qu'avec toi et Lilian, mon bonheur continuera.

Finalement, ils n'ont peut-être pas contribué directement à la thèse, mais ils ont toujours été avec moi depuis le début : Maman, mon frère Mickaël avec sa fille Maïwenn et sa femme Manuëla, Titine, Benoît, Fredouille, Dieu, Pétronille, Chat-bite (ou Samy pour les moins intimes), Sabrina, Mic-Mic, Tchoff, Rodolphe, Sandra, Cédric, Anne, Emmanuel, Marie, Lydie et j'en oublie certainement. A tous, merci d'avoir contribué à faire de moi ce que je suis aujourd'hui.

A toi Papa qui est parti beaucoup trop tôt.

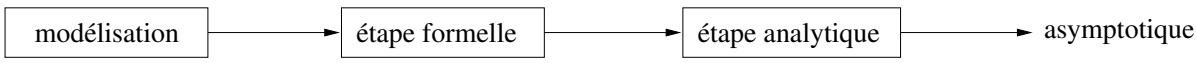
Table des matières

Introduction générale

Cette thèse adopte le point de vue de l'analyse d'algorithmes, dans deux domaines algorithmiques a priori bien distincts, qui sont l'arithmétique et la fouille de données. En algorithmique, nous analysons le comportement probabiliste des paramètres essentiels à des algorithmes de type Euclide, qui calculent le pgcd, aussi bien sur les entiers que sur les polynômes, et nous exhibons en particulier des lois limites gaussiennes. En fouille de données, nous analysons des phénomènes probabilistes qui apparaissent dans des bases de données, liés en particulier au nombre moyen de propriétés partagées par les objets d'une telle base.

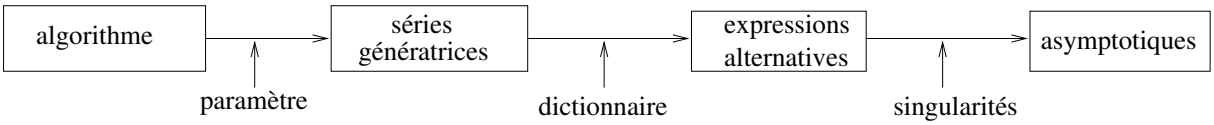
Analyse d'algorithmes. Fondée dans les années 60 par Knuth, l'analyse d'algorithmes cherche à décrire plus précisément le comportement "générique" des algorithmes et de leurs structures de données associées. On parle souvent d'analyse en moyenne, par opposition à l'analyse dans le pire des cas. Alors que l'analyse dans le pire des cas isole le plus mauvais comportement [correspondant souvent à un cas pathologique], l'analyse en moyenne prend en compte l'ensemble des entrées possibles, pour en déduire le comportement moyen. Une telle analyse permet d'appréhender de manière assez réaliste le comportement des algorithmes en pratique, et permet en retour de les optimiser : choix entre deux algorithmes, réglage de paramètres, etc. Souvent, l'analyse en moyenne, quand elle réussit, est seulement une première étape dans le processus complet d'analyse, qui trouve son aboutissement final dans l'analyse en distribution. En effet, même si la valeur moyenne d'un paramètre est plus "significative" que son pire des cas, cette valeur moyenne ne donne pas nécessairement non plus beaucoup d'informations sur la réalité. Il faut la compléter par des études supplémentaires, qui donnent des informations sur la dispersion des valeurs autour de la moyenne (comme l'écart-type), et plus généralement, par une étude en "distribution" qui vise à décrire la répartition des valeurs possibles des divers paramètres.

L'analyse d'un algorithme se décrit selon trois principales étapes.



La première étape vise à modéliser : modéliser tout d'abord les entrées de l'algorithme, mais aussi l'algorithme lui-même. C'est une étape délicate : un modèle trop simple n'est pas assez réaliste, et un modèle trop complexe n'est pas accessible à l'analyse. Il faut donc trouver un compromis .... Nous serons confrontés à cette difficulté en fouille des données (voir chapitres 6 à 8). La modélisation de l'algorithme est également essentielle ; elle vise à isoler les paramètres fondamentaux, qui portent l'information recherchée. A la fin de l'étape de modélisation, nous avons ainsi isolé un ensemble de données, avec une distribution associée, et des paramètres, sur cet ensemble, qui mesurent la qualité de l'algorithme, et dont nous voulons étudier certaines caractéristiques probabilistes. Comme par ailleurs, on est surtout intéressé à ce qui se passe pour des données de grandes tailles, et qu'on ne peut espérer des résultats exacts (qui d'ailleurs ne seraient peut-être pas utilisables), l'analyse d'algorithmes vise à caractériser l'asymptotique précise du comportement probabiliste des algorithmes.

Combinatoire analytique. C'est alors, dans la deuxième et la troisième étape, qu'entre en scène la combinatoire analytique. Le livre fondateur de Knuth [Knu97, Knu98a, Knu98b] a d'abord élaboré les premières méthodes du domaine, où les résolutions de récurrences jouaient un rôle central. Depuis, Philippe Flajolet et son école [FS96, FS99] ont défini la méthodologie générale de la combinatoire analytique, où l'outil principal est la série génératrice, avec ses propriétés à la fois formelles et analytiques. On peut alors préciser le précédent schéma comme suit :



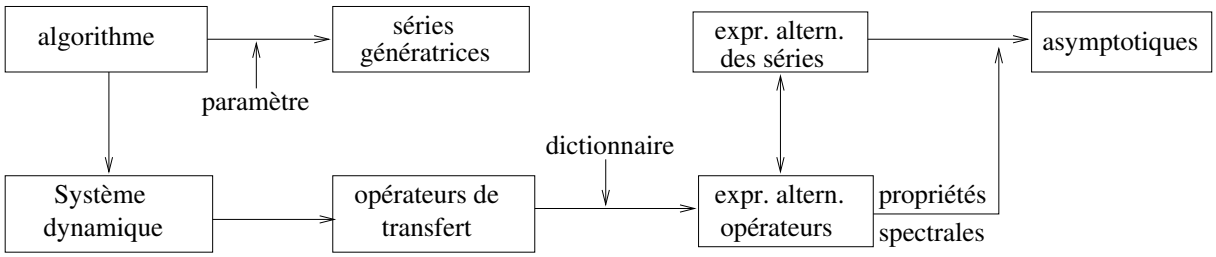
La combinatoire analytique associe au paramètre étudié une série génératrice, dont les coefficients sont étroitement reliés aux caractéristiques probabilistes du paramètre. L'objectif est donc d'extraire les coefficients. Les extractions suivent toujours le même principe : c'est la position et la nature de la singularité dominante qui déterminent l'asymptotique des coefficients. Une formule alternative des séries qui met en évidence les singularités s'avère donc indispensable. Dans cette optique, la combinatoire analytique consiste en deux principales étapes, qui utilisent chacune un dictionnaire. L'étape algébrique traduit les opérations de l'algorithme en opérations algébriques sur les séries génératrices [considérées à ce moment-là comme des séries formelles]. Elle utilise à cette fin un dictionnaire algébrique. A la fin de cette étape, on a trouvé une expression alternative pour les caractéristiques probabilistes des paramètres centraux de l'algorithme (espérance, variance, etc.). Finalement, l'étape analytique utilise l'expression alternative fournie par l'étape précédente, la considère désormais comme une fonction de la variable complexe, détermine la position et la nature de la singularité dominante (celle qui a le plus petit module), et utilise un dictionnaire analytique pour en déduire l'asymptotique des coefficients.

La combinatoire analytique a des applications très variées en théorie des langages, en chimie avec les structures moléculaires, en théorie des nombres avec les partitions d'entiers ou les facteurs des polynômes, en pattern matching avec le dénombrement de motifs particuliers, etc .... Ici, dans cette thèse, nous emploierons cette méthode en arithmétique, dans le cas des polynômes (voir chapitre 2). Cette méthode trouve naturellement son champ dans ce domaine d'application. Les résultats que nous obtenons font certainement partie, pour nombre d'entre eux, du non dit de la littérature scientifique. Mais, ils n'apparaissent pas en tant que tels dans la littérature, et surtout, comme nous le verrons, ils "montrent le chemin" pour l'analyse de l'arithmétique entière, qui s'avère elle beaucoup plus complexe, et nécessite d'autres outils.

Analyse dynamique. La combinatoire analytique adopte somme toute un point de vue assez *statique*, puisqu'elle traduit les opérations de l'algorithme en opérations sur les séries génératrices. Ce n'est plus possible si le cours de l'algorithme modifie trop profondément la distribution des données. Il faut alors trouver un autre objet, qui soit "compatible" avec l'évolution de l'algorithme, et qui traduise cette évolution. Quand il existe un système dynamique sous-jacent à l'algorithme, cet objet est l'opérateur de transfert (ou opérateur de Ruelle) [Rue78] associé au système dynamique, et nous allons donc l'utiliser en lui donnant un rôle (non classique en systèmes dynamiques) d'opérateur générateur. C'est ce qui fournit le cadre de l'analyse dynamique, née du mariage de deux domaines : l'analyse d'algorithmes et les systèmes dynamiques. Ce domaine pourrait aussi bien s'appeler combinatoire dynamique, car il reprend les mêmes principes que la combinatoire analytique, en y intégrant le mouvement apporté par l'opérateur de transfert. Cette méthodologie se développe depuis une dizaine d'années sous l'impulsion de Brigitte Vallée. On peut trouver une présentation générale de ce domaine dans deux articles. L'article

[CMV03] explique la méthode dans le cadre de la théorie de l'information, tandis que l'article [Val06] est consacré à la description de la méthode dans le cas des algorithmes d'Euclide.

L'analyse dynamique emprunte alors un chemin détourné, qu'on peut décrire comme suit :



L'opérateur de transfert décrit l'évolution des données. C'est un objet qui est un condensé de toutes les propriétés du système dynamique, et on peut le modifier pour qu'il décrive en même temps l'évolution d'un paramètre lié à l'évolution des données. Ces opérateurs modifiés jouent alors le rôle d'opérateurs générateurs, car ces opérateurs engendrent alors eux-mêmes les séries génératrices. Puis, on utilise deux dictionnaires qui travaillent sur ces opérateurs de transfert. Le dictionnaire algébrique traduit les opérations de l'algorithme en opérations sur les opérateurs générateurs. Puis le dictionnaire analytique relie les propriétés du système aux propriétés analytiques des opérateurs (notamment les propriétés de leur spectre). Ce sont les propriétés de la valeur propre dominante des opérateurs, qui, une fois analysées et traduites sur les séries, conduiront aux asymptotiques recherchées.

Cette méthodologie a débuté naturellement dans le cadre de l'analyse en moyenne, et a permis d'obtenir de nombreux résultats novateurs dans deux domaines algorithmiques : l'arithmétique et la théorie de l'information. Cette méthode a d'abord été élaborée pour répondre aux besoins de l'analyse des algorithmes d'Euclide, et a permis d'obtenir nombre de résultats dans ce domaine ([AV00, BV05, DV04, DMDV05, LV06b, Val98b]). Puis, Brigitte Vallée a transporté ce point de vue en théorie de l'information, et y a introduit dans [Val01] le modèle des sources dynamiques, qui sont des sources produites par un système dynamique. Ce modèle a permis par la suite des analyses à la fois générales et précises en théorie de l'information : description fine des paramètres des principales structures de données utilisées, analyse précise de plusieurs problèmes de motifs dans les textes produits par ces sources dynamiques [CFV01, CFV01, BV02, BNV01, Bou01]. Puis, il y a trois ans, Viviane Baladi et Brigitte Vallée [BV04, BV05] ont défini les bases de l'analyse en distribution, qui s'avère ici beaucoup plus difficile, et ont élaboré un cadre général pour cette démarche en arithmétique. Tout récemment, la même méthode voit ses premières applications en théorie de l'information.

Cette thèse est un exemple d'application de ces méthodes "dynamiques", que nous utilisons nous aussi dans ses deux domaines privilégiés d'application, l'arithmétique et la théorie de l'information. En arithmétique, nous étendons les méthodes d'analyse en distribution, pour obtenir des résultats sur le comportement probabiliste de toute une classe de paramètres des algorithmes d'Euclide sur les entiers (chapitres 3 à 5). En fouille de données, nous modélisons une base de données comme un objet de la théorie de l'information, et nous pouvons alors étendre le concept de source dynamique en un concept de "base de données dynamique". Nous appliquons alors le champ de l'analyse dynamique (en moyenne) à ce nouveau domaine d'application (chapitres 7 et 8).

Combinatoire analytique et analyse dynamique. Chaque méthode semble donc avoir son champ spécifique d'applications : –algorithmes non corrélés pour la première, – algorithmes construits sur un système dynamique sous-jacent pour la seconde. L'algorithme d'Euclide sur les polynômes

est bien particulier, puisqu'il peut prétendre à être dans chacun des ces deux champs. C'est l'occasion de comparer sur un exemple précis les deux méthodes, et de relier séries génératrices, et opérateurs de transfert (chapitre 2, section 3.6).

Arithmétique. Nous appelons algorithme d'Euclide tout algorithme qui calcule le pgcd (plus grand commun diviseur) par une suite de divisions. L'algorithme d'Euclide (sur les entiers) est souvent considéré comme le plus vieil algorithme au monde. Décrit aux environs de -300 dans "Les éléments" d'Euclide, il a connu de nombreuses évolutions et admet maintenant plusieurs versions à la fois sur les entiers et sur les polynômes [Ste67, Knu98a, BK85, Leh38, Sch71, Nak81]. Les algorithmes d'Euclide jouent un rôle central dans tous les systèmes de calcul formel. D'après Jebelean [Jeb95], certaines applications arithmétiques utilisent 60% de leur temps total de calcul en calculs de pgcd et ce chiffre monte à 80% pour des calculs de bases de Gröbner. En outre, les algorithmes euclidiens ne servent pas seulement à calculer le pgcd ; ils sont indispensables en arithmétique modulaire, puisqu'ils servent à calculer l'inverse modulaire, et sont ainsi centraux en cryptographie. La factorisation de polynômes utilise le calcul de pgcd comme une brique de base,... Analyser et mieux comprendre ces algorithmes est donc une entreprise essentielle pour l'algorithmique arithmétique.

Dans cette thèse, nous étudions deux types différents d'algorithmes d'Euclide. Le premier est l'algorithme d'Euclide lui-même, dans sa version classique. Le second est un algorithme d'Euclide dit "rapide", proposé par Knuth [Knu71] en 1971 et amélioré par Schönhage [Sch71] la même année. L'algorithme classique utilise une succession de divisions euclidiennes pour calculer le pgcd. Comme on ne connaît pas d'algorithme plus efficace que la division naïve pour effectuer des divisions sur de grands entiers ou de grands polynômes, l'algorithme d'Euclide classique a une complexité binaire quadratique dans le pire des cas. L'algorithme de Knuth-Schönhage, lui, applique récursivement une idée de Lehmer [Leh38] qui remplace de grandes divisions par des petites divisions et des grandes multiplications. On tire alors profit de l'existence de multiplications rapides (Katatsuba, FFT), de sorte que l'algorithme de Knuth-Schönhage est sous-quadratique (dans le pire des cas).

Les analyses de l'algorithme classique débutent en 1845 avec Lamé [Lam45] qui démontre que dans le pire des cas, le nombre de divisions (ou étapes) est linéaire en la taille des entrées. Dès que l'analyse en moyenne fait son entrée en algorithmique, autour des années soixante-dix, Heilbronn [Hei69] et Dixon [Dix70] prouvent avec des méthodes différentes que le nombre moyen d'étapes est aussi linéaire. Finalement, Hensley [Hen94] effectue en 1994 l'analyse complète du nombre d'étapes de l'algorithme d'Euclide sur les nombres, et exhibe une loi limite gaussienne. Un résultat similaire est obtenu sur les polynômes par les auteurs de [KK88], par des méthodes de dénombrement direct.

L'inconvénient majeur de ces premières analyses est leur spécificité – spécificité du paramètre "nombre de divisions" et spécificité du type d'algorithme d'Euclide considéré. La méthode d'analyse dynamique s'applique, elle, à une large classe d'algorithmes et de paramètres [Val97b, Val98a, AV00, Val00b, BDV02, DV04, DMDV05] et permet d'atteindre l'analyse de paramètres très divers : nombre d'étapes, nombre de quotients prenant une certaine valeur, taille binaire (ou en bits) totale de tous les quotients, complexité binaire (i.e. le nombre d'opérations sur les bits effectuées par l'algorithme), taille du reste à une fraction de l'exécution, ... Tous ces paramètres sont maintenant bien décrits en moyenne, et ce, pour toute une classe d'algorithmes euclidiens. Par exemple, Akhavi et Vallée [AV00] ont fait une analyse fine de la complexité binaire moyenne.
Une fois les analyses en moyenne effectuées, il est naturel de se tourner vers l'analyse en distribution, qui s'avère ici aussi beaucoup plus difficile. La méthodologie élaborée par Baladi et

Vallée [BV05, BV04] leur permet d'analyser en distribution, et pour une classe d'algorithmes d'Euclide, toute une classe de paramètres, qu'on appelle les coûts additifs et à croissance modérée. Elles exhibent pour ces paramètres une loi limite gaussienne. Les paramètres ainsi étudiés sont divers : nombre d'étapes, nombre de quotients prenant une certaines valeur, taille totale de tous les quotients, ... Mais les analyses de Baladi et Vallée ne permettent d'atteindre, ni la complexité binaire ni la taille des restes à une fraction de l'exécution[1].

Contributions de la thèse en arithmétique. Ici, nous étudions les algorithmes d'Euclide, à la fois sur les entiers et les polynômes, et visons à décrire le comportement probabiliste fin de ces algorithmes en obtenant des lois limite gaussiennes. L'analyse polynomiale est plus facile que celle sur les entiers puisque l'on peut utiliser les outils de la combinatoire analytique tandis que pour l'analyse entière, la combinatoire analytique s'avère infructueuse. Nous cherchons pourtant à adopter une démarche parallèle pour les analyses de ces algorithmes d'Euclide dans chacun de ces cadres – puisque c'est de fait le MEME algorithme– et, même si les outils d'analyse sont différents, nous montrons que les méthodes de combinatoire analytique utilisées sur les polynômes peuvent être vues comme un cas particulier d'analyse dynamique (chapitre 3).

Nous nous intéressons d'abord aux algorithmes d'Euclide classiques, à la fois sur les polynômes et les entiers, et nous étudions trois paramètres : – complexité binaire de l'algorithme standard, qui calcule (seulement) le pgcd – complexité binaire de l'algorithme étendu, qui calcule aussi les coefficients de Bezout, et qui est utile pour calculer les inverses modulaires – taille du reste à une fraction de l'exécution, qui décrit l'état de l'algorithme d'Euclide à une fraction de l'exécution et qui intervient dans l'analyse des algorithmes rapides. Une analyse en distribution de ces trois paramètres pour chacun de ces deux algorithmes devrait produire en tout SIX lois limites gaussiennes. De fait, nous en obtenons seulement ... CINQ, car nous échouons à obtenir une loi limite gaussienne pour la complexité binaire de l'algorithme standard sur les entiers.

Dans les deux cas – entiers ou polynômes–, les preuves sont fondées sur le même principe. Nous effectuons tout d'abord une décomposition des coûts à étudier, qui fait apparaître plusieurs familles de coûts : coûts à croissance modérée, coûts à croissance intermédiaire, coûts terminaux, coût de type longueur de cheminement (voir les définitions 3, 4 et le théorème 11). Nous analysons aussi ces familles de coûts, pour lesquelles nous obtenons des résultats généraux, qui ont leur intérêt propre.

Nous effectuons enfin l'analyse en moyenne d'une variante de l'algorithme de Knuth-Schönhage, lorsque la multiplication rapide utilisée n'est pas trop ... rapide ! Quand l'algorithme $\mathcal{KS}$ utilise une multiplication rapide de complexité $\Theta(n^\alpha)$ avec $\alpha \in]3/2, 2[$, nous montrons que la complexité binaire moyenne de $\mathcal{KS}$ est d'ordre $\Theta(n^\alpha)$ (théorème 8). Ceci constitue le premier résultat d'analyse en moyenne sur un algorithme sous-quadratique et est fondé sur une une étude fine de la "régularité" de l'algorithme (théorème 6).

Dans toutes les analyses précédentes sur les entiers, les termes dominants des asymptotiques font intervenir des constantes qui sont liées aux propriétés spectrales des opérateurs de transfert. Le statut algorithmique de ces constantes, et leur calcul éventuel, sont alors des questions très naturelles. Certaines de ces constantes admettent des expressions closes, mais ce ne semble pas être le cas pour d'autres. Dans le chapitre 5, nous étudions en particulier trois constantes : la constante de Gauss-Kuz'min-Wirsing (liée au système dynamique des fractions continues), la constante de Hensley (constante dominante de la variance du nombre d'étapes) et les dimensions de Hausdorff d'espaces de Cantor sur les fractions continues. Nous montrons que ces trois constantes sont calculables en temps polynomial (théorème 17). Nous utilisons pour cela un algorithme proposé par Daudé, Flajolet et Vallée [DFV97], que ces auteurs avaient déjà utilisé

[1] Ces deux paramètres ne sont pas additifs (voir le théorème BV dans le premier chapitre).

pour calculer des constantes analogues [DFV97, Val98b], et qui semblait très bien fonctionner. Cependant, l'algorithme n'était pas prouvé, et c'est aussi ce que nous faisons dans ce chapitre.

Fouille de données. L'extraction de connaissances dans une base de données (ECBD), ou fouille de données, constitue notre deuxième champ d'étude. Avec le développement des techniques informatiques, la quantité d'informations contenues dans les bases de données s'accroît continûment, et l'automatisation du traitement de ces données devient de ce fait indispensable. La fouille de données s'attache à trouver des informations significatives dans une base de données. Les applications sont multiples : retrouver les gènes responsables d'une maladie, caractériser et détecter des intrusions dans un système sécurisé, concevoir des logiciels d'aide à la décision, etc. La fouille de données s'organise généralement selon deux phases : une phase de prétraitement, qui "formate" la base de données et une phase d'extraction proprement dite, qui vise à trouver les informations dans la base ainsi formatée. Le prétraitement demande beaucoup d'énergie puisque chaque étape doit être validée par un expert (qui peut revenir sur ses choix) alors que l'extraction des motifs est automatique. Nous analysons la base de données une fois formatée, qui peut être alors représentée comme un tableau de 0 et de 1. Les lignes représentent les objets, et les colonnes représentent les propriétés de ces objets, appelés encore attributs. Un motif est alors un ensemble de propriétés. Dans ce contexte, l'information significative est représentée par des motifs qui apparaissent fréquemment et c'est le nombre *moyen* de ces motifs dits fréquents que nous cherchons à analyser.

Dans une base de données, il y deux types de motifs intéressants : les motifs γ-fréquents qui, apparaissent au moins γ fois dans la base [étant donné un seuil γ fixé par l'utilisateur], et les motifs γ-fermés qui sont une représentation condensée des motifs γ-fréquents. Le nombre de motifs fréquents (fermés) est un paramètre qui peut varier beaucoup, puisque sa valeur maximale est exponentielle en le nombre d'attributs, tandis que sa valeur minimale est nulle. Dans [GGV01], Geerts, Goethals et Van den Bussche obtiennent une borne supérieure pour le nombre de motifs fréquents. Cette borne supérieure, fondée sur la décomposition d'entiers en somme de coefficients binomiaux, est peu maniable, et il est difficile d'en extraire un ordre de grandeur asymptotique. A notre connaissance, il n'existe pas d'autres analyses dans le pire des cas pour le nombre de ces motifs, motifs fréquents ou motifs fermés.

Contributions de la thèse en fouille de données. L'analyse en moyenne est très peu présente dans le domaine de la fouille de données. Lorsque Agrawal et al. introduisent pour la première fois l'algorithme Apriori [AMS$^+$96], ils montrent que les motifs sont en moyenne très courts. Toivonen utilise des résultats probabilistes pour justifier sa méthode d'échantillonnage [Toi96]. La première véritable analyse en moyenne vient peut-être de Purdom et al. [PGG04] qui analysent le taux d'échec d'un algorithme d'extraction de motifs.

Dans cette thèse, nous cherchons à effectuer une analyse en moyenne du nombre de ces motifs, motifs fréquents ou motifs fermés. Même si ce n'est pas le point de vue généralement admis en fouille de données, ce domaine de la fouille de données fait partie, de notre point de vue, du domaine général de la théorie de l'information. La théorie de l'information vise à étudier les sources, qui sont juste des mécanismes qui produisent des symboles. Une base de données peut toujours être décrite par une unique source, qui produit en "parallèle" les textes correspondant à chaque ligne (chaque objet) de la base. Dans ce contexte, le nombre de motifs fréquents (ou fermés) est exactement le nombre de motifs présents simultanément dans plusieurs textes produits par la même source. Bien sûr, les propriétés de ces bases vont être très liées aux propriétés de la source elle-même, et aux hypothèses faites sur l'indépendance des "mots –ligne" produits. Et ici, nous faisons toujours l'hypothèse que les mots–ligne sont indépendants.
Nous analysons donc le nombre de motifs fréquents (ou fermés), et ce, dans trois situations défi-

nies par la nature du seuil. Nous considérons d'abord un seuil γ qui est une fonction linéaire du nombre d'objets (hypothèse 6). Sous une hypothèse très naturelle sur la source (hypothèse 8), qui exprime que la probabilité d'apparition d'un motif décroît exponentiellement avec sa longueur, nous montrons que le nombre de motifs γ-fréquents est polynomial en le nombre d'attributs [de colonnes] (théorème 20). C'est un résultat intéressant, qui exhibe un cas moyen bien différent du pire des cas. Il explique le bon fonctionnement pratique des algorithmes pour des seuils raisonnables, alors que le pire des cas indiquerait toujours un comportement exponentiel.
Nous considérons ensuite un seuil dit intermédiaire (hypothèse 7). Sous une condition sur la source émettrice plus forte que la précédente (hypothèse 9), nous montrons que le nombre moyen de motifs γ-fréquents est équivalent au nombre moyen de motifs γ-fermés. Ce résultat est bien conforme aux observations faites sur des bases de données faiblement corrélées (du type panier de la ménagère). Bien entendu, il n'est plus vrai avec des bases fortement corrélées.
Nous considérons enfin le cas d'un seuil fixe (hypothèse 5). Nous faisons dans ce cas une hypothèse plus élaborée sur la source, fondée sur une estimation précise du nombre moyen de motifs partagés par γ mots produits par la source (hypothèse 10). Sous cette condition, nous montrons que le nombre moyen de motifs fréquents est exponentiel en la taille de la base. Un seuil fixe signifie en pratique un seuil très petit devant le nombre de lignes et les expériences montrent bien l'explosion du nombre de motifs fréquents associés à ce type de seuil.

Les résultats que nous obtenons sont a priori très généraux, puisqu'ils sont valables pour toute base dont les mots-ligne sont produits indépendamment par la même source, vérifiant les trois hypothèses décrites ci-dessus. Les trois hypothèses que nous avons faites sur de telles sources sont-elles légitimes ? Il est naturel de s'interroger sur l'existence de sources qui vérifient ces hypothèses, qui soient les plus générales possibles, mais qui soient aussi susceptibles d'un traitement "analytique". Nous avons décidé de faire l'analyse dans le cas où la source émettrice est une source dynamique, c'est-à-dire une source créée par un système dynamique, dont nous avons déjà parlé. Ce modèle représente aussi le compromis intéressant que nous recherchions, car ces sources vérifient les hypothèses recherchées 8, 9 et 10. Nos résultats s'appliquent donc à un grand nombre de modèles de bases de données.

Publications liées à la thèse. En arithmétique, les résultats des chapitres 1 à 4 sur l'algorithme d'Euclide classique ont été obtenus en collaboration avec Brigitte Vallée (GREYC, Caen). Une version courte a été acceptée et publiée à la conférence internationale LATIN'06 [LV06b]. Les mêmes résultats font aujourd'hui l'objet d'un article invité à la revue Algorithmica [LV06a]. L'analyse des algorithmes d'Euclide rapides a été effectuée en collaboration avec Benoît Daireaux (GREYC, Caen), Véronique Maume-Deschamps (Institut Mathématiques de Bourgogne, Dijon) et Brigitte Vallée. Une version courte est actuellement soumise à la conférence internationale ESA'06 [DLMDV06]. Le chapitre sur les calculs de constantes a fait l'objet d'un article dans les actes de la conférence internationale ANALCO'04, satellite de la conférence SODA [Lho04].

En fouille de données, l'analyse du nombre de motifs fréquents et fermés a été effectuée en collaboration avec François Rioult et Arnaud Soulet du GREYC. Ce travail a fait l'objet de deux publications : une publication à la conférence française sur l'apprentissage CAP'05 [LRS05a] et une autre publication à la conférence internationale sur la fouille de données ICDM'05 [LRS05a].

Structure de la thèse. Le manuscrit s'organise autour des deux principaux thèmes algorithmiques et une partie de la thèse est consacrée à chacun d'eux : l'arithmétique (partie A) et la fouille de données (Partie B).

La partie A rassemble les chapitres 1 à 5. Le chapitre 1 présente les algorithmes d'Euclide (classiques, étendus, rapides) et décrit les principaux paramètres qui seront analysés ici : complexités binaires, taille des restes, etc. On y annonce les principaux résultats. Le chapitre 2 présente l'analyse complète des algorithmes d'Euclide sur les polynômes, en adoptant les méthodes de combinatoire analytique. Le chapitre 3 décrit les principales étapes de l'analyse dynamique, utilisée pour obtenir les résultats dans le cas des entiers. Finalement, le chapitre 5 décrit les résultats de calcul de constantes.

La partie B rassemble les chapitres 6 à 8. Le chapitre 6 est une introduction générale à la fouille de données, puis le chapitre 7 présente la modélisation par les sources (avec leurs hypothèses associées) et décrit les principaux résultats. Finalement, le chapitre 8 montre que nos résultats s'appliquent dans le modèle général des bases de données associées à des sources dynamiques.

Dans tous les cas, nos résultats originaux sont numérotés (avec des nombres), alors que les autres résultats utilisés (mais dus à d'autres auteurs) sont numérotés par des des lettres.

Partie A :
Analyse des algorithmes d'Euclide

Introduction de la Partie A

Selon Knuth, *l'algorithme d'Euclide est le grand-père de tous les algorithmes puisque c'est le plus ancien algorithme non-trivial ayant survécu jusqu'à ce jour.* L'algorithme du pgcd est une brique de base essentielle dans les systèmes de calcul formel. Des expériences menées par Jebelean montrent que le calcul de pgcd représente 60% du temps total pour certaines applications sur de grands entiers, et que ce chiffre s'élève à 80% dans les calcul de bases de Gröbner. L'algorithme d'Euclide est aussi indispensable en arithmétique modulaire, et est ainsi très présent en cryptographie à clé publique.

Les algorithmes d'Euclide. Décrit autour de -300 dans "Les Éléments", l'algorithme d'Euclide admet maintenant plusieurs versions, y compris sur les polynômes. Les premières (sur les entiers) utilisent toutes une succession de divisions de la forme $u = qv + \epsilon 2^k r$ où les couples (u, v) sont remplacés après les divisions par les couples (v, r), le dernier reste non nul r étant le pgcd. Parmi ces algorithmes, on trouve les algorithmes d'Euclide Classique, Centré, Par-excès ou α-euclidiens [Nak81, BDV02] qui imposent des conditions sur le reste r. Les algorithmes Pair ou Impair imposent des conditions de parité sur les quotients q. Les algorithmes Binaire de Stein [Ste67, Knu98a] et Plus-Moins de Brent et Kung [BK85] utilisent une stratégie qui élimine les bits de poids faibles. Finalement, il existe des algorithmes de pgcd qui sont uniquement fondés sur des décisions sur les bits de poids faible [SZ04]. Cette liste est bien entendu non exhaustive. Sur les polynômes, il n'existe de fait qu'une version de l'algorithme d'Euclide classique qui élimine les monômes de plus grand degré, puisque la version opérant sur les monômes de plus petits degré est complètement symétrique de la première. En utilisant une idée de Lehmer [Leh38] [qui remplace de grandes divisions d'entiers par de grandes multiplications et des petites divisions], Knuth [Knu71] décrit le premier algorithme de type diviser pour régner, amélioré ensuite par Schönhage [Sch71]. Stehlé et Zimmermann [SZ04] utilisent la même démarche pour obtenir une version sur les "bits de poids faibles". Ces algorithmes profitent de l'efficacité des multiplications rapides (FFT ou multiplication de Karatsuba) pour atteindre une meilleure complexité que les algorithmes précédents. Il existe également une version de ces algorithmes "rapides" sur les polynômes.

Le cadre de l'étude et les résultats. Dans cette partie, nous étudions deux algorithmes d'Euclide : l'algorithme d'Euclide classique (à la fois sur les polynômes et les entiers, et avec ses deux versions, standard et étendu), et l'algorithme rapide de Knuth-Schönhage, uniquement sur les entiers[2]. Nous souhaitons étudier le paramètre le plus important de ces algorithmes, la complexité binaire. La complexité binaire (ou complexité en bits) est le nombre total d'opérations sur les bits nécessaires à l'algorithme pour obtenir le résultat. C'est aussi le paramètre le plus précis pour décrire la complexité d'un algorithme arithmétique. Nous analysons également un autre paramètre important dans la compréhension de l'algorithme d'Euclide classique : la taille du reste à une fraction de l'exécution. Ce paramètre décrit la taille des objets manipulés à un

[2] l'analyse sur les polynômes ne pose pas de difficultés particulières. Seul le temps nous a manqué.

instant donné de l'exécution de l'algorithme d'Euclide classique, et il a un intérêt tout particulier dans l'analyse de l'algorithme de Lehmer-Euclide [Leh38, DV04].

Organisation de la Partie A. Cette partie est organisée en chapitres, de la manière suivante.

Chapitre 1. Nous y présentons les algorithmes d'Euclide classiques (standard et étendus) à la fois sur les entiers et sur les polynômes. L'algorithme de Knuth-Schönhage est également décrit ainsi que la variante étudiée. L'analyse de cet algorithme est fondée sur une décomposition de l'algorithme classique en diverses "phases", regroupant des fractions d'exécution et correspondant à ce nous appelons des algorithmes interrompus. Ensuite, nous présentons, de manière uniforme sur les polynômes et les entiers, les principaux paramètres étudiés : complexité binaire standard et étendue, taille des continuants. L'analyse de l'algorithme de Knuth-Schönhage se réduit essentiellement à l'analyse de paramètres sur les algorithmes interrompus, qui sont aussi introduits. Nous énonçons tous les résultats obtenus sur ces paramètres fondamentaux.

Chapitre 2. C'est un chapitre consacré à l'analyse sur les polynômes. Nous y analysons d'abord les deux complexités binaires (standard et étendue), et montrons qu'elles suivent une loi limite gaussienne d'espérance quadratique et de variance cubique en la taille des entrées (théorèmes 1 et 4). Nous utilisons les méthodes de combinatoire analytique, en nous appuyant sur un principe de décomposition, qui fait intervenir plusieurs familles de coûts : des coûts principaux, pour lesquels nous exhibons une loi limite gaussienne, et des coûts secondaires, pour lesquels nous obtenons un phénomène de concentration. Les coûts qui jouent un rôle principal sont de deux types [coûts additifs à croissance modérée et coût de type longueur de cheminement] et nous montrons donc leur comportement asymptotiquement gaussien (théorèmes 9 et 11). Les coûts secondaires regroupent des coûts assez divers, dont nous analysons la variance, afin de montrer qu'elle est d'un ordre de grandeur inférieur à celle des coûts principaux (théorème 10). Nous obtenons aussi une loi limite gaussienne pour la taille du reste à une fraction (fixe) de l'exécution (théorème 5).

Chapitre 3. Ce chapitre est consacré à l'analyse sur les entiers. C'est un chapitre long qui décrit aussi la structure générale d'une analyse dynamique. Nous y présentons donc d'abord les objets dynamiques : système dynamique des fractions continues, et opérateurs de transfert. Mais nous suivons aussi la même démarche générale que sur les polynômes et utilisons un principe de décomposition, avec des coûts principaux et secondaires. Les décompositions obtenues font apparaître, comme pour les polynômes, des coûts additifs à croissance modérée (dont la loi limite gaussienne a déjà été obtenue par Baladi et Vallée). Les coûts secondaires sont de même type que sur les polynômes et nous montrons sans difficulté, avec des outils dynamiques cette fois, qu'ils sont concentrés. Nous obtenons donc deux lois limites gaussiennes pour la complexité binaire étendue et pour la taille du reste à une fraction (fixe) de l'exécution (théorèmes 3 et 5). Nous échouons dans l'obtention d'une loi limite gaussienne pour la complexité binaire standard, mais nous obtenons des résultats sur la variance de ce coût (théorème 2), résultats complétés par une conjecture. L'analyse de la complexité binaire de l'algorithme de Knuth-Schönhage est fondée sur l'analyse du comportement probabiliste fin des algorithmes interrompus. Ce comportement probabiliste est relié à un phénomène de régularité de l'algorithme classique : quelle est la taille du reste à une fraction de l'exécution, quand cette fraction tend vers 0 avec la taille des données ? [cette fraction δ n'est donc plus fixe, comme dans l'étude précédente]. Nous attachons aussi un paramètre à cette étude ainsi que la série génératrice correspondante.

Finalement, nous associons à chaque paramètre une série génératrice, puis, grâce à la structure de l'algorithme, nous obtenons une expression alternative de ces séries génératrices en termes d'opérateurs de transfert. Les propriétés analytiques de ces opérateurs seront démontrées au chapitre suivant, mais nous expliquons dans le présent chapitre comment ces propriétés peuvent

être "transférées" sur les coefficients de ces séries. Suivant la démarche de Baladi et Vallée, nous mettons d'abord en évidence l'importance d'une propriété US vérifiée par ces séries [Existence d'une bande verticale avec un unique pôle et un bon comportement sur la droite verticale à gauche]. A partir de cette propriété US qui sera démontrée au prochain chapitre, nous pouvons extraire les coefficients en utilisant la formule de Perron, et obtenir les résultats cherchés.

Chapitre 4. Ce chapitre est un chapitre d'analyse fonctionnelle, où nous obtenons les propriétés analytiques des opérateurs, qui ont déjà été utilisés dans le chapitre précédent. Trois opérateurs jouent ici un rôle essentiel, puisque ce sont eux qui interviennent dans le chapitre précédent, en relation étroite avec les séries génératrices. Chacun de ces opérateurs fait intervenir de manière simple l'opérateur de transfert du système dynamique. Le premier opérateur en est le "vrai" quasi-inverse, tandis que les deux autres opérateurs en sont des perturbations. Pour ces trois opérateurs, nous exhibons une propriété de type US, [existence d'une bande verticale, avec un unique pôle à l'intérieur, combinée à une borne sur la droite verticale à gauche de la bande]. La propriété US pour le quasi-inverse a déjà été prouvée par Baladi et Vallée [BV04], et nous cherchons à étendre leur démarche, pour l'appliquer à nos pseudo-quasi inverses. Cette démarche générale s'étend bien aux pseudo-quasi-inverses, loin de l'axe réel. Mais c'est la zone autour de l'axe réel qui est la plus délicate à traiter, à cause d'une accumulation possible de pôles à laquelle il faut faire attention. C'est cette accumulation possible qui limite d'ailleurs les théorèmes 5 et 6 au cas δ rationnel, et qui limite les analyses de l'algorithme "rapide" au cas où les multiplications utilisées ne sont pas trop rapides.

Chapitre 5. Ce chapitre est consacré au calcul des constantes "spectrales" qui interviennent dans les analyses. Certaines de ces constantes admettent une formule close. En revanche, la constante du terme dominant de la variance de la complexité binaire étendue n'est pas explicite et s'exprime en fonction de la valeur propre de l'opérateur de transfert. Daudé, Flajolet et Vallée [DFV97] ont proposé un algorithme qui calcule des valeurs numériques approchées pour les valeurs propres de certains opérateurs de transfert. Ils ont constaté son bon comportement pratique, mais n'ont pas prouvé, ni sa validité, ni sa complexité. Dans ce chapitre, nous prouvons ces propriétés, dans le cas où l'opérateur de transfert est associé à un système dynamique dont les branches possèdent des propriétés fortes de contraction. Nous prouvons ainsi que beaucoup de constantes "spectrales" qui interviennent dans les analyses de type fraction continue sont calculables en temps polynomial : la constante de Gauss-Kuz'min-Wirsing, la constante de Hensley (constante du terme dominant de la variance du nombre d'étapes), et les dimensions de Hausdorff d'espaces de Cantor sur les fractions continues.

Chapitre 1

Algorithmes d'Euclide

Sommaire

1.1 Introduction

L'algorithme d'Euclide est souvent considéré comme le plus vieil algorithme au monde et il reste aujourd'hui un des algorithmes de base pour de nombreuses applications : simplification de calculs, cryptographie à clé publique, tests de primalité, factorisation de polynômes, ... L'algorithme d'Euclide admet maintenant plusieurs versions, y compris sur les polynômes. Citons par exemple les algorithmes Pair, Impair, Centré, α-euclidiens [Nak81, BDV02], l'algorithme Binaire de Stein [Ste67, Knu98a], l'algorithme Plus-Moins de Brent et Kung [BK85], ... Tous ces algorithmes ont une complexité binaire quadratique en la taille des entrées. En utilisant de manière récursive une idée de Lehmer [Leh38], qui remplace de grandes divisions d'entiers par de grandes multiplications et des petites divisions, Knuth [Knu71] propose le premier algorithme de type diviser pour régner, amélioré la même année par Schönhage [Sch71]. Avec une multiplication

rapide comme la FFT, un algorithme quasi-linéaire est obtenu et constitue encore aujourd'hui, l'algorithme le plus rapide (pour de très grands entiers). Depuis, Stehlé et Zimmermann [SZ04] ont proposé une version similaire mais sur les "bits de poids faibles".

Les analyses des algorithmes euclidiens ont tout d'abord concerné le nombre de divisions nécessaires pour trouver le pgcd. Pour les polynômes, les auteurs de [KK88] ont démontré que le nombre de divisions suit une loi binomiale. Pour les entiers, l'algorithme d'Euclide a été analysé dans le pire des cas par Lamé [Lam45] autour de 1850 et en moyenne par Heilbronn [Hei69] et Dixon [Dix70] autour de 1970. Finalement la loi limite gaussienne du nombre de divisions a été démontrée par Hensley [Hen94] en 1994. Le même paramètre pour les algorithmes Centré, Soustractif et Par-excès a respectivement été étudié (en moyenne) par Rieger [Rie78], Knuth et Yao [YK75] et Vardi. L'algorithme Binaire a quant à lui été analysé par Brent [Bre76] sous certaines conjectures et par Vallée sans les conjectures [Val98a]. Dans ces analyses (exceptées celles de Brent et Vallée), les méthodes sont très diverses et il n'est pas clair qu'elles s'adaptent à d'autres paramètres ou algorithmes. Par exemple et pour un même résultat, Dixon a opté pour une méthode probabiliste alors que celle de Heilbronn est de nature combinatoire.

Depuis une dizaine d'années, le groupe de Caen a développé une méthode générale, appelée analyse dynamique, qui n'est spécifique ni aux algorithmes, ni au paramètre "nombre de divisions". Les premières analyses dynamiques d'algorithmes euclidiens (sur les entiers) ont été des analyses en moyenne sur des paramètres comme le nombre de divisions, le nombre de quotients prenant une certaine valeur, la complexité binaire, les continuants, le nombre total de bits pour coder tous les quotients, etc. En 2002, une percée significative a été réalisée avec la première analyse dynamique en distribution des algorithmes Standard, Centré et Impair pour des coûts dits additifs et à croissance modérée. En particulier, Baladi et Vallée [BV05, BV04] ont montré que ces coûts satisfont une loi limite gaussienne. Le nombre de divisions, le nombre de quotients prenant une certaine valeur, le nombre total de bits pour coder tous les quotients sont des coûts à croissance modérée mais pas la complexité binaire.

La complexité binaire d'un algorithme est le nombre d'opérations sur les bits effectuées par l'algorithme pour obtenir le résultat. Il est très difficile de calculer exactement la complexité binaire car la manière d'implémenter l'algorithme et la machine utilisée influencent énormément cette complexité. Des simplifications sont alors faites pour les analyses mais les paramètres obtenus restent ceux qui décrivent le mieux la complexité de l'algorithme.
Dans [Val00a, Val03], la complexité binaire moyenne de plusieurs algorithmes euclidiens sur les entiers est analysée. Cette analyse conduit à une classification des algorithmes en deux catégories. Tout d'abord les algorithmes cubiques qui admettent une complexité binaire moyenne d'ordre cubique en la taille des entrées. Les algorithmes Pair, Par-excès ou Soustractif sont des algorithmes cubiques. Les *algorithmes quadratiques* ont une complexité binaire moyenne d'ordre quadratique en la taille des entrées. Les algorithmes d'Euclide Standard, Centré ou Impair sont quadratiques ainsi que l'algorithme Binaire. Nous considérons une troisième classe dite des algorithmes *rapides*. Elle est uniquement composée des algorithmes de type diviser pour régner de Knuth-Schönhage ou Stehlé-Zimmermann dont la complexité binaire est sous-quadratique.
La classification précédente est également valable pour des entrées polynomiales. Il est bien connu que l'algorithme d'Euclide standard sur les polynômes a une complexité dans le pire des cas quadratique et il est admis[3] qu'il en est de même pour sa complexité moyenne. D'un autre coté, l'algorithme de Knuth-Schönhage sur les polynômes profite également de la FFT ou des autres multiplications rapides ce qui en fait un algorithme sous-quadratique (dans le pire des cas). Sur

[3] Même si cela n'a jamais été publié, cela a sûrement été donné un jour comme exercice à des étudiants de master !

les entiers, il existe une autre classification des algorithmes d'Euclide selon qu'ils agissent sur les bits de poids forts (algorithmes *Most Significant Bits*), sur les bits de poids faibles (algorithmes *Least Significant Bits*) ou les deux (algorithmes *Mixtes*). Nous renvoyons à [Val06] pour un tour d'horizon des algorithmes euclidiens.

Les analyses probabilistes précédentes sur la complexité binaire sont des analyses qui concernent uniquement les deux premiers moments (on parle d'analyse en moyenne). L'ultime étape, mais aussi la plus difficile, lorsque l'on pratique des analyses probabilistes est l'analyse en distribution.

Dans cette thèse, nous présentons la première analyse en distribution de la complexité binaire des algorithmes d'Euclide classiques et étendus [qui calculent les coefficients de Bezout] lorsqu'ils agissent sur les entiers ou sur les polynômes. Nous abordons aussi pour la première fois une analyse en moyenne d'un algorithme rapide. Cet algorithme est une version légèrement modifiée de l'algorithme de Knuth-Schönhage qui utilise des multiplications de type Karatsuba (la version originale utilise la FFT). L'analyse de l'algorithme modifié passe également par l'analyse d'algorithmes dit interrompus et qui correspondent à des fractions de l'exécution de l'algorithme classique.
(1) Tout d'abord, nous montrons que la complexité binaire de l'algorithme d'Euclide étendu [qui calcule les coefficients de Bezout] satisfait une loi limite gaussienne à la fois sur les polynômes et les entiers (théorèmes 3 et 4). La loi limite des complexités binaires vient donc s'ajouter à celles déjà connues [BV04] des coûts additifs à croissance modérée.
(2) Ensuite, nous obtenons que la complexité binaire de l'algorithme d'Euclide classique *sur les polynômes* satisfait aussi une loi limite gaussienne (théorème 1). Toutefois, nous ne sommes pas arrivés à cette conclusion sur les entiers mais nous améliorons l'ordre de grandeur connue pour la variance et nous donnons des termes d'erreur (théorème 2).
(3) Notre troisième résultat est une analyse fine et en moyenne de la complexité binaire d'une version de l'algorithme rapide de Knuth-Schönhage sur les entiers. Cette version diffère sur deux points avec l'originale. Tout d'abord le seuil pour stopper les appels récursifs n'est plus fixe mais dépend de la taille des entrées. Ensuite, la multiplication rapide n'est plus la FFT, mais toutes les multiplications dont la complexité est en $O(n^\alpha)$ pour des entiers de taille n avec $\alpha > 3/2$. Pour ces versions, nous montrons que la complexité moyenne est sous-quadratique (théorème 8). C'est la première fois qu'un algorithme euclidien rapide est analysé en moyenne. Daireaux, Maume-Deschamps et Vallée [DMDV05] ont analysé en moyenne l'algorithme de Stehlé-Zimmermann mais ils ont considéré la multiplication classique. L'analyse de l'algorithme de Knuth-Schönhage est essentiellement basée sur l'analyse d'algorithmes dits interrompus qui correspondent à des portions de l'algorithme d'Euclide. Le nombre d'étapes mais aussi la taille binaire des données de sorties de ces algorithmes sont présentées au théorème 6.
(4) Les algorithmes interrompus sont liés à l'évolution des restes successifs dans l'exécution de l'algorithme d'Euclide. Notre dernier résultat montre que la taille binaire du reste situé à une fraction (rationnelle) de l'exécution, aussi appelé continuant, suit une loi normale (théorème 5). Ceci est la version discrète d'un résultat bien connu de Philipp [Phi70] et amélioré par Vallée dans [Val97b].

Plan. Dans la première section, nous décrivons les algorithmes d'Euclide classiques et étendus, l'algorithme de Knuth-Schönhage et les algorithmes dits interrompus. Dans les trois parties qui suivent, nous définissons les paramètres importants des trois types d'algorithmes et nous en donnons les résultats principaux. Finalement, nous concluons.

1.2 Présentations des algorithmes

Afin de ne pas traiter séparément le cas des polynômes et le cas des entiers, la première section introduit des notations communes aux deux contextes. Avec ces notations, les paramètres admettent des définitions identiques dans les deux cas. Les sections suivantes présentent dans l'ordre les algorithmes d'Euclide classiques et étendus, les algorithmes interrompus et les algorithmes de Knuth-Schönhage.

1.2.1 Notations communes aux deux contextes

Nous étudions les algorithmes à la fois sur les entiers et sur les polynômes. L'ensemble des entiers naturels est noté $\mathbb{N}$ alors que l'anneau $\mathbb{F}_q[X]$ désigne l'ensemble des polynômes à une indéterminée sur l'unique corps $\mathbb{F}_q$ à q éléments. Dans la suite, $\mathbb{A}$ désigne soit $\mathbb{N}$, soit l'anneau $\mathbb{F}_q[X]$.
Le degré d'un polynôme non-nul u est noté $\deg u$. Pour $u = 0$, nous posons $\deg u = -\infty$. Sur les entiers positifs, nous considérons la valeur absolue usuelle $||v|| = v$ tandis que sur les polynômes, nous considérons la valeur absolue ultramétrique définie par $||v|| := q^{\deg v}$ et $||0|| = 0$. La taille d'un entier non-nul, notée $\ell(v)$, est la taille binaire de v ; elle est égale à $\lfloor \lg v \rfloor + 1$ où lg est le logarithme en base 2 et $\lfloor x \rfloor$ est la partie entière de x. La taille $\ell(v)$ d'un polynôme v non nul est le nombre de coefficients du polynôme, i.e., $1 + \deg v$. Un polynôme est dit unitaire si son coefficient dominant est 1. Dans le cas des entiers, les éléments unitaires sont par définition tous les éléments non-nuls de $\mathbb{N}$.
L'ensemble Ω des entrées possibles des algorithmes d'Euclide est

$$\Omega = \{(u, v) \in \mathbb{A}^2;\ 0 \leq ||v|| < ||u||,\ u \text{ unitaire}\}. \tag{1.1}$$

Pour tout élément (u, v) de Ω, la taille de la paire (u, v) est simplement la taille $\ell(u)$ de u et la norme de cette paire, notée $||(u, v)||$, est la norme $||u||$ de u.

1.2.2 Algorithmes d'Euclide classiques et étendus

Soit u et v deux éléments de $\mathbb{A}$ avec $||v|| < ||u||$ et $v \neq 0$. La division euclidienne de u par v est donnée par

$$u = m \cdot v + r \qquad \text{où} \qquad ||r|| < ||v|| \qquad \text{et} \qquad m \in \mathbb{A}. \tag{1.2}$$

L'entier m (resp. r) s'appelle le quotient (resp. le reste) de la division euclidienne. Cette division admet la représentation matricielle

$$\begin{pmatrix} v \\ u \end{pmatrix} = M_{[m]} \begin{pmatrix} r \\ v \end{pmatrix}, \qquad \text{où} \qquad M_{[m]} = \begin{pmatrix} 0 & 1 \\ 1 & m \end{pmatrix}.$$

L'algorithme d'Euclide effectue des divisions successives jusqu'à ce que le reste soit nul. Le dernier reste non nul est alors le pgcd. Si pour une entrée $(u, v) = (v_0, v_1)$, l'exécution est composée des p divisions suivantes,

$$v_0 = v_1 \cdot m_1 + v_2, \quad v_1 = v_2 \cdot m_2 + v_3, \quad \ldots, \quad v_{p-1} = v_p \cdot m_p + 0, \tag{1.3}$$

alors v_p est le pgcd de (v_0, v_1) et nous avons la relation matricielle,

$$\begin{pmatrix} v_1 \\ v_0 \end{pmatrix} = M_{(i)} \begin{pmatrix} v_{i+1} \\ v_i \end{pmatrix} \qquad \text{avec} \qquad M_{(i)} := M_{[m_1]} M_{[m_2]} \ldots M_{[m_i]}. \tag{1.4}$$

Dans la suite, nous considérons une *tranche* de l'algorithme d'Euclide, entre les indices i et j, appelée algorithme d'Euclide interrompu $\mathcal{E}_{(i,j)}$, qui commence avec la paire (a_i, a_{i+1}) en entrée et calcule la suite de divisions (1.3) entre les étapes i et $j-1$. La sortie de cet algorithme est composée de la paire (a_j, a_{j+1}) et de la matrice $M_{(i,j)}$ correspondant à la tranche d'exécution,

$$M_{(i,j)} := M_{[m_{i+1}]} M_{[m_2]} \dots M_{[m_j]}. \tag{1.5}$$

La taille binaire de la matrice $\mathcal{M}_{(i,j)}$ est notée $\ell_{(i,j)}$ et satisfait

$$\ell(v_{i+1}) - \ell(v_j) - 2 \le \ell_{(i,j)} \le \ell(v_i) - \ell(v_{j+1}) + 2. \tag{1.6}$$

Comme toutes les matrices $M_{[m]}$ sont inversibles, les matrices $M_{(i)}$ le sont aussi et admettent comme inverse

$$M_{(i)}^{-1} = \begin{pmatrix} a_{i+1} & b_{i+1} \\ a_i & b_i \end{pmatrix}, \quad \text{avec} \quad \begin{cases} a_{i+1} = a_{i-1} - m_i \cdot a_i, & a_0 = 0, \quad a_1 = 1, \\ b_{i+1} = b_{i-1} - m_i \cdot b_i, & b_0 = 1, \quad b_1 = 0. \end{cases} \tag{1.7}$$

En particulier pour $i = p$, les coefficients $a_p = a$ et $b_p = b$ sont appelés les coefficients de Bezout et satisfont la formule bien connue,

$$a \cdot v_1 + b \cdot v_0 = \mathrm{pgcd}(v_0, v_1).$$

L'algorithme d'Euclide étendu calcule à la fois le pgcd et le coefficient de Bezout a. Pour cela, il effectue les calculs suivants en plus des divisions,

$$a_0 = 0, \quad a_1 = 1, \qquad a_{i+1} = a_{i-1} - m_i \cdot a_i, \quad \text{pour } i = 1 \dots p-1 \tag{1.8}$$

($a_p = a$). Lorsque v_1 et v_0 sont premiers entre eux, le coefficient a correspond à l'inverse modulaire de v_1 modulo v_0. L'algorithme d'Euclide étendu offre donc un moyen automatique de calculer des inverses modulaires ce qui est par exemple utile à la génération de clés publiques et privées du protocole RSA.

1.2.3 Algorithmes interrompus

Nous introduisons maintenant des algorithmes dits interrompus qui sont adaptés à la compréhension mais aussi à l'analyse des algorithmes de Knuth-Schönhage. Ces algorithmes ont été introduits la première fois dans [DV04] pour l'analyse de l'algorithme de Lehmer-Euclide [Leh38]. Pour deux indices i, j qui satisfont $0 \le i \le j \le p$, l'exécution de l'algorithme d'Euclide entre les étapes i et j correspond à l'algorithme interrompu $\mathcal{E}_{(i,j)}$. Il existe deux types d'indices (i, j), selon qu'ils dépendent directement du nombre d'étapes p, ou de la taille $n = \ell(v_0)$ de l'entrée. Nous notons Ω_n l'ensemble des entrées de taille n,

$$\Omega_n := \{(v_0, v_1); \quad 0 < v_1 < v_0, \quad \ell(v_0) = n\}.$$

Dans ce qui suit, γ, δ satisfont les conditions suivantes : $\gamma \in [0, 1], 0 < \delta < 1 - \gamma$.

Algorithme $\mathcal{E}_{[\gamma,\delta]}$. Sur une entrée (v_0, v_1) de Ω_n, l'algorithme $\mathcal{E}_{[\gamma,\delta]}$ commence à la k^e itération de l'algorithme d'Euclide, dès que la taille du reste v_k a diminué de $\gamma \cdot n$, i.e., $\ell(v_k) \le (1-\gamma) \cdot n$, et s'arrête lorsque la taille du reste v_i a encore diminué de $\delta \cdot n$, i.e., $\ell(v_i) \le (1-\gamma-\delta) \cdot n$. Autrement dit, nous avons $\mathcal{E}_{[\gamma,\delta]} = \mathcal{E}_{(i,j)}$ avec

$$i := \min\{k;\ \ell(v_k) \le n - \lfloor \gamma\, n \rfloor\}, \quad j := \min\{k;\ \ell(v_k) \le n - \lfloor \gamma\, n \rfloor - \lfloor \delta\, n \rfloor\}.$$

Algorithme $\mathcal{E}_\delta(u,v)$	Algorithme $\widehat{\mathcal{E}}_\delta(u,v)$
$n := \ell(u)$	$n := \ell(u)$
$i := 1$	$i := 1$
$v_0 = u,\ v_1 = v$	$v_0 = u,\ v_1 = v$
$M_0 = I$	$M_0 = I$
Tant que $\ell(v_i) > (1-\delta)\cdot n$	Tant que $\ell(v_i) > (1-\delta)\cdot n$
$m_i := \lfloor v_{i-1}/v_i \rfloor$	$m_i := \lfloor v_{i-1}/v_i \rfloor$
$v_{i+1} := v_{i-1} - m_i v_i$	$v_{i+1} := v_{i-1} - m_i v_i$
$M_i := M_{i-1}\cdot M_{[m_i]}$	$M_i := M_{i-1}\cdot M_{[m_i]}$
$i++$	$i++$
Retourner (v_{i-1}, v_i, M_{i-1})	Retourner $(v_{i-3}, v_{i-2}, M_{i-3})$

FIG. 1.1 – Algorithmes interrompus $\mathcal{E}_\delta$ et $\widehat{\mathcal{E}}_\delta$.

L'algorithme $\widehat{\mathcal{E}}_{[\gamma,\delta]}$ correspond au même algorithme sauf que les trois dernières étapes ne sont pas effectuées. $\mathcal{E}_{[\gamma,\delta]}$ et $\widehat{\mathcal{E}}_{[\gamma,\delta]}$ sont des algorithmes purement théoriques sauf si le point de départ est celui de l'algorithme d'Euclide, i.e. $\gamma = 0$. Dans ce cas, nous obtenons les algorithmes $\mathcal{E}_\delta$ et $\widehat{\mathcal{E}}_\delta$ décrits à la figure 1.1

Nous verrons que les algorithmes $\widehat{\mathcal{E}}_{[\gamma,\delta]}$ et $\widehat{\mathcal{E}}_{1/2}$ sont particulièrement bien adaptés pour décrire la procédure récursive de l'algorithme de Knuth-Schönhage.

Algorithme $\mathcal{E}_{<\gamma,\delta>}$. Il est difficile d'analyser directement les algorithmes interrompus $\mathcal{E}_{[\gamma,\delta]}$ car on ne sait ni à quelle étape ils commencent, ni à quelle étape ils terminent. C'est pourquoi, nous introduisons l'algorithme $\mathcal{E}_{<\gamma,\delta>}$. Sur une entrée (v_0, v_1) de Ω_n, l'algorithme $\mathcal{E}_{<\gamma,\delta>}$ commence à l'étape $\lfloor \gamma\cdot p \rfloor$ (où p est la variable aléatoire du nombre total d'étapes) et s'arrête $\lfloor \delta\cdot p \rfloor$ itérations plus tard. Autrement dit, $\mathcal{E}_{<\gamma,\delta>} = \mathcal{E}_{(i,j)}$ avec

$$i := \lfloor \gamma\, p \rfloor, \quad j := \lfloor \gamma\, p \rfloor + \lfloor \delta\, p \rfloor.$$

Dans notre analyse, nous allons montrer que les algorithmes $\mathcal{E}_{<\gamma,\delta>}$ et $\mathcal{E}_{[\gamma,\delta]}$ sont très proches l'un de l'autre d'un point de vue probabiliste. Pour démontrer cette similitude, nous utilisons un algorithme théorique intermédiaire $\mathcal{E}_{<\gamma,\delta]}$.

Algorithme $\mathcal{E}_{<\gamma,\delta]}$. Sur une entrée (v_0, v_1) de Ω_n, l'algorithme $\mathcal{E}_{<\gamma,\delta]}$ commence à l'étape $k = \lfloor \gamma\cdot P \rfloor$ et s'arrête lorsque la taille du reste $v_{\lfloor \gamma\cdot P\rfloor + i}$ a diminué de $\delta\cdot n$ par rapport à celle de v_k, i.e., $\ell(v_{\lfloor \gamma\cdot P\rfloor+i}) \le \ell(v_k) - \delta_n$. Les deux algorithmes $\mathcal{E}_{<\gamma,\delta]}$ et $\mathcal{E}_{(i,j)}$ vérifient $\mathcal{E}_{<\gamma,\delta]} = \mathcal{E}_{(i,j)}$ avec

$$i := \lfloor \gamma\, p \rfloor, \quad j := \min\{k;\ \ell(v_k) \le n - \lfloor \gamma\, n \rfloor - \lfloor \delta\, n \rfloor\}.$$

Distance entre les algorithmes interrompus. Nous souhaitons comparer les algorithmes interrompus $\mathcal{E}_{[\gamma,\delta]}$ [qui interviennent de manière naturelle pour décrire l'algorithme de Knuth-Schönhage] avec l'algorithme $\mathcal{E}_{<\gamma,\delta>}$ [qui est plus régulier puisque nous connaissons exactement son nombre d'étapes]. Nous introduisons alors une distance entre deux algorithmes interrompus $\mathcal{E}_{(i,j)}$ et $\mathcal{E}_{(i',j')}$ définie par la relation

$$d(\mathcal{E}_{(i,j)}, \mathcal{E}_{(i',j')}) = \max(|i-i'|, |j-j'|),$$

et nous allons montrer que la distance entre $\mathcal{E}_{[\gamma,\delta]}$ et $\mathcal{E}_{<\gamma,\delta>}$ est petite avec une très grande probabilité. Cela prouvera une grande régularité pour l'algorithme $\mathcal{E}_{[\gamma,\delta]}$.

1.2.4 Algorithme de Knuth-Schönhage $\mathcal{KS}$: version originale

Nous nous limitons ici au cas des entiers. L'algorithme de Knuth-Schönhage [Knu71, Sch71] utilise de manière récursive une idée de Lehmer [Leh38] qui est de remplacer de grandes divisions par de grandes multiplications et des petites divisions. La complexité d'une division est d'ordre $\Theta(n^2)$ sur des entiers de taille n alors qu'il existe des multiplications de complexité $O(n^\alpha)$ (Karatsuba ou Tom-Cook) avec $\alpha < 2$ ainsi que des multiplications quasi-linéaires comme la FFT (complexité en $O(n \log n \log\log n)$). Nous commençons cette section avec l'idée de Lehmer, puis nous introduisons l'algorithme de Knuth-Schönhage et sa fonction principale $\mathcal{HG}$ (pour H-Gcd ou Half-Gcd). Cependant, nous n'étudions pas la version originale de l'algorithme de Knuth-Schönhage, et la version analysée est présentée à la section 1.2.5.

1.2.4.1 Critère de Jebelean et procédé de Lehmer

Dans le procédé de Lehmer, les petits entiers utilisés pour faire les petites divisions correspondent aux grands entiers dont un certain nombre de bits de poids faible ont été supprimés. Ensuite, le critère de Jebelean relie les quotients issus des grands entiers aux quotients issus des entiers tronqués. Nous définissons tout d'abord la fonction de troncature.

Définition 1 *Considérons une paire (v_0, v_1) avec $v_0 > v_1 > 0$ et $n := \ell(v_0)$. Pour $m < n$, la paire $(\overline{v}_0, \overline{v}_1)$ définie par $\overline{v}_0 := \lfloor v_0 \cdot 2^{m-n} \rfloor, \overline{v}_1 := \lfloor v_1 \cdot 2^{m-n} \rfloor$ est notée $T_m(v_0, v_1)$.*
La fonction T_m supprime les $(n-m)$ bits de poids faible et est appelée la troncature d'ordre m. Notons que l'entier $\overline{v}_0$ vérifie $\ell(\overline{v}_0) = m$.

Le critère de Jebelean montre que, pour un ordre de troncature donné, les premiers quotients issus des grands entiers sont exactement les quotients issus des entiers tronqués.

Critère 1 (Jebelean) *Pour une paire (v_0, v_1) avec $v_0 > v_1 > 0$ et $n := \ell(v_0)$, considérons la paire $(\overline{v}_0, \overline{v}_1) = T_m(v_0, v_1)$, et la séquence de restes $(\overline{v}_i)$ de l'algorithme d'Euclide sur l'entrée $(\overline{v}_0, \overline{v}_1)$. Nous notons k le plus petit entier i pour lequel $\overline{v}_i$ satisfait $\ell(\overline{v}_i) \leq \lceil m/2 \rceil$. Alors la séquence des quotients m_i de l'algorithme d'Euclide sur (v_0, v_1) coïncide avec la séquence des quotients $\overline{m}_i$ de l'algorithme d'Euclide sur $(\overline{v}_0, \overline{v}_1)$ pour $i \leq k-3$.*

Pour une entrée (v_0, v_1), le critère de Jebelean dit qu'il existe un entier k tel que les $k-3$ premiers quotients issus de (v_0, v_1) sont les mêmes que les $k-3$ premiers quotients issus des entiers tronqués $T_m(v_0, v_1)$. Ces quotients étant égaux, la matrice $M_{(k-3)}$ produite par l'algorithme d'Euclide sur l'entrée (v_0, v_1) est aussi celle produite par les $k-3$ premières étapes de l'algorithme d'Euclide sur $T_m(v_0, v_1)$. Il est ainsi possible de calculer, à moindre coût et à partir des entiers tronqués, la matrice $M_{(k-3)}$ de l'algorithme sur les grands entiers. En effectuant le produit de $M_{(k-3)}^{-1}$ et (v_0, v_1) avec des multiplications rapides, une paire (v_{k-3}, v_{k-2}), de deux restes successifs issus de (v_0, v_1), est obtenue. L'algorithme de Lehmer-Euclide [Leh38] recommence ce processus avec la nouvelle paire (v_{k-3}, v_{k-2}) et continue jusqu'à un certain seuil où il bascule vers l'algorithme d'Euclide classique (cf. figure 1.2).

L'ordre de troncature utilisé par Lehmer est $n/2$ et chaque étape correspond en fait à une exécution de l'algorithme interrompu $\widehat{\mathcal{E}}_{1/2}$ sur les entiers tronqués. En pratique, l'algorithme $\mathcal{E}_{1/2}$ est d'abord utilisé et produit une matrice $\overline{M}_{(k)}$. Ensuite, l'algorithme recule de 3 étapes pour retrouver la matrice $\overline{M}_{(k-3)} = M_{(k-3)}$ et qui correspond à une matrice qui aurait été calculée sur les grands entiers. La matrice $\overline{M}_{(k)}$ transforme la paire $T_m(v_0, v_1)$, de taille m, en une paire

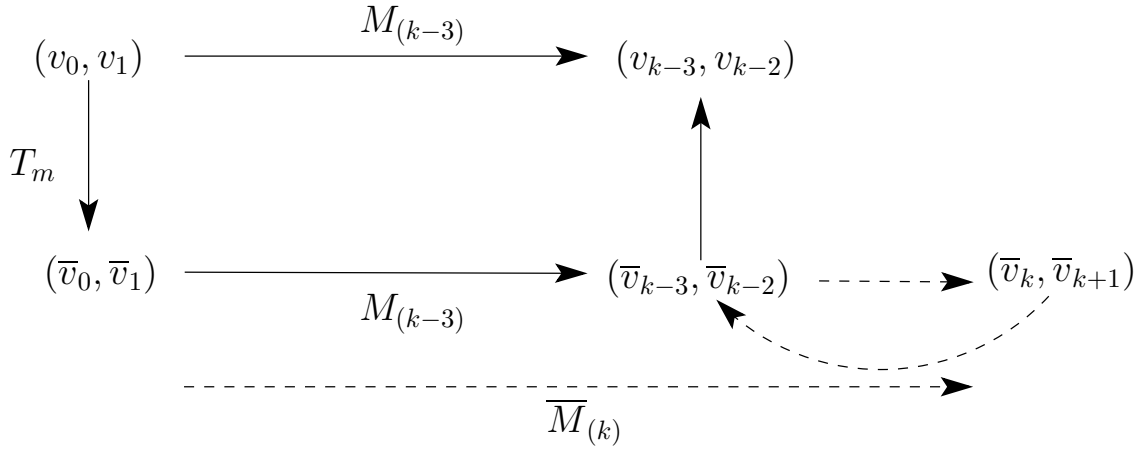


FIG. 1.2 – Une étape de l'algorithme de Lehmer-Euclide.

$(\overline{v}_k, \overline{v}_{k+1})$ de taille environ $m/2$ (la condition d'arrêt de $\mathcal{E}_{1/2}$ étant d'avoir perdu la moitié des bits initiaux). La taille binaire de $\overline{M}_{(k)}$ est alors environ $m/2$. Nous espérons (nous préciserons notre espoir plus tard) que le recul de trois étapes ne modifie que très peu la taille des entiers. Sous cette hypothèse, la matrice $M_{(k-3)}$ est aussi de taille environ $m/2$, et la paire (v_{k-3}, v_{k-2}) a une taille d'environ $n - (m/2) = 3n/4$. En admettant les "environs", le procédé de Lehmer permet de perdre $n/4$ bits en utilisant uniquement $n/2$ bits.

Contrairement au procédé de Lehmer, qui utilise explicitement l'algorithme $\widehat{\mathcal{E}}_{1/2}$, l'algorithme de Knuth-Schönhage utilise une fonction récursive $\mathcal{HG}$ dont le résultat est le même que celui retourné par $\widehat{\mathcal{E}}_{1/2}$ mais dont l'approche de type diviser-pour-régner est plus efficace.

1.2.4.2 Algorithme Half-Gcd $\mathcal{HG}$

La figure 1.3 donne le code de la fonction Half-Gcd $\mathcal{HG}$ et de l'algorithme de Knuth-Schönhage $\mathcal{KS}$. La fonction $\mathcal{HG}$ retourne le même résultat que la fonction $\widehat{\mathcal{E}}_{1/2}$ (d'où son nom) et nous décrivons maintenant ses étapes principales.

Lignes 1 à 7. Le premier appel récursif à la ligne 5 est fait sur des entiers de taille $n/2$ et retourne une matrice dont les coefficients sont de taille environ $n/4$ [le environ sera justifié plus tard]. Les entiers (u_1, v_1) à la ligne 6 ont alors une taille d'environ $3n/4$. En appliquant le critère de Jebelean, on montre que (u_1, v_1) sont des restes successifs dans la suite des restes issus de (u, v). La fonction $Adjust_1$ à la ligne 7 assure essentiellement la condition $\ell(u_1) < 3n/4$. Pour cela, des tests et certain nombre de divisions euclidiennes sont effectuées.

Jusque là, l'exécution de la fonction $\mathcal{HG}$ correspond à l'exécution de l'algorithme interrompu $\mathcal{E}_{1/4}$ sur l'entrée (u, v).

Lignes 8 à 11. Le second appel récursif à la ligne 9 utilise aussi des entiers de tailles environ $n/2$. La matrice M_2 est de taille environ $n/4$ et u_2 satisfait $\ell(u_2) \approx n/2$. Finalement, la fonction $Adjust_2$ vérifie si les conditions du critère de Jebelean sont vérifiées et fait ou défait certaines divisions euclidiennes.

La deuxième partie de la fonction $\mathcal{HG}$ correspond de fait à l'algorithme $\widehat{\mathcal{E}}_{[1/4,1/4]}$.

Pour une description précise de l'algorithme et des fonctions $Adjust$, nous renvoyons à [Yap96]. Pour une entrée de taille n, nous supposons que le coût des deux fonctions $Adjust$ est d'ordre au

Algorithme $\mathcal{HG}(u, v)$
  $n := \ell(u)$
  $S :=$seuil en fonction de n
  Retourner $\mathcal{HG}(u, v, S)$

Algorithme $\mathcal{HG}(u, v, S)$
1 $n := \ell(u)$
2 Si $n \leq S$ alors Retourner $\widehat{\mathcal{E}}_{1/2}(u, v)$
3 $m := \lfloor n/2 \rfloor$; $k = n - m$
4 $(\overline{u}, \overline{v}) := T_m(u, v)$
5 $(s_1, t_1, M_1) := \mathcal{HG}(\overline{u}, \overline{v}, S)$
6 $(u_1, v_1) = M_1^{-1}(u, v)$
7 $Adjust_1(u_1, v_1, M_1)$
8 $(\overline{u}_1, \overline{v}_1) := T_k(u_1, v_1)$
9 $(s_2, t_2, M_2) := \mathcal{HG}(\overline{u}_1, \overline{v}_1, S)$
10 $(u_2, v_2) = M_2^{-1}(u_1, v_1)$
11 $Adjust_2(u_2, v_2, M_2)$
12 Retourner $(u_2, v_2, M_1 \cdot M_2)$

Algorithme $\mathcal{KS}(u, v)$
  Tant que $\ell(u) > S$
    $n := \ell(u)$
    $m := \lfloor n/2 \rfloor$
    $(\overline{u}, \overline{v}) := T_m(u, v)$
    $(s, t, M) := \mathcal{HG}(\overline{u}, \overline{v})$
    $(u, v) := M^{-1}(u, v)$
  Fin Tant que
  Retourner $Eucl - Classique(u, v)$

FIG. 1.3 – Algorithme de Knuth-Schönhage.

plus $o(\mu(n))$, où $\mu(n)$ désigne le coût de la multiplication utilisée sur des entiers de taille n. Les étapes 6, 10 et 12 effectuent un certain nombre K de multiplications sur des entiers de taille au plus n. La complexité binaire $B(n)$ de l'algorithme $\mathcal{HG}$ sur des entrées de taille n satisfait alors

$$B(n) = 2B(\frac{n}{2}) + K \cdot \mu(n) + o(\mu(n)).$$

Si la FFT est utilisée, avec $\mu(n) = O(n \log n \log\log n)$, la complexité binaire totale est alors de l'ordre de $O(n \log^2 n \log\log n)$. Avec une multiplication de complexité $\mu(n) = \Theta(n^\alpha)$ avec $1 < \alpha < 2$ (par exemple Karatsuba ou Tom-Cook), le coût total est $B(n) = \Theta(n^\alpha)$.

Choix du seuil S. Dans la version originale de l'algorithme $\mathcal{KS}$, le seuil S est fixé expérimentalement et dépend de l'algorithme utilisé en-dessous de ce seuil. La version que nous proposons plus loin fait intervenir un seuil qui dépend de la taille initiale des entrées.

1.2.5 Versions paramétrées de $\mathcal{KS}$ et $\mathcal{HG}$

La version que nous étudions de l'algorithme de Knuth-Schönhage est différente sur deux points de la version originale.

La multiplication. La première différence se fait au niveau des étapes 6, 10 et 12. Ces étapes correspondent à la multiplication de deux matrices ou à la multiplication d'une matrice par un vecteur. Ces multiplications mettent en jeu de grands entiers et dans sa version originale, l'algorithme de Knuth-Schönhage utilise la transformée de Fourier rapide. Pour des entiers de taille n, la FFT a une complexité quasi-linéaire en $O(n \log n \log\log n)$. Les techniques d'analyse que nous utilisons ne nous ont pas permis d'aborder des multiplications plus rapides que celles en $O(n^{3/2})$. La multiplication de Karatsuba (complexité en $O(n^{\lg 3})$ avec $\lg 3 \approx 1.58$) entre dans le

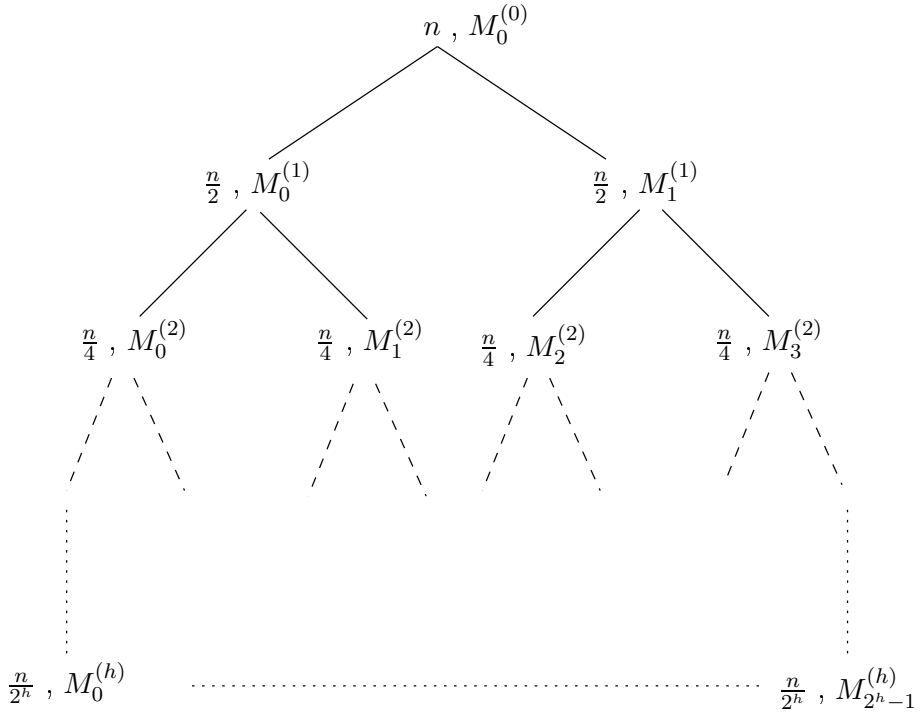


FIG. 1.4 – Arbre des appels récursifs de la fonction $\mathcal{HG}$ sur une entrée de taille binaire n. A chaque nœud, le premier élément est la taille de l'entrée "attendue" et le second est la matrice de sortie.

cadre tandis que la multiplication de Tom-Cook (complexité en $O(n^\alpha)$ avec $\alpha \approx 1.46$) n'y rentre plus.

Le seuil S. La version originale de $\mathcal{KS}$ utilise un seuil fixe S. Il est aussi possible de choisir un seuil S qui dépend de la taille n des entrées de manière à ce que le coût total de tous les algorithmes interrompus $\widehat{\mathcal{E}}_{1/2}$ en fin de récursion soit plus petit que le coût total de toutes les multiplications effectuées avant la fin de la récursion. La figure 1.4 représente les appels récursifs de la fonction $\mathcal{HG}$. Nous avons disposé sur chaque nœud la taille *attendue* des entrées. Pour un nœud de profondeur k, l'entrée est environ de taille $n/2^k$. Le seuil S vérifie aussi $n/2^h \approx S$ où h est la hauteur de l'arbre et le nombre de feuilles est d'environ 2^h. Comme l'algorithme interrompu $\widehat{\mathcal{E}}_{1/2}$ est d'ordre quadratique, le coût total des feuilles est de l'ordre de $2^h(n/2^h)^2$. Maintenant, à la profondeur i, l'algorithme $\mathcal{HG}$ effectuent un nombre fini de multiplications (lignes 6,10 et 12) sur des entiers de taille au plus $n/2^i$. Si ces multiplications ont une complexité de la forme $\mu(n) \sim An^\alpha$ [c'est cette situation qui nous intéressera par la suite], alors le coût total des produits est de l'ordre de

$$A\sum_{i=1}^{h-1} 2^i \mu(\frac{n}{2^i}) \approx An^\alpha \frac{2^{1-\alpha}}{2^{\alpha-1}-1}.$$

Le coût total des algorithmes interrompus et le coût total des multiplications sont du même ordre de grandeur dès que la hauteur h satisfait $h \sim (2-\alpha)\lg n$.

Algorithmes $\mathcal{HG}_\alpha$. Nous désignons par $\mathcal{HG}_\alpha$ l'algorithme $\mathcal{HG}$ pour lequel le niveau de récursion satisfait $h = \lfloor (2-\alpha+r)\lg n \rfloor$, avec un petit réel r défini plus tard. Ainsi, le coût aux feuilles devient négligeable devant le coût total de tous les nœuds et le coût total est alors de l'ordre

de $\Theta(n^\alpha)$. Notre objectif dans cette partie est de montrer que l'heuristique que nous venons d'utiliser est en moyenne exacte. L'algorithme de Knuth-Schönhage associé à la fonction $\mathcal{HG}_\alpha$ est noté $\mathcal{KS}_\alpha$.

1.3 Paramètres des algorithmes d'Euclide classiques et étendus

Dans cette section nous introduisons la complexité binaire des algorithmes d'Euclide classiques et étendus ainsi que les continuants à une fraction de l'exécution. Pour certains de ces paramètres, nous montrons qu'ils admettent une loi limite gaussienne. La première section décrit cette notion et le modèle probabiliste choisi. Ensuite, nous traitons séparément les complexités binaires classiques (sur les polynômes et sur les entiers), les complexités binaires étendues et la taille des continuants à une fraction de l'exécution. Finalement, nous présentons les différents paramètres des algorithmes interrompus et nous montrons que leur analyse est suffisante pour établir la complexité moyenne des algorithmes $\mathcal{KS}_\alpha$.

1.3.1 Modèle probabiliste et loi gaussienne

Que cela soit sur les polynômes ou sur les entiers, Ω_n désigne l'ensemble des couples $(u, v) \in \Omega$ tel que $\ell(u) = n$,

$$\Omega_n := \{(u, v) \in \Omega;\ \ell(u) = n\}.$$

Chaque ensemble Ω_n est muni de la distribution uniforme notée $\mathbb{P}_n$. L'espérance et la variance d'un paramètre R sur Ω_n sont respectivement notés $\mathrm{E}_n[R]$ et $\mathrm{V}_n[R]$. Dans cette thèse, nous montrons qu'un certains nombre de paramètres satisfont une loi limite gaussienne.

Définition 2 *Un paramètre R sur Ω satisfait une loi limite gaussienne s'il existe trois suites $(e_n)_n$, $(v_n)_n$ et $(r_n)_n$ telles que sur chaque Ω_n ont ait*

$$\mathbb{P}_n\left[a \leq \frac{R - e_n}{\sqrt{v_n}} \leq b\right] = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt + O(r_n), \qquad (a, b) \in \mathbb{R}^2, \quad r_n \to 0.$$

La suite $(r_n)_n$ est appelée la vitesse de convergence de R vers la loi limite et est également notée $r_n[R]$. Les suites e_n et v_n sont équivalentes à l'espérance et la variance de R sur Ω_n,

$$e_n \sim \mathrm{E}_n[R], \qquad v_n \sim \mathrm{V}_n[R].$$

Les lois limites gaussiennes se rencontrent souvent dans les analyses probabilistes. En particulier, lorsqu'un paramètre est la somme d'expériences aléatoires identiques et indépendantes (ex : nombre de piles obtenus après n lancés d'une pièce), il admet une loi limite gaussienne. L'algorithme d'Euclide classique répète plusieurs fois des divisions euclidiennes mais celles-ci ne sont pas indépendantes. La situation est alors plus compliquée et même si des loi limites sont attendues, elles restent difficiles à démontrer.

Dans le cas des entiers, Baladi et Vallée [BV05, BV04] ont montré que les coûts additifs à croissance modérée satisfont une loi limite gaussienne. Ces paramètres sont associés à un coût élémentaire sur les quotients et sont de la forme

$$C(u, v) := \sum_{i=1}^{p} c(m_i).$$

Lorsque $c(m) = O(\log m)$, les coûts c et C sont dits à croissance modérée. Cette classe de coûts contient de nombreux paramètres naturels comme le nombre d'étapes p (pour $c = 1$), le nombre d'occurrences d'un quotient m_0 (pour $c(m) = [[m = m_0]]$), la taille totale pour coder tous les quotients, ...Les résultats de [BV05, BV04] s'expriment par exemple de la manière suivante.

Théorème BV (Baladi-Vallée) *Considérons un coût additif C relatif à un coût élémentaire c à croissance modérée.*
(i) Sur l'ensemble Ω_n des entrées de taille n, le coût C suit asymptotiquement une loi limite gaussienne dont l'espérance, la variance et la vitesse de convergence sont

$$\mathrm{E}_n[C] = \mu(c) \cdot n + \mu_1(c) + O(2^{-n\gamma}), \quad \mathrm{V}_n[C] = \rho(c) \cdot n + \rho_1(c) + O(2^{-n\gamma}), \quad r_n[C] = O(n^{-1/2}).$$

Ici γ est une constante strictement positive qui ne dépend pas de c.
(ii) Les constantes $\mu(c)$ et $\rho(c)$ sont bien définies et sont liées à la valeur propre dominante $\lambda(s, w)$ de l'opérateur de transfert

$$\mathbf{G}_{s,w,[c]}[f](x) = \sum_{m \geq 1} \frac{e^{wc(m)}}{(m+x)^{2s}} f(\frac{1}{m+x}) \tag{1.9}$$

agissant sur $C^1([0,1])$. Plus précisément,

$$\mu(c) = \frac{2\log 2}{|\lambda'_s(1,0)|}\lambda'_w(1,0),$$

$$\rho(c) = \frac{2\log 2}{|\lambda'_s(1,0)|^3}[\lambda'^2_s(1,0) \cdot \lambda''_{w^2}(1,0) - 2\lambda'_w(1,0) \cdot \lambda'_s(1,0) \cdot \lambda''_{sw}(1,0) + \lambda'^2_w(1,0) \cdot \lambda''_{s^2}(1,0)].$$

La calculabilité des constantes $\mu(c)$ et $\rho(c)$ est différente. Très souvent, la première dérivée de $\lambda(s, w)$ est connue en $(1, 0)$. Par exemple, lorsque c est la taille binaire ℓ,

$$\lambda'_s(1,0) = -\frac{\pi^2}{6\log 2}, \quad \lambda'_w(1,0) = \frac{2}{\log 2}\log\prod_{i=0}^{\infty}(1+\frac{1}{2^i}), \quad \mu(l) = \frac{12\log 2}{\pi^2}\log\prod_{i=0}^{\infty}(1+\frac{1}{2^i}).$$

D'un autre coté, il n'existe pas encore de formule close de $\rho(\ell)$ mais au chapitre 5, nous présentons une méthode générale pour approcher sa valeur et nous avons obtenu l'estimation

$$\rho(\ell) \approx 0.091.$$

Lors de l'analyse des algorithmes d'Euclide sur les polynômes (Chapitre 2), l'équivalent du théorème BV sur les polynômes est démontré (théorème 9).

1.3.2 Complexité binaire classique

Chaque itération de l'algorithme classique se compose d'une division euclidienne de la forme $u = m \cdot v + r$ suivie d'un échange. La complexité binaire naïve pour effectuer la division est $\ell(v)\ell(m)$. Le coût de l'échange se fait en temps constant (affectation entre pointeurs) et peut être considéré comme négligeable.
La complexité binaire B de l'algorithme d'Euclide classique est la somme des coûts binaires de chaque division et avec les notations de la section 1.2.2, elle est donnée par formule,

$$B(v_0, v_1) := \sum_{k=1}^{p} \ell(v_k)\ell(m_k). \tag{1.10}$$

Dans le pire des cas, la complexité binaire de l'algorithme d'Euclide sur les polynômes est d'ordre quadratique et nous caractérisons maintenant sa loi limite.

Théorème 1 (Complexité binaire classique (polynômes)) *La complexité binaire B de l'algorithme d'Euclide classique sur l'ensemble Ω_n suit une loi limite gaussienne dont l'espérance, la variance et la vitesse de convergence satisfont*

$$\mathrm{E}_n[B] = \frac{2q-1}{2q} n^2 + O(n) \qquad \mathrm{V}_n[B] = \frac{q-1}{3q^2} n^3 + O(n^2), \qquad r_n[B] = O(n^{-1/2}).$$

En utilisant les techniques de séries génératrices présentées au chapitre 2, nous obtenons les développements précis de l'espérance et de la variance de B sur Ω_n,

$$\mathrm{E}_n[B] = \frac{2q-1}{2q} n^2 - \frac{2q^2-q+1}{2(q-1)q} n + \frac{q}{(q-1)^2} + O(n^2 q^{-n}),$$

$$\mathrm{V}_n[B] = \frac{q-1}{3q^2} n^3 + \frac{q^2+2q-1}{2q^2(q-1)} n^2 + \frac{q^4+2q^3+12q^2-4q+1}{6(q-1)^3 q^2} n - \frac{q(q^2+5q+1)}{(q-1)^4} + O(n^4 q^{-n}).$$

Dans [Val03], Vallée a développé une méthodologie générale pour effectuer l'analyse dynamique en moyenne de toute une classe de paramètres dans le cas des entiers. La complexité binaire B fait partie de cette classe et Akhavi et Vallée [AV00] ont établit (avec cette méthodologie) que l'espérance de B est en moyenne quadratique en n sur Ω_n et que le second moment est asymptotiquement équivalent au carré de l'espérance. Ceci implique pour la variance

$$\mathrm{V}_n[B] = \mathrm{E}_n[B^2] - \mathrm{E}_n[B]^2 = o(n^4).$$

Nous ne sommes pas parvenu à démontrer la loi limite gaussienne du paramètre B dans le cas des entiers, mais nous améliorons les résultats précédents.

Théorème 2 (Complexité binaire classique (entiers)) *(i) Sur l'ensemble Ω_n des entrées de taille n, l'espérance et la variance de B satisfont*

$$\mathrm{E}_n[B] = \frac{1}{2}\mu(\ell) \cdot n^2 \left[1 + O(n^{-1})\right] \qquad \mathrm{V}_n[B] = \rho_0 \cdot n^3 \left[1 + O(n^{-1})\right]$$

avec ρ_0 une constante positive ou nulle liée au spectre de l'opérateur de transfert pondéré $\mathbf{G}_{s,w,[c]}$ (voir formule 1.9).
(ii) Sous une conjecture (C), la constante ρ_0 est non nulle et est reliée à la constante $\rho(\ell)$ du théorèmes BV via l'égalité $\rho_0 = \rho(\ell)/3$.

D'une manière générale, les algorithmes d'Euclide sur les polynômes sont plus simples à étudier que leurs homologues sur les entiers. Une des raisons est que sur les polynômes, tout se passe comme si les divisions successives de l'algorithme d'Euclide étaient *indépendantes* alors que sur les entiers, ces divisions sont très corrélées. Nous expliquons l'hypothèse $\rho_0 = \rho(\ell)/3$ à la section 1.3.4.

1.3.3 Complexité binaire des algorithmes interrompus

L'algorithme d'Euclide interrompu $\mathcal{E}_{(i,j)}$ correspond à l'exécution de l'algorithme d'Euclide classique entre les étapes i et j. Comme pour l'algorithme classique, la complexité binaire $B_{(i,j)}$ de l'algorithme d'Euclide interrompu $\mathcal{E}_{(i,j)}$ est définie par

$$B_{(i,j)}(v_0, v_1) := \sum_{k=i+1}^{j} \ell(v_k)\ell(m_k). \tag{1.11}$$

Avec les relations $\ell(v_k) \leq \ell(v_{k+1})$ et $m_k \leq v_{k-1}/v_k$, la complexité binaire $B_{(i,j)}$ satisfait

$$B_{(i,j)} \leq \ell(v_{i+1}) \cdot [|j-i+1| + \ell(v_i) - \ell(v_j)]. \tag{1.12}$$

De fait, les principaux paramètres des algorithmes interrompus $\mathcal{E}_{(i,j)}$, i.e. les complexités binaires $B_{(i,j)}$ et les tailles $\ell_{(i,j)}$ des matrices $M_{(i,j)}$ (voir formule 1.6), font essentiellement intervenir les différences $\ell(A_t) - \ell(A_u)$ avec $t \in \{i, i+1\}$ et $u \in \{j, j+1\}$. C'est pourquoi, l'étude de l'évolution de la taille des restes (voir section 1.3.5) sera très importante dans cette thèse.

1.3.4 Complexité binaire étendue

En plus du pgcd, l'algorithme d'Euclide étendu calcule le coefficient de Bezout a en évaluant la séquence 1.8. Le surcoût associé au calcul de chaque a_i provient essentiellement de la multiplication de a_i par le quotient m_i dont la complexité binaire naïve est $\ell(a_i)\ell(m_i)$. La complexité binaire E de l'algorithme étendu satisfait alors,

$$E(v_0, v_1) := \ell(v_p)\ell(m_p) + \sum_{i=1}^{p-1} (\ell(v_i) + \ell(a_i))\ell(m_i). \tag{1.13}$$

La complexité binaire étendue E est plus régulière que la complexité binaire classique B. En effet, que cela soit sur les polynômes ou sur les entiers, la somme $\ell(v_i) + \ell(a_i)$ est quasi-constante et égale à la taille n de l'entrée. Ainsi, la complexité binaire étendue est quasiment de la forme $n \cdot C$ où C est un coût additif à croissance modérée. En utilisant la loi limite gaussienne de C, nous obtenons la loi limite gaussienne de B sur les polynômes et sur les entiers.

Théorème 3 (Complexité binaire étendue (entiers)) *Sur l'ensemble Ω_n des entiers de taille n, la complexité binaire E de l'algorithme d'Euclide étendu suit une loi limite gaussienne, d'espérance, de variance et de vitesse de convergence*

$$\mathrm{E}_n[E] = \mu(l) \cdot n^2[1 + O(n^{-1})], \quad \mathrm{V}_n[E] = \rho(\ell) \cdot n^3[1 + O(n^{-1})], \quad r_n[E] = O(n^{-1/3}).$$

Avec la méthodologie développée dans [Val03], il était seulement connues que $\mathrm{E}_n[E] \sim \mu(l) \cdot n^2$ et $\mathrm{V}_n[E] = o(n^4)$. Un résultat identique est montré sur les polynômes.

Théorème 4 (Complexité binaire étendue (polynômes)) *Sur l'ensemble Ω_n des polynômes de taille n, la complexité binaire E de l'algorithme d'Euclide étendu admet une loi limite gaussienne, d'espérance, de variance et de vitesse de convergence*

$$\mathrm{E}_n[E] = \frac{2q-1}{q} \cdot n^2[1 + O(n^{-1})], \quad \mathrm{V}_n[E] = \frac{q-1}{q^2} \cdot n^3[1 + O(n^{-1})], \quad r_n[E] = O(n^{-1/2}).$$

Les techniques de séries génératrices décrites au chapitre 2 mènent aux asymptotiques exactes de l'espérance et de la variance de E sur Ω_n,

$$\begin{aligned}
\mathrm{E}_n[E] &= \frac{2q-1}{q} n^2 - \frac{4q}{q-1} n + \frac{2q^3+8q^2-4q+1}{q(q-1)^2} + O(n^2 q^{-n}), \\
\mathrm{V}_n[E] &= \frac{q-1}{q^2} n^3 + \frac{11q^3-11q^2+3q-1}{(q-1)^2 q^2} n^2 - \frac{21q^4-6q^3-8q^2-4q+1}{(q-1)^3 q^2} n \\
&\quad - \frac{22q^5-20q^4+23q^3-1-4q^2+5q}{(q-1)^4 q^2} + O(n^4 q^{-n}).
\end{aligned}$$

Nous observons sur les polynômes que la variance de la complexité binaire étendue est, pour le terme dominant, trois fois la variance de la complexité binaire classique. C'est la raison pour laquelle nous avons conjecturé que la constante dominante $\rho(\ell)$ de la variance de la complexité binaire étendue sur les entiers est trois fois celle obtenue pour la complexité classique sur les entiers.

1.3.5 Continuant à une fraction de l'exécution

Les restes successifs dans l'exécution de l'algorithme d'Euclide classique sont aussi appelés continuants. Le résultat suivant montre que la taille du continuant à une fraction rationnelle de l'exécution admet une loi limite gaussienne.

Théorème 5 (Loi limite pour les continuants) *Soit δ un rationnel dans $]0,1[$. Sur l'ensemble Ω_n des entrées de taille n, la variable aléatoire $L^{[\delta]}$ égale à la taille du reste $v_{\lfloor \delta P \rfloor}$,*

$$L^{[\delta]} = \ell(v_{\lfloor \delta P \rfloor}),$$

admet une loi limite gaussienne d'espérance et de variance linéaire en n

$$\mathrm{E}_n[L^{[\delta]}] = (1-\delta)\cdot n\left[1+O(n^{-1})\right] \qquad \mathrm{V}_n[L^{[\delta]}] = \rho_{[\delta]}\cdot n\left[1+O(n^{-1})\right]$$

et de vitesse de convergence

$$\begin{aligned} r_n[L^{[\delta]}] &= O(n^{-1/3}) \qquad \textit{pour les entiers,}\\ r_n[L^{[\delta]}] &= O(n^{-1/2}) \qquad \textit{pour les polynômes.}\end{aligned}$$

Dans le cas des entiers, $\rho_{[\delta]}$ est associé aux dérivées de la fonction $\Lambda(s) = \log\lambda(s)$, où $\lambda(s)$ est la valeur propre dominante de l'opérateur de transfert $\mathbf{G}_s = \mathbf{G}_{s,0,[c]}$ (qui ne dépend pas de c). En particulier,

$$\rho_{[\delta]} = \delta(1-\delta)\frac{|\Lambda''(1)|}{|\Lambda'(1)|} \approx 1.4531\delta(1-\delta).$$

Dans le cas des polynômes, $\rho_{[\delta]}$ est connue et satisfait

$$\rho_{[\delta]} = \frac{\delta(1-\delta)}{q-1}.$$

Le théorème précédent montre la régularité de la décroissance des continuants. Il constitue aussi une version discrète d'un résultat bien connu de Philipp [Phi70] et amélioré par Vallée dans [Val97b].

1.4 Paramètres des algorithmes interrompus

1.4.1 Régularité des algorithmes interrompus

Le théorème 5 montre que si P est la variable aléatoire du nombre d'étapes, le reste après $\delta\cdot P$ étapes a perdu $\delta\cdot n$ bits par rapport à l'entrée initiale. Nous allons montrer avec les algorithmes interrompus que cette régularité reste vrai quel que soit le point de départ dans l'exécution (au milieu, au tiers, ...). Nous allons aussi considérer des rationnels δ avec un dénominateur croissant

(il était fixe dans le théorème 5). Mais cette régularité ne peut être démontrée lorsque δ est trop petit. En effet, il existe des divisions qui font intervenir de grands quotients et dans ce cas, il y a une perte rapide de bits. Il faut alors attendre un certain temps avant que l'algorithme ne compense ce phénomène.

Nous notons P, $P_{[\gamma,\delta]}$ et $P_{<\gamma,\delta]}$ le nombre d'itérations de l'algorithme d'Euclide classique et des algorithmes interrompus $\mathcal{E}_{[\gamma,\delta]}$ et $\mathcal{E}_{<\gamma,\delta]}$. Remarquons que le nombre d'étapes $P_{<\gamma,\delta>}$ de $\mathcal{E}_{<\gamma,\delta>}$ est connu est vaut $\delta \cdot P$. De même, $\ell_{<\gamma,\delta>}$, $\ell_{[\gamma,\delta]}$ et $\ell_{<\gamma,\delta]}$ désignent les tailles binaires des matrices produites par les algorithmes $\mathcal{E}_{<\gamma,\delta>}$, $\mathcal{E}_{[\gamma,\delta]}$ et $\mathcal{E}_{<\gamma,\delta]}$.
Le théorème suivant montre que ces paramètres sont en moyenne proches.

Théorème 6 *Il existe une constante $K < 1$, et un entier n_0, tels que, pour tout $n \geq n_0$, pour toute suite $D(n)$ qui satisfait*

$$D(n) = \sqrt{\frac{n}{b(n)}}, \qquad \textit{avec} \quad \lim_n b(n) = +\infty,$$

et pour tout rationnel $\delta(n), \gamma(n)$ de $[0,1]$, avec dénominateur $D(n)$, il existe un sous-ensemble $\mathcal{O}_n(\gamma,\delta)$ de Ω_n [appelé l'ensemble ordinaire relatif à (γ,δ)], avec $\mathbb{P}_n[\mathcal{O}_n(\gamma,\delta)] \geq 1 - K^{b(n)}$, sur lequel nous avons :

(i) les deux variables $P_{<\gamma,\delta]}$ et $\lfloor \delta P \rfloor$ sont proches :

$$\left| P_{<\gamma,\delta]} - \lfloor \delta P \rfloor \right| \leq \frac{P}{D(n)};$$

(ii) les deux algorithmes interrompus $\mathcal{E}_{[\gamma,\delta]}$ et $\mathcal{E}_{<\gamma,\delta>}$ sont proches :

$$d(\mathcal{E}_{<\gamma,\delta>}, \mathcal{E}_{[\gamma,\delta]}) \leq \frac{P}{D(n)};$$

(iii) la taille binaire de la matrice produite par l'algorithme interrompu $\mathcal{E}_{<\gamma,\delta>}$ a la taille "attendue" : pour tout β, avec $1 \leq \beta \leq 2$, on a

$$\left| \ell^{\beta}_{<\gamma,\delta>} - (\delta n)^{\beta} \right| \leq \frac{n}{D(n)} \cdot (\delta n)^{\beta-1};$$

(iv) la taille binaire de la matrice produite par l'algorithme interrompu $\mathcal{E}_{[\gamma,\delta]}$ a la taille "attendue" : pour tout β, avec $1 \leq \beta \leq 2$, on a

$$\left| \ell^{\beta}_{[\gamma,\delta]} - (\delta n)^{\beta} \right| \leq \frac{n}{D(n)} \cdot (\delta n)^{\beta-1}.$$

Preuve. La partie 3 est consacrée à l'analyse des différents paramètres dans le cas des entiers. En particulier, on y trouve les preuves des assertions (i) et (iii). Nous décrivons maintenant comment obtenir les assertions (ii) et (iv).

Assertion (ii). Par définition des algorithmes interrompus, l'algorithme $\mathcal{E}_{[0,\gamma+\delta]}$ est juste la concaténation $\mathcal{E}_{[0,\gamma]} \cdot \mathcal{E}_{[\gamma,\delta]}$. Alors, en fixant $\epsilon = 1/D$, l'assertion (i) entraîne que l'indice de départ de $\mathcal{E}_{[\gamma,\delta]}$ appartient à l'intervalle $[(\gamma-\epsilon)P, (\gamma+\epsilon)P]$ (quand $(v_0, v_1) \in \mathcal{O}_n(0,\delta)$) alors que l'indice de fin de $\mathcal{E}_{[\gamma,\delta]}$ appartient à l'intervalle $[(\gamma+\delta-\epsilon)P, (\gamma+\delta+\epsilon)P]$ (quand $(v_0, v_1) \in \mathcal{O}_n(0,\gamma+\delta)$). Cela prouve (ii) sur $\mathcal{O}_n(0,\gamma+\delta) \cap \mathcal{O}_n(0,\delta)$.

Assertion (iv). De (ii), on tire les inclusions suivantes :

$$\mathcal{E}_{<\gamma+\epsilon,\, \delta-2\epsilon>} \subset \mathcal{E}_{[\gamma,\delta]} \subset \mathcal{E}_{<\gamma-\epsilon,\, \delta+2\epsilon>}$$

Comme la taille de la matrice produite par un algorithme interrompu est une fonction décroissante [i.e., si $\mathcal{E}_{(i,j)} \subset \mathcal{E}_{(i',j')}$ alors $\ell_{(i,j)} \leq \ell_{(i',j')}$], l'assertion (iii) entraîne (iv). ■

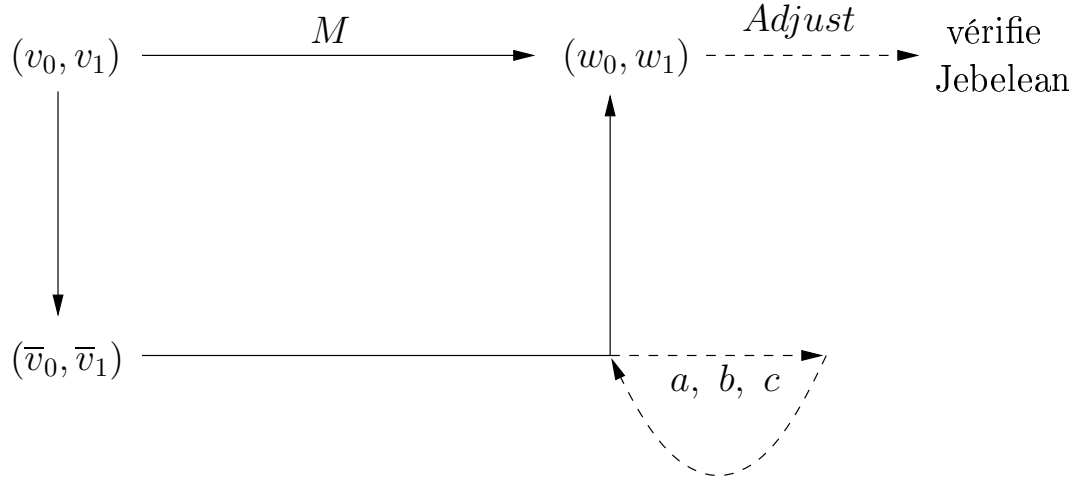
1.5 Complexité binaire de l'algorithme $\mathcal{KS}_\alpha$

1.5.1 Fonctions *Adjust* : le grain de sable dans l'analyse

Les fonctions *Adjust* ont essentiellement pour rôle de vérifier le critère de Jebelean sur les grands entiers après les appels récursifs aux fonctions $\mathcal{HG}_\alpha$. Plus précisément, sur une entrée (v_0, v_1), l'algorithme $\mathcal{HG}_\alpha$ construit la troncature $(\overline{v}_0, \overline{v}_1) = T_m(v_0, v_1)$ avec $m \approx \ell(v_0)/2$. Ensuite, la fonction $\mathcal{HG}_\alpha$ appliquée à $(\overline{v}_0, \overline{v}_1)$ construit une matrice M commune aux exécutions de l'algorithme d'Euclide sur la paire initiale et la paire tronquée. Cette matrice correspond à celle qui aurait été calculée avec l'algorithme $\widehat{\mathcal{E}}_{1/2}$ sur $(\overline{v}_0, \overline{v}_1)$. Nous notons $\overline{M}$ la matrice qui aurait été calculée par l'algorithme $\mathcal{E}_{1/2}$ sur $(\overline{v}_0, \overline{v}_1)$. Par définition, l'algorithme $\widehat{\mathcal{E}}_{1/2}$ comporte trois étapes en moins que l'algorithme $\mathcal{E}_{1/2}$. Il existe alors trois entiers a, b et c tels que

$$M = \overline{M} M_{[c]}^{-1} M_{[b]}^{-1} M_{[a]}^{-1}.$$

Si n est la taille de v_0, alors $\overline{v}_0$ est de taille $n/2$ et la matrice $\overline{M}$ sera a peu près de taille $n/4$. En revanche, nous ne disposons d'aucune information sur les quotients a, b et c, et de fait, nous n'avons aucune idée de la taille de M. Comme la taille de M est inconnue, la taille des entiers (w_0, w_1), définis par ${}^t(v_0, v_1) = M^t(w_0, w_1)$, et appartenant à la suite des restes issus de (v_0, v_1), n'est pas connue. Ne connaissant pas la taille de (w_0, w_1), nous ne pouvons pas estimer le coût des fonctions *Adjust*.



On s'attend tout de même à ce que la matrice M ait la bonne taille, c'est à dire environ $n/4$, ou de manière équivalente, que les quotients a, b et c ne soient pas trop grands. Cette attente est formulée par l'hypothèse des petits reculs qui va suivre. Mais avant, quelques notations sont nécessaires.

Chaque appel à la fonction $\mathcal{HG}$ avec des éléments de tailles plus grandes que le seuil S implique deux appels récursifs à la fonction $\mathcal{HG}$. Les appels récursifs successifs se représentent donc sous la forme d'un arbre binaire (cf. figure 1.4) où la racine est l'appel initial à la fonction $\mathcal{HG}$ et où le sous-arbre gauche (resp. droit) est l'arbre issu du premier (second) appel à la fonction $\mathcal{HG}$. La matrice calculée par l'appel à la fonction $\mathcal{HG}$, située au j^e noeud du niveau i, est notée $M_j^{(i)}$. A ce même noeud, les deux fonctions *Adjust* correspondent à deux matrices $U_j^{(i)}$ et $V_j^{(i)}$. Par construction, nous avons l'égalité

$$M_j^{(i)} = M_{2j}^{(i+1)} U_j^{(i)} M_{2j+1}^{(i+1)} V_j^{(i)}. \tag{1.14}$$

Hypothèse 1 (hypothèse des petits reculs) *Nous définissons $R(i,j)$ la somme des tailles binaires des matrices d'ajustement $U_j^{(i)}$ et $V_j^{(i)}$,*

$$R(i,j) = \ell(U_j^{(i)}) + \ell(V_j^{(i)}).$$

Il existe une suite $\epsilon = \epsilon(n)$ avec $n^2\epsilon(n) = O(\mu(n)/n^r)$ ($r > 0$ sera précisé plus tard) et $\mu(n)$ la complexité binaire de la multiplication sur les entiers de taille n, et il existe un réel positif $K < 1$, tels que, pour tout $i \leq h$ (h est la hauteur de l'arbre des appels récursifs), pour tout $j < 2^i - 1$, la somme $R(i,j)$ des tailles binaires des deux matrices associées aux fonctions Adjust au j^e noeud du niveau i vérifie

$$\mathbb{P}_n\left[R(i,j) \geq \epsilon(\frac{n}{2^{i+2}}) \cdot \frac{n}{2^{i+2}}\right] = O(K^{n^r}).$$

L'hypothèse des petits reculs entraîne l'existence un ensemble ordinaire $\overline{\mathcal{O}}_n$ de probabilité proche de 1 sur lequel tous les $R(i,j)$ vérifient

$$R(i,j) \leq \epsilon(\frac{n}{2^{i+2}}) \cdot \frac{n}{2^{i+2}}.$$

Le nombre d'étapes maximum pour obtenir une matrice de taille $R(i,j)$ s'obtient lorsque tous les quotients sont 1. Un petit calcul simple avec la suite de Fibonacci montre que le nombre d'étapes est au maximum d'ordre $O(R(i,j))$ avec une constante universelle dans le O (qui fait intervenir le nombre d'or). En appliquant la formule 1.12 (page 28), et sachant que le niveau i fait intervenir des entiers de taille $O(n/2^{i+2})$, la complexité $A(i,j)$ des fonctions *Adjust* au noeud (i,j) vérifie

$$A(i,j) = O\left(\epsilon(\frac{n}{2^{i+2}}) \cdot \left(\frac{n}{2^{i+2}}\right)^2\right).$$

La condition sur ϵ et la somme sur tous les noeuds des $A(i,j)$ conduisent à la proposition suivante.

Proposition 1 *Sous l'hypothèse des petits reculs, la contribution totale des fonctions Adjust à l'algorithme $\mathcal{HG}_\alpha$ sur l'ensemble ordinaire $\overline{\mathcal{O}}_n$ est de l'ordre de $O(\mu(n)/n^r)$.*

1.5.2 Algorithme de Knuth-Schönhage et algorithmes interrompus

Une première manière de procéder à l'analyse de la fonction $\mathcal{HG}_\alpha$ est d'étudier l'évolution des données à chaque noeud de l'arbre des appels récursifs. Cela suppose de connaître l'effet des troncatures sur la distribution des données, puis comment cette distribution est modifiée par les appels récursifs et les fonctions d'ajustement. On voit clairement la complexité d'une telle approche.

La proposition fondamentale suivante identifie les appels aux fonctions $\mathcal{HG}_\alpha$ à une fraction de l'exécution sur les grands entiers initiaux. En appliquant le théorème 6 page 30 sur la régularité des algorithmes interrompus, nous obtenons immédiatement une description très précise de la taille des matrices $M_j^{(i)}$ produites par les fonctions $\mathcal{HG}_\alpha$ aux noeuds de l'arbre. Nous introduisons l'algorithme interrompu $\overline{\mathcal{E}}_{[a,b]}$ qui commence à la plus petite étape i telle que $\ell(v_i) \geq a$ et termine à la plus grande étape k telle que $\ell(v_k) \leq b$.

Proposition 2 *Considérons le j^e noeud du niveau i dans l'arbre des appels récursifs. Supposons que le niveau i satisfait $i < \eta \lg n$ pour une constante $\eta < 1$. Sous l'hypothèse des petits reculs et*

pour une entrée appartenant à l'ensemble ordinaire $\overline{\mathcal{O}}_n$, la matrice $M_j^{(i)}$ coïncide avec la matrice qui aurait été calculée par un algorithme interrompu $\overline{\mathcal{E}}_{[a,b]}$ sur les entiers initiaux avec

$$a := \frac{jn}{2^{i+1}} + O\left(\frac{n}{2^{i+1}}\epsilon(\frac{n}{2^{i+1}})\right), \quad b := \frac{jn}{2^{i+1}} + O\left(\frac{n}{2^{i+1}}\epsilon(\frac{n}{2^{i+1}})\right)$$

où la constante dans le O est uniforme pour tous les noeuds.

Cette proposition se démontre très rapidement par récurrence en utilisant les relations matricielles 1.14, l'hypothèse des petits reculs sur les matrices d'ajustement et le fait que la première partie de la fonction $\mathcal{HG}_\alpha$ (lignes 1 à 7 de la figure 1.3) correspond à un algorithme $\mathcal{E}_{1/4}$ alors que la deuxième partie (lignes 8 à 11) correspond à l'algorithme $\widehat{\mathcal{E}}_{1/4,1/4}$.

La proposition est essentielle dans nos analyses puisqu'elle ramène l'analyse de la fonction $\mathcal{HG}_\alpha$ à celle des algorithmes interrompus.

1.5.3 Régularité de l'arbre des appels récursifs

Nous combinons maintenant la proposition 2 avec le théorème 6 page 30 sur les algorithmes interrompus pour obtenir une description complète de la taille des matrices $M_j^{(i)}$.

La proposition 2 indique que cette matrice est aussi celle calculée par l'algorithme interrompu $\overline{\mathcal{E}}_{[a,b]}$, avec $a \approx j/2^{i+1}$ et $b \approx (j+1)/2^{i+1}$.
Nous considérons maintenant l'algorithme $\mathcal{HG}_\alpha$, avec un paramètre α de la forme $\alpha = 3/2 + 3r$, avec $r > 0$. Nous choisissons la profondeur de récursion $h = \lceil (2 - \alpha + r) \lg n \rceil = \lceil (1/2 - 2r) \lg n \rceil$. Alors, tous les paramètres δ qui interviennent dans les algorithmes interrompus ont un dénominateur divisible par $2^h = n^{1/2-2r}$, et le nombre de paramètres δ différents est d'ordre $\Theta(2^h)$. Nous choisissons le dénominateur D de la forme $D = \Theta(2^h \cdot n^r) = \Theta(n^{1/2-r})$, et $b(n) = \Theta(n^{2r})$. En appliquant la propriété (iv) du théorème 6 avec ces paramètres particuliers, nous définissons l'ensemble ordinaire $\mathcal{O}_n$ par

$$\mathcal{O}_n := \overline{\mathcal{O}}_n \bigcap \left[\bigcap_{\substack{i \le h \\ 0 \le j < 2^i}} \mathcal{O}_n(\frac{j}{2^i}, \frac{1}{2^i}) \right].$$

Cet ensemble vérifie

$$\mathbb{P}_n[\mathcal{O}_n] \ge 1 - \sqrt{n} K^{b(n)} \ge 1 - K^{n^r},$$

et l'assertion (iv) du théorème 6 précise les tailles des matrices qui interviennent. Finalement, nous avons prouvé le théorème suivant.

Théorème 7 (Régularité de l'arbre des appels récursifs) *Considérons le j^e nœud du niveau i dans l'arbre des appels récursifs de l'algorithme $\mathcal{HG}_\alpha$. Supposons que i satisfait $i < [2 - \alpha - r] \lg n$ pour un r positif. Sous l'hypothèse des petits reculs, la matrice $M_i^{(j)}$ calculée par l'algorithme $\mathcal{HG}_\alpha$ au j^e nœud du niveau i $(0 \le j < 2^i)$ a pour taille binaire $\ell_j^{(i)}$ qui satisfait, sur l'ensemble Ω_n et pour tout $\beta > 1$,*

$$\mathbb{P}_n\left[\left|(\ell_i^{(j)})^\beta - \left(\frac{1}{2^{i+1}}n\right)^\beta\right| > n^{(1/2)+r} \cdot \left(\frac{1}{2^{i+1}}n\right)^{\beta-1}\right] = O(K^{-n^r})$$

Le théorème 7 décrit complètement la taille des éléments qui interviennent dans les calculs des fonctions $\mathcal{HG}_\alpha$. Ces tailles sont conformes aux tailles attendues (sous l'hypothèse des petits reculs), ce qui montre la régularité de l'arbre des appels récursifs. Nous disposons donc de toutes les informations pour étudier la complexité binaire moyenne de l'algorithme de Knuth-Schönhage paramétré $\mathcal{KS}_\alpha$.

1.5.4 Complexités binaires de $\mathcal{HG}_\alpha$ et $\mathcal{KS}_\alpha$

Nous supposons qu'il existe deux constantes positives A_1 et A_2 telles que la complexité moyenne $\mu(n)$ de la multiplication sur des entiers de taille n satisfait

$$A_1 \cdot n^\alpha \leq \mu(n) \leq A_2 \cdot n^\alpha.$$

Le raisonnement exact qui va suivre montre comment à partir du théorème 7, nous obtenons la complexité binaire moyenne de l'algorithme $\mathcal{HG}_\alpha$. Ce raisonnement s'inspire très largement des travaux de Daireaux et Vallée [DV04] sur l'analyse de l'algorithme de Lehmer-Euclide.

Étape 1 : L'ensemble des entrées Ω_n de taille n est scindé en l'ensemble régulier $\mathcal{O}_n$ où toutes les matrices $M_j^{(i)}$ ont la taille attendue et son complémentaire. L'ensemble $\mathcal{O}_n$ est de probabilité $1 - n^{2-\alpha+r} \cdot O(K^{-n^r}) = 1 - O(K^{-n^{r/2}})$. Par suite, la probabilité de l'ensemble complémentaire, aussi appelé ensemble des exceptions, est $O(K^{-n^{r/2}})$.

Étape 2 : L'apport de l'ensemble des exceptions à la complexité binaire moyenne est négligeable. En effet, dans le pire des cas, la complexité binaire de $\mathcal{HG}_\alpha$ est de l'ordre de $O(n^\alpha)$. L'apport moyen de l'ensemble des exceptions est donc majoré par $O(n^\alpha \cdot K^{-n^{r/2}})$ qui tend vers 0.

Étape 3 : L'étape 2 montre que nous pouvons nous restreindre uniquement sur les entrées régulières de $\mathcal{O}_n$. Sur ces entrées, l'arbre des appels récursifs est complet et de hauteur $h = \lfloor (2-\alpha+r)\lg n \rfloor$. A chaque feuille, l'algorithme $\widehat{\mathcal{E}}_{1/2}$ est exécuté avec des entrées de taille au plus $n/2^h$. La complexité binaire de $\widehat{\mathcal{E}}_{1/2}$ est asymptotiquement quadratique et équivalente à $A_3 n^2$ (pour une constante A_3) sur des entrées de taille n. Le coût total aux feuilles est alors majoré par $A_3 2^h (n/2^h)^2 = O(n^{\alpha-r})$.

Étape 4 : L'étape 3 donne une borne supérieure du coût total au niveau des feuilles. Nous considérons maintenant le coût total de toutes les multiplications effectuées lors d'un appel. Plaçons nous au j^e nœud du niveau i de l'arbre. La taille des entrées est alors $m = n/2^i$. La ligne 6 effectue un produit d'une matrice 2×2 dont les coefficients sont de taille $\ell_{2j}^{(i+1)} = (m/4)(1 \pm \epsilon_i)$ avec un couple d'entiers de taille m. Quatre multiplications de ce type sont effectuées et quitte à couper les entiers de taille m en quatre morceaux de taille $\ell_{2j}^{(i+1)}$, le coût total V_1 de ce produit de matrice satisfait

$$16A_1 \left(\frac{1}{2^{i+2}}\right)^\alpha (1-\epsilon_i) \leq V_1 \leq 16A_2 \left(\frac{1}{2^{i+2}}\right)^\alpha (1+\epsilon_i).$$

Avec une implémentation plus fine, il est possible d'utiliser le couple (s_1, t_1) calculé à la ligne 5 et d'effectuer le produit de matrice avec ce couple d'entiers. Si M_1 est la matrice calculée à la ligne 5, alors $(u_1, v_1) = M_1(s_1, t_1)$ soit (s_1, t_1) sont des entiers de taille $m/2 + \ell_{2j}^{(i+1)} = (3m/4)(1 \pm \epsilon_i)$. Le coût amélioré $\overline{V}_1$ issue de cette optimisation satisfait alors

$$12A_1 \left(\frac{1}{2^{i+2}}\right)^\alpha (1-\epsilon_i) \leq \overline{V}_1 \leq 12A_2 \left(\frac{1}{2^{i+2}}\right)^\alpha (1+\epsilon_i).$$

En suivant la même démarche, le coût amélioré $\overline{V}_2$ de la multiplication de la ligne 10 satisfait

$$8A_1\left(\frac{1}{2^{i+2}}\right)^\alpha(1-\epsilon_i) \le \overline{V}_2 \le 8A_2\left(\frac{1}{2^{i+2}}\right)^\alpha(1+\epsilon_i).$$

La multiplication à la ligne 12 se fait entre deux matrices dont les coefficients sont de tailles respectives $\ell_{2j}^{(i+1)}$ et $\ell_{2j+1}^{(i+1)}$. Huit multiplications d'entiers sont calculées soit le coût $\overline{V}_3$ satisfait

$$8A_1\left(\frac{1}{2^{i+2}}\right)^\alpha(1-\epsilon_i) \le \overline{V}_3 \le 8A_2\left(\frac{1}{2^{i+2}}\right)^\alpha(1+\epsilon_i).$$

Au total et à chaque nœud du niveau i, le coût V de toutes les multiplications est encadré par

$$28A_1\left(\frac{1}{2^{i+2}}\right)^\alpha(1-\epsilon_i) \le V \le 28A_2\left(\frac{1}{2^{i+2}}\right)^\alpha(1+\epsilon_i).$$

Étape 5 : Ceci est la dernière étape. Considérons maintenant le coût total H_α de l'algorithme $\mathcal{HG}_\alpha$. H_α est donné par la somme de tous les coûts aux feuilles, plus le coût de toutes les multiplications des nœuds internes, plus le coût total de toutes les fonctions d'ajustement. Sous l'hypothèse 1 des petits reculs, la complexité totale de toutes ces fonctions est d'ordre $O(\mu(n)/n^r) = O(n^{3/2+r})$. L'étape 3 montre que le coût total aux feuilles est de l'ordre de $O(n^{\alpha-r})$. Finalement, en sommant sur tous les nœuds internes à droite et à gauche de l'encadrement, le coût total H_α satisfait

$$A_1[1-\epsilon_h] + O(n^{-r}+n^{1-\alpha}) \le \frac{\mathrm{E}_n[H_\alpha]}{7n^\alpha} \cdot \frac{2^{\alpha-1}-1}{2^{1-\alpha}} \le A_2[1+\epsilon_h] + O(n^{-r}+n^{1-\alpha})$$

où $\epsilon_h = O(n^{3/2-\alpha})$. Nous en déduisons le théorème suivant.

Théorème 8 *Supposons que $\alpha = 3/2+3r$ avec $r>0$, et notons H_α (resp K_α) la complexité binaire totale de l'algorithme $\mathcal{HG}_\alpha$ (resp $\mathcal{KS}_\alpha$), où le niveau de récursion est stoppé à la hauteur $h = \lfloor(2-\alpha+r)\lg n\rfloor$, où la multiplication de deux entiers de n bits admet une complexité binaire moyenne de l'ordre $\mu(n) = \Theta(n^\alpha)$. Sur des entiers de taille n, et sous l'hypothèse 1 des petits reculs, les variables aléatoires H_α et K_α satisfont*

$$\mathrm{E}_n[H_\alpha] = \theta(n^\alpha)\cdot[1+O(n^{-r})], \qquad \mathrm{E}_n[K_\alpha] = \theta(n^\alpha)\cdot[1+O(n^{-r})].$$

Si le coût moyen de la multiplication satisfait $A_1\cdot n^\alpha \le \mu(n) \le A_2\cdot n^\alpha$, alors

$$A_1[1+O(n^{-r})] \le \frac{\mathrm{E}_n[H_\alpha]}{7n^\alpha} \cdot \frac{2^{\alpha-1}-1}{2^{1-\alpha}} \le A_2[1+O(n^{-r})]$$

$$A_1[1+O(n^{-r})] \le \frac{\mathrm{E}_n[K_\alpha]}{7n^\alpha} \cdot \frac{2^{\alpha-1}-1}{2^{1-\alpha}} \le A_2[1+O(n^{-r})]$$

Le théorème 8 ne fonctionne que pour des multiplications dont α est plus grand que $3/2$. La multiplication de Karatsuba satisfait cette hypothèse puisque dans ce cas $\alpha = \log 3 \approx 1.58$. En revanche, le coefficient α de la multiplication de Tom-Cook satisfait $\alpha \approx 1.46 < 3/2$. Pour obtenir la complexité de $\mathcal{KS}_\alpha$, il suffit d'utiliser la relation $K_\alpha(n) = H_\alpha(n/2) + K_\alpha(n/2)$.

1.6 Conclusion

Dans cette partie, nous avons décrit les algorithmes d'Euclide classiques et étendus, à la fois sur les entiers et les polynômes. Pour ces algorithmes, nous avons introduit les complexités binaires associées. Nous prouvons dans les prochains chapitres que toutes ces complexités admettent une loi limite gaussienne exceptée celle de l'algorithme d'Euclide classique sur les entiers. Dans ce dernier cas, nous améliorons les résultats connus sur l'espérance et la variance. Pour tous ces paramètres, l'espérance est d'ordre quadratique et la variance est d'ordre au plus cubique. Le comportement moyen des complexités binaires est connu depuis les travaux de [AV00]. Mais pour la première fois, des résultats concernant leur distribution sont prouvés.

Nous avons également décrit une version légèrement modifiée de l'algorithme de Knuth-Schönhage. Pour ces versions mais aussi pour la première fois, une analyse en moyenne de la complexité binaire d'un algorithme rapide est réalisée (modulo une hypothèse sur les fonctions *Adjust*). Cette analyse est basée sur une étude fine des algorithmes interrompus qui correspondent à une fraction de l'exécution de l'algorithme classique. Nous montrons que sur des portions pas trop petites, les algorithmes d'Euclide sont très réguliers.

Les prochains chapitres abordent la preuve de ces résultats. Nous commençons par le cas le plus simple avec les polynômes (chapitre 2) et nous traitons ensuite les entiers (chapitre 3). La différence entre les polynômes et les entiers réside surtout dans les techniques d'analyses. Sur les polynômes, la combinatoire analytique avec les séries génératrices ordinaires est suffisante. Pour les entiers, la combinatoire analytique ne s'applique plus et nous utilisons l'analyse dynamique et les opérateurs de transfert. Les aspects techniques liés à l'analyse des opérateurs de transfert sont traités au chapitre 4. Finalement Nous terminons la partie sur les algorithmes euclidiens avec le calcul de constantes (chapitre 5) qui apparaissent dans les analyses en moyenne comme la constante $\rho(\ell)$ des théorèmes BV et 3.

Chapitre 2

Le cas des polynômes

Sommaire

2.1 Introduction

L'algorithme d'Euclide sur les polynômes constitue une brique de base essentielle pour tous langages de calcul formel comme MAPLE, MATHEMATICA, etc. Cet algorithme permet de simplifier les calculs intermédiaires, de calculer des inverses modulaires, de construire des bases de Gröbner, de calculer des développements en fraction continue, etc. Les algorithmes de factorisation de polynômes utilisent également plusieurs calculs de pgcd pour décomposer en facteurs irréductibles [FGP01].

Outre ces applications, l'algorithme d'Euclide sur les polynômes a un comportement proche de son homologue sur les entiers. Mais contrairement à ce dernier, il est relativement plus simple

à analyser. Il est souvent utile (mais pas obligatoire!) de prouver un résultat sur les polynômes avant de s'attaquer aux entiers. C'est la stratégie que nous avons adopté pour la conjecture (C) du théorème 2.

Dans le pire des cas, le nombre de divisions calculées par l'algorithme d'Euclide est linéaire en la taille des entrées. L'analyse en distribution a été effectuée par Knopfmacher J. et Knopfmacher A. [KK88] avec un dénombrement direct et par Flajolet [Fla06] avec des techniques de combinatoire analytique. Ils ont montré que le nombre de divisions suit asymptotiquement une loi binomiale. D'un autre coté, Friesen et Hensley [FH 9] ont obtenu des propriétés de larges déviations sur ce même paramètre et sur les entrées dont le degré des quotients est borné. Ces entrées trouvent notamment des applications dans la conception et l'analyse de générateurs pseudo-aléatoires [Rue85, Rue86, Nie87, Nie88]. Dans le pire des cas, la complexité binaire est quadratique et il semble généralement admis qu'il en est de même avec la complexité binaire moyenne.

Dans cette partie nous analysons la loi limite de la complexité binaire des algorithmes classique et étendu ainsi que la taille du continuant à une fraction de l'exécution, et nous prouvons les résultats des théorèmes 1, 4 et 5. Les techniques utilisées sont celles de la combinatoire analytique dont les outils principaux sont les séries génératrices. L'analyse de la taille du continuant à une fraction de l'exécution est une application directe de ces techniques et du théorème des quasi-puissances de Hwang [Hwa94]. Les analyses des complexités binaires sont aussi basées sur un principe de décomposition (voir proposition 3). Ce principe donne des conditions suffisantes sur une somme de paramètres afin qu'elle admette une loi limite gaussienne. Les complexités binaires sont décomposées en plusieurs familles de paramètres dont les coûts à croissance modérée, les coûts à croissance intermédiaire, les coûts terminaux et un paramètre N du type longueur de cheminement. Nous adaptons au cadre des polynômes les résultats de Baladi et Vallée [BV05, BV04] concernant la loi limite gaussienne des coûts à croissance modérée (théorème 9). La loi limite gaussienne du paramètre N est également démontrée. Ensuite, nous analysons les deux premiers moments des coûts additifs à croissance intermédiaire et des coûts terminaux (théorème 10).

Plan. La première section présente le principe de décomposition et les décompositions des complexités binaires. La section suivante introduit rapidement quelques éléments de combinatoire analytique comme les séries génératrices ordinaires et les dictionnaires associés. Finalement, aux sections 2.4 et 2.5, les méthodes de combinatoire analytique sont appliquées et les théorèmes sur les polynômes sont prouvés.

2.2 Principe de décomposition et décomposition

Le point de départ de l'analyse des complexités binaires est une décomposition des coûts en deux parties : une partie principale qui apporte le comportement gaussien et une seconde partie plus concentrée que la première. Si cette décomposition a lieu, nous montrons à la section 2.2.1 qu'elle entraîne la loi limite gaussienne du paramètre étudié. Chaque analyse des complexités binaires suit donc le même schéma. Tout d'abord, une décomposition est proposée. Ensuite il est montré que la première partie suit asymptotiquement une loi gaussienne de variance v_n. Puis il est montré qu'il existe une suite α_n qui tend vers 0 et telle que la variance de chaque coût qui compose la seconde partie soit d'ordre au plus $\alpha_n v_n$. La décomposition est alors valide et entraîne la loi limite gaussienne du paramètre étudié.

2.2.1 Principe de décomposition

Le principe de décomposition est donné par la proposition suivante.

Proposition 3 *Soit deux paramètres X et Y définis sur $\Omega = \cup_n \Omega_n$. Supposons que X admet une loi limite gaussienne avec une vitesse de convergence $r_n[X]$ et que les variances de X et Y satisfont $\mathrm{V}_n[Y] = \alpha_n \mathrm{V}_n[X]$, avec $\alpha_n \to 0$. Alors, la variable aléatoire $X+Y$ suit asymptotiquement une loi limite gaussienne d'espérance, de variance et de vitesse de convergence données par*

$$\mathrm{E}_n[X+Y] = \mathrm{E}_n[X] + \mathrm{E}_n[Y], \quad \mathrm{V}_n[X+Y] = \mathrm{V}_n[X](1+O(\alpha_n^{1/2})), \quad r_n[X+Y] = r_n[X] + O(\alpha_n^{1/3}).$$

En pratique, $r_n[X]$ sera toujours $n^{-1/2}$. Si α_n est de l'ordre de n^{-1} alors la vitesse de convergence est en $n^{-1/3}$. Cette vitesse de convergence est non-optimale dans le sens où l'on attend plutôt une vitesse de l'ordre de $n^{-1/2}$. Si α_n est de l'ordre de n^{-2}, alors la vitesse de convergence est optimale en $n^{-1/2}$. Dans nos études, la variable aléatoire X admet toujours une variance d'ordre cubique. Pour obtenir une vitesse non optimale en $n^{-1/3}$, il suffit de montrer que la variance de Y est d'ordre au plus n^2. En revanche pour obtenir la vitesse optimale, la variance de Y doit être au plus d'ordre $n^{3/2}$. Les moments de tous les paramètres étudiés sont polynomiaux en n et l'erreur dans la variance peut être remplacée par $O(1/n)$.

Preuve. Fixons $\overline{X}_n$ et $\overline{Y}_n$ les variables aléatoires

$$\overline{X}_n = \frac{X - \mathrm{E}_n[X]}{\mathrm{V}_n[X]^{1/2}}, \qquad \overline{Y}_n = \frac{Y - \mathrm{E}_n[Y]}{\mathrm{V}_n[X]^{1/2}}.$$

Alors, la variable aléatoire $\overline{X}_n + \overline{Y}_n$ satisfait

$$\mathbb{P}_n[\overline{X}_n + \overline{Y}_n \leq a] = \mathbb{P}_n[[\overline{X}_n + \overline{Y}_n \leq a] \cap [|\overline{Y}_n| \leq \epsilon_n]] + \mathbb{P}_n[[\overline{X}_n + \overline{Y}_n \leq a] \cap [|\overline{Y}_n| > \epsilon_n]].$$

Le deuxième terme est plus petit que $\mathbb{P}_n[|\overline{Y}_n| > \epsilon_n]$, et avec l'inégalité de Markov, il est d'ordre $O(\alpha_n \cdot \epsilon_n^{-2})$. Maintenant, pour le premier terme, nous avons

$$\mathbb{P}_n[\overline{X}_n \leq a - \epsilon_n] \leq \mathbb{P}_n[[\overline{X}_n + \overline{Y}_n \leq a] \cap [|\overline{Y}_n| \leq \epsilon_n]] \leq \mathbb{P}_n[\overline{X}_n \leq a + \epsilon_n]$$

et les termes à gauche et à droite sont bornés par

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a \pm \epsilon_n} e^{-t^2/2} dt + O(r_n[X]) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-t^2/2} dt + O(r_n[X] + \epsilon_n).$$

Finalement, la vitesse de convergence est d'ordre $O(r_n[X] + \epsilon_n + \alpha_n \cdot \epsilon_n^{-2})$ et le choix optimal $\epsilon_n^3 = \alpha_n$ donne le résultat. Si $\sigma_n(X)$ est l'écart type de X, alors le terme d'erreur de la variance vient de l'encadrement

$$\sigma_n(X) - \sigma_n(Y) \leq \sigma_n(X+Y) \leq \sigma_n(X) + \sigma_n(Y)$$

et de l'hypothèse $\sigma_n(Y) = \alpha_n^{1/2} \sigma_n(X)$. ■

2.2.2 Décompositions des complexités binaires

2.2.2.1 Relations sur les degrés

Nous proposons maintenant des décompositions pour les complexités binaires qui satisfont le principe de décomposition. Ces décompositions sont basées sur des relations simples entre le degré des continuants a_i et v_i et le degré des quotients successifs. Le même type de propriétés existe sur les entiers (voir les formules 3.3 et 3.4) mais la situation est un peu plus complexe.

Lemme 1 *Le degré des continuants v_i et a_i s'exprime en fonction des degrés des quotients m_i de la manière suivante,*

$$\deg v_i = \deg v_p + \sum_{u=i+1}^{p} \deg m_u \qquad et \qquad \deg a_i = \sum_{u=1}^{i-1} \deg m_u. \tag{2.1}$$

En particulier, $\ell(v_i) + \ell(a_i)$ vérifie l'identité

$$\ell(v_i) + \ell(a_i) = 2 + \ell(v_0) - \ell(m_i). \tag{2.2}$$

Sur Ω_n, la taille de v_0 est constante et égale à n et $\ell(v_i) + \ell(a_i)$ est *presque* constant. Il existe une relation similaire entre les continuants et les quotients sur les entiers (voir formules 3.3 et 3.4), mais des phénomènes de distorsion apparaissent.

2.2.3 Décomposition de la complexité binaire étendue

La complexité binaire de l'algorithme d'Euclide étendu est définie par

$$E(v_0, v_1) = \ell(v_p)\ell(m_p) + \sum_{i=1}^{p-1} \ell(m_p)(\ell(v_i) + \ell(a_i)).$$

En utilisant la formule 2.2 sur les degrés, pour (v_0, v_1) dans Ω_n, la complexité E s'écrit aussi

$$E = (n+2) \cdot Y_0 + Y_1 - Y_2$$

avec

$$Y_0(v_0, v_1) = \sum_{i=1}^{p} \ell(m_i), \qquad Y_1(v_0, v_1) = \ell(m_p)\ell(v_p) - \ell(m_p) - \ell(m_p)^2, \qquad Y_2(v_0, v_1) = \sum_{i=1}^{p} \ell(m_i)^2.$$

Les coûts Y_0 et Y_2 sont appelés des coûts additifs car ce sont la somme d'un coût élémentaire sur tous les quotients. Selon les propriétés du coût élémentaire, les coûts additifs pourront être linéaires, à croissance modérée ou à croissance intermédiaire. Le coût Y_1 est appelé un coût terminal puisqu'il ne fait intervenir que les éléments de la dernière étape (le quotient m_p et le pgcd v_p).

Définition 3 *(i) Les coûts terminaux Y sont tous les coûts polynomiaux en la la taille de v_p et m_p et sont de la forme $Y(v_0, v_1) = O((\ell(m_p) + \ell(v_p))^k)$ avec k un entier positif fixé, et si $p = 0$ alors $Y = 0$.*
(ii) Si $c : \mathbb{F}_q[X] \to \mathbb{R}^+$ est un coût élémentaire non-nul sur les quotients, le coût additif C associé à c est défini par

$$C(v_0, v_1) = \sum_{i=1}^{p} c(m_i).$$

1. *Un coût additif est dit linéaire si le coût élémentaire c est de la forme $c(m) = \alpha \cdot \deg m$ où α est une constante positive.*
2. *Un coût additif est dit à croissance modérée si le coût élémentaire c satisfait $c(m) = O(\deg m)$.*
3. *Un coût additif est dit à croissance intermédiaire s'il n'est pas à croissance modérée et si le coût élémentaire c satisfait $c(m) = O(\deg^k m)$ pour un entier k fixé.*

Avec ces définitions, le coût Y_0 est à croissance modérée alors que le coût Y_2 est à croissance intermédiaire et les deux ne sont pas linéaires. Les coûts additifs linéaires sont très particuliers puisqu'ils se ramènent à des coûts terminaux. Si c est de la forme $c(m) = \alpha \cdot \deg m$ avec α une constante, la formule 2.1 sur les degrés entraîne la relation

$$C(v_0, v_1) = \sum_{i=1}^{p} \alpha \cdot \deg m_i = \alpha \cdot (\deg v_0 - \deg v_p).$$

Sur Ω_n, $\deg v_0 = (n-1)$ et le coût C est donc à une constante près un coût terminal.

Nous souhaitons montrer que la complexité binaire E admet une loi limite gaussienne en appliquant le principe de décomposition. Comme sur les entiers, nous allons montrer que les coûts à croissance modérée admettent une loi limite gaussienne d'espérance et de variance linéaires en n. Ainsi, $(n+1) \cdot Y_0$ admet une loi limite gaussienne d'espérance quadratique et de variance cubique et constitue la partie principale de la décomposition.

Théorème 9 (loi gaussienne des coûts à croissance modérée) *Soit C un coût additif à croissance modérée et non-linéaire associé au coût élémentaire c. Alors C satisfait sur Ω_n une loi limite gaussienne de paramètres*

$$\mathrm{E}_n[C] = \Lambda'(0) \cdot n(1 + O(n^{-1})), \qquad \mathrm{V}_n[C] = \Lambda''(0) \cdot n(1 + O(n^{-1})), \qquad r_n = O(n^{-1/2})$$

où $\Sigma(w) = -\log \sigma(w)$ et $\sigma(w)$ est la fonction définie dans un voisinage de 0 par $G(\sigma(w), w) = 1$ avec

$$G(z, w) = \sum_{m : \deg m \geq 1} e^{wc(m)} z^{\deg m}. \tag{2.3}$$

Pour satisfaire le principe de décomposition, il reste à montrer que Y_1 et Y_2 sont plus concentrés que $(n+1) \cdot Y_0$.

Théorème 10 (Coûts à croissance intermédiaire et coûts terminaux) *(i) Soit C un coût additif à croissance intermédiaire (ou modérée) associé au coût élémentaire c. Si G est la série génératrice 2.3, alors les moments de C satisfont sur Ω_n*

$$\mathrm{E}_n[C] = \frac{q-1}{q} \frac{dG}{dw}(q^{-2}, 0) \cdot n(1 + O(n^{-1})), \qquad \mathrm{V}_n[C] = O(n).$$

(ii) Soit Y un coût terminal. Alors tous les moments de Y sont d'ordre constant sur Ω_n.

En admettant le théorème précédent, E satisfait le principe de décomposition et admet donc une loi limite gaussienne.

2.2.4 Décomposition de la complexité binaire classique

Nous abordons maintenant la décomposition de la complexité binaire classique. En écrivant $\ell(m_i)\ell(v_i) = (1 + \deg m_i)(1 + \deg v_i)$, le coût B satisfait

$$\begin{aligned} B(v_0, v_1) &= p + \sum_{i=1}^{p} \deg v_i + \sum_{i=1}^{p} \deg m_i \deg v_i + \deg v_0 - \deg v_p \\ &= \sum_{i=1}^{p} \deg v_{i-1} + \sum_{i=1}^{p} \deg m_i \deg v_i + p - \deg v_p. \end{aligned} \tag{2.4}$$

Le nombre d'étapes p est un coût à croissance modérée associé au coût primitif $c = 1$ et sa variance est d'ordre linéaire si l'on admet le théorème 9. De même, le paramètre $\deg v_p$ est un coût terminal et ses moments sont d'ordre constant. La seconde somme se simplifie. Afin d'alléger les notations, nous posons $m_{p+1} = v_p$. Avec les égalités 2.1 sur les degrés, la seconde somme devient

$$\sum_{i=1}^{p} \deg m_i \deg v_i = \sum_{i=1}^{p} \deg m_i \sum_{u=i+1}^{p+1} \deg m_u$$

La somme à droite est composée des doubles produits issus du développement de $(\sum_{i=1}^{p+1} \deg m_i)^2$. Or les égalités 2.1 appliquées à v_0 montrent que ce carré n'est autre que $\deg^2 v_0$. La somme initiale satisfait alors

$$\sum_{i=1}^{p} \deg m_i \deg v_i = \frac{1}{2} \deg^2 v_0 - \frac{1}{2} \deg^2 v_p - \frac{1}{2} \sum_{i=1}^{p} \deg^2 m_i.$$

Sur Ω_n, $\deg^2 v_0 = (n-1)^2$ est constant, $\deg^2 v_p$ est un coût terminal et la dernière somme est un coût additif à croissance intermédiaire. En admettant le théorème 10, nous obtenons que le paramètre $\sum \deg m_i \deg v_i$ est un coût concentré de variance d'ordre au plus linéaire (puisqu'il est la somme de paramètres de variances au plus linéaires). Pour démontrer que la décomposition 2.4 suit le principe de décomposition, il reste à prouver que le paramètre $\sum_{i=1}^{p} \deg v_{i-1}$ suit une loi limite gaussienne avec une variance adéquate. C'est l'objet du théorème suivant.

Théorème 11 *Le paramètre N le paramètre définit sur Ω par*

$$N(v_0, v_1) = \sum_{i=1}^{p} \deg v_{i-1}.$$

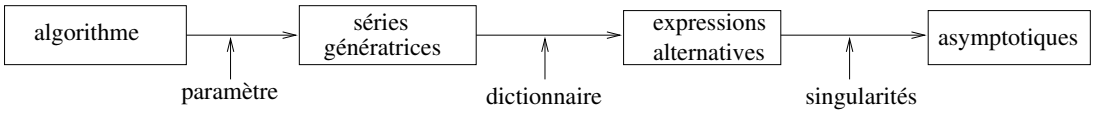
admet une loi limite gaussienne d'espérance, de variance et de vitesse de convergence

$$\mathrm{E}_n[N] = \frac{q-1}{2q} \cdot n^2 + O(n), \qquad \mathrm{V}_n[N] = \frac{q-1}{3q^2} \cdot n^3 + O(n^2), \qquad r_n = O(n^{-1/2}).$$

Ce théorème, les théorèmes 9et 10 et le principe de décomposition implique la loi limite gaussienne pour la complexité binaire classique.

2.3 Combinatoire analytique

L'étude des complexités binaires se ramène à l'analyse probabiliste des coûts additifs à croissance modérée ou intermédiaire, aux coûts terminaux ainsi qu'au coût N. Les relations additives entre les continuants et le degrés des quotients font que la combinatoire analytique est bien adaptée à l'étude de ces coûts. A la première étape, une ou plusieurs séries génératrices sont associées au paramètre étudié. Ces séries renferment toutes les informations du paramètre pour toutes les entrées possibles. Ensuite, les moments du paramètre sont reliés aux coefficients de la série. L'objectif est donc d'extraire ces coefficients. L'extraction des coefficients suit toujours le même principe : c'est la position et la nature des singularités dominantes qui déterminent l'asymptotique des coefficients et de fait, l'asymptotique des paramètres. La difficulté est de trouver une expression alternative des séries qui met en évidence les singularités. Des dictionnaires existent et transfèrent certaines propriétés de décomposition des données en une décomposition sur les séries.



Nous expliquons maintenant les idées principales de cette démarche. Le contexte des polynômes fait que nous abordons peut-être la partie la plus simple de cette discipline. Pour une introduction plus complète, nous renvoyons au livre de Flajolet et Sedgewick [FS99].

2.3.1 Séries génératrices ordinaires

Dans cette section, Ω désigne un ensemble muni d'une fonction de taille $\|\cdot\| : \Omega \to \mathbb{N}$ et Ω_n est l'ensemble des éléments de Ω de taille n. Dans toute la suite, R est un paramètre sur Ω.
La série génératrice bivariée associée à R est définie par

$$S_R(z,w) = \sum_{x\in\Omega} e^{wR(x)} z^{\|x\|},$$

où la variable z marque la taille de l'entrée et w marque le paramètre. La série $S_R^{[k]}$ du moment d'ordre k est la dérivée k^e de $S_R(z,w)$ en $w=0$,

$$S_R^{[k]}(z) := \frac{d^k S_R}{d_w^k}(z,0) = \sum_{x\in\Omega} R(x)^k z^{\|x\|}.$$

Comme son nom l'indique, la série du moment d'ordre k est adaptée à l'analyse du k^e moment $\mathrm{E}_n[R^k]$ de R. En effet, avec la distribution uniforme sur Ω_n, le k^e moment satisfait

$$\mathrm{E}_n[R^k] := \frac{1}{|\Omega_n|}\sum_{x\in\Omega_n} R(x)^k = \frac{[z^n]S_R^{[k]}(z)}{|\Omega_n|}.$$

Ici $[z^n]f$ signifie le coefficient de z^n dans f.
Un outil essentiel pour l'analyse en distribution de R est la suite de fonctions caractéristiques $\phi_{R,n}$ associée au paramètre R et définie par

$$\phi_{R,n}(it) := \mathrm{E}_n[e^{itR}].$$

Une loi de probabilité est complètement définie par sa fonction caractéristique et le Théorème des Quasi-puissances de Hwang [Hwa94, Hwa98] donne une condition suffisante sur la suite de fonctions caractéristiques (ou plutôt sur $\phi_{R,n}(w)$ afin que R admette une loi limite gaussienne. De plus, si une formule explicite de la fonction caractéristique est connue, les dérivations successives de $\phi_{R,n}$ conduisent à des formules explicites de tous les moments,

$$\mathrm{E}_n[R^k] = \frac{\partial^k}{\partial w^k}\phi_{R,n}(w)|_{w=0}.$$

De manière analogue au moment d'ordre k, les fonctions caractéristiques s'expriment avec les coefficients de la série bivariée et satisfont

$$\phi_{R,n}(it) = \frac{[z^{n-1}]S_R(z,it)}{|\Omega_n|} = \frac{[z^{n-1}]S_R(z,it)}{[z^{n-1}]S_R(z,0)}.$$

Que cela soit pour les moments ou pour la loi limite, nous sommes amenés à extraire les coefficients des séries. L'extracteur pour les séries ordinaires est le théorème de Cauchy qui relie les singularités dominantes à l'asymptotique des coefficients. Pour localiser les singularités dominantes, il est nécessaire de trouver une formule alternative aux séries. Ces formules s'obtiennent avec des dictionnaires.

2.3.2 Dictionnaires sur les séries génératrices

Les dictionnaires transfèrent une décomposition sur les entrées en une décomposition sur les séries génératrices. Il existe un dictionnaire pour chaque type de séries génératrices (ordinaire, exponentielle, de Dirichlet,...). Nous présentons celui des séries ordinaires.

Pour $i = 1, 2$, nous posons Ω_i un ensemble muni d'une fonction de taille $\|\cdot\|_i$ et R_i un paramètre sur Ω_i. Nous décrivons maintenant un ensemble de décompositions de Ω en fonction des Ω_i ainsi que les décompositions associées sur les séries.
Union disjointe. Si Ω est l'union disjointe de Ω_1 et Ω_2, et si la fonction de taille et le paramètre R vérifient

$$\|v\| = \begin{cases} \|v\|_1 & \text{si } v \in \Omega_1 \\ \|v\|_2 & \text{si } v \in \Omega_2 \end{cases} \qquad R(v) = \begin{cases} R_1(v) & \text{si } v \in \Omega_1 \\ R_2(v) & \text{si } v \in \Omega_2 \end{cases}$$

alors les séries génératrices $S_R(z,w)$ et $S_R^{[k]}(z)$ sont la somme des séries génératrices de R_1 sur Ω_1 et R_2 sur Ω_2,

$$S_R(z,w) = S_{R_1}(z,w) + S_{R_2}(z,w), \qquad S_R^{[k]}(z) = S_{R_1}^{[k]}(z,w) + S_{R_2}^{[k]}(z,w),$$

Ce premier résultat montre qu'une union disjointe se traduit par une somme sur les séries.
Produit cartésien. Si Ω est (en bijection avec) le produit cartésien de $\Omega_1 \times \Omega_2$, et si la fonction de taille et le paramètre R sont additifs, c'est à dire s'ils vérifient

$$\|v\| = \|v_1\|_1 + \|v_2\|_2, \qquad R(v) = R_1(v_1) + R_2(v_1) \qquad \text{dès que } v = (v_1, v_2)$$

alors la série génératrice bivariée $S_R(z,w)$ est le produit des séries génératrices bivariées de R_1 sur Ω_1 et R_2 sur Ω_2,

$$S_R(z,w) = S_{R_1}(z,w)S_{R_2}(z,w)$$

Si le paramètre R n'est plus additif mais multiplicatif, i.e. s'il vérifie

$$R(v) = R_1(v_1) \times R_2(v_1) \qquad \text{dès que } v = (v_1, v_2),$$

Ω	taille $\|v\|$	coût R	Série génératrice
$\Omega_1 \cup \Omega_2$	$\|v\| = \|v\|_1$ ou $\|v\|_2$	$R(v) = R_1(v)$ ou $R_1(v)$	$S_R = S_{R_1} + S_{R_2}$
$\Omega_1 \times \Omega_2$	$\|v\| = \|v_1\|_1 + \|v_1\|_2$	$R(v) = R_1(v_1).R_2(v_2)$	$S_{R_1}(z)S_{R_2}(z)$
		$R(v) = R_1(v_1) + R_2(v_2)$	$S_{R_1}(z,w)S_{R_2}(z,w)$
$\Omega_1^\star$	$\|v\| = \|v_1\|_1 + \ldots + \|v_n\|_1$	$R(v) = R_1(v_1)\ldots R_1(v_n)$	$S_R^{[k]}(z) = (1 - S_{R_1}^{[k]}(z))^{-1}$
		$R(v) = R_1(v_1) + \ldots + R_1(v_k)$	$S_R(z,w) = (1 - S_{R_1}(z,w))^{-1}$

FIG. 2.1 – Dictionnaire pour les séries génératrices

alors la série du moment d'ordre k de R est encore un produit :

$$S_R^{[k]}(z) = S_{R_1}^{[k]}(z)S_{R_2}^{[k]}(z).$$

Dans tous les cas, un produit cartésien se traduit en un produit.
Suites finies. Nous terminons avec une dernière décomposition qui est un mélange des deux précédentes. Supposons que Ω est (en bijection avec) l'ensemble $\Omega_1^\star$ des suites finies (ou vides) sur Ω_1, c'est à dire

$$\Omega = \bigcup_{n\geq 0} \Omega_1^n.$$

Si la fonction de taille et le paramètre R sont additifs i.e. s'ils vérifient

$$\|v\| = \|v_1\|_1 + \ldots + \|v_n\|_1, \qquad R(v) = R_1(v_1) + \ldots + R_1(v_n) \qquad \text{dès que } v = (v_1, \ldots, v_n)$$

alors en appliquant les deux décompositions précédentes, la série génératrice bivariée satisfait

$$S_R(z,w) = \sum_{n\geq 0} S_{R_1}(z,w)^n = \frac{1}{1 - S_{R_1}(z,w)}.$$

De la même manière, si le coût R est multiplicatif en R_1, nous obtenons

$$S_R^{[k]}(z) = \sum_{n\geq 0} S_{R_1}^{[k]}(z)^n = \frac{1}{1 - S_{R_1}^{[k]}(z)}.$$

Les suites finies (ou séquences) se transforme en quasi-inverse sur les séries.

Lors des analyses, nous exploiterons souvent la bijection entre les entrées Ω et un ensemble du type $\mathcal{G}^\star \times \mathcal{U}$ avec un paramètre R de type additif. La série $S_R(z,w)$ est alors de la forme

$$S_R(z,w) = \frac{1}{1 - G(z,w)} U(z,w),$$

où G et U sont les séries génératrices sur les ensembles $\mathcal{G}$ et $\mathcal{U}$. Le tableau 2.1 résume toutes les décompositions.

2.3.3 Extraction des coefficients

L'étape précédente se préoccupait de trouver une formule alternative aux séries génératrices en utilisant un dictionnaire. Les grandeurs étudiées font intervenir les coefficients de ces séries et l'extraction des coefficients est alors indispensable. Il existe un principe invariant pour l'extraction

de coefficients : *la nature et la localisation des singularités dominantes déterminent la croissance des coefficients.*

Au cours de cette thèse, les séries auront toujours un unique pôle dominant. Pour extraire les coefficients, la formule de Cauchy appliquée à un cercle contenant le pôle dominant est alors suffisante. Cependant, il arrive que le *pôle dominant* (on parle alors de singularité) soit l'extrémité d'une demi-droite (ex : 1 avec la fonction $f(z) = -\log(1-z)$). Dans ce cas, la formule de Cauchy avec des contours de Hankel mènent à l'asymptotique des coefficients. Il existe toute une zoologie de contours qui correspondent chacun à des contextes différents et mènent à des asymptotiques différentes. Nous ne les abordons pas ici. La proposition suivante résume le cadre d'application des extractions que nous rencontrerons.

Proposition 4 *Soit $A(z)$ une série génératrice définie sur un disque centré D_r, de rayon r et contenant $1/\sigma$. Si $A(1/\sigma) \neq 0$, alors le coefficient de z^n dans la série $f(z)$ avec*

$$f(z) = \frac{A(z)}{(1 - z\sigma)^{k+1}}$$

vérifie l'asymptotique

$$[z^n]f(z) = A(1/\sigma)\binom{n+k}{k}\sigma^n(1 + O(n^{-1}))$$

Il est tout à fait possible d'obtenir plus de termes dans le développement asymptotiques. Les termes supplémentaires font intervenir les dérivées de la fonction A en $z = \sigma$. Nous n'en avons toutefois pas besoin.

2.3.4 Loi limite gaussienne

Les moments se déduisent facilement de l'étape d'extraction. En revanche, pour l'analyse en distribution, il est nécessaire de transférer les renseignements sur la fonction caractéristique $\phi_{R,n}$ pour en déduire la loi limite.

Le théorème des quasi-puissances de Hwang [Hwa94] donne des propriétés suffisantes sur la série génératrice des moments $\phi_{R,n}(w) = \mathrm{E}_n[e^{wR}]$ pour que le paramètre R admette une loi limite gaussienne.

Théorème H (Quasi-puissances de Hwang) *Supposons que la série des moments $\mathrm{E}_n[e^{wR}]$ d'un paramètre R sur une suite d'espaces probabilistes $(\Omega_n, \mathbb{P}_n)$ soit analytique en w dans un voisinage $\mathcal{W}$ de $w = 0$ et vérifie*

$$\mathrm{E}_n[e^{wR}] = \exp[\beta_n U(w) + V(w)]\left(1 + O(\kappa_n^{-1})\right)$$

avec $\beta_n, \kappa_n \to \infty$, $U(w)$, $V(w)$ analytiques sur $\mathcal{W}$ et l'erreur dans le O uniforme sur $\mathcal{W}$. Supposons de plus que $U''(w) \neq 0$. Alors, le paramètre R admet une loi limite gaussienne de paramètres

$$\mathrm{E}_n[R] = \beta_n U'(0) + V'(0) + O(\kappa_n^{-1}), \quad \mathrm{V}_n[R] = \beta_n U''(0) + V''(0) + O(\kappa_n^{-1}), \quad r_n = O(\kappa_n^{-1} + \beta_n^{-1/2}).$$

La preuve de ce résultat utilise l'inégalité de Berry-Esseen qui relie les fonctions caractéristiques aux fonctions de répartition tout en mettant en évidence une vitesse de convergence.
La thèse de Hwang [Hwa94] regroupe de nombreux exemples où le théorème des quasi-puissances s'applique. Le livre de Flajolet et Sedgewick reprend ces exemples et bien d'autres encore. Ils

exhibent également une méthode générale qui à partir des séries génératrices bivariées permet de déduire la loi limite gaussienne (voir théorème IX.8 de [FS99]). Dans notre contexte, cette méthode s'exprime de la manière suivante.

Théorème FS *Soit R un paramètre sur l'ensemble Ω (muni d'une fonction de taille $\|\cdot\|$) et posons $S_R(z,w)$ la série génératrice bivariée de R,*

$$S_R(z,w) = \sum_{x\in\Omega} e^{wR(x)} z^{\|x\|}.$$

Supposons que $S_R(z,w)$ s'écrive sous la forme

$$S_R(z,w) = \frac{A(z,w)}{1-B(z,w)}$$

où les fonctions A et B vérifient :

1. *$A(z,w)$ et $B(z,w)$ sont analytiques pour z dans un disque centré D_r de rayon r et w dans un voisinage complexe $\mathcal{W}$ de 0.*
2. *Il existe un unique $\sigma \in D_r$ tel que $B(\sigma,0) = 1$ et $A(\sigma,0) \neq 0$. En particulier σ est un unique pôle simple de $S_R(z,0)$.*
3. *Les dérivées premières de B en σ sont non-nulles, i.e., $\partial_z B(\sigma,0)\cdot\partial_w B(\sigma,0) \neq 0$, ce qui assure l'existence d'une fonction non constante $\sigma(w)$ telle que $B(\sigma(w),w) = 1$ et $\sigma(0) = \sigma$,*
4. *Condition de variation : la fonction $\Sigma(w) = \log\sigma(w)$ admet une dérivée seconde non nulle,i.e., $\Sigma''(0) \neq 0$.*

Alors, le paramètre R admet une loi limite gaussienne d'espérance, de variance et de vitesse de convergence

$$\mathrm{E}_n[R] = -\Sigma'(0)\cdot n + O(1), \qquad \mathrm{V}_n[R] = -\Sigma''(0)\cdot n + O(1), \qquad r_n = O(n^{-1/2}).$$

Le théorème des quasi-puissances de Hwang est un peu plus précis en fait puisqu'il donne le développement de l'espérance et la variance jusqu'aux termes constants [Hwa94]. Les deux premiers points assurent l'existence d'un unique pôle simple σ dans le disque D_r. Par perturbation analytique, la troisième condition implique que la série $S_R(z,w)$ admet un unique pôle simple dans D_r en $z = \sigma(w)$ pour w dans un voisinage de 0. Dans ce cas, la série des moments $\mathrm{E}_n[e^{wR}]$ satisfait

$$\mathrm{E}_n[e^{wR}] = \frac{[z^{n-1}]S_R(z,w)}{[z^{n-1}]S_R(z,0)} = \exp(n(\log\sigma - \log\sigma(w)) + \log K(w))\,(1+O(\Theta^n))$$

où $\Theta < 1$ et $K(w) = A(\sigma(w),w)/(\sigma(w)B'_z(\sigma(w),w))$. Le théorème des quasi-puissances s'applique alors puisque $\Sigma''(0) \neq 0$ ce qui permet de conclure sur la loi gaussienne.

2.4 Analyse des coûts principaux

Nous disposons maintenant de tous les outils pour étudier les familles de coûts introduits aux théorèmes 9 et 11. Nous commençons par l'analyse des coûts à croissance modérée pour lesquels nous désirons montrer la loi limite gaussienne. Nous étudions ensuite le coût principal N et la taille du continuant à une fraction de l'exécution.

2.4.1 Séries liées aux ensembles d'intérêts

L'algorithme d'Euclide sur les polynômes permet de construire la bijection entre Ω et l'ensemble $\mathcal{G}^\star \times \mathcal{U}$ où $\mathcal{G}$ est l'ensemble des polynômes de degré au moins 1 (i.e. l'ensemble des quotients possibles) et $\mathcal{U}$ est l'ensemble des polynômes unitaires (i.e. l'ensemble des pgcd possibles),

$$\Omega \approx \mathcal{G}^\star \times \mathcal{U}, \quad \mathcal{G} = \{m \in \mathbb{F}_q[X];\ \deg m \geq 1\}, \quad \mathcal{G}^\star = \cup_{n\geq 0}\mathcal{G}^n, \quad \mathcal{U} = \{v \in \mathbb{F}_q[X];\ v \text{ unitaire}\}. \tag{2.5}$$

Remarquons que sur les entiers, nous avons le même type de bijection (voir formule 3.2). C'est à partir de cette décomposition et du dictionnaire que l'analyse des différents paramètres s'effectuera. Pour que cette décomposition soit utile, il faut déterminer les séries génératrices sur les ensembles $\mathcal{G}$ et $\mathcal{U}$.

Dans toute la suite, nous fixons c un coût élémentaire défini sur les ensembles $\mathcal{G}$ et $\mathcal{U}$. La série génératrice bivariée liée au coût c et à l'ensemble $\mathcal{G}$ (resp. $\mathcal{U}$) est donnée par

$$G(z,w) = \sum_{m\in\mathcal{G}} e^{wc(m)} z^{\deg m}, \qquad U(z,w) = \sum_{m\in\mathcal{U}} e^{wc(m)} z^{\deg m}.$$

Pour $w = 0$, ces séries sont indépendantes de c. Elle sont respectivement notées $G(z)$ et $U(z)$ et vérifient,

$$G(z) = \frac{(q-1)qz}{1-qz}, \qquad U(z) = \frac{1}{1-qz}.$$

Les quasi-inverses $(1-G(z,w))^{-1}$ et $(1-G(z))^{-1}$ joueront un rôle essentiel puisque ce sont eux qui apporteront les pôles dominants. En particulier, $(1-G(z))^{-1}$ vérifie

$$(1-G(z))^{-1} = \frac{1}{1-G(z)} = \frac{1-qz}{1-q^2z}$$

et admet un unique pôle simple dominant en $z = 1/q^2$.

2.4.2 Coûts à croissance modérée

Dans cette section, nous fixons c un coût élémentaire sur les quotients et nous notons C le coût additif à croissance modérée associé,

$$C(v_0, v_1) = \sum_{i=1}^{p} c(m_i).$$

Nous souhaitons montrer que C admet une loi limite gaussienne (théorème 9). L'ensemble Ω est en bijection avec $\mathcal{G}^\star \times \mathcal{U}$. Le paramètre C est additif en c et en appliquant le dictionnaire, la série génératrice bivariée de C satisfait

$$S_C(z,w) = \frac{U(z)}{1-G(z,w)} = \frac{1}{1-qz}\frac{1}{1-G(z,w)}, \tag{2.6}$$

où $G(z,w)$ est la série génératrice bivariée sur l'ensemble $\mathcal{G}$ associé au coût c.
Cette formule est bien entendue valable pour tout coût additif. Elle est à rapprocher avec celle des coûts additifs sur les entiers (voir formule 3.14). Nous montrons maintenant que la série $S_C(z,w)$ satisfait les conditions du théorème FS.

La série $U(z)$ est analytique sur le disque centré $D_{1/q}$ de rayon $1/q$. Comme $c(m) = O(\ell(m))$, pour tout $\epsilon > 0$, il existe un voisinage $\mathcal{W}$ de 0 et un disque centré D_ϵ de rayon $(1/q) - \epsilon$ tels que

la série bivariée $G(z,w)$ est analytique sur $D_\epsilon \times \mathcal{W}$. Pour $\sigma = 1/q^2$, $G(\sigma,0) = 1$ et $G(\sigma)U(\sigma) \neq 0$. En particulier, σ est l'unique pôle simple de $S_R(z,0)$ dans D_ϵ. Les dérivées partielles par rapport à z et w de $G(z,w)$ sont non-nulles (et mêmes strictement positives) dès que le coût élémentaire c est non nul. Il existe alors une fonction $\sigma(w)$ définie sur un voisinage de $w = 0$ telle que $G(\sigma(w),w) = 1$ et $z = \sigma(w)$ est l'unique pôle de $S_C(z,w)$ dans le disque D_ϵ. Pour utiliser le théorème FS, il reste à prouver que $\Sigma(w) = \log\sigma(w)$ admet une dérivée seconde non-nulle lorsque le coût n'est pas linéaire. En supposant ce dernier point, nous obtenons la loi limite gaussienne des coûts additifs à croissance modérée.

Le lemme suivant donne une condition nécessaire et suffisante pour que la dérivée seconde $\Sigma''(0)$ soit nulle.

Lemme 2 *Soit $\sigma(w)$ la fonction définie par $G(\sigma(w),w) = 1$ pour w dans un voisinage de 0 et Σ la fonction $\Sigma = \log\sigma$. Alors, $\Sigma''(0) = 0$ si et seulement si C est un coût additif linéaire.*

Preuve. La série $G(\sigma(w),w)$ s'écrit sous la forme

$$G(\sigma(w),w) = \sum_{m\in\mathcal{G}} e^{f_m(w)} \qquad \text{avec} \qquad f_m(w) = wc(m) + \Sigma(w)\deg m.$$

Par définition de $\sigma(w)$, $G(\sigma(w),w) = 1$ et en dérivant deux fois par rapport à w, nous obtenons

$$0 = \sum_{m\in\mathcal{G}} (f_m''(w) + f_m'(w)^2)e^{f_m(w)} \tag{2.7}$$

Si $\Sigma''(0) = 0$, alors $f_m''(0) = \Sigma''(0)\deg m = 0$ pour tout m dans $\mathcal{G}$ et l'égalité 2.7 montre que $f_m'(0) = 0$ pour tout $m \in \mathcal{G}$. Comme $f_m'(0) = c(m) + \Sigma'(0)\deg m$, $c(m) = -\Sigma'(0)\deg m$ et est linéaire en $\deg m$. Vérifions tout de même que $\Sigma'(0)$ n'est pas nul. En dérivant une fois la relation 2.7, nous obtenons

$$\Sigma'(0) = -q^2\frac{\partial_w G(q^{-2},0)}{\partial_z G(q^{-2},0)}$$

qui est strictement négatif si c est non nul.
Réciproquement, si c est de la forme $c(m) = \alpha\cdot\deg m$, la série $G(z,w)$ admet une formule explicite :

$$G(z,w) = \frac{(q-1)qze^{\alpha w}}{1-qze^{\alpha w}}.$$

La fonction $\sigma(w)$ est alors donnée par $\sigma(w) = 1/(q^2e^{\alpha w})$ soit $\Sigma(w) = -2\log q - w\alpha$ et $\Sigma''(0) = 0$. Ceci finit de montrer l'équivalence. ■

Comme conséquence du lemme, nous obtenons que les coûts à croissance modérée non-linéaires satisfont une loi limite gaussienne avec une espérance et une variance linéaires en n.
Ce résultat est l'équivalent sur les polynômes des résultats de Baladi-Vallée ([BV05, BV04] et théorème BV) dont la preuve se limite aux entiers. La partie principale de la décomposition de la complexité binaire étendue est de la forme $(n+2)Y_0$ où Y_0 est le coût additif à croissance modérée associé au coût élémentaire $c = \ell$. Ainsi, $(n+2)Y_0$ suit une loi limite gaussienne d'espérance quadratique et de variance cubique en n.

2.4.3 Un cas particulier de coût modéré

Le coût à croissance modérée Y_0 qui intervient dans la décomposition de la complexité binaire étendue est associé au coût élémentaire $c(m) = \ell(m)$. La série génératrice $G(z, w)$ est vérifie alors

$$G(z,w) = \sum_{m\in\mathcal{G}} z^{\deg m} e^{w\ell(m)} = \frac{(q-1)qze^{2w}}{1-qze^w}.$$

L'équation $G(\sigma(w), w) = 1$ a pour solution

$$\sigma(w) = \frac{1}{qe^w((q-1)e^w+1)}$$

et les deux premières dérivées de $\Sigma = \log\sigma$ en 0 sont

$$\Sigma'(0) = -\frac{2q-1}{q}, \qquad \text{et} \qquad \Sigma''(0) = -\frac{q-1}{q^2}.$$

Le coût Y_0 suit donc une loi gaussienne de paramètres

$$\mathrm{E}_n[Y_0] = \frac{2q-1}{q}\cdot n + O(1), \qquad \mathrm{V}_n[Y_0] = \frac{q-1}{q^2}\cdot n + O(1), \qquad r_n[Y_0] = O(n^{-1/2})$$

et si l'on admet le théorème 10 sur les coûts concentrés, la complexité binaire étendue admet une loi limite gaussienne de paramètres

$$\mathrm{E}_n[E] = \frac{2q-1}{q}\cdot n^2 + O(n), \qquad \mathrm{V}_n[E] = \frac{q-1}{q^2}\cdot n^3 + O(n^2), \qquad r_n[E] = O(n^{-1/2}).$$

2.4.4 Le coût N est gaussien

Le coût principal dans la décomposition de la complexité binaire classique fait intervenir le coût N et nous voulons montrer qu'il admet une loi limite gaussienne. Par définition, la forme générale de N est

$$N(v_0, v_1) = \sum_{i=1}^{p} \deg v_{i-1}.$$

En utilisant la formule 2.1 sur les degrés des v_i et en inversant les deux signes sommes, une formule alternative de N en fonction des degrés des quotients est

$$N(v_0, v_1) = \sum_{u=1}^{p+1} u \deg m_u, \quad \text{avec} \quad m_{p+1} = v_p.$$

Nous utilisons une fois de plus la bijection $\Omega \approx \mathcal{G}^\star \times \mathcal{U}$ et le dictionnaire pour exprimer la série génératrice bivariée de N. La bijection se récrit

$$\Omega \approx \mathcal{U} + \mathcal{G}\times\mathcal{U} + \mathcal{G}\times\mathcal{G}\times\mathcal{U} + \dots.$$

Nous notons c_j le coût $c_j(m) = j\cdot\deg m$. Pour un ensemble de la forme $\mathcal{G}\times\ldots\times\mathcal{G}\times\mathcal{U} = \mathcal{G}^p\times\mathcal{U}$, le coût sur le j^e ensemble $\mathcal{G}$ est fixé à c_j et le coût sur $\mathcal{U}$ est fixé à c_{p+1}. Ainsi défini, les séries génératrices $G_j(z, w)$ et $U_j(z, w)$ associées au coût c_j sur les ensembles $\mathcal{G}$ et $\mathcal{U}$ satisfont

$$U_j(z,w) = U(z, j\cdot w) = \frac{1}{1-qze^{jw}}, \qquad G_j(z,w) = G(z, j\cdot w) = \frac{(q-1)qze^{jw}}{1-qze^{jw}}.$$

Le dictionnaire s'applique alors au paramètre N et nous obtenons l'expression

$$\begin{aligned} S_N(z,w) &= U(z,w) + \sum_{p\geq 1} G(z,w)G(z,2w)\dots G(z,pw)U(z,(p+1)w) \\ &= \Phi(-(q-1),-qz,e^w) \end{aligned}$$

où la fonction Φ est définie par

$$\Phi(u,\xi,t) = \sum_{p\geq 0} t^{p(p+1)/2} \frac{\xi^p u^p}{\prod_{j=1}^{p+1}(1+\xi t^j)}.$$

Φ satisfait l'identité $\Phi(u,\xi,t) = 1 - t\xi(1-u)\Phi(tu,\xi,t)$. Cette identité relève du domaine des q-calculs et mène par induction à la formule

$$\Phi(u,\xi,t) = 1 + \sum_{n\geq 1}(-1)^n (t\xi)^n \prod_{j=0}^{n-1}(1-ut^j).$$

Appliquée à la série $S_N(z,w)$, on obtient une formule alternative qui fait clairement apparaître les coefficients,

$$S_N(z,w) = 1 + \sum_{n\geq 1} z^n (qe^w)^n \prod_{j=0}^{n-1}(1+(q-1)e^{wj}).$$

La série génératrice des moments du coût N sur Ω_n est alors égale à

$$\mathrm{E}_n[\exp(wN)] = \frac{[z^{n-1}]S_N(z,w)}{|\Omega_n|} = e^{(n-1)w} \prod_{j=0}^{n-2}\left[\frac{1+(q-1)e^{wj}}{q}\right].$$

Le paramètre N suit asymptotiquement une loi gaussienne dès que le paramètre $\overline{N} = N/(n-1)$ satisfait aussi une loi gaussienne. Maintenant, la série génératrice des moments de $\overline{N}$ satisfait $\mathrm{E}_n[\exp(w\overline{N})] = \exp \delta_n(w)$ avec

$$\delta_n(w) = w + \sum_{j=0}^{n-2} f_w(\frac{j}{n-1})$$

où f_w est la fonction

$$f_w(t) = \log \frac{1+(q-1)e^{wt}}{q}.$$

La formule d'Euler-Mac-Laurin transforme une somme en une intégrale et appliquée à $\delta_n(w)$, on obtient

$$\delta_n(w) = n.U(w) + V(w) + O(n^{-1})$$

où V est une fonction analytique indépendante de n dans un voisinage de 0 et

$$U(w) = \int_0^1 f(t)dt = \int_0^1 \log \frac{1+(q-1)e^{wt}}{q}.$$

La série des moments vérifie alors

$$\mathrm{E}_n[e^{w\overline{N}}] = \exp(n\cdot U(w) + V(w))(1+O(n^{-1}))$$

avec un terme d'erreur uniforme en w. Le théorème des quasi-puissances de Hwang s'applique à $\overline{N}$ et en multipliant par $(n-1)$, entraîne la loi limite gaussienne de N de paramètres

$$\mathrm{E}_n[N] = U'(0)\cdot n^2 + O(n) = \frac{q-1}{2q}\cdot n^2 + O(n), \quad \mathrm{V}_n[N] = U''(0)\cdot n^3 + O(n^2) = \frac{q-1}{3q^2}\cdot n^3 + O(n^2),$$

et de vitesse de convergence $r_n[N] = O(n^{-1/2})$.

En admettant le théorème 9, nous pouvons calculer les constantes dominantes qui interviennent dans la loi limite de la complexité binaire classique. Pour l'espérance, il faut aussi tenir compte de $\frac{1}{2}\deg^2 v_0/2 = (n-1)^2/2$ qui intervient dans la décomposition alors que pour v_n, seule la variance du coût gaussien N intervient. La complexité binaire classique B suit alors une loi limite gaussienne de paramètres

$$\mathrm{E}_n[B] = \frac{2q-1}{2q}\cdot n^2 + O(n), \qquad \mathrm{V}_n[B] = \frac{q-1}{3q^2}\cdot n^3 + O(n^2), \qquad r_n[B] = O(n^{-1/2}).$$

La série $S_N(z,w)$ satisfait une équation fonctionnelle de la forme

$$S_N(z,w) = U(e^w z, w) + G(e^w z, w) S_N(e^w z, w).$$

Ce type d'équation se rencontre par exemple lors de l'analyse de paramètres de type longueur de cheminement.

2.4.5 Continuant à une fraction de l'exécution

La série génératrice de $L^{[\delta]}$ est notée simplement $S_{[\delta]}$. Avec la bijection $\Omega \approx \mathcal{G}^\star \times \mathcal{U}$, les relations sur les degrés et les dictionnaires, la série bivariée $S_{[\delta]}$ admet comme forme alternative

$$S_{[\delta]}(z,w) = U(ze^w)\cdot \sum_{p\geq 0} G(z)^{\lfloor \delta p\rfloor} G(ze^w)^{p-\lfloor \delta p\rfloor}.$$

Maintenant, si δ est le rationnel $\delta = c/(c+d)$, alors en écrivant $p = k(c+d)+j$, la série se récrit

$$S_{[\delta]}(z,w) = U(ze^w)\cdot \left[\sum_{j=0}^{c+d-1} G(ze^w)^{j-\lfloor \delta j\rfloor} G(z)^{\lfloor \delta j\rfloor}\right] \cdot \frac{1}{1-G(z)^c G(ze^w)^d}. \tag{2.8}$$

Il est alors facile de vérifier que $S_{[\delta]}$ vérifie toutes les conditions d'application du théorème FS. Notons toutefois une petite difficulté. Pour $w=0$, les valeurs de z pour lesquelles $G(z)^c G(ze^w)^d = G(z)^{c+d}$ vaut 1 sont

$$z = \frac{e^{2ik\pi/(c+d)}}{q-1+qe^{2ik\pi/(c+d)}}, \qquad k\in\mathbb{N}.$$

Le pôle dominant est $z = 1/q^2$ et les pôles sous-dominant s'obtiennent avec $k = \pm 1$. Plus le dénominateur $(c+d)$ de δ est grand, plus la distance entre le pôle dominant et le pôle sous-dominant tend à s'annuler. Le même phénomène apparaît sur les entiers et c'est la raison pour laquelle nous ne pouvons analyser l'algorithme de Knuth-Schönhage avec une multiplication d'ordre plus petit que $O(n^{3/2})$.

En dérivant deux fois la relation $\delta\log G(\sigma(w)) + (1-\delta)\log G(\sigma(w)e^w) = 0$ par rapport à w, les deux premières dérivées de $\sigma(w)$ apparaissent et en prenant $w=0$, nous obtenons

$$\sigma'(0) = \frac{\delta-1}{q^2}, \qquad \sigma''(0) = \frac{(\delta-1)(q(1-\delta)-1)}{q^2(q-1)}.$$

Les deux premières dérivées de $\Sigma(w) = \log \sigma(w)$ vérifient alors

$$\Sigma'(0) = -(1-\delta), \qquad \Sigma''(0) = -\frac{\delta(1-\delta)}{q-1}.$$

Ceci termine la preuve du théorème 5.

2.5 Analyse des coûts concentrés

La partie précédente regroupe l'analyse de tous les coûts principaux des décompositions. Dans cette partie, nous traitons les autres coûts. Nous montrons que les coûts à croissance intermédiaire et les coûts terminaux ont une espérance et une variance d'ordre constant ou linéaire en n. Nous établissons ainsi le théorème 10 et les résultats généraux sur les complexités binaires classique et étendue.

2.5.1 Coûts additifs à croissance intermédiaire

Nous fixons un coût élémentaire c tel que le coût additif associé C est à croissance intermédiaire. Pour les coûts additifs, le dictionnaire s'applique directement. La série génératrice bivariée admet la forme générale

$$S_C(z,w) = \frac{U(z)}{1-G(z,w)} \tag{2.9}$$

où $G(z,w)$ est la série associée au coût élémentaire c. Les séries des moments d'ordre 1 ou 2 sont les dérivées par rapport à w de $S_C(z,w)$ prises en $w=0$, soit

$$S_C^{[1]} = \frac{U \cdot G_{[c]}}{(1-G)^2}, \qquad S_C^{[2]} = \frac{U \cdot G_{[c^2]}}{(1-G)^2} + 2\frac{U \cdot G_{[c]}^2}{(1-G)^3} \tag{2.10}$$

avec $G_{[c^i]}$ les séries

$$G_{[c^i]} = \sum_{m \in \mathcal{G}} c(m)^i z^{\deg m}.$$

Une fois de plus, ces séries sont très similaires à celles obtenues pour les entiers (cf. formule 3.14 et proposition 9). En remplaçant $U(z)$ et $G(z)$ par leurs expressions exactes dans les séries $S_C^{[j]}$, nous obtenons les formules alternatives

$$S_C^{[1]}(z) = \frac{(1-qz)\cdot G_{[c]}(z)}{(1-q^2z)^2}, \qquad S_C^{[2]}(z) = \frac{(1-qz)\cdot G_{[c^2]}(z)}{(1-q^2z)^2} + 2\frac{(1-qz)^2 G_{[c]}^2(z)}{(1-q^2z)^3}.$$

La série $S_C^{[j]}$ admet un pôle d'ordre $j+1$ en $z = q^{-2}$. En appliquant la formule d'extraction 4 puis la normalisation par le cardinal de Ω_n, les deux premiers moments de C vérifient

$$\mathrm{E}_n[C^i] = \frac{[z^{n-1}]S_C^{[i]}(z)}{|\Omega_n|} = \left(\frac{q-1}{q} G_{[c]}(q^{-2}) \cdot n\right)^i (1 + O(n^{-1})).$$

En particulier, le terme en n^2 de la variance s'annule. La variance des coûts additifs à croissance intermédiaire satisfait

$$\mathrm{V}_n[C] = \mathrm{E}_n[C^2] - \mathrm{E}_n[C]^2 = O(n).$$

Ceci établit le premier résultat du théorème 10 concernant les coûts à croissance intermédiaire. Ces derniers sont donc plus concentrés que les parties principales qui sont toutes de variances cubiques.

2.5.2 Coûts terminaux

Les coûts terminaux sont tous les coûts de la forme $O((\ell(m_p)+\ell(v_p))^k)$ pour un entier k fixé. Si T dénote le coût terminal $T(v_0,v_1)=\ell(m_p)+\ell(v_p)$, les coûts terminaux ont tout leur moment d'ordre constant si et seulement si tous les moments de T sont constants. Une fois de plus, le dictionnaire appliqué à T et à la bijection $\Omega = \mathcal{U} + \mathcal{G}^\star \times \mathcal{G} \times \mathcal{U}$ conduit à la série génératrice bivariée

$$S_T(z,w) = \frac{U(z,w)G(z,w)}{1-G(z)} + U(z)$$

avec $G(z,w)$ et $U(z,w)$ sont les séries bivariées associées au coût élémentaire c sur $\mathcal{G}$ et $\mathcal{U}$. Par dérivation, la série du moment d'ordre k satisfait

$$S_T^{[k]}(z) = \frac{1}{1-G(z)}\left(\frac{d}{dw^k}U(z,w)G(z,w)\right)|_{w=0} = \frac{1-qz}{1-q^2z}\left(\frac{d}{dw^k}U(z,w)G(z,w)\right)|_{w=0}.$$

Remarquons encore une fois la similarité des formules entre le cas polynomial et le cas entier (cf. proposition 10).
Comme $U(z,w)$ et $G(z,w)$ sont analytiques dans un disque contenant strictement $1/q^2$ pour w suffisamment petit, il en est de même de la dérivée k^e. Ainsi, les séries $S_T^{[k]}$ admettent un unique pôle simple dominant en $z=1/q^2$. La proposition 4 entraîne que le coefficient de z^n dans $S_T^{[k]}$ satisfait

$$[z^n]S_T^{[k]}(z) = q^{2n}\left(\frac{d}{dw^k}U(z,w)G(z,w)\right)|_{w=0,z=q^{-2}}(1+O(n^{-1})).$$

Avec la normalisation par le cardinal de Ω_n, nous déduisons que tous les moments de T sont d'ordre constant. Ceci démontre le deuxième point du théorème 10.

A ce stade de notre analyse, nous avons complètement démontré les théorèmes 9 et 10. Or ces théorèmes sont suffisants pour valider la décomposition de la complexité binaire de l'algorithme étendu. Le principe de décomposition s'applique et nous obtenons la loi limite gaussienne de E (théorème 4).

2.6 Développements précis des moments des complexités binaires

Jusqu'à présent, nous nous sommes contentés des termes dominants dans les développements asymptotiques des complexités binaires. Dans cette partie, nous montrons comment avec les séries génératrices, nous obtenons une asymptotique plus précise.

2.6.1 Cas de l'algorithme classique

La complexité de l'algorithme classique est donnée par

$$B(v_0,v_1) = \sum_{i=1}^{p} \ell(m_i)\ell(v_i).$$

Nous posons Δ_z l'opérateur qui agit sur les séries génératrices en z par

$$\Delta_z(F(z)) = (zF(z))'.$$

L'intérêt de cet opérateur est qu'il fait apparaître la taille des polynômes. En effet, si F est de la forme

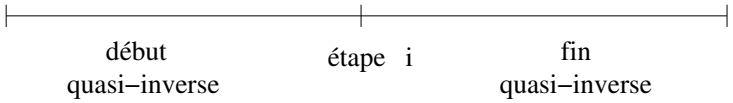
$$F(z) = \sum_m z^{\deg m} \qquad \text{alors} \qquad \Delta_z(F(z)) = \sum_m \ell(m) z^{\deg m},$$

où la somme se fait sur un ensemble de polynômes.

Avec la bijection $\Omega \approx \mathcal{G}^\star \times \mathcal{U}$ et les relations 2.1, la série génératrice du premier moment de B est donnée par

$$S_B^{[1]} = \frac{1}{1-G}\Delta_z(G)\Delta_z(\frac{U}{1-G}) = \frac{1-qz}{1-q^2z} \cdot \frac{q(q-1)(2z-qz^2)}{(1-qz)^2} \cdot \frac{1}{(1-q^2z)^2}. \tag{2.11}$$

Intuitivement, le premier Δ_z génère $\ell(m_i)$ alors que le second génère $\ell(v_i)$. Une autre manière de voir est de découper l'exécution de l'algorithme d'Euclide en trois parties : le début qui correspond au premier quasi-inverse, l'étape i correspondant au $G(z)$ dans le premier Δ_z et la fin de l'exécution qui correspond à la série dans le deuxième Δ_z.



La série $S_B^{[1]}$ s'écrit aussi sous la forme

$$S_B^{[1]}(z) = \frac{A_1(z)}{(1-q^2z)^3}$$

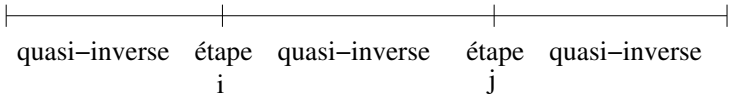
avec A_1 analytique sur le disque de rayon $1/q$ et $A_1(q^{-2}) = (2q-1)/q$. Le pôle dominant $z = 1/q^2$ est d'ordre 3. La proposition 4 s'applique et après normalisation par le cardinal de Ω_n, entraîne que l'espérance est d'ordre quadratique. Une extraction plus fine effectuée avec MAPLE conduit au résultat plus précis

$$\mathrm{E}_n[B] = \frac{2q-1}{2q}n^2 - \frac{2q^2-q+1}{2(q-1)q}n + \frac{q}{(q-1)^2} + O(n^2q^{-n}).$$

Le carré de la complexité binaire satisfait

$$B^2(v_0, v_1) = \sum_{i=1}^{p} \ell(v_i)^2\ell(m_i)^2 + 2\sum_{1\le i<j\le p} \ell(m_i)\ell(v_i)\ell(v_j)\ell(m_j).$$

La série génératrice du premier moment pour la première somme s'obtient de manière similaire à celle obtenue pour B en découpant les exécutions en trois parties. Pour la seconde somme, il faut cette fois découper l'exécution en 5 parties : deux pour les étapes i et j et les trois autres situées avant, au milieu et après i et j.



Toujours en utilisant la bijection et les relations sur les degrés, la série génératrice du second moment de B satisfait

$$\begin{aligned} S_B^{[2]} &= \frac{1}{1-G}\Delta_z^2(G)\Delta_z^2(\frac{U}{1-G}) \\ &\quad +2\frac{1}{1-G}\Delta_z(G)\Delta_z\left(\frac{1}{1-G}\Delta_z(G)\Delta_z(\frac{U}{1-G})\right). \end{aligned} \tag{2.12}$$

La série $S_B^{[2]}$ s'écrit aussi sous la forme

$$S_B^{[2]}(z) = \frac{A_2(z)}{(1-q^2z)^5}$$

avec A_2 analytique sur le disque de rayon $1/q$ et $A_2(q^{-2}) \neq 0$. Le pôle dominant $z = 1/q^2$ est d'ordre 5. La proposition 4 s'applique et après normalisation par le cardinal de Ω_n, entraîne que le second moment est d'ordre $\Theta(n^4)$. Une extraction plus fine effectuée avec MAPLE conduit à un résultat plus précis et par différence, la variance de B satisfait

$$\mathrm{V}_n[B] = \frac{q-1}{3q^2}n^3 + \frac{q^2+2q-1}{2(q-1)q^2}n^2 + \frac{q^4+2q^3+12q^2-4q+1}{6(q-1)^3q^2}n - \frac{q(1+5q+q^2)}{(q-1)^4} + O(n^4q^{-n}). \tag{2.13}$$

2.6.2 Cas de l'algorithme étendu

La complexité de l'algorithme d'Euclide étendu s'écrit sous la forme $E = B + X$ où X est le coût

$$X(v_0, v_1) = \sum_{i=1}^{p-1} \ell(m_i)\ell(a_i).$$

Le coût X est très similaire à la complexité binaire classique B. En particulier, les mêmes idées s'appliquent pour le calcul des moments et les séries génératrice pour les deux premiers moments sont données par

$$\begin{aligned} S_X^{[1]} &= \Delta_z(\frac{1}{1-G})\Delta_z(G)\frac{UG}{1-G} \quad \text{et} \\ S_X^{[2]} &= \Delta_z^2(\frac{1}{1-G})\Delta_z^2(G)\frac{UG}{1-G} \\ &+ 2\Delta_z(\Delta_z(\frac{1}{1-G})\Delta_z(G)\frac{1}{1-G})\Delta_z(G)\frac{UG}{1-G}. \end{aligned}$$

Il faut noter la symétrie avec les séries 2.11 et 2.12 associées à B. Au lieu de dériver *à la fin* pour générer les $\ell(v_i)$, la dérivation se fait au début pour générer les $\ell(a_i)$. La présence de G au coté de U est due au fait que la somme dans la définition de X est limitée à $p-1$. Le calcul de ces séries montre qu'elles sont de la forme

$$S_X^{[i]}(z) = \frac{A_i}{(1-q^2z)^{2i+1}},$$

où A_i est analytique sur le disque de rayon $1/q$ et $A_i(1/q^2) \neq 0$. Combinées avec la proposition 4 et la normalisation par le cardinal de Ω_n, l'espérance et la variance de X satisfont

$$\begin{aligned} \mathrm{E}_n[X] &= \frac{2q-1}{2q}n^2 - \frac{(2q+1)(3q-1)}{2(q-1)q}n + \frac{2q^3+7q^2-4q+1}{q(q-1)^2} + O(n^2q^{-n}) \qquad \text{et} \\ \mathrm{V}_n[X] &= \frac{q-1}{3q^2}n^3 + \frac{15q^3-13q^2+3q-1}{2(q-1)^2q^2}n^2 - \frac{191q^4-50q^3-60q^2+4q-1}{6(q-1)^3q^2}n \\ &+ \frac{27q^5+31q^4-36q^3+6q^2-5q+1}{q^2(q-1)^4} + O(n^4q^{-n}). \end{aligned} \tag{2.14}$$

L'espérance de E est la somme des espérances de X et de B soit

$$\mathrm{E}_n[E] = \frac{2q-1}{q}n^2 - \frac{4q}{q-1}n + \frac{2q^3+8q^2-4q+1}{q(q-1)^2} + O(n^2q^{-n}).$$

La variance de E fait intervenir la covariance de X et de B et vérifie

$$\mathrm{V}_n[E] = \mathrm{V}_n[B] + \mathrm{V}_n[X] + 2\mathrm{cov}_n(X,B) \quad \text{où} \quad \mathrm{cov}_n(X,B) = \mathrm{E}_n[BX] - \mathrm{E}_n[B]\mathrm{E}_n[X].$$

Tous les éléments sont connus sauf l'espérance de $B \cdot X$. Le coût $B \cdot X$ est donnée par la formule

$$\begin{aligned} B(v_0,v_1)\cdot X(v_0,v_1) &= \sum_{i=1}^{p-1} \ell(m_i)^2\ell(v_i)\ell(a_i) + \sum_{1\le j<i\le p} \ell(m_i)\ell(m_j)\ell(v_i)\ell(a_j) \\ &+ \sum_{1\le i<j\le p-1} \ell(m_i)\ell(m_j)\ell(v_i)\ell(a_j). \end{aligned}$$

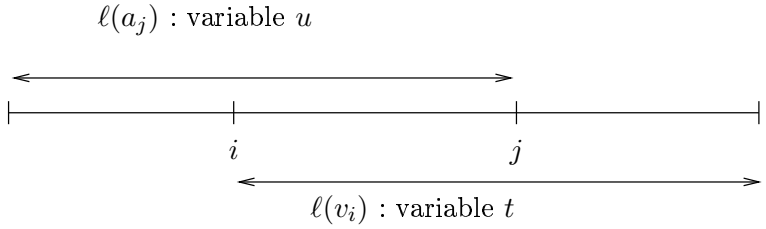
Pour la première somme, il suffit de couper les exécutions en trois parties comme précédemment. La série génératrice pour cette partie est

$$\Delta_z(\frac{1}{1-G})\Delta_z^2(G)\Delta_z(\frac{UG}{1-G}) := \frac{A_3(z)}{(1-q^2z)^4}$$

où A_3 a les mêmes propriétés que A_1 et A_2. Le premier Δ_z génère $\ell(a_i)$, le second génère $\ell(m_i)^2$ et le troisième génère $\ell(v_i)$.
Pour la seconde somme, il suffit de décomposer en cinq parties comme les fois précédentes et la série génératrice associée vérifie alors

$$\Delta_z(\frac{1}{1-G})\Delta_z(G)\frac{1}{1-G}\Delta_z(G)\Delta_z(\frac{U}{1-G}) := \frac{A_4(z)}{(1-q^2z)^5}.$$

où A_4 a les mêmes propriétés que A_1 et A_2. Intuitivement, les deux premiers Δ_z génèrent $\ell(m_j)\ell(a_j)$ alors que les deux derniers génèrent $\ell(m_i)\ell(v_i)$.
Finalement, la dernière somme est plus difficile à traiter. Pour générer $\ell(v_i)$, les étapes $i+1$ jusqu'à p sont nécessaires alors que pour générer $\ell(a_j)$, les étapes 1 à $j-1$ sont utilisées. Si $i<j$, il peut y avoir chevauchement des deux zones et nous ne pouvons utiliser deux fois l'opérateur de dérivation Δ_z sur deux zones non incluses l'une dans l'autre mais qui se chevauchent. L'idée est de faire intervenir des variables intermédiaires u et t marquant les deux zones et ayant exactement le même rôle que z.



Avec cette décomposition, la série génératrice est de la forme

$$\Delta_t\Delta_u\left(\frac{1}{1-G(uz)}\Delta_z(G(uz))\frac{1}{1-G(ztu)}\Delta_z(G(zt))\frac{U(zt)G(zt)}{1-G(zt)}\right)|_{t=u=1} := \frac{A_5(z)}{(1-q^2z)^5}.$$

En combinant toutes ces séries, la série du premier moment de $X \cdot B$ admet un unique pôle d'ordre 5 en $z = 1/q^2$. Après extraction avec la proposition 4 puis normalisation par le cardinal de Ω_n, l'ordre de l'espérance de $X \cdot B$ est $\Theta(n^4)$. Une extraction plus fine avec MAPLE mène au développement précis de $\mathrm{E}_n[X \cdot B]$ et de la covariance de X et de B,

$$\begin{aligned}\mathrm{cov}_n(X, B) &= \frac{q-1}{6q^2}n^3 + \frac{3q^2-2q+1}{2(q-1)q^2}n^2 + \frac{2(8q^4-2q^3-3q^2+4q-1)}{3(q-1)^3q^2}n \\ &\quad -\frac{24q^3+3q^2-7q+1}{(q-1)^4} + O(n^4q^{-n}).\end{aligned} \tag{2.15}$$

La variance de E est maintenant calculable et vérifie

$$\begin{aligned}\mathrm{V}_n[E] &= \frac{q-1}{q^2}n^3 + \frac{11q^3-11q^2+3q-1}{(q-1)^2q^2}n^2 - \frac{21q^4-6q^3-8q^2-4q+1}{(q-1)^3q^2}n \\ &\quad -\frac{22q^5-20q^4+23q^3-1-4q^2+5q}{(q-1)^4q^2} + O(n^4q^{-n}).\end{aligned}$$

2.7 Conclusion

Dans cette partie, nous avons montré que la complexité binaire des algorithmes d'Euclide classique et étendu ainsi que la taille du continuant à une fraction rationnelle de l'exécution admettent toutes une loi limite gaussienne. L'analyse des complexités binaires est basée sur une décomposition en plusieurs famille de coûts plus simples à étudier. Les techniques employées pour analyser ces familles sont celles de la combinatoire analytique et des séries génératrices. L'analyse dans le cas des polynômes est facilitée par le fait que les séries commutent et que tous les paramètres s'expriment en fonction des degrés des quotients. Pour les entiers, nous verrons que les séries s'expriment avec des opérateurs qui ne commutent pas.

L'analyse sur les polynômes est très similaire à celle qui sera faite sur les entiers. Nous retrouverons le principe de décomposition et les mêmes familles de coûts à analyser (excepté le coût N). Les séries génératrices sur les polynômes ont aussi des équivalents en terme d'opérateurs de transfert. Mais les séries génératrices sur les entiers sont d'une part, des séries de Dirichlet qui sont plus difficiles à manipuler. D'autre part, les séries génératrices s'expriment avec des opérateurs et les propriétés analytiques des opérateurs sont plus difficiles à obtenir que celles des séries ordinaires.

Dans ce travail, nous nous sommes intéressés au cas particulier des polynômes à une indéterminé sur un corps fini. Dans l'avenir, il serait intéressant d'aborder l'algorithme d'Euclide pour des polynômes à plusieurs indéterminés sur d'autres types d'ensembles.

Chapitre 3

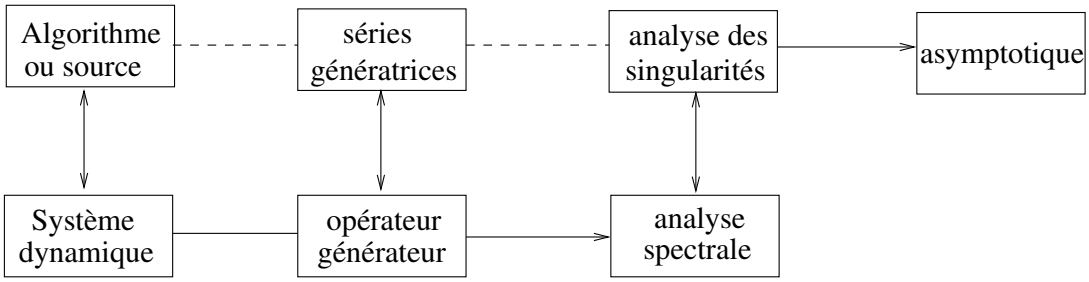
Le cas des entiers

Sommaire

3.1 Introduction

Au chapitre précédent, nous avons analysé la distribution des complexités binaires des algorithmes d'Euclide classique et étendu sur les polynômes. Nous avons également étudié la loi limite des continuants à une fraction de l'exécution. Ce chapitre aborde les mêmes paramètres mais dans le cadre des entiers. La complexité moyenne de l'algorithme de Knuth-Schönhage paramétré $\mathcal{KS}_\alpha$ est aussi traitée à travers l'analyse des algorithmes interrompus. Les mêmes idées que sur les polynômes s'appliquent dans le cadre des entiers. En particulier, nous proposons des décompositions pour la complexité binaire étendue et les continuants qui satisfont le principe de décomposition. Nous sommes alors ramenés à étudier les mêmes familles de coûts mais sur les entiers. La méthode utilisée sur les polynômes était la combinatoire analytique. Cette méthode était adaptée car tous les paramètres s'exprimaient "simplement" en fonction des degrés des quotients et du pgcd. Malheureusement, ce n'est plus le cas sur les entiers et le dictionnaire sur les séries génératrices ne s'applique plus. Il n'est alors plus possible d'obtenir directement une formule alternative des séries.

L'analyse dynamique utilise un chemin détourné pour trouver une forme alternative aux séries.



Tout d'abord, l'algorithme est vu comme un système dynamique agissant sur les données. L'outil classique des systèmes dynamiques pour étudier l'évolution des données est l'opérateur de transfert (ou opérateur de Ruelle) [Rue78] qui dépend d'un paramètre complexe s. L'idée originale de l'analyse dynamique est d'utiliser les opérateurs de transfert (et leurs descendances) pour générer les séries génératrices. Ensuite, les actions de l'algorithme se traduisent à travers des dictionnaires en actions sur les opérateurs (et donc sur les séries). Ensuite, l'analyse des singularités des séries pour extraire les coefficients passe par une analyse des singularités des opérateurs. Il est classique en analyse fonctionnelle de relier les singularités des opérateurs à *leur* spectre. Une propriété fondamentale des opérateurs de transfert est qu'ils admettent une unique valeur propre dominante, notée $\lambda(s)$, isolée du reste du spectre par un saut spectral. Selon la valeur de $\lambda(s)$, les opérateurs générateurs admettent (ou n'admettent pas) un pôle en s. L'étude des singularités des séries se ramène donc à une analyse spectrale. Une fois les singularités déterminées, l'étape d'extraction des coefficients s'applique et nous obtenons les asymptotiques.

L'analyse dynamique est née du mariage entre deux domaines : l'analyse (en moyenne) d'algorithmes et les systèmes dynamiques. Cette méthode, initiée par Brigitte Vallée à Caen, se développe depuis une dizaine d'années et a contribué à résoudre de nouveaux problèmes sur les mots et les algorithmes arithmétiques. Au chapitre 8, nous appliquons pour la première fois les techniques d'analyse dynamique en fouille de données. Pour quelques analyses dynamiques sur les mots, nous renvoyons à [CFV01, BNV01, Bou01, BV02]. Les premières analyses en moyenne des

algorithmes euclidiens étaient soit spécifiques aux algorithmes, soit spécifiques aux paramètres étudiés. L'analyse dynamique s'applique à de nombreux algorithmes ainsi qu'à de nombreux paramètres [Bre76, Val98a, Val00b, Val03, BDV02, DV04, DMDV05]. Les premières analyses dynamiques étaient des analyses en moyenne, mais en 2002, Baladi et Vallée [BV04, BV05] ont effectué une percée significative. Elles ont adapté la méthode pour réaliser la première analyse dynamique en distribution des coûts additifs à croissance modérée. Comme sur les polynômes, cette famille de coûts regroupe plusieurs paramètres naturels comme le nombre de divisions, le nombre de quotients égaux à 1, la taille totale de tous les quotients, etc. Mais ni la complexité binaire, ni le continuant à une fraction de l'exécution ne sont additifs à croissance modérée.

Notre objectif dans ce chapitre est de donner les grandes étapes pour démontrer les théorèmes 2, 3, 5 et 6 qui concernent respectivement la complexité binaire classique, étendue, le continuant à une fraction de l'exécution et les paramètres des algorithmes interrompus. Nous avons vu au premier chapitre que l'analyse des algorithmes interrompus est suffisante pour déterminer la complexité binaire moyenne des algorithmes sous-quadratiques $\mathcal{KS}_\alpha$. C'est la première fois que la distribution de la complexité binaire est traitée. C'est aussi la première fois qu'une analyse en moyenne d'un algorithme sous-quadratique est réalisée.

Plan. Nous suivons les étapes d'une analyse dynamique. La première section présente le système dynamique associé à l'algorithme d'Euclide et les décompositions des complexités binaires. Nous ramenons aussi l'analyse des algorithmes interrompus à un unique paramètre M. La deuxième section introduit les séries génératrices et les opérateurs de transfert. Ensuite, les séries génératrices des paramètres sont exprimées avec les opérateurs. La troisième étape d'analyse spectrale et analytique des opérateurs est laissée au prochain chapitre. Toutefois, à la section 3.4, les propriétés analytiques des séries sont expliquées. La dernière section aborde l'étape d'extraction des coefficients et les asymptotiques.

3.2 Système dynamique et décompositions

Les décompositions sur les polynômes étaient principalement dues à la relation simple entre le degré des continuants et le degré des quotients. Cette relation admet un équivalent sur les entiers qui fait intervenir les branches inverses du système dynamique associé à l'algorithme d'Euclide. Avant de proposer les décompositions, nous présentons ce système dit des fractions continues.

3.2.1 Système dynamique des fractions continues

La première étape de l'analyse dynamique est d'associer un système dynamique à l'algorithme d'Euclide.

Sur une entrée (v_0, v_1) l'algorithme d'Euclide classique effectue une séquence de divisions euclidiennes de la forme

$$v_0 = m_1 v_1 + v_2, \quad v_1 = m_2 v_2 + v_3, \quad \dots, \quad v_{p-1} = m_p v_p + 0.$$

Si au lieu d'observer la suite des couples (v_i, v_{i+1}), on observe les rationnels v_{i+1}/v_i, alors il existe une fonction T telle que

$$T\left(\frac{v_1}{v_0}\right) = \frac{v_2}{v_1}, \quad \dots, \quad T\left(\frac{v_{i+1}}{v_i}\right) = \frac{v_{i+2}}{v_{i+1}}, \quad \dots, \quad T\left(\frac{v_p}{v_{p-1}}\right) = 0.$$

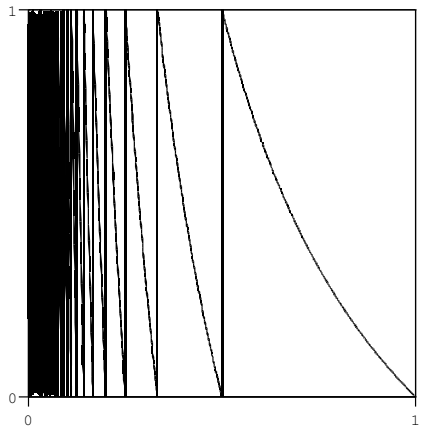


FIG. 3.1 – Système dynamique des fractions continues

La fonction T est appelée fonction de décalage ou shift, agit sur l'intervalle $[0,1]$ et est définie par

$$T(x) = \frac{1}{x} - \left\lfloor \frac{1}{x} \right\rfloor, \qquad T(0) = 0.$$

La paire (I,T) forme le système dynamique dit des fractions continues. Un système dynamique (de l'intervalle) est une paire $(\tilde{I},\tilde{T})$ formée par un intervalle $\tilde{I}$ et une fonction $\tilde{T} : \tilde{I} \to \tilde{I}$ qui est C^2 par morceaux sur une partition finie ou dénombrable de $\tilde{I}$ et telle que chaque morceaux est strictement monotone. Une représentation graphique de T est donnée à la figure 3.1. La fonction T est surjective par morceaux sur les intervalles $I_m =]\frac{1}{m+1}, \frac{1}{m}]$. De plus, les bijections inverses $h_{[m]}$ de T restreintes à I_m sont des homographies définies par

$$h_{[m]} : [0,1[\to I_m, \qquad h_{[m]}(x) = \frac{1}{m+x}, \qquad \forall m \geq 1.$$

Une bijection inverse est appelée branche inverse et $\mathcal{H}$ désigne l'ensemble des branches inverses,

$$\mathcal{H} = \{h_{[m]} : x \to \frac{1}{m+x};\ m \geq 1\}.$$

Une branche inverse de profondeur n est la composée de n branches inverses. Nous notons $\mathcal{H}^n$ l'ensemble des branches de profondeur n et $\mathcal{H}^\star = \cup_{n\geq 0}\mathcal{H}^n$ le semi-groupe engendré par $\mathcal{H}$,

$$\mathcal{H}^n = \{h_1 \circ h_2 \circ \ldots \circ h_n;\ h_i \in \mathcal{H}, \forall i\}, \quad \mathcal{H}^\star = \cup_{n\geq 0}\mathcal{H}^n, \quad \mathcal{H}^0 = \{id\}.$$

Par construction, $v_i/v_{i-1} = h_{[m_i]}(v_{i+1}/v_i)$ et par induction, nous obtenons le développement en fraction continue du rationnel v_i/v_{i-1},

$$\frac{v_i}{v_{i-1}} = h_{[m_i]} \circ \ldots \circ h_{[m_p]}(0) = \cfrac{1}{m_i + \cfrac{1}{m_{i+1} + \cfrac{1}{\ddots \cfrac{}{m_{p-1} + \cfrac{1}{m_p}}}}}. \tag{3.1}$$

Le dernier quotient est par hypothèse différent de 1 ce qui nous fait introduire l'ensemble $\mathcal{F} = \mathcal{H}\backslash\{h_{[1]}\}$. Comme sur les polynômes, l'algorithme d'Euclide construit la bijection

$$\Omega \approx (\mathcal{H}^0 + \mathcal{H}^\star\mathcal{F}) \times \mathbb{N}^\star. \tag{3.2}$$

$(\mathcal{H}^0 + \mathcal{H}^\star\mathcal{F})$ est l'ensemble des suites vides (symbolisée par $\mathcal{H}^0$) ou contenant au moins un quotient $(\mathcal{H}^\star\mathcal{F})$ et $\mathbb{N}^\star$ est l'ensemble des pgcd possibles. La présence de $\mathcal{H}^0$ est due aux entrées (u, v) avec $v = 0$. Dans ce cas, il n'y a pas de développement en fractions continues et le pgcd est u. Cette bijection est à rapprocher avec celle obtenue sur les polynômes (formule 2.5). La bijection est un peu plus simple car sur les polynômes, $\mathcal{F} = \mathcal{G}$.

Pour une branche inverse $h = h_1 \circ \ldots \circ h_p$ de profondeur p, $e_i(h)$ désigne la branche inverse de fin (au rang i) composée des $p - i$ dernières branches inverses,

$$e_i(h) = h_{i+1} \circ \ldots \circ h_p, \qquad i \leq p - 1.$$

De la même manière, la branche inverse du début $b_i(h)$ (au rang i) est la branche inverse composée des $i - 1$ premières branches,

$$b_i(h) = h_1 \circ \ldots \circ h_{i-1}, \qquad i \geq 2.$$

Lorsqu'il n'y aura pas d'ambiguïté, les branches de fin et de début seront notées e_i et b_i. Les branches inverses de début et de fin de h sont liées aux continuants v_i et a_i. Nous avons déjà vu que les restes successifs v_i vérifient

$$\frac{v_{i+1}}{v_i} = e_i(0)$$

et comme les branches inverses $h_{[m]}$ sont des homographies de déterminant -1, si $D[x]$ désigne le dénominateur de x pour un rationnel x, les v_i vérifient également

$$D\left[\frac{v_{i+1}}{v_i}\right] = D[e_i(0)] = |e_i'(0)|^{-1/2} \quad \text{et} \quad v_i = \operatorname{pgcd}(v_0, v_1)|e_i'(0)|^{-1/2}. \tag{3.3}$$

Une propriété symétrique existe avec les continuants a_i mais avec les branches du début,

$$|a_i| = |b_i'(0)|^{-1/2}. \tag{3.4}$$

Les égalités 3.3 et 3.4 sont équivalentes aux relations sur les degrés introduites pour les polynômes (voir lemme 1). Elles auront un rôle essentiel pour l'analyse de tous les coûts de cette partie, à commencer pour la décomposition de tous les paramètres.

3.2.2 Décomposition de la complexité étendue

La décomposition $h = b_i \circ e_{i-1} = b_i \circ h_i \circ e_i$ entraîne

$$v_0^{-2} = v_p^{-2} \cdot |h'(0)| = v_p^{-2} \cdot |b_i'(e_{i-1}(0))| \cdot |e_{i-1}'(0)| = v_p^{-2} \cdot |b_i'(e_{i-1}(0))| \cdot |h_i'(e_i(0))| \cdot |e_i'(0)|,$$

si bien que le terme $\ell(v_i) + \ell(a_i) + \lg m_i$ se décompose en

$$\ell(v_i) + \ell(s_i) + \lg m_i = \ell(v_0) + 1 + f + d_i + f_i, \tag{3.5}$$

$$\text{avec} \quad d_i := d_i^{(1)} + d_i^{(2)}, \qquad d_i^{(1)} := \lg\left|\frac{h_i'(0)}{h_i'(e_i(0))}\right|, \qquad d_i^{(2)} := \lg\left|\frac{b_i'(e_{i-1}(0))}{b_i'(0)}\right|, \tag{3.6}$$

$$f := \{\lg v_0\}, \qquad f_i := -\{\lg v_i\} - \{\lg s_i\}.$$

Avec cette décomposition, nous obtenons une décomposition de la complexité binaire étendue.

Proposition 5 (Décomposition complexité étendue) *La complexité binaire E de l'algorithme d'Euclide étendu se décompose en*

$$D = (\ell + 1) \cdot Z + Y \qquad \textit{avec} \qquad Y = -Y_1 + O(Y_2) + Y_3 + Y_4 + Y_5.$$

Ici ℓ est la taille des entrées définie par $\ell(u, v) = \ell(u) = \ell(v_0)$ et

$$Z = \sum_{i=1}^{p-1} \ell(m_i), \qquad Y_1 = \sum_{i=1}^{p-1} \ell(m_i) \cdot \lg m_i, \qquad Y_2 = (\ell(m_p) + \lg v_p)^2, \tag{3.7}$$

$$Y_3 = f \cdot \sum_{i=1}^{p-1} \ell(m_i), \qquad Y_4 = \sum_{i=1}^{p-1} d_i \cdot \ell(m_i) \qquad Y_5 = \sum_{i=1}^{p-1} f_i \cdot \ell(m_i). \tag{3.8}$$

Sur les polynômes, la décomposition de D s'écrivait (à peu près) sous la forme $(\ell+1) \cdot Z - Y_1 + O(Y_2)$. Les termes correctifs supplémentaires dans le cas des entiers viennent notamment de la différence entre la taille binaire et le logarithme des entiers.

Les coûts Z et Y_1 sont appelés des coûts additifs car ce sont la somme d'un coût élémentaire sur tous les quotients. Comme sur les polynômes, les coûts additifs pourront être à croissance modérée ou à croissance intermédiaire. Le coût Y_2 est appelé un coût terminal puisqu'il ne fait intervenir que les éléments de la dernière étape (le quotient m_p et le pgcd v_p).
Comme les branches inverses h sont toutes de la forme

$$h(x) = (\alpha x + \beta)/(\gamma x + \delta), \qquad \text{avec } \alpha, \beta, \gamma, \delta \text{ des entiers premiers entre eux et } 0 < \gamma \leq \delta,$$

la distorsion est bornée, c'est-à-dire, pour tout $x, y \in [0, 1]$, pour toute branche inverse $h \in \mathcal{H}^\star$, on a $|h'(x)| \leq 4|h'(y)|$. Chaque coût $d_i^{(j)}$ est alors uniformément borné par une constante d et il en est de même pour f et les f_i. Il vient alors que pour $i = 3, 4, 5$, les coûts Y_i vérifient $Y_i = O(Z)$. Les analyses des paramètres Y_3, Y_4 et Y_5 se réduisent à l'analyse de Z. Pour Y_1 et Y_2, nous reprenons la terminologie des polynômes.

Définition 4 *(i) Les coûts terminaux Y sont tous les coûts polynomiaux en la la taille de v_p et m_p, de la forme $T(v_0, v_1) = O((\ell(m_p) + \ell(v_p))^k)$ avec k un entier fixé et $T = 0$ si le nombre d'étapes est nul.*
(ii) Si $c : \mathbb{N}^\star \to \mathbb{R}^+$ est un coût élémentaire non-nul sur les quotients, le coût additif associé à c est

$$C(v_0, v_1) = \sum_{i=1}^{p-1} c(m_i),$$

1. *Un coût additif est dit à croissance modérée si le coût élémentaire c satisfait $c(m) = O(\ell(m))$.*
2. *Un coût additif est dit à croissance intermédiaire s'il n'est pas à croissance modérée et si le coût élémentaire c satisfait $c(m) = O(\ell(m)^k)$ pour un entier k fixé.*

Avec ces définitions, le coût Z est à croissance modérée et Y_1 est à croissance intermédiaire. Nous souhaitons montrer que la complexité binaire E admet une loi limite gaussienne en appliquant le principe de décomposition. Contrairement aux polynômes, il est déjà connu que les coûts à croissance modérée admettent une loi limite gaussienne d'espérance et de variance linéaires en n

(théorème BV). Comme $\ell+1$ est constant et égal à $n+1$ sur Ω_n, le théorème BV appliqué à Z entraîne que $(\ell+1)\cdot Z$ suit une loi limite gaussienne de paramètres

$$\mathrm{E}_n[(\ell+1)\cdot Z] = \mu(\ell)\cdot n^2 + O(n), \quad \mathrm{V}_n[(\ell+1)\cdot Z] = \rho(\ell)\cdot n^3 + O(n^2), \quad r_n[(\ell+1)\cdot Z] = O(n^{-1/2}).$$

Maintenant, pour $i = 3, 4, 5$, on a $Y_i = O(Z)$. Les espérances et les variances des Y_i vérifient alors les relations triviales

$$\mathrm{E}_n[Y_i] = O(\mathrm{E}_n[Z]) = O(n), \quad \text{et} \quad \mathrm{V}_n[Y_i] \le \mathrm{E}_n[Y_i^2] = O(\mathrm{E}_n[Z^2]) = O(n^2).$$

Les Y_i $(i = 3, 4, 5)$ sont concentrés par rapport à $(\ell+1)\cdot Z$. Pour satisfaire le principe de décomposition, il faut également montrer la concentration des coûts Y_1 et Y_2.

Théorème 12 *(i) Soit C un coût additif à croissance intermédiaire (ou modérée) associé au coût élémentaire c. Alors les moments de C satisfont sur Ω_n*

$$\mathrm{E}_n[C] = \frac{12\log 2}{\pi^2}\left(\sum_{m\ge 1} c(m)\log(1+\frac{1}{m(m+2)})\right)\cdot n(1+O(n^{-1})), \qquad \mathrm{V}_n[C] = O(n).$$

(ii) Soit Y un coût terminal. Alors tous les moments de Y sont d'ordre constant sur Ω_n.

Avec ce théorème, le principe de décomposition s'applique à la complexité binaire étendue. L'espérance et la variance sont asymptotiquement l'espérance et la variance de $(\ell+1)\cdot Z$, et la vitesse de convergence est $O(n^{-1/3})$ car certains coûts concentrés ont une variance d'ordre quadratique. Si le théorème 12 est admis, la loi limite gaussienne de la complexité binaire étendue est prouvée (théorème 3).

3.2.3 Variance de la complexité binaire classique

Nous décrivons ici la conjecture (C) du théorème 2 qui implique l'égalité suivante sur les variances,

$$3\mathrm{V}_n[B] = \mathrm{V}_n[E] + O(n^2).$$

Pour cela, nous introduisons les continuants approchés t_i et w_i définis par

$$t_i := b_i'(e_{i-1}(0))^{-1/2} \qquad \text{et} \qquad w_i := v_i/v_p = e_i'(0)^{-1/2}.$$

Nous considérons deux nouveaux coûts

$$A(v_0, v_1) := \sum_{i=1}^{p} \ell(m_i)\lg w_i, \qquad \overline{A}(v_0, v_1) := \sum_{i=1}^{p} \ell(m_i)\lg t_i,$$

qui font intervenir des logarithmes à la place des tailles binaires. Ils seront facilement générés avec les opérateurs de transfert et ils sont aussi très proches des complexités binaires classiques et étendues puisque l'on a

$$B = A + Z\cdot\lg v_p + O(Z) \qquad \text{et} \qquad E = (A+\overline{A}) + Z\cdot\lg v_p + O(Z).$$

Les variances de $O(Z)$ et $Z\cdot\lg v_p$ sont toutes les deux d'ordre au plus quadratique[4]. Nous pouvons donc affirmer trois choses :

[4] $\mathrm{V}_n[Z\cdot\lg v_p] \le \mathrm{E}_n[Z^2\cdot\lg^2 v_p] \le \mathrm{E}_n[Z^4]^{1/2}\mathrm{E}_n[\lg^4 v_p]^{1/2} = O(n^2)$.

1. La complexité binaire étendue E suit une loi limite gaussienne de paramètre $[\Theta(n^2), \Theta(n^3), O(n^{-1/3})]$ ssi $A + \overline{A}$ suit une loi limite gaussienne de paramètre $[\Theta(n^2), \Theta(n^3), O(n^{-1/3})]$.
2. La complexité binaire classique B suit une loi limite gaussienne de paramètre $[\Theta(n^2), \Theta(n^3), O(n^{-1/3})]$ ssi A suit une loi limite gaussienne de paramètre $[\Theta(n^2), \Theta(n^3), O(n^{-1/3})]$.
3. On a l'équivalence $\mathrm{V}_n[E] = 3\mathrm{V}_n[B] + O(n^2)$ ssi $\mathrm{V}_n[A + \overline{A}] = 3\mathrm{V}_n[A] + O(n^2)$.

En supposant le théorème 12 démontré, la complexité binaire E suit une loi limite gaussienne dont la variance est de la forme $\mathrm{V}_n[E] = \rho_0 \cdot n^3 + O(n^2)$. Sous l'hypothèse du théorème 12, la variance de $A + \overline{A}$ est donc de la forme $\mathrm{V}_n[A + \overline{A}] = \rho_0 \cdot n^3 + O(n^2)$.

Nous présentons maintenant deux conjectures (G) et (C). La conjecture (C) permet de prouver l'égalité $\mathrm{V}_n[E] = 3\mathrm{V}_n[B] + O(n^2)$ sur les variances. Montrer la conjecture (G) serait un premier pas (non suffisant) pour obtenir la loi limite gaussienne de la complexité binaire classique B.

3.2.3.1 Conjecture (G) et loi limite de B

Nous souhaitons obtenir une décomposition similaire à celle obtenue sur les polynômes pour la complexité binaire classique (voir formule 2.4 et théorème 11 page 42). Mais la décomposition sera plus difficile que sur les polynômes. Cela est du au fait que il n'est plus possible d'exprimer les tailles binaires avec seulement les paramètres $\ell(m_i)$. Nous introduisons donc de nouveaux paramètres, le coût $r = \lg v_p + \lg v_0 - \ell(v_0)$ et la séquence θ_i que nous définissons maintenant. Nous notons x_i le rationnel v_{i+1}/v_i [qui est aussi égal au rationnel w_{i+1}/w_i. La relation $w_{i-1} = m_i w_i + w_{i+1}$ entraîne que, pour tout i avec $1 \leq i \leq p$,

$$\lg w_{i-1} - \lg w_i = \lg(m_1 + x_i) = \ell(m_i) + \theta_i, \quad \text{avec} \quad \theta_i = \lg \frac{m_i + x_i}{\tilde{m}_i}, \tag{3.9}$$

et $\tilde{m}_i$ la plus petite puissance de deux au moins égale à m_i. Remarquons que θ_i satisfait $-1 \leq \theta_i \leq 1$. La relation 3.9 remplace la relation polynomiale $\ell(v_{i-1}) - \ell(v_i) = \ell(m_i) - 1$, et, dans le cas entier, la suite θ_i joue le même rôle que la suite constante et égale à -1 dans le cas polynomial. Alors, pour tout i, $0 \leq i \leq p-1$, on a

$$\lg w_i = \sum_{j=i+1}^{p} [\ell(m_j) + \theta_j],$$

et, avec la définition de r, la relation précédente pour $i = 0$ entraîne sur ω_n la décomposition :

$$\sum_{i=1}^{p} [\ell(m_i) + \theta_i] = \lg w_0 = n - r. \tag{3.10}$$

Le coût A peut alors s'écrire

$$A(v_0, v_1) = \sum_{i=1}^{p} ([\ell(m_i) + \theta_i] - \theta_i) \left(\sum_{j>i} [\ell(m_j) + \theta_j] \right) = \frac{1}{2}(n-r)^2 - \frac{1}{2}(n-r) - \sum_{i=1}^{p} \theta_i \lg w_i.$$

Nous venons de prouver la première assertion de la proposition suivante.

Proposition 6 *(a) Sur Ω_n, la complexité binaire approchée A de l'algorithme d'Euclide classique vérifie la décomposition*

$$A = \left[\frac{n^2}{2} - \underline{N} \right] - n[r + \frac{1}{2}] + r \quad \textit{avec} \quad \underline{N} = \sum_{i=1}^{p} \theta_i \lg w_i.$$

(b) De plus, le coût Θ égal à la somme de tous les θ_i satisfait

$$\Theta := \sum_{i=1}^{p} \theta_i = [n - Z] - r,$$

et Θ suit une loi limite gaussienne de paramètres $[O(n), O(n), O(n^{-1/3})]$

Le point (b) est une conséquence directe de la relation 3.10 et du fait que le coût Z est un coût additif à croissance modérée et qu'il satisfait une loi limite gaussienne (théorème BV).

La conjecture (G) est lié au coût $\underline{N}$. Le paramètre $\underline{N}$ admet comme forme alternative

$$\underline{N} = \sum_{i=1}^{p} (\sum_{j<i} \theta_i) \cdot \lg(m_i + x_i).$$

Notons la similitude avec le coût N pour les polynômes (théorème 11 page 42). Comme le coût Θ est gaussien avec de paramètre $[O(n), O(n), O(n^{-1/3})]$, le facteur $(\sum_{j<i} \theta_i)$ est probablement proche de son espérance, d'ordre i. De fait, une première étape afin de prouver que $\underline{N}$ est gaussien, est de prouver que les versions lissées, obtenues en remplaçant les θ_i par une constante, et données par

$$N^{(v)} = \sum_{i=1}^{p} \lg w_i, \qquad \hat{N}^{(v)} = \sum_{i=1}^{p} \lg v_i \qquad \text{ou} N^{(m)} = \sum_{i=1}^{p} i \cdot \ell(m_i)$$

sont asymptotiquement gaussien. Comme $\hat{N}^{(v)} = N^{(v)} + p \lg v_p$, il est suffisant d'étudier $N^{(v)}$ et $N^{(m)}$ (car $\mathrm{V}_n[p \lg v_p] = O(n^2)$). Nous ne savons pas prouver ces lois limites gaussiennes, mais nous avons accès à des formes alternatives pour les séries génératrices associées. Notre première conjecture (G) est la suivante.

Conjecture (G). *Les coûts $N^{(v)}$ et $N^{(m)}$ admettent des lois limites gaussiennes.*

3.2.3.2 Conjecture (C) et variance de B

Même si la première conjecture est démontrée, elle ne donne pas une estimation de la variance de B (ou A). La conjecture (C) traite directement de la constante dominante de $\mathrm{V}_n[B]$ et veut la relier avec celle de $\mathrm{V}_n[E]$. Dans le cas polynomial, nous savons que $\mathrm{V}_n[B] \sim \mathrm{V}_n[E]/3$, et nous supposons que la même relation existe sur les entiers. Il n'est pas facile d'utiliser le paramètre $\underline{N}$ pour montrer cette relation. Cependant, les paramètres A et $\overline{A}$ sont plus faciles à manipuler. De plus nous disposons des relations suivantes

$$\mathrm{V}_n[B] = \mathrm{V}_n[A] + O(n^2) \qquad \text{et} \qquad \mathrm{V}_n[D] = \mathrm{V}_n[A + \overline{A}] + O(n^2).$$

A cela, s'ajoute la proposition suivante.

Proposition 7 *Les variances $\mathrm{V}_n[A]$ et $\mathrm{V}_n[\overline{A}]$ sont équivalentes et satisfont $\mathrm{V}_n[A] = \mathrm{V}_n[\overline{A}] + O(n^2)$.*

Preuve. Pour comparer la variance de A et $\overline{A}$, le coût $\hat{A}$ défini par

$$\hat{A}(v_0, v_1) := \sum_{i=1}^{p} \ell(m_i) \cdot \lg s_i$$

sera utile. En fait, A et $\hat{A}$ sont étroitement liés via l'opération miroir, que nous décrivons maintenant. A un élément $h = h_1 \circ \ldots \circ h_p$ de $\mathcal{H}^\star$, nous associons son miroir $\hat{h}$, qui est formé des mêmes branches inverses mais dans l'ordre opposé, $\hat{h} = h_p \circ \ldots \circ h_1$. Alors, $h(0)$ et $\hat{h}(0)$ sont deux rationnels avec le même dénominateur et nous notons par un chapeau l'opération miroir induite sur les pairs (u, v) d'entiers premiers entre eux. Maintenant, $\hat{A}$ est le miroir de A : cela signifie que $\hat{A}(v_0, v_1) = A(v_0, \hat{v_1})$ et cela implique les égalités $\mathrm{E}_n[A] = \mathrm{E}_n[\hat{A}]$ et $\mathrm{V}_n[A] = \mathrm{V}_n[\hat{A}]$.
D'un autre coté, nous avons $\hat{A} - \overline{A} = \sum_{i=1}^{p} d_i^{(2)}$ soit $\mathrm{V}_n[\hat{A} - \overline{A}] = O(n^2)$. Si les variances de $\hat{A}$ et $\overline{A}$ sont d'ordre $\Theta(n^3)$, cela implique $\mathrm{V}_n[\overline{A}] = \mathrm{V}_n[\hat{A}](1 + O(n^{-1}))$. Sinon, les variances sont d'ordre au plus quadratique et nous avons toujours la relation

$$\mathrm{V}_n[\overline{A}] = \mathrm{V}_n[\hat{A}] + O(n^2).$$

Ceci termine la preuve.■

La proposition entraîne immédiatement les égalités

$$4\mathrm{V}_n[A] = 2\mathrm{V}_n[A] + 2\mathrm{V}_n[\overline{A}] + O(n^2) = \mathrm{V}_n[A + \overline{A}] + \mathrm{V}_n[A - \overline{A}] + O(n^2)$$

et en particulier, nous obtenons

$$12(\mathrm{V}_n[B] - \frac{1}{3}\mathrm{V}_n[E]) = \mathrm{V}_n[A + \overline{A}] - 3\mathrm{V}_n[A - \overline{A}] + O(n^2).$$

La conjecture (C) suivante est équivalente à l'assertion $\mathrm{V}_n[B] \sim (1/3)\mathrm{V}_n[E]$.

Conjecture (C)**.** *La variance* $\mathrm{V}_n[A]$ *est non nulle et les deux termes* $3\mathrm{V}_n[A - \overline{A}]$ *et* $\mathrm{V}_n[A + \overline{A}]$ *ont le même terme dominant d'ordre* n^3*.*

3.2.4 Continuant à une fraction de l'exécution

Contrairement au polynôme, l'analyse du paramètre $L^{[\delta]}$ nécessite une décomposition. Par définition, $L^{[\delta]}$ est la taille binaire du continuant $v_{\lfloor \delta p \rfloor}$. Sur les polynômes, la taille vérifie $\ell = 1 + \deg$ et le degré des continuants s'exprimait simplement avec le degré des quotients. Sur les entiers, la taille fait intervenir des parties entières que nous ne savons pas générer. C'est la raison pour laquelle nous décomposons $L^{[\delta]}$ de manière à les faire disparaître.

Proposition 8 (Décomposition de $L^{[\delta]}$**)** *Le paramètre* $L^{[\delta]}$ *se décompose en* $L^{[\delta]} = \tilde{L}^{[\delta]} + Y_6$ *avec* $\tilde{L}^{[\delta]}$ *et* Y_6 *les paramètres*

$$\tilde{L}^{[\delta]}(v_0, v_1) = \log v_{\lfloor \delta p \rfloor}, \qquad Y_6(v_0, v_1) = 1 - \{\log v_{\lfloor \delta p \rfloor}\}.$$

Y_6 est borné par 1 soit l'espérance et la variance associées sont d'ordre constant. Pour démontrer le théorème 5, il suffit que $\tilde{L}^{[\delta]}$ suit une loi limite gaussienne d'espérance et de variance identiques à celles données dans le théorème.

3.2.5 Algorithmes interrompus

Avec le principe de décomposition, l'analyse des complexités binaires des algorithmes d'Euclide classique et étendu se résume à l'analyse de trois familles de coûts : les coûts à croissance intermédiaire, les coûts terminaux et les coûts A et $\overline{A}$. Pour la complexité binaire de $\mathcal{KS}$, il suffit d'établir le théorème 6 qui, combiné avec la proposition 2, implique le théorème 7 qui lui même suffit à démontrer le résultat final sur la complexité binaire de $\mathcal{HG}_\alpha$ (théorème 8).

Le théorème 6 met en jeu des évènements du type $[P_{<\gamma,\delta]} > i]$ ou $[P_{<\gamma,\delta]} \leq i]$. L'algorithme $\mathcal{E}_{<\gamma,\delta]}$ s'arrête dès que le reste $v_{\lfloor \gamma P \rfloor + i}$ satisfait $\ell(v_{\lfloor \gamma P \rfloor + i}) \leq \ell(v_{\lfloor \gamma P \rfloor}) - \delta n$. En particulier, les évènements précédents satisfont

$$[P_{<\gamma,\delta]} > i] \subset \left[\frac{4 v_{\lfloor \gamma P \rfloor + i}}{v_{\lfloor \gamma P \rfloor} \cdot v_0^{-\delta}} > 1 \right], \qquad [P_{<\gamma,\delta]} \leq i] \subset \left[\frac{v_{\lfloor \gamma P \rfloor + i}}{2 v_{\lfloor \gamma P \rfloor} \cdot v_0^{-\delta}} > 1 \right]$$

et l'inégalité de Markov entraîne

$$\mathbb{P}_n \left[P_{<\gamma,\delta]} > i \right] \leq \mathrm{E}_n \left[\left(\frac{4 v_{\lfloor \gamma P \rfloor + i}}{v_{\lfloor \gamma P \rfloor} \cdot v_0^{-\delta}} \right)^{2w} \right], \ \forall w > 0, \tag{3.11}$$

$$\mathbb{P}_n \left[P_{<\gamma,\delta]} \leq i \right] \leq \mathrm{E}_n \left[\left(\frac{v_{\lfloor \gamma P \rfloor + i}}{2 v_{\lfloor \gamma P \rfloor} \cdot v_0^{-\delta}} \right)^{2w} \right], \ \forall w < 0. \tag{3.12}$$

Dans notre cas, l'entier i est remplacé par $\lfloor (\delta + \epsilon) P \rfloor$ et les relations 3.11 et 3.12 montrent que l'étude des évènements initiaux se ramène à l'analyse de l'espérance de la variable M^{2w} avec M donné par

$$M(v_0, v_1) = \frac{v_{\lfloor \gamma P \rfloor + \lfloor \delta' P \rfloor}}{v_{\lfloor \gamma P \rfloor} \cdot v_0^{-\delta' + \epsilon}}, \qquad \text{et} \quad \delta' = \delta + \epsilon. \tag{3.13}$$

Nous venons de ramener l'étude des deux premiers résultats du théorème 6 à l'analyse en moyenne du paramètre M. Il nous faut maintenant traiter les évènements $[\ell_{<\gamma,\delta>} \geq (\delta + \epsilon) n]$ et $[\ell_{<\gamma,\delta>} \leq (\delta + \epsilon) n]$. La relation matricielle 1.4 montre que la taille de la matrice produite par $\mathcal{E}_{<\gamma,\delta>}$ satisfait $|\ell_{<\gamma,\delta>} - \ell(v_{\lfloor \gamma P \rfloor + \lfloor \delta P \rfloor}) + \ell(v_{\lfloor \gamma P \rfloor})| \leq 2$. Nous sommes donc ramener à étudier les mêmes évènements que précédemment.

3.2.6 État d'avancement de l'analyse

A la fin de chaque étape des analyses nous ferons un petit récapitulatif comme celui-ci. Cette première section correspondait à la première étape d'une analyse dynamique. Nous avons construit un système dynamique associé à l'algorithme d'Euclide sur les entiers. A partir des branches inverses de ce système dynamique, nous avons proposé des décompositions pour les continuants et les complexités binaires afin de ramener leurs analyses à des paramètres plus simples. Nous avons aussi réduit l'analyse des algorithmes interrompus à un paramètre M. Pour établir tous les résultats, nous devons donc analyser le paramètre M, mais aussi les coûts additifs à croissance intermédiaire, les coûts terminaux, le paramètre $\tilde{L}^{[\delta]}$ et les coûts A et $\overline{A}$.

3.3 Séries génératrices

La deuxième étape d'une analyse dynamique associe à chaque série génératrice une expression alternative en fonction d'opérateurs. Nous avons vu que les continuants s'expriment naturellement en fonction des dénominateurs des branches inverses (voir formules 3.3 et 3.4). Les séries génératrices de Dirichlet sont alors plus adaptées pour les analyses. Nous commençons cette section avec une description des séries génératrices de Dirichlet. Nous introduisons ensuite l'outil classique des systèmes dynamiques, à savoir les opérateurs de transfert. La section 3.3.3 décrit les dictionnaires sur les opérateurs et toutes les sections suivantes traitent des formules alternatives des séries en fonction des opérateurs.

3.3.1 Séries génératrices de Dirichlet

Nous fixons R un paramètre sur $\Omega = \cup_n \Omega_n$. La série génératrice bivariée (de Dirichlet) associée à R est définie par

$$S_R(s,w) = \sum_{(u,v)\in\Omega} \frac{e^{wR(u,v)}}{u^s},$$

où la variable s marque l'entrée et la variable w est liée à la valeur du paramètre. Si $r_n(w)$ désigne la somme cumulée des $e^{wR(u,v)}$ avec $u = n$, la série bivariée s'écrit sous la forme d'une vrai série de Dirichlet

$$S_R(s,w) = \sum_{n\geq 1} \frac{r_n(w)}{n^s}, \qquad r_n(w) = \sum_{(n,v)\in\Omega} e^{wR(n,v)}.$$

La série $S_R^{[k]}$ du moment d'ordre k est définie comme la dérivée k^e par rapport à w de $S_R(s,w)$ en $w = 0$,

$$S_R^{[k]}(s) := \frac{d^k S_R}{d_w^k}(z,0) = \sum_{(u,v)\in\Omega} \frac{R(u,v)^k}{u^s}.$$

Comme la série bivariée, si $r_n^{[k]}$ désigne la somme cumulée des $R(u,v)^k$ avec $u = n$, $S_R^{[k]}$ s'écrit sous la forme d'une vrai série de Dirichlet

$$S_R^{[k]}(s) = \sum_{n\geq 1} \frac{r_n^{[k]}}{n^s}, \qquad r_n^{[k]} = \sum_{(n,v)\in\Omega} R(n,v)^k.$$

Comme pour les séries ordinaires, la série du moment d'ordre k est adaptée à l'analyse du k^e moment $\mathrm{E}_n[R^k]$ de R. L'ensemble Ω_n regroupe toutes les entrées de taille n, c'est à dire l'ensemble des couples d'entiers (u,v) avec $u > v$ et $\ell(u) = n$. Mais

$$\ell(u) = n \quad \text{équivaut à} \quad 2^{n-1} \leq u \leq 2^n - 1.$$

La probabilité sur Ω_n étant uniforme, le k^e moment satisfait

$$\mathrm{E}_n[R^k] := \frac{\psi_n^{[k]}}{|\Omega_n|} \quad \text{avec} \quad \psi_n = \sum_{u=2^{n-1}}^{2^n-1} r_u^{[k]} = \sum_{u=2^{n-1}}^{2^n-1} [u^{-s}] S_R^{[k]}(s).$$

Ici $[u^{-s}]f$ signifie ici le coefficient de u^{-s} dans f.
De manière analogue au moment d'ordre k, la série des moments s'exprime en fonction des coefficients de la série bivariée,

$$\mathrm{E}_n[e^{wR}] = \frac{\psi_n(w)}{|\Omega_n|} \quad \text{avec} \quad \psi_n = \sum_{u=2^{n-1}}^{2^n-1} r_u(w) = \sum_{u=2^{n-1}}^{2^n-1} [u^{-s}] S_R(s,w).$$

Contrairement aux séries ordinaires, nous sommes amenés à extraire plusieurs coefficients en même temps. Il existe deux types d'extracteurs pour les séries de Dirichlet : les théorèmes taubériens qui ne donnent pas de terme d'erreur, et la formule de Perron qui donne des termes d'erreurs. Les termes d'erreurs sont importants dans nos analyses, surtout pour les variances et les lois limites gaussiennes. Nous utiliserons donc la formule de Perron. En contrepartie, la formule de Perron *nécessite* des informations plus fortes sur les pôles de la série. Nous abordons ce problème dans la section 3.5.

3.3.2 Opérateurs de transfert

Un sujet d'étude principal en théorie des systèmes dynamiques est l'étude des trajectoires d'un point x sous l'action de la fonction de décalage. Pour les algorithmes d'Euclide, nous sommes intéressés par des trajectoires particulières, celles des rationnels. Ces trajectoires rencontrent toutes 0 et ne semblent pas du tout typiques. Mais nous allons les comparer aux trajectoires plus génériques. Le comportement des trajectoires génériques d'un système dynamique est décrit par l'évolution des densités. L'ensemble I est muni d'une densité initiale $f = f_0$ et chaque itération de T modifie cette densité. Les densités successives f_1, f_2, ... décrivent l'évolution globale du système. Il existe un opérateur, appelé transformateur de densité ou opérateur de Perron-Frobenius, qui vérifie $\mathbf{G}[f_n] = f_{n+1}$ et plus généralement $f_n = \mathbf{G}^n[f_0]$. Pour le système dynamique des fractions continues, cet opérateur est défini par

$$\mathbf{G}[f](x) := \sum_{h\in\mathcal{H}} |h'(x)| \cdot f\circ h(x) = \sum_{m\geq 1} \frac{1}{(m+x)^2} \cdot f(\frac{1}{m+x}).$$

Il est très utile d'ajouter un paramètre supplémentaire s afin de générer les séries de Dirichlet. Cela définit l'opérateur de transfert $\mathbf{G}_s$ (où opérateur de Ruelle [Rue78]),

$$\mathbf{G}_s[f](x) := \sum_{h\in\mathcal{H}} |h'(x)|^s \cdot f\circ h(x) = \sum_{m\geq 1} \frac{1}{(m+x)^{2s}} \cdot f(\frac{1}{m+x}).$$

Nous considérons également une dernière généralisation de l'opérateur de transfert afin de traiter les coûts relatifs à un coût élémentaire c. L'opérateur de transfert pondéré $\mathbf{G}_{s,w,[c]}$ relatif au coût c est

$$\mathbf{G}_{s,w,[c]}[f](x) = \sum_{h\in\mathcal{H}} e^{wc(h)}|h'(x)|^s \cdot f\circ h(x) = \sum_{m\geq 1} \frac{e^{wc(m)}}{(m+x)^{2s}} \cdot f(\frac{1}{m+x}).$$

Dans cette définition, nous identifions le coût $c(m)$ pour un entier m, avec le coût $c(h_{[m]})$ sur la branche inverse associée à m. Pour $w = 0$, l'opérateur pondéré est l'opérateur de transfert qui ne dépend pas du coût élémentaire c. Maintenant, si c est étendu par additivité à l'ensemble des branches inverses $\mathcal{H}^\star$, le n^e itéré de $\mathbf{G}_{s,w,[c]}$ et le quasi-inverse $(\mathbf{I} - \mathbf{G}_{s,w,[c]})^{-1}$ vérifient

$$\mathbf{G}^n_{s,w,[c]}[f] = \sum_{h\in\mathcal{H}^n} e^{wc(h)}|h'(x)|^s \cdot f\circ h(x),$$

$$(\mathbf{I} - \mathbf{G}_{s,w,[c]})^{-1} = \sum_{h\in\mathcal{H}^\star} e^{wc(h)}|h'(x)|^s \cdot f\circ h(x).$$

Nous terminons avec l'opérateur du premier moment $\mathbf{G}^{[c]}_s$ associé au paramètre c, défini comme la dérivée par rapport à w en 0 de l'opérateur pondéré $\mathbf{G}_{s,w,[c]}$,

$$\mathbf{G}^{[c]}_s[f] := \left(\frac{d}{dw}\mathbf{G}_{s,w,[c]}[f]\right)|_{w=0} = \sum_{h\in\mathcal{H}^n} c(h)|h'(x)|^s \cdot f\circ h(x).$$

Tous les opérateurs précédents s'expriment sur l'ensemble de branche inverses $\mathcal{H}$. Il est bien entendu possible d'étendre ces définitions à un sous ensemble de $\mathcal{H}$. Par exemple pour l'ensemble $\mathcal{F}$, l'opérateur pondéré $\mathbf{F}_{s,w,[c]}$ associé s'écrit

$$\mathbf{F}_{s,w,[c]}[f] := \sum_{h\in\mathcal{F}} e^{wc(h)}|h'(x)|^s \cdot f\circ h(x).$$

3.3.3 Dictionnaire sur les opérateurs

Pour $i = 1, 2, 3$, $\mathcal{H}_i$ désigne un ensemble de branches inverses muni d'un paramètre c_i sur ces branches inverses. Il existe un dictionnaire sur les opérateurs qui est similaire à celui des séries ordinaires.

Union disjointe. Si $\mathcal{H}_1$ est l'union disjointe $\mathcal{H}_2 \cup \mathcal{H}_3$ ou l'union est disjointe et si le paramètre R_1 vérifie

$$R_1 = \mathbf{1}_{\mathcal{H}_2} R_2 + \mathbf{1}_{\mathcal{H}_3} R_3$$

alors les opérateurs de transfert sur sur $\mathcal{H}_1$ s'écrivent

$$\mathbf{G}_{s,w,[c_1]} = \mathbf{G}_{s,w,[c_2]} + \mathbf{G}_{s,w,[c_3]}, \qquad \mathbf{G}_s^{[c_1]} = \mathbf{G}_s^{[c_2]} + \mathbf{G}_s^{[c_3]}.$$

Comme sur les polynômes, une union disjointe se traduit par une somme sur les opérateurs.
Produit cartésien. Si $\mathcal{H}_1 = \mathcal{H}_2 \mathcal{H}_3$ où les éléments de $\mathcal{H}_2 \mathcal{H}_3$ sont de la forme $h_2 \circ h_3$ avec $h_i \in \mathcal{H}_i$ et si les coûts primitifs c_i sur $\mathcal{H}_i$ vérifient $c_1 = c_2 + c_3$, alors

$$\mathbf{G}_{s,w,[c_1]} = \mathbf{G}s, w, [c_3] \circ \mathbf{G}_{s,w,[c_2]}.$$

Notons ici l'inversion de l'ordre des opérateurs par rapport à l'ordre de composition des branches. Le produit cartésien avec des coûts additifs se traduit par une composition des opérateurs pondérés. Pour des coûts multiplicatifs, le produit cartésien entraîne également une composition mais des opérateurs du premier moment. Si $c_1 = c_2 \cdot c_3$, alors

$$\mathbf{G}_s^{[c_1]} = \mathbf{G}_s^{[c_3]} \circ \mathbf{G}_s^{[c_2]}.$$

Bien entendu, ces relations sont à rapprocher avec celles obtenues sur les séries génératrices ordinaires (voir figure 2.1). Un principe général apparaît : le produit sur les séries ordinaires est remplacé par une composition sur les opérateurs.
Suites finies. Si $\mathcal{H}_1 = \mathcal{H}_2^\star$ et si les coûts primitifs sont additifs, i.e.,

$$c_1(h) = c_2(h_1) + \ldots + c_2(h_p), \qquad \text{dès que} \qquad h = h_1 \circ \ldots \circ h_p,$$

alors les opérateurs pondérés vérifient

$$\mathbf{G}_{s,w,[c_1]} = (\mathbf{I} - \mathbf{G}_{s,w,[c_2]})^{-1}.$$

De même, si les coûts sont multiplicatifs, i.e.,

$$c_1(h) = c_2(h_1) \times \ldots \times c_2(h_p), \qquad \text{dès que} \qquad h = h_1 \circ \ldots \circ h_p,$$

alors les opérateurs du premier moment vérifient

$$\mathbf{G}_s^{[c_1]} = (\mathbf{I} - \mathbf{G}_s^{[c_2]})^{-1}.$$

Sur les polynômes, les suites finies conduisaient également à des quasi-inverses. La figure 3.2 résume ces décompositions.

$\mathcal{H}_1$	coût c_1	opérateur
$\mathcal{H}_2 \cup \mathcal{H}_3$	$c_1 = c_2$ ou c_3	$\mathbf{G}_{s,w,[c_1]} = \mathbf{G}s, w, [c_3] + \mathbf{G}_{s,w,[c_2]}$
$\mathcal{H}_2\mathcal{H}_3$	$c_1 = c_2 \cdot c_3$	$\mathbf{G}_s^{[c_1]} = \mathbf{G}_s^{[c_3]} \circ \mathbf{G}_s^{[c_2]}$
	$c_1 = c_2 + c_3$	$\mathbf{G}_{s,w,[c_1]} = \mathbf{G}_{s,w,[c_3]} \circ \mathbf{G}_{s,w,[c_2]}$
$\mathcal{H}_2^\star$	$c_1 = c_2 \times \ldots \times c_2$	$\mathbf{G}_s^{[c_1]} = (\mathbf{I} - \mathbf{G}_s^{[c_2]})^{-1}$
	$c_1 = c_2 + \ldots + c_2$	$\mathbf{G}_{s,w,[c_1]} = (\mathbf{I} - \mathbf{G}_{s,w,[c_2]})^{-1}$

FIG. 3.2 – Dictionnaire pour les opérateurs de transfert

3.3.4 Développements propre et impropres

Nous avons vu que l'algorithme d'Euclide construit, sur une entrée (u, v), le développement en fraction du rationnel v/u. Ce développement est dit propre puisque le dernier quotient est par construction différent de 1. Il existe également un développement impropre où le dernier quotient est 1. Plus précisément, si $v/u = h_{[m_1]} \circ \ldots \circ h_{[m_p]}(0)$ est le développement propre $(m_p > 1)$, alors le développement impropre est donné par $v/u = h_{[m_1]} \circ \ldots \circ h_{[m_{p-1}]} \circ h_{[m_p-1]} \circ h_{[1]}(0)$, ou écrit autrement

$$\frac{v}{u} = \cfrac{1}{m_1 + \cfrac{1}{m_1 + \cfrac{1}{\ddots \cfrac{}{m_{p-1} + \cfrac{1}{m_p}}}}} = \cfrac{1}{m_1 + \cfrac{1}{m_1 + \cfrac{1}{\ddots \cfrac{}{m_{p-1} + \cfrac{1}{m_p - 1 + \cfrac{1}{1}}}}}}.$$

Tous les coûts que nous utilisons s'étendent naturellement au développement impropre et au lieu de ne considérer que le développement propre, nous considérerons dans la suite les deux développements. Cela a deux conséquences. Tout d'abord, au lieu de travailler sur l'ensemble $\Omega = (\mathcal{H}^0 + \mathcal{H}^\star\mathcal{F}) \times \mathbb{N}^\star$, nous travaillons sur l'ensemble $\tilde{\Omega} = \mathcal{H}^\star \times \mathbb{N}^\star$. Ensuite, pour tout paramètre X que nous avons défini auparavant, nous n'étudions plus X mais son extension $\tilde{X}$ à $\tilde{\Omega}$. En fait, nous pouvons montrer que pour tous les coûts qui nous intéressent, le changement d'espace probabiliste et de paramètre n'influence en rien les résultats annoncés. Les résultats annoncés sur X sont également vrai pour $\tilde{X}$ et réciproquement, les résultats démontrés pour $\tilde{X}$ sont également vrai pour X.

Fixons X un paramètre sur Ω et $\tilde{X}$ un prolongement de X sur $\tilde{\Omega}$. Pour tout élément x de Ω, l'élément *impropre* de $\tilde{\Omega}$ associé à x est noté $\tilde{x}$. Alors, le premier moment de X satisfait

$$\mathrm{E}_n[X] = \tilde{\mathrm{E}}_n[\tilde{X}] + \frac{1}{2|\Omega_n|} \sum_{x \in \Omega_n} \tilde{X}(\tilde{x}) - X(x).$$

Nous somme donc amené à étudier le paramètre $\tilde{X}(\tilde{x}) - X(x)$ sur Ω_n.
Dans le cas où X est un coût additif associé au coût primitif c, nous avons

$$\tilde{X}(\tilde{x}) = c(m_p - 1) + c(1) + \sum_{i=1}^{p-1} c(m_i)$$

avec $(m_1, \ldots, m_p)$ les quotients du développement propre associé à l'entrée x. En particulier $\tilde{X}(\tilde{x}) - X(x) = c(m_p - 1) + c(1) - c(m_p)$ qui a un premier moment d'ordre constant (cela

se démontre comme les coûts terminaux). Ce paramètre n'intervient donc pas dans le terme dominant.
Il en sera de même pour tous les paramètres abordés. L'erreur commise en considérant le coût sur $\tilde{\Omega}$ n'aura aucune incidence sur les termes étudiés. En revanche, considérer $\tilde{\Omega}$ simplifie les notations puisque l'on évite le cas particulier du dernier quotient (qui doit être différent de 1).

3.3.5 Série pour les coûts additifs

Nous relions maintenant les séries génératrices avec les opérateurs de transfert en utilisant les dictionnaires.

Nous fixons un coût primitif c et C le coût additif associé. Comme $\tilde{\Omega}$ est en bijection avec $\mathcal{H}^\star \times \mathbb{N}^\star$, à un couple (u, v) correspond une unique branche inverse h de $\mathcal{H}^\star$ et son pgcd $d \in \mathbb{N}^\star$. Avec la relation 3.3, l'entier u vérifie $u = d|h'(0)|^{-1/2}$. Par définition, le coût additif C ne fait pas intervenir le pgcd. Nous posons alors $C(u, v) = C(h)$ avec h la branche inverse associée à (u, v). La série génératrice bivariée de C devient

$$S(2s, w) = \sum_{(u,v)\in\Omega} \frac{e^{wC(u,v)}}{u^s} = \sum_{d\geq 1} \sum_{h\in\mathcal{H}^\star} \frac{e^{wC(h)}}{d^{2s}} |h'(0)|^s = \zeta(2s)(\mathbf{I} - \mathbf{G}_{s,w,[c]})^{-1}[\mathbf{1}](0), \qquad (3.14)$$

où ζ est la fonction ζ de Riemann.
Notre objectif est d'étudier la variance des coûts additifs à croissance intermédiaire. Pour cela, il faut analyser les deux premiers moments. Les séries des moments d'ordre 1 et 2 s'obtiennent en dérivant une ou deux fois par rapport à w la série $S(s, w)$. La dérivation d'un quasi-inverse est formellement définie par

$$\frac{d}{dw}(\mathbf{I} - \mathbf{G}_{s,w,[c]})^{-1} = (\mathbf{I} - \mathbf{G}_{s,w,[c]})^{-1} \circ \left(\frac{d}{dw}\mathbf{G}_{s,w,[c]}\right) \circ (\mathbf{I} - \mathbf{G}_{s,w,[c]})^{-1}$$

Après dérivations, nous obtenons la proposition suivante.

Proposition 9 *Les séries des deux premiers moments d'un coût additif C, associé au coût élémentaire c sont données par*

$$S^{[1]}(2s) = \zeta(2s)(\mathbf{I} - \mathbf{G}_s)^{-1} \circ \mathbf{G}_s^{[c]} \circ (\mathbf{I} - \mathbf{G}_s)^{-1}[\mathbf{1}](0),$$

$$\begin{aligned} S^{[2]}(2s) &= \zeta(2s)(\mathbf{I} - \mathbf{G}_s)^{-1} \circ \mathbf{G}_s^{[c]} \circ (\mathbf{I} - \mathbf{G}_s)^{-1}[\mathbf{1}](0) \\ &\quad +2\zeta(2s) \circ (\mathbf{I} - \mathbf{G}_s)^{-1} \circ \mathbf{G}_s^{[c]} \circ (\mathbf{I} - \mathbf{G}_s)^{-1} \circ \mathbf{G}_s^{[c]} \circ (\mathbf{I} - \mathbf{G}_s)^{-1}[\mathbf{1}](0) \end{aligned}$$

Nous verrons par la suite que chaque quasi-inverse admet un pôle simple en $s = 1$. En particulier, la série du premier moment admet un pôle double en $s = 2$ ce qui donne une asymptotique linéaire du premier moment. De même, la série du second moment s'exprime avec trois quasi-inverses et admet un pôle triple en $s = 2$ ce qui implique une asymptotique quadratique du second moment. La section 3.4 décrit précisément les propriétés analytiques de ces séries.

3.3.6 Série pour les coûts terminaux

Les coûts terminaux s'expriment de manière polynomiale en $\ell(m_p)$ et $\ell(v_p)$ et si $p = 0$, alors ils sont nuls. Nous notons W le coût terminal $W(v_0, v_1) = \ell(m_p) + \ell(v_p)$. Si T est un coût terminal, alors il existe k tel que $T = O(W^k)$. Les moments d'un coût terminal sont d'ordre constant si et

seulement si tous les moments de W sont d'ordres constants. Puisque W est additif en la taille du pgcd et du dernier quotient, la stratégie employée avec les coûts additifs s'applique et nous obtenons la proposition suivante.

Proposition 10 *La série de Dirichlet bivariée W admet une expression alternative en fonction des opérateurs de la forme*

$$S_W(2s, w) = \zeta(2s) + \zeta(2s, w, [\ell])\mathbf{G}_{s,w,[\ell]} \circ (\mathbf{I} - \mathbf{G}_s)^{-1}[\mathbf{1}](0)$$

où $\zeta(s)$ est la fonction ζ de Riemann et $\zeta(2s, w, [\ell])$ est la fonction ζ pondérée

$$\zeta(2s, w, [\ell]) = \sum_{d \geq 1} \frac{e^{w\ell(d)}}{d^{2s}}.$$

La série du moment d'ordre k s'obtient en dérivant k fois par rapport à w et vérifie

$$S_W^{[k]}(2s) = \left(\frac{d^k}{dw^k} (\zeta(2s, w, [\ell])\mathbf{G}_{s,w,[\ell]})|_{w=0} \right) \circ (\mathbf{I} - \mathbf{G}_s)^{-1}[\mathbf{1}](0)$$

Comme les séries de tous les moments font intervenir un unique quasi-inverse, elles admettent un unique pôle simple dominant en $s = 2$ ce qui entraîne que les moments sont tous d'ordre constant. Les propriétés analytiques seront exposées précisément à la section 3.4.

3.3.7 Série pour le continuant à une fraction de l'exécution

Nous nous intéressons maintenant au coût $\tilde{L}^{[\delta]}$ défini pour un rationnel δ par

$$\tilde{L}^{[\delta]}(v_0, v_1) = \log v_{\lfloor \delta p \rfloor}.$$

Nous souhaitons montrer que ce paramètre suit une loi normale avec les mêmes moments que le paramètre $L^{[\delta]}$ (cf. théorème 5).
Soit h la branche inverse de profondeur p associée à une entrée (v_0, v_1) et d le pgcd. La branche h peut s'écrire sous la forme $h = h_b \circ h_e$ où h_e est la branche de fin $e_{\lfloor \delta p \rfloor}(h)$. Nous avons alors la relation,

$$\frac{e^{2w \log v_{\lfloor \delta p \rfloor}}}{v_0^{2s}} = \frac{v_{\lfloor \delta p \rfloor}^{2w}}{v_0^{2s}} = \frac{1}{d^{2s-2w}} |h'(0)|^s |h'_e(0)|^{-w}. \tag{3.15}$$

où d est le pgcd de (v_0, v_1). La branche h_e est une branche inverse de profondeur $p - \lfloor \delta p \rfloor$ appartenant à l'ensemble $\mathcal{H}^{p-1-\lfloor \delta p \rfloor}\mathcal{F}$. Si $\mathbf{G}_{s,h}$ désigne l'opérateur composante associée à h, i.e.,

$$\mathbf{G}_{s,h}[f] = |h'|^s f \circ h,$$

alors l'égalité 3.15 devient,

$$\frac{e^{2w \log v_{\lfloor \delta p \rfloor}}}{v_0^{2s}} = \frac{1}{d^{2s-2w}} \mathbf{G}_{s-w,h_e} \circ \mathbf{G}_{s,h_b}[\mathbf{1}](0).$$

En sommant sur tous les pgcd possibles, sur toutes les profondeurs p et toutes les branches h_e et h_b possibles, la série génératrice des moments de $\tilde{L}^{[\delta]}$ satisfait

$$S_{\tilde{L}^{[\delta]}}(2s, 2w) = \zeta(2s - 2w) \sum_{p \geq 0} \mathbf{G}_{s-w}^{p-\lfloor \delta p \rfloor} \circ \mathbf{G}_s^{\lfloor \delta p \rfloor}[\mathbf{1}](0).$$

Nous pourrions nous contenter de cette formule mais elle admet une forme plus simple si le rationnel δ est écrit sous la forme $\delta = c/(c+d)$. Dans ce cas, tout entier p s'écrit sous la forme $p = k(c+d) + r$ avec $0 \leq r < c+d$. La proposition suivante s'obtient en remplaçant la somme sur p par une somme sur k et r.

Proposition 11 *La série de Dirichlet bivariée associée au paramètre $\tilde{L}^{[\delta]}$ admet la formule alternative*

$$S_{L^{[\delta]}}(2s, 2w) = \zeta(2s-2w) \sum_{r=0}^{c+d-1} \mathbf{G}_{s-w}^{r-\lfloor \delta r \rfloor} \circ \mathbb{G}_{s,w} \circ \mathbf{G}_s^{\lfloor \delta r \rfloor}[\mathbf{1}](0)$$

avec $\mathbb{G}_{s,w}$ le pseudo-quasi-inverse

$$\mathbb{G}_{s,w} = \sum_{k \geq 0} \mathbf{G}_{s-w}^{kd} \circ \mathbf{G}_s^{kc}.$$

Le terme de pseudo quasi-inverse vient du fait que pour $w = 0$, on retrouve le vrai quasi-inverse. Si $\mathbf{G}_{s-w}$ et $\mathbf{G}_s$ commutaient, là encore nous aurions un véritable quasi-inverse. Ce n'est malheureusement pas le cas. Toutefois, on s'attend à ce que le pseudo quasi-inverse admette les mêmes propriétés analytiques que le quasi-inverse. En particulier, nous verrons qu'ils apportent aussi un unique pôle dominant.

Remarquons que sur les polynômes, les séries génératrices $G(ze^w)$ et $G(z)$ commutent, ce qui a conduit à un vrai quasi-inverse (cf. formule 2.8).

3.3.8 Série pour M

A la section 3.2.5, nous avons montré comment le théorème 6 se ramène à l'analyse en moyenne du paramètre M^{2w} avec

$$M(v_0, v_1) = \frac{v_{\lfloor \gamma p \rfloor + \lfloor \delta p \rfloor}}{v_{\lfloor \gamma p \rfloor} \cdot v_0^{-\delta+\epsilon}}.$$

Dans la situation de la section précédente, les branches inverses étaient coupées en deux car seul le reste $v_{\lfloor \gamma p \rfloor}$ était abordé (outre v_0). Cette fois, deux continuants sont à traiter : $v_{\lfloor \gamma p \rfloor}$ et $v_{\lfloor \gamma p \rfloor + \lfloor \delta p \rfloor}$. Les branches inverses sont alors décomposées en trois parties.

Fixons (v_0, v_1) une entrée de Ω sur laquelle l'algorithme d'Euclide classique effectue p divisions. Il existe une unique branche inverse h et un unique entier positif d associés à (v_0, v_1). La branche h se décompose en trois branches g, r, k de profondeurs respectives $\lfloor \gamma p \rfloor$, $\lfloor \delta p \rfloor$ et $p - \lfloor \gamma p \rfloor - \lfloor \delta p \rfloor$ et telles que $h = g \circ r \circ k$. De plus, les continuants satisfont

$$(g \circ r \circ k)'(0) = v_0^{-2}, \qquad (r \circ k)'(0) = v_{\lfloor \gamma p \rfloor}^{-2}, \qquad k'(0) = v_{\lfloor \gamma p \rfloor + \lfloor \delta p \rfloor}^{-2}.$$

Le terme général de la série du premier moment de M^{2w} s'écrit alors

$$\frac{M^{2w}(v_0, v_1)}{v_0^{2s}} = \frac{1}{d^{2s^-}} \cdot \frac{v_{\lfloor \gamma p \rfloor + \lfloor \delta p \rfloor}^{2w}}{v_{\lfloor \gamma p \rfloor}^{2w} \cdot v_0^{2s-2(\delta-\epsilon)w}} = k'(0)^{s^-} \cdot r'[k(0)]^{s^+} \cdot g'[r \circ k(0)]^{s^-},$$

avec $s^- := s - (\delta - \epsilon)w$, $s^+ := s + (1 - \delta + \epsilon)w$. La proposition suivante s'obtient en sommant sur toutes les branches g, r, k et sur tous les pgcd d.

Proposition 12 *La série $S_M^{[1]}$ du premier moment de M s'écrit $S_M^{[1]}(2s) = \zeta(2s^-)\mathbb{H}_{s,w}[\mathbf{1}](0)$ avec $\mathbb{H}_{s,w}$ le pseudo quasi-inverse*

$$\mathbb{H}_{s,w} = \sum_{p\geq 0} \mathbf{G}_{s^-}^{p-\lfloor \gamma p\rfloor - \lfloor \delta p\rfloor} \circ \mathbf{G}_{s^+}^{\lfloor \delta p\rfloor} \circ \mathbf{G}_{s^-}^{\lfloor \gamma p\rfloor}.$$

Pour $w = 0$, nous obtenons un vrai quasi-inverse. Les propriétés analytiques de $\mathbb{H}$ sont proches de celles de $\mathbb{G}$. Toutefois, l'analyse pour $\mathbb{G}$ se fait avec un rationnel constant contrairement à celle de $\mathbb{H}$ où nous imposerons des rationnels avec de grands dénominateurs.

3.3.9 Séries pour $A - \overline{A}$ et conjecture (C)

Ici, il n'est pas possible d'obtenir une expression alternative pour la série génératrice bivariée $S_R(s,w)$ ($R = A$ ou $\overline{A}$), mais c'est possible pour les séries $S_R^{[i]}(s)$ avec $i = 1, 2$.

Nous définissons d'abord la notation "crochet" qui sera très utile dans cette section.

Définition 5 *Considérons l'algèbre $\mathcal{A}$ générée par les deux opérateurs suivants : la dérivation par rapport à s, notée Δ, et l'opération de pondération, notée par $W_{[c]}$ qui pondère chaque composante d'un opérateur de transfert avec le coût $c(h)$. Par exemple, $W_{[c]}\mathbf{G}_s = \mathbf{G}_s^{[c]}$. Fixons k éléments $A_1, A_2, \ldots, A_k$ de $\mathcal{A}$. La fonction $[A_1, A_2, \ldots, A_k](s)$ est la fonction de variable s égale à*

$$\zeta(2s)(I - \mathbf{G}_s)^{-1} \circ A_1\mathbf{G}_s \circ (I - \mathbf{G}_s)^{-1} \circ A_2\mathbf{G}_s \circ \ldots \circ A_k\mathbf{G}_s \circ (I - \mathbf{G}_s)^{-1}[1](0)$$

Un crochet contenant k éléments est dit d'ordre k.

En particulier, les séries génératrices d'un coût additif C se simplifient en

$$S_C^{[1]} = [W], \qquad S_C^{[2]} = [W^2] + 2[W, W].$$

Nous étudions maintenant la conjecture (C) et nous obtenons une expression précise de la série de Dirichlet $S_{A-\bar{A}}^{[2]}$.

Proposition 13 *La partie $\widetilde{S}_{A-\bar{A}}^{[2]}$ de la série $S_{A-\bar{A}}^{[2]}$ qui fait seulement intervenir tous les crochets d'ordre au moins 3 peut être écrit comme la somme de deux séries $\Gamma_1(s)$ (qui s'exprime avec tous les crochets d'ordre 4), et $\Gamma_2(s)$ (qui s'exprime avec tous les crochets d'ordre 3), où*

$$\Gamma_1(s) := [\Delta, \Delta, W, W] + [W, W, \Delta, \Delta] - [W, \Delta, \Delta, W] - [\Delta, W, W, \Delta],$$

et

$$2\Gamma_2(s) = [\Delta, \Delta, W^2] + [W^2, \Delta, \Delta] - [\Delta, W^2, \Delta] + [\Delta^2, W, W] + [W, W, \Delta^2] - [W, \Delta^2, W] +$$
$$+ [\Delta, \Delta W, W] + [W, \Delta W, \Delta] - [\Delta, W, \Delta W] - [W, \Delta, \Delta W] - [\Delta W, \Delta, W] - [\Delta W, W, \Delta]$$

Preuve. Nous commençons avec le premier moment, puis nous traitons le moment d'ordre 2.

Moments d'ordre 1. Nous étudions d'abord les coûts élémentaires $[\ell(m_i) \cdot w_i^{2w}]$, $[\ell(m_i) \cdot \bar{a}_i^{2w}]$ pour un petit w. Les séries de Dirichlet correspondantes sont

$$\zeta(2s) \cdot \sum_{p\geq i} \mathbf{G}_{s-w}^{p-i} \circ \mathbf{G}_{s,[\ell]} \circ \mathbf{G}_s^{i-1}[1](0), \qquad \zeta(2s) \cdot \sum_{p\geq i} \mathbf{G}_s^{p-i} \circ \mathbf{G}_{s,[\ell]} \circ \mathbf{G}_{s-w}^{i-1}[1](0).$$

Maintenant, les séries de Dirichlet $(2\log 2)S_A^{[1]}(s), (2\log 2)S_{\bar{A}}^{[1]}(s)$, sont obtenues en prenant la somme sur tous les indices i entre 1 et p, et en prenant la dérivée par rapport à w (en $w=0$). Après la première étape [i.e., en prenant la somme sur tous les indices i],

$$\zeta(2s)\,(I-\mathbf{G}_{s-w})^{-1}\circ\mathbf{G}_s^{[\ell]}\circ(I-\mathbf{G}_s)^{-1}[1](0), \qquad \zeta(2s)\,(I-\mathbf{G}_s)^{-1}\circ\mathbf{G}_s^{[\ell]}\circ(I-\mathbf{G}_{s-w})^{-1}[1](0)$$

et, après la deuxième étape

$$(2\log 2)\cdot S_A^{[1]}=[\Delta,W], \qquad (2\log 2)\cdot S_{\bar{A}}^{[1]}=[W,\Delta] \tag{3.16}$$

Moments d'ordre 2. Pour les trois moments d'ordre 2, $\mathrm{E}_n[A^2], \mathrm{E}_n[\bar{A}^2], \mathrm{E}_n[A\cdot\bar{A}]$, nous traitons d'abord le coût élémentaire $[\ell(m_i)\cdot\ell(m_j)\cdot u_i^{2w}\cdot {u_j}^{2t}]$, pour $u_k\in\{v_k,\bar{a}_k\}$ et des indices fixes i,j avec $1\le i,j\le p$. Il y a deux cas, $i=j$ et $i\neq j$. Tout d'abord, nous prenons la somme sur toutes les paires i,j avec i,j entre 1 et p et tous les p possibles. Nous obtenons une expression alternative pour la série de Dirichlet correspondante (première étape) et ensuite, nous prenons la dérivée par rapport à t,w, en $w=0,t=0$ (deuxième étape). Nous obtenons la série de Dirichlet avec un facteur multiplicatif de $4\log^2 2$.

Comme nous sommes intéressés uniquement par les deux termes dominants dans l'asymptotique, nous n'avons pas besoin des termes avec moins trois quasi-inverses (il y en a 5 au maximum). La série de Dirichlet qui prend en compte uniquement les crochets d'ordre au moins 3 seront notées avec un tilde.

Coût A^2. La série de Dirichlet pour le coût élémentaire $[\ell(m_i)\cdot\ell(m_j)\cdot v_i^{2w}\cdot {v_j}^{2t}]$ avec $j>i$ satisfait

$$\sum_{p\ge j}\mathbf{G}_{s-w-t}^{p-j}\circ\mathbf{G}_{s-w}^{[\ell]}\circ\mathbf{G}_{s-w}^{j-i-1}\circ\mathbf{G}_s^{[\ell]}\circ\mathbf{G}_s^{i-1}[1](0),$$

soit après la première étape pour $i\neq j$,

$$2(I-\mathbf{G}_{s-w-t})^{-1}\circ\mathbf{G}_{s-w}^{[\ell]}\circ(I-\mathbf{G}_{s-w})^{-1}\circ\mathbf{G}_s^{[\ell]}\circ(I-\mathbf{G}_s)^{-1},$$

et finalement après la deuxième étape

$$4[\Delta,\Delta,W,W]+2[\Delta,W,\Delta,W]+2[\Delta^2,W,W]+2[\Delta,\Delta W,W].$$

Ensuite, pour $i=j$, les mêmes idées s'appliquent avec l'opérateur

$$(I-\mathbf{G}_{s-w})^{-1}\circ\mathbf{G}_s^{[\ell^2]}\circ(I-\mathbf{G}_s)^{-1}$$

avec deux dérivations successives par rapport à w (en $w=0$). Cela conduit au terme $2[\Delta,\Delta,W^2]$. Finalement, la séries avec uniquement les crochets d'ordres 3 et 4 vérifie

$$(2\log^2 2)\cdot\widetilde{S}_A^{[2]}=2[\Delta,\Delta,W,W]+[\Delta,W,\Delta,W]+[\Delta^2,W,W]+[\Delta,\Delta W,W]+[\Delta,\Delta,W^2]. \tag{3.17}$$

Coût $\bar{A}^2$. Nous commençons avec le coût élémentaire $[\ell(m_i)\cdot\ell(m_j)\cdot\bar{s}_i^{2w}\cdot{\bar{s}_j}^{2t}]$, et, pour $j>i$, nous obtenons

$$\sum_{p\ge j}\mathbf{G}_s^{p-j}\circ\mathbf{G}_s^{[\ell]}\circ\mathbf{G}_{s-w}^{j-i-1}\circ\mathbf{G}_s^{[\ell]}\circ\mathbf{G}_{s-w-t}^{i-1}[1](\eta).$$

Ensuite, nous prenons la somme sur toutes les paires i,j avec $i\neq j$ et i,j entre 1 et p, et nous dérivons par rapport à t,w (en $w=0,t=0$). Nous obtenons après la première étape [pour A^2]

$$2(I-\mathbf{G}_s)^{-1}\circ\mathbf{G}_s^{[\ell]}\circ(I-\mathbf{G}_{s-w})^{-1}\circ\mathbf{G}_{s-w}^{[\ell]}\circ(I-\mathbf{G}_{s-w-t})^{-1},$$

et après la deuxième étape,

$$4[W,W,\Delta,\Delta]+2[W,\Delta,W,\Delta]+2[W,W,\Delta^2]+2[W,\Delta W,\Delta].$$

Après, pour $i=j$, les mêmes idées s'appliquent avec l'opérateur

$$(I-\mathbf{G}_s)^{-1}\circ\mathbf{G}_s^{[\ell^2]}\circ(I-\mathbf{G}_{s-w})^{-1}$$

et deux dérivées successives par rapport à w. Le terme obtenu est alors $2[W^2,\Delta,\Delta]$. Finalement, pour le coût $\bar{A}^2$,

$$(2\log^2 2)\cdot\widetilde{S}_{\bar{A}}^{[2]}=2[W,W,\Delta,\Delta]+[W,\Delta,W,\Delta]+[W,W,\Delta^2]+[W,\Delta W,\Delta]+[W^2,\Delta,\Delta]. \quad (3.18)$$

Coût $A\bar{A}$. Nous étudions le coût élémentaire $[\ell(m_i)\cdot\ell(m_j)\cdot v_i^{2w}\cdot\bar{s_j}^{2t}]$, pour les trois cas différents $j>i, j<i$ et $j=i$. Pour $j>i$, nous obtenons,

$$\sum_{p\geq j}\mathbf{G}_{s-w}^{p-i}\circ\mathbf{G}_{s-w}^{[\ell]}\circ\mathbf{G}_{s-w-t}^{j-i-1}\circ\mathbf{G}_{s-t}^{[\ell]}\circ\mathbf{G}_{s-t}^{j-1}[1](\eta),$$

et après la première étape

$$(I-\mathbf{G}_{s-w})^{-1}\circ\mathbf{G}_{s-w}^{[\ell]}\circ(I-\mathbf{G}_{s-w-t})^{-1}\circ\mathbf{G}_{s-t}^{[\ell]}\circ(I-\mathbf{G}_{s-t})^{-1}[1](0).$$

Pour $j<i$, nous obtenons,

$$\sum_{p\geq i}\mathbf{G}_{s-w}^{p-i}\circ\mathbf{G}_s^{[\ell]}\circ\mathbf{G}_s^{i-j-1}\circ\mathbf{G}_s^{[\ell]}\circ\mathbf{G}_{s-t}^{i-1}[1](\eta),$$

et après la première étape,

$$(I-\mathbf{G}_{s-w})^{-1}\circ\mathbf{G}_s^{[\ell]}\circ(I-\mathbf{G}_s)^{-1}\circ\mathbf{G}_s^{[\ell]}\circ(I-\mathbf{G}_{s-t})^{-1}[1](0)$$

Finalement, pour $i=j$, les mêmes idées s'appliquent à l'opérateur

$$(I-\mathbf{G}_{s-w})^{-1}\circ\mathbf{G}_s^{[\ell^2]}\circ(I-\mathbf{G}_{s-t})^{-1}$$

et les deux dérivées successives par rapport à w et t (en $w,t=0$) et conduisent au terme $[\Delta,W^2,\Delta]$. La série de Dirichlet de $A\bar{A}$ avec les crochets d'ordre 3 ou 4 est alors,

$$\begin{aligned}(4\log^2 2)\cdot\widetilde{S}_{A\bar{A}}^{[1]}&=2[W,\Delta,\Delta,W]+2[\Delta,W,W,\Delta]+[W,\Delta,W,\Delta]+[\Delta,W,\Delta,W]+\\&+[W,\Delta^2,W]+[\Delta,W,\Delta W]+[W,\Delta,\Delta W]+[\Delta W,\Delta,W]+[\Delta W,W,\Delta]+[\Delta,W^2,\Delta].\end{aligned} \quad (3.19)$$

Avec une combinaison des trois séries relatives à $A^2,\bar{A}^2,A\bar{A}$, nous obtenons le résultat de la proposition. ■

Au passage, nous avons aussi calculé les séries génératrices des deux premiers moments de A (formules 3.16 et 3.17). Nous rappelons que si nous montrons que l'espérance et la variance de A sont respectivement d'ordre quadratique et d'ordre au plus cubique, alors nous aurons démontré la première partie du théorème 2 sur la complexité binaire classique.

3.3.10 Conjecture (G)

La conjecture (G) suppose que les paramètres $N^{(v)}$ et $N^{(m)}$ définis par

$$N^{(v)} = \sum_{i=1}^{p} \lg w_i, \qquad \text{et} N^{(m)} = \sum_{i=1}^{p} i \cdot \ell(m_i)$$

suivent une loi limite gaussienne. Nous ne savons pas démontrer ces lois limites mais il existe des formules alternatives aux séries génératrices bivariées associées à ces paramètres. Si $S_{(v)}(s,w)$ et $S_{(m)}(s,w)$ sont ces deux séries, nous avons

$$S_{(v)}(2s,2w) = \zeta(2s) \cdot \left(\mathbf{I} + \sum_{p\geq 1} \mathbf{G}_{s-pw} \circ \mathbf{G}_{s-(p-1)w} \circ \ldots \circ \mathbf{G}_{s-w} \right) [\mathbf{1}](0),$$

et

$$S_{(v)}(2s,2w) = \zeta(2s) \cdot \left(\mathbf{I} + \sum_{p\geq 1} \mathbf{G}_{s,pw} \circ \mathbf{G}_{s,(p-1)w} \circ \ldots \circ \mathbf{G}_{s,w} \right) [\mathbf{1}](0)$$

où $\mathbf{G}_{s,w} = \mathbf{G}_{s,w,[\ell]}$. Notons que le premier opérateur $\mathbb{S}_{s,w}$ dans les premières parenthèses vérifie l'équation fonctionnelle

$$\mathbb{S}_{s,w} = \mathbf{I} + \mathbb{S}_{s-w,w} \circ \mathbf{G}_{s-w}$$

alors que le deuxième opérateur ne semble pas avoir d'équation similaire. Les deux opérateurs présents dans les deux séries peuvent aussi être appelés des pseudo quasi-inverse. Mais nous ne sommes pas parvenu à analyser leurs propriétés analytiques.

3.3.11 État d'avancement de l'analyse

Nous en avons terminé avec la deuxième étape de l'analyse dynamique. Pour tous les paramètres d'intérêt, c'est à dire les coûts additifs, les coûts terminaux, le logarithme du continuant à une fraction de l'exécution $(\tilde{L}^{[\delta]})$, les coûts M, A et $\overline{A}$, nous avons déterminé une formule alternative des séries en fonction des opérateurs de transfert. Ces expressions font intervenir des quasi-inverses ainsi que des pseudo quasi-inverses. Nous avons également introduit les crochets afin d'écrire simplement les formes alternatives.

3.4 Propriétés analytique des séries

L'étape suivante dans une analyse dynamique est l'étude des propriétés analytiques des opérateurs de transfert, et principalement des (pseudo) quasi-inverses. Ensuite, ces propriétés analytiques sont transférées sur les séries génératrices. L'analyse des opérateurs est une tache technique que nous traitons complètement au prochain chapitre. Nous donnons ici uniquement les conséquences de ces analyses sur les séries.

3.4.1 Propriétés $US(s)$ et $US(s,w)$

Une série de Dirichlet satisfait une propriété de type US si elle admet un unique pôle dans un demi plan de la forme $\Re(s) > \sigma_0$ et si sur l'axe $\Re(s) = \sigma_0$, la série admet une borne à la Dolgopyat [Dol98],

$$|f(\sigma_0 + it)| \leq \kappa \cdot \max(1, \cdot t^{\xi}).$$

Ce type de propriété est difficile à démontrer pour les séries génératrices de Dirichlet. C'est par exemple ce qui est utilisé pour le théorème des nombres premiers. La propriété équivalente pour les séries ordinaires est d'avoir un disque contenant un unique pôle, ce qui est beaucoup plus simple à obtenir.

Définition 6 (Propriétés US) *(i) Une série de Dirichlet univariée $f(s)$ satisfait la propriété $US(s)$ s'il existe $\alpha > 0$, $\xi \in [0, 1[$, deux constantes $\kappa > 0$ et $t_0 > 0$ tels que*

1. *f admet un unique pôle $s = \sigma$ dans le demi-plan $\Re(s) > \sigma - \alpha$,*
2. *sur la ligne verticale $\Re(s) = \sigma - \alpha$, la série est uniformément bornée*

$$|f(\sigma - \alpha + it)| \leq \kappa \cdot \max(1, |t|^{\xi}), \qquad \forall t \in \mathbb{R},$$

3. *loin de l'axe réel, la série est uniformément bornée sur la bande $|\Re(s) - \sigma| \leq \alpha$ et $|\Im(s)| > t_0$,*

$$|f(s)| \leq \kappa \cdot \max(1, |t|^{\xi}), \qquad \forall s,\ |\Im(s)| > t_0.$$

(ii) Une série de Dirichlet bivariée $f(s, w)$ satisfait la propriété $US(s, w)$ s'il existe un voisinage complexe $\mathcal{W}$ de $w = 0$, un voisinage réel Σ de la forme $\Sigma =]\sigma_0 - \alpha, \sigma_0 + \alpha[$ avec $\alpha > 0$, des constante $\xi \in [0, 1[$, $t_0 > 0$ et $\kappa > 0$ tels que

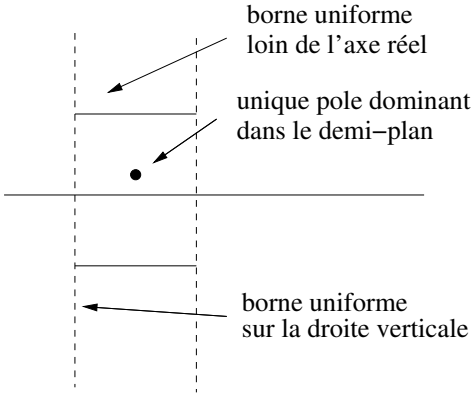
1. *$f(s, w)$ admet un unique pôle simple en $s = \sigma(w)$ avec $\Re(\sigma(w)) \subset \Sigma$ pour tout $w \in \mathcal{W}$,*
2. *sur la ligne verticale $\Re(s) = \sigma_0 - \alpha$, la série est uniformément bornée*

$$|f(\sigma_0 - \alpha + it, w)| \leq \kappa \cdot \max(1, |t|^{\xi}), \qquad \forall t \in \mathbb{R}, \forall w \in \mathcal{W},$$

3. *loin de l'axe réel, la série est uniformément bornée sur la bande $\Re(s) \in \Sigma$ et $|\Im(s)| > t_0$.*

$$|f(s, w)| \leq \kappa \cdot \max(1, |t|^{\xi}), \qquad \forall |\Im(s)| > t_0, \forall w \in \mathcal{W}.$$

Outre la condition sur le pôle simple, la propriété $US(s, w)$ est plus forte que la propriété $US(s)$ car la borne est uniforme sur le voisinage complexe $\mathcal{W}$. Au prochain chapitre, nous allons montrer que les (pseudo) quasi-inverses satisfont une propriété de type US qu'ils transmettent aux séries. La figure suivante résume une propriété US.



Nous pouvons maintenant décrire les propriétés analytiques des séries.

3.4.2 Coût additifs, coûts terminaux, coûts A et $\overline{A}$

Les séries des coûts additifs, des coûts terminaux et des coûts A et $\overline{A}$ ont un point commun : elles font toutes intervenir le quasi-inverse $(\mathbf{I} - \mathbf{G}_s)^{-1}$. Or, Dolgopyat [Dol98] puis Baladi et Vallée [BV05, BV04] ont montré que le quasi-inverse satisfait la propriété $US(s)$ avec un pôle simple en $s = 1$ (voir prochain chapitre). Cette propriété se transfère bien entendu sur les séries. Nous obtenons alors la proposition suivante.

Proposition 14 *(i) [coûts additifs] Pour un coût additif à croissance intermédiaire, la série $S^{[i]}(2s)$ du moment d'ordre i pour $i = 1, 2$ (donnée par la proposition 9) satisfait la propriété $US(s)$ et admet un unique pôle d'ordre $i+1$ en $s = 1$.*
(ii) [coûts terminaux] Pour le coût terminal W, la série $S_W^{[k]}(2s)$ (voir proposition 10) du moment d'ordre k, satisfait la propriété $US(s)$ et admet un unique pôle simple en $s = 1$.
(iii) [coûts A et $\overline{A}$] Les séries $S_A^{[1]}(2s)$, $S_{\bar{A}}^{[1]}(2s)$, $S_A^{[2]}(2s)$, $S_{\bar{A}}^{[2]}(2s)$, $S_{A\bar{A}}^{[1]}(2s)$ et $S_{A-\bar{A}}^{[2]}(2s)$ (voir section 3.3.9) satisfont la propriété $US(s)$ et admettent un unique pôle en $s = 1$ d'ordre 3 pour les séries $S_A^{[1]}(2s)$ et, $S_{\bar{A}}^{[1]}(2s)$, et d'ordre 5 pour les autres.

3.4.3 Paramètre $\tilde{L}^{[\delta]}$

La série de Dirichlet bivariée associée à $\tilde{L}^{[\delta]}$ fait intervenir le pseudo quasi-inverse $\mathbb{G}_{s,w}$. Pour $w = 0$, le pseudo quasi-inverse est un vrai quasi-inverse. Par perturbation, on s'attend à ce qu'il vérifie une propriété $US(s, w)$ et qu'il en soit de même avec la série de Dirichlet. La propriété $US(s, w)$ pour le pseudo quasi-inverse sera démontrée au prochain chapitre.

Proposition 15 *La série génératrice bivariée $S_{[\delta]}(2s, 2w)$ satisfait la propriété $US(s, w)$. De plus, si $s = \sigma(w)$ est l'unique pôle dominant pour $\Re(s) > \sigma_0$ et $w \in \mathcal{W}$, alors*

$$\sigma(0) = 1, \qquad \sigma'(0) = 1 - \delta \qquad et \qquad \sigma''(0) = \delta(1-\delta)\frac{\Lambda''(1)}{\Lambda'(1)},$$

où $\Lambda(s) = \log \lambda(s)$ avec $\lambda(s)$ l'unique valeur propre dominante de l'opérateur $\mathbf{G}_s$.

3.4.4 Paramètre M

Contrairement au pseudo quasi-inverse $\mathbb{G}_{s,w}$, l'opérateur $\mathbb{H}_{s,w}$ sera utilisé avec des rationnels δ et γ dont le dénominateur tend vers l'infini. Si l'on se rappelle l'équivalent de $\mathbb{G}_{s,w}$ sur les polynômes (donnée par la formule 2.8), nous avons vu qu'il admettait des pôles sous-dominants de plus en plus proche du pôle dominant dès que le dénominateur grandit. Le même phénomène apparaît avec les pseudo-quasi-inverses $\mathbb{G}_{s,w}$ et $\mathbb{H}_{s,w}$. Comme l'opérateur $\mathbb{H}_{s,w}$ est utilisé avec des dénominateurs grands, nous devons préciser la condition $US(s, w)$ en donnant des informations sur l'évolution des voisinages mais aussi sur la borne à gauche.

Proposition 16 *Il existe huit constantes $K_0, K_1, K_2, K_3, K_4', K_5', K_6', K_7'$ vérifiant la propriété suivante : pour tout entier $D \geq 1$, notons $\mathcal{S}$ la bande de la forme $|\Re s - 1| \leq K_1/D^2$, et $\mathcal{W}$ le voisinage* réel *de $w = 0$ de la forme $|w| \leq K_2/D$. Alors, pour tout rationnel $\delta \in]0, 1[$ de dénominateur D, pour tout ϵ avec $1/D \leq |\epsilon| \leq K_3/D$, la série $S_M(2s, 2w)$ relative au paramètre M^{2w}, avec*

$$M(a_0, a_1) := \frac{a_{\lfloor \gamma p \rfloor + \lfloor \delta p \rfloor}}{a_{\lfloor \gamma p \rfloor} \cdot a_0^{(-\delta+\epsilon)}}$$

satisfait les propriétés suivantes :

(i) Pour tout $w \in \mathcal{W}$, la série $s \to S_M(2s, 2w)$ admet un unique pôle $s = \sigma(w)$ dans la bande $\mathcal{S}$. Ce pôle est d'ordre 1.

(ii) Pour tout $w \in \mathcal{W}$, on a $|\sigma(w) - 1| \leq K_4'/(2D^2)$. De plus, il existe $w_0 \in \mathcal{W}$ du même signe que ϵ, pour lequel $\sigma(w_0) - 1 < 0$ et $|\sigma(w_0) - 1| \geq \epsilon^2/(4K)$.

(iii) Sur $\mathcal{W}$, le résidu $E(w)$ de $S_M(2s, 2w)$ en $s = \sigma(w)$ satisfait $K_5' \leq |E(w)| \leq K_6'$.

(iv) Sur la ligne verticale $\Re s = 1 - K_1/D^2$, et pour tout $w \in \mathcal{W}$, la série $S_M(2s, 2w)$ satisfait

$$|S_M(2s, 2w)| \leq K_7' \cdot D^2 \max(1, \Im s)^\xi.$$

(v) Il existe $t_0 > 0$ tel que pour $s \in \mathcal{S}$ avec $|\Re(s)| > t_0$, la série vérifie

$$|S_M(2s, 2w)| \leq K_7' \cdot D^2 (\Im s)^\xi.$$

En d'autres termes, lorsque le dénominateur D est fixé, la série $S_M(2s, 2w)$ satisfait la propriété $US(s, w)$ avec un voisinage $\mathcal{W}$ de taille $O(1/D)$ et Σ_0 de l'ordre de $1/D^2$.

3.4.5 État d'avancement de l'analyse

Dorénavant, nous connaissons toutes les propriétés analytiques dominantes des séries génératrices de tous les paramètres. En particulier, toutes ces séries admettent une bande sans pôle à gauche du pôle dominant. Cette propriété découle directement des propriétés analytiques des (pseudo-)quasi-inverses que nous établirons au prochain chapitre.

Pour faire un parallèle avec les séries ordinaires, c'est comme si nous venions d'établir que les séries admettent un unique pôle dans un disque. Pour les séries génératrices ordinaires, il est alors très facile d'extraire les coefficients avec des termes d'erreurs. Pour les séries de Dirichlet, l'extraction est un peu plus complexe mais le principe reste le même.

3.5 Extractions et asymptotiques

Maintenant que les propriétés analytiques des séries sont connues, nous pouvons passer à la dernière étape d'une analyse en moyenne : l'extraction des coefficients des séries et le calcul des asymptotiques.

L'extraction se base essentiellement sur la localisation du pôle dominant et ses propriétés. Pour extraire des coefficients dans une série de Dirichlet, nous disposons de deux outils : la formule de Perron et le théorème taubérien de Delange [Del54]. Le théorème de Delange nécessite des informations uniquement sur la droite verticale qui contient le pôle mais il ne donne qu'une asymptotique sans terme d'erreur. La formule de Perron, appliquée à une série de Dirichlet qui admet une propriété de type US, fournit un développement asymptotique précis. Jusqu'au travail très récent de Baladi et Vallée [BV04], il n'était pas prouvé que les opérateurs admettaient une propriété US. Les analyses dynamiques (sur les algorithmes euclidiens) se bornaient donc au théorème taubérien de Delange et à des analyses en moyenne.

Nous commençons cette section avec la formule de Perron et son utilisation dans les extractions. Ensuite, nous appliquons cette formule pour le calcul de tous les moments de tous les paramètres.

3.5.1 Formule de Perron

La formule de Perron d'ordre 2 relie les coefficients d'une série de Dirichlet avec une intégrale complexe sur le demi-plan de convergence.

Formule de Perron *Soit* $F(2s) := \sum_{n\geq 1} a_n/n^{2s}$ *une série de Dirichlet convergente sur* $\Re(s) > \sigma_0$ *et fixons* $D > \max(0, \sigma_0)$. *La formule de Perron d'ordre 2 [EE85] est donnée par*

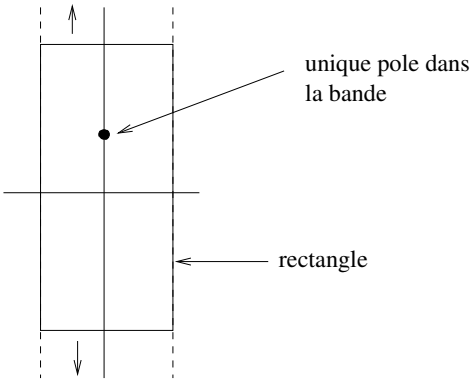
$$\Psi(T) := \sum_{n\leq T}\sum_{j=1}^{n} a_j = \frac{1}{2i\pi}\int_{D-i\infty}^{D+i\infty} F(s)\frac{T^{2s+1}}{s(2s+1)}ds.$$

Les formules sur les moments font intervenir les sommes

$$\psi(n) = \sum_{j=1}^{n} a_j$$

et plus particulièrement, $\psi(2^n-1) - \psi(2^{n-1})$. Il n'est pas possible de retrouver directement ces grandeurs, mais Baladi et Vallée [BV04] et Cesaratto [Ces06] ont décrit une démarche générale pour retrouver les moments des paramètres.

Etape 1 : extraction par la formule de Perron. Si $F(2s)$ satisfait la propriété $US(s)$, alors F admet un unique pôle $s = \sigma_0$ dans une bande de la forme $\Re(s) \in]\sigma_0 - \alpha, \sigma_0 + \alpha[$ et est uniformément bornés sur la ligne verticale gauche et loin de l'axe réel. En intégrant sur un rectangle $\mathcal{R}$ délimité par les droites $\Re(s) = \sigma_0 \pm \alpha$ et $|\Im(s)| = t$, et en faisant tendre t vers l'infini,



l'intégrale à droite est exactement la formule de Perron, l'intégrale à gauche est bornée par $O(T^{2\sigma_0+1-2\alpha})$ à cause de la propriété $US(s)$ et pour les mêmes raisons, les intégrales sur les lignes horizontales tendent vers 0. L'intégrale sur le rectangle est également donnée par le résidu selon le théorème de Cauchy. Au final, l'étape d'extraction donne

$$\Psi(T) = \mathrm{Res}(F(2s)\frac{T^{2s+1}}{s(2s+1)}, s = \sigma_0) + O(T^{2\sigma_0+1-2\alpha}).$$

où le résidu est de la forme $T^{2\sigma_0+1}P(\log T)(1+O(T^{-1}))$ avec P un polynôme de degré k si le pôle est d'ordre $k+1$. Avec les séries génératrices bivariées, la démarche est la même et nous obtenons une asymptotique des $\Psi(T)$ avec un terme d'erreur uniforme en w. Le fait que le terme d'erreur soit uniforme en w est important pour appliquer le théorème des quasi-puissances de Hwang.

Etape 2 : passage par les distributions lissées. Les $\Psi(T)$ ne sont pas adaptés pour le calcul des moments des paramètres avec la distribution uniforme sur les Ω_n. En revanche, ils correspondent à une autre distribution $\overline{\mathbb{P}}_n$ sur un ensemble légèrement différent $\overline{\Omega}_n$ (voir [BV04, Ces06]). Sur cet ensemble, les paramètres ont un sens et donc une espérance, une variance, ... qui se calculent avec les $\Psi(T)$.

Etape 3 : transfert des résultats sur Ω_n. Les espérances et les variances pour la distribution lisse étant connues, il est possible à travers les lemmes 10 à 14 de [BV04], de les transférer sur Ω_n avec la distribution uniforme. Ce transfert nécessite toutefois que dans le pire des cas, le paramètre est au plus d'ordre polynomial en la taille des entrées, ce qui est évidemment le cas de tous nos paramètres.

3.5.2 Application aux paramètres d'intérêt

Toutes les séries admettent une propriété de type US. Il est alors possible d'appliquer la formule de Perron pour extraire les coefficients.

3.5.2.1 Constantes dominantes

Nous commençons par un lemme qui donne les constantes dominantes d'un crochet. Nous rappelons que les crochets interviennent à la fois pour les coûts additifs, les coûts terminaux et les coûts A et $\overline{A}$.

Lemme 3 (Equivalent autour de $s = 1$ des crochets) *Toute série de Dirichlet de la forme $[A_1, A_2, \dots A_k]$, où chaque A_i opère sur les opérateurs de transfert et tel que $A_i\mathbf{G}[\psi]$ est intégrable, a un pôle d'ordre $k+1$ en $s = 1$, et elle admet un développement de la forme*

$$[A_1, A_2, \dots, A_k](s) = \sum_{p=0}^{k} \frac{a_{k-p}}{|\lambda'(1)|^p} \cdot \frac{1}{(s-1)^{p+1}} + O(1)$$

$$\textit{avec} \quad a_0 = \prod_{i=1}^{k} I[A_i\mathbf{G}] \qquad \textit{où} \quad I[\mathbf{H}] := \int_I \mathbf{H}[\varphi](t)dt \quad \textit{et} \quad \varphi(x) := \frac{1}{\log 2}\frac{1}{1+x}. \tag{3.20}$$

On constate que la constante dominante ne dépend que des opérateurs $A_1, A_2, \dots, A_k$ et non de l'ordre dans lequel ils sont disposés . Cette propriété remarquable entraînera des annulations dans les termes dominants.

Cette propriété vient d'un équivalent en $s = 1$ du quasi-inverse,

$$(\mathbf{I} - \mathbf{G}_s)^{-1}[f] \sim \frac{1}{s-1} \cdot \frac{-1}{\lambda'(s)}\varphi \int_I f(t)dt. \tag{3.21}$$

L'opérateur $\mathbf{G}_s$ admet une unique valeur propre dominante $\lambda(s)$ séparée du reste du spectre par un saut spectral. Cette propriété induit une décomposition de l'opérateur de transfert et par itération du quasi-inverse,

$$\mathbf{G}_s = \lambda(s)\mathbf{P}_s + \mathbf{N}_s, \qquad (\mathbf{I} - \mathbf{G}_s)^{-1} = \frac{1}{1-\lambda(s)}\mathbf{P}_s + (\mathbf{I} - \mathbf{N}_s)^{-1}$$

où $\mathbf{P}_s$ est un projecteur qui commute avec $\mathbf{N}_s$ et $\mathbf{N}_s$ a un rayon spectral strictement plus petit que $|\lambda(s)|$. Avec cette décomposition, le quasi-inverse admet un pôle dominant lorsque $\lambda(s) = 1$, ce qui est le cas en $s = 1$. Maintenant, en $s = 1$, l'opérateur $\mathbf{P}_1$ est connue et satisfait

$$\mathbf{P}_1[f] \sim \varphi \int_I f(t)dt.$$

L'équivalent du quasi-inverse vient alors immédiatement. Remarquons aussi que la relation $\zeta(2) = \pi^2/6$ et $\lambda'(1) = -\pi^2/(6\log 2)$ intervient dans le développement du crochet pour enlever un $\lambda'(1)$.

Comme chaque quasi-inverse satisfait une propriété du type $US(s)$, il est possible d'appliquer aux crochets la formule de Perron et la méthodologie développée par Baladi-Vallée et Cesaratto [BV04, Ces06] afin d'extraire l'asymptotique des coefficients des séries. Ceci est l'objet du lemme suivant.

Lemme 4 (Développements asymptotiques) *Soit R un coût pour lequel la série de Dirichlet $S_R(2s)$ admet une formule qui fait intervenir un crochet de la forme $[A_1, A_2, \ldots, A_k]$. Alors, ce crochet contribue à l'espérance $\mathrm{E}_n[R]$ avec un terme*

$$\left(\sum_{p=0}^{k} \frac{(2\log 2)^p}{p!|\lambda'(1)|^p} \cdot a_{k-p} n^p\right) \cdot (1 + O(2^{-n\beta})),$$

pour un $\beta > 0$. De plus, la constante dominante a_0 vérifie

$$a_0 = \prod_{i=1}^{k} I[A_i \mathbf{G}].$$

Nous disposons maintenant de tous les outils pour extraire les coefficients des séries.

3.5.2.2 Coûts additifs à croissance intermédiaire

Les séries génératrices pour un coût additif sont données par

$$S_C^{[1]}(2s) = [W_{[c]}], \quad \text{et} \quad S_C^{[2]}(2s) = 2[W_{[c]}, W_{[c]}] + [W_{[c^2]}].$$

Le lemme 4 montre que les deux premiers moments des coûts additifs vérifient

$$\mathrm{E}_n[C^i] = 2 \cdot d_0^{[i]} (n\log 2)^i (1 + O(n^{-1})) = \frac{2^i I[W_{[c]}\mathbf{G}_s]^i}{|\lambda'(1)|^i} (n\log 2)^i (1 + O(n^{-1})).$$

Le moment d'ordre 2 est donc équivalent au carré du moment d'ordre 1 ce qui montre que la variance est d'ordre au plus linéaire.
Ceci termine la preuve du théorème 12 pour les coûts additifs.■

3.5.2.3 Coûts terminaux

Nous adoptons ici la même méthode que pour les coûts additifs. La série génératrice du moment d'ordre k du coût terminal W est de la forme

$$S_W^{[k]}(2s) = \mathbf{H}_s \circ [\,] \quad \text{où} \quad \mathbf{H}_s = \frac{d^k}{dw^k}(\zeta(2s, w, [\ell])\mathbf{G}_{s,w,[\ell]})|_{w=0}.$$

Le lemme 4 entraîne que tous les moments de W satisfont

$$\mathrm{E}_n[W^i] = 2 \cdot b_0^{[k]}(1 + O(n^{-1})) = 2\mathbf{H}_s[\varphi](0)(1 + O(n^{-1})).$$

ce qui montre que tous les moment du coût terminal W sont d'ordre constant.
Ceci termine la preuve du théorème 12 pour les coûts terminaux.■

3.5.2.4 Complexité binaire classique et conjecture (C)

Nous prouvons ici le théorème 2 concernant la complexité binaire de l'algorithme d'Euclide classique. A la section 3.3.9, nous avons réduit l'analyse de B à l'analyse du coût A puisque nous avons les deux points suivants

$$\mathrm{E}_n[B] = \mathrm{E}_n[A] + O(n), \qquad \mathrm{V}_n[B] = O(\mathrm{V}_n[A] + n^2).$$

De plus, si $\mathrm{V}_n[A]$ est d'ordre cubique, alors $\mathrm{V}_n[B]$ est aussi d'ordre cubique. Ces propriétés sont dues à la décomposition

$$B = A + \lg v_p \cdot Z + O(Z) \quad \text{avec} \quad Z(v_0, v_1) = \sum_{i=1}^{p} \ell(m_i).$$

Dans cette décomposition, $\lg v_p \cdot Z$ et $O(Z)$ ont une espérance linéaire et une variance au plus quadratique.
Nous prouvons tout d'abord le théorème 2 (i). De nombreux crochets interviennent dans les séries génératrices de A et $\overline{A}$. Pour chacun de ces crochets, nous devons calculer le terme dominant (voir sous dominant pour la conjecture (C)) qui apparaît finalement dans l'asymptotique des moments . Cette constante se calcule avec le lemme 4 qui traite de l'asymptotique des coefficients des crochets.

Les séries génératrices des deux premiers moments de A sont données par les formules 3.16 page 78 et 3.17 page 78. Le lemme 4 entraîne les estimations suivantes pour $\mathrm{E}_n[A], \mathrm{E}_n[A^2]$,

$$(\log 2) \cdot \mathrm{E}_n[A] = \frac{1}{2} I[\Delta\mathbf{G}] \cdot I[W\mathbf{G}] \cdot \left(\frac{2^2}{2!} \cdot \frac{(\log 2)^2}{|\lambda'(1)|^2} \right) n^2 + O(n),$$

$$(\log^2 2) \cdot \mathrm{E}_n[A^2] = \frac{3}{2} I[\Delta\mathbf{G}]^2 \cdot I[W\mathbf{G}]^2 \cdot \left(\frac{2^4}{4!} \cdot \frac{(\log 2)^4}{|\lambda'(1)|^4} \right) n^4 + O(n^3).$$

Nous remarquons que les termes dominants (d'ordre 4) sont les mêmes dans $\mathrm{E}_n[A]^2$ et $\mathrm{E}_n[A^2]$, ce qui prouve une estimation pour la variance de A de la forme

$$\mathrm{V}_n[A] = \tau \cdot n^3 + O(n^2).$$

Ceci termine le premier point du théorème 2.

La conjecture (C) sur les paramètres A et $\bar{A}$ est donnée par

$$(C) \qquad \mathrm{V}_n[A - \bar{A}] - \frac{1}{3}\mathrm{V}_n[A + \bar{A}] = O(n^2).$$

Pour montrer la conjecture (C), nous devons prouver que $\tau = (1/3)\rho(\ell)$ où $\rho(\ell)$ est la constante qui apparaît dans le terme dominant de la variance du théorème 3 page 28.

Lemme 5 *Soit* $\mathbf{Q}$ *l'opérateur de Porter défini comme le terme constant du quasi-inverse* $(\mathbf{I} - \mathbf{G}_S)^{-1}$ *lors du développement en* $s = 1$,

$$(I - \mathbf{G}_s)^{-1} = \frac{1}{s-1} \frac{\mathbf{P}}{|\lambda'(1)|} + \mathbf{Q} + O(s-1)$$

avec $\mathbf{P}[f] = \varphi \int_0^1 f(t)dt$. *Alors, la conjecture* (C) *est vérifiée si pour tout* $X, Y \in \{\Delta, W\}$, *on a*

$$I[X\mathbf{G} \circ \mathbf{Q} \circ Y\mathbf{G}] = \int_I (X\mathbf{G})[Y\varphi](t)dt - I[X\mathbf{G}] \int_I [Y\varphi](t)dt.$$

Preuve. Avec la proposition 13 et le lemme 3, les deux séries Γ_1 et Γ_2 ont un pôle d'ordre au plus 4 en $s = 1$ et peuvent s'écrire sous la forme

$$\Gamma_i(s) = \frac{1}{\log 2 \cdot |\lambda'(1)|^4} \frac{1}{(s-1)^4} \gamma_i + O(\frac{1}{(s-1)^3}).$$

Forme explicite pour les constantes γ_i. Le coefficient dominant γ_2 est égal à

$$2\gamma_2 = I[\Delta\mathbf{G}]^2 \cdot I[W^2\mathbf{G}] + I[\Delta^2\mathbf{G}] \cdot I[W\mathbf{G}]^2 - 2I[\Delta\mathbf{G}] \cdot I[W\mathbf{G}] \cdot I[\Delta W\mathbf{G}].$$

Le coefficient dominant γ_1 s'exprime avec le terme sous-dominant de $\Gamma_1(s)$. En $s = 1$, les développement des trois opérateurs $(I - \mathbf{G}_s)^{-1}, \Delta\mathbf{G}_s, W\mathbf{G}_s$ font respectivement intervenir les trois opérateurs $\mathbf{Q}, \Delta^2\mathbf{G}, \Delta W\mathbf{G}$, sous la forme

$$(I - \mathbf{G}_s)^{-1} = \frac{1}{s-1} \frac{\mathbf{P}}{|\lambda'(1)|} + \mathbf{Q} + O(s-1),$$

$$\Delta\mathbf{G}_s = \Delta\mathbf{G} + (s-1)\Delta^2\mathbf{G} + O((s-1)^2) \qquad W\mathbf{G}_s = W\mathbf{G} + (s-1)\Delta W\mathbf{G} + O((s-1)^2).$$

La constante sous-dominante de la série Γ_1 est obtenue en remplaçant dans chacun des quatre termes de Γ, un des neuf facteurs par sa constante sous-dominante. Cependant, tous les termes obtenus en remplaçant $\Delta\mathbf{G}_s$ ou $W\mathbf{G}_s$ par leur terme sous-dominant disparaissent. C'est la même chose avec les termes qui contiennent l'opérateur $\mathbf{Q}$ au début ou à la fin. Alors, la constante sous-dominante γ_1 de Γ_1 s'exprime via les intégrales de la forme $I[\mathbf{H}]$ (définie en 3.20) comme une somme de quatre termes principaux,

$$\gamma_1 = \sum_{\substack{X,Y \in \{\Delta, W\} \\ X' \neq X, Y' \neq Y}} (-1)^{[X=Y]} \cdot I[X\mathbf{G}] \cdot I[Y\mathbf{G}] \cdot I[X'\mathbf{G} \circ \mathbf{Q} \circ Y'\mathbf{G}].$$

Finalement, le coefficient $\gamma := \gamma_1 + \gamma_2$ satisfait

$$2\gamma = I[\Delta\mathbf{G}]^2 \left(I[W^2\mathbf{G}] + 2I[W\mathbf{G} \circ \mathbf{Q} \circ W\mathbf{G}]\right) + I[W\mathbf{G}]^2 \left(I[\Delta^2\mathbf{G}] + 2I[\Delta\mathbf{G} \circ \mathbf{Q} \circ \Delta\mathbf{G}]\right)$$
$$-2I[W\mathbf{G}] \cdot I[\Delta\mathbf{G}] \left(I[\Delta W\mathbf{G}] + I[\Delta\mathbf{G} \circ \mathbf{Q} \circ W\mathbf{G}] + I[W\mathbf{G} \circ \mathbf{Q} \circ \Delta\mathbf{G}]\right).$$

Ensuite, le lemme 4 sur l'asymptotique des crochets implique

$$(\log^2 2) \cdot \mathrm{E}_n[(A - \bar{A})^2] = \frac{2^3 (\log 2)^3}{3! |\lambda'(1)|^3} \cdot \gamma n^3 + O(n^2) \quad \text{soit} \quad \mathrm{E}_n[(A - \bar{A})^2] = \frac{2 \log 2}{3 |\lambda'(1)|^3} \cdot (2\gamma) \cdot n^3 + O(n^2).$$

Forme explicite de la constante $\rho(\ell)$. Pour la conjecture (C), nous comparons la constante 2γ avec la constante $\rho(\ell)$ qui apparaît dans la variance de la complexité binaire étendue,

$$\mathrm{V}_n[A + \bar{A}] = \frac{2 \log 2}{|\lambda'(1)|^3} \cdot \rho(\ell) \cdot n^3 + O(n^2).$$

Une expression alternative pour $\rho(\ell)$ est (voir [BV05, BV04])

$$\rho(\ell) = \lambda'^2_s(1,0) \cdot \lambda''_{w^2}(1,0) - 2\lambda'_w(1,0) \cdot \lambda'_s(1,0) \cdot \lambda''_{sw}(1,0) + \lambda'^2_w(1,0) \cdot \lambda''_{s^2}(1,0).$$

Maintenant, pour $(X, Y) \in \{\Delta, W\}$, il existe des relations entre $I[X\mathbf{G}]$ et $X\lambda$, et entre $I[XY\mathbf{G}]$ et $XY\lambda$, obtenues en prenant les dérivées de la relation $\mathbf{G}_{s,w}[\varphi_{s,w}] = \lambda_{s,w}\varphi_{s,w}$,

$$I[X\mathbf{G}] = (X\lambda)$$

$$I[XY\mathbf{G}]+\int_I (X\mathbf{G})[Y\varphi](t)dt+\int_I (Y\mathbf{G})[X\varphi](t)dt-I[X\mathbf{G}]\int_I [Y\varphi](t)dt-I[Y\mathbf{G}]\int_I [X\varphi](t)dt = (XY\lambda).$$

Alors, si pour tout $X, Y \in \{\Delta, W\}$, on a

$$I[X\mathbf{G}\circ\mathbf{Q}\circ Y\mathbf{G}] = \int_I (X\mathbf{G})[Y\varphi](t)dt - I[X\mathbf{G}]\int_I [Y\varphi](t)dt,$$

alors la conjecture (C) est vérifiée. ■

3.5.2.5 Continuant à une fraction de l'exécution

Selon la proposition 15, la série de Dirichlet bivariée associée au paramètre $\tilde{L}^{[\delta]}$ vérifie la propriété $US(s,w)$, et admet un unique pôle simple en $s=\sigma(w)$ avec $\sigma(0)=1$. En appliquant la formule de Perron à la série $S_{[\delta]}(2s,2w)$, les sommes partielles $\Psi(T)$ vérifient

$$\Psi(T) := \Psi_{2w}(T) = \frac{E(w)}{\sigma(w)(2\sigma(w)+1)}T^{2\sigma(w)+1}(1+O(T^{-2\alpha})),$$

où le terme d'erreur est uniforme en w et $E(w)$ est le résidu de $S(2s,2w)$ en $s=\sigma(w)$. Les sommes partielles s'écrivent également sous la forme $\Psi_{2w}(T)=\exp(U(w)\log T+V(w))$ avec

$$U(w)=(2\sigma(w)+1), \quad V(w)=\log E(w)-\log\sigma(w)(2\sigma(w)+1).$$

Une fois de plus, le quotient $\Psi_w(n)/\Psi_0(n)$ correspond à la série génératrice des moments du paramètre $\tilde{L}^{[\delta]}$ mais avec une distribution lisse. Selon la proposition 15, $\sigma''(0)\neq 0$ et le théorème des quasi-puissances de Hwang (théorème H) s'applique dans le cadre des distributions lisses. Le paramètre $\tilde{L}^{[\delta]}$ admet alors une loi limite gaussienne pour la distribution lisse. Après les étapes de délissage présentées dans [BV05, BV04, Ces06], la même loi est satisfaite par $\tilde{L}^{[\delta]}$ sur Ω_n et le calcul des constantes conduit au théorème 5.

3.5.2.6 Paramètres des algorithmes interrompus

En appliquant la formule de Perron dans une bande satisfaisant les conditions de la proposition 16, puis en effectuant les étapes de lissage et de dé-lissage, la série des moments de la variable M^{2w} satisfait

$$\mathrm{E}_n[M^{2w}] = \frac{\Psi_n(w)}{\Psi_n(0)} \qquad \text{avec} \quad \Psi_n(w) = E(w)\cdot 2^{n\rho(w)} + O(D^2)\cdot 2^{n(1-\frac{B}{D^2})}$$

si bien que

$$\mathrm{E}_n[M^{2w}] = \frac{E(w)}{E(0)}2^{n(\rho(w)-1)} + O(D^2)\cdot 2^{-nB/D^2}.$$

Nous choisissons maintenant w_0 comme dans la proposition 16, et nous appliquons les inégalités de Markov (voir inégalités 3.11, 3.12). Alors, lorsque δ et ϵ sont rationnels avec un dénominateur D, les probabilités d'intérêt satisfont

$$\mathbb{P}_n\left[\left|P_{<\gamma,\delta]} - \delta P\right| \geq \epsilon P\right] \leq C_1\cdot 2^{-n/(LD^2)} + C_2\cdot D^2\cdot 2^{-nB/D^2},$$

pour des constantes C_1, C_2, L, B. Si maintenant le dénominateur D vérifie l'hypothèse du théorème 6, c'est à dire $D=n^{1/2}/b(n)$ avec $b(n)\to\infty$, alors

$$\mathbb{P}_n\left[\left|P_{<\gamma,\delta]} - \delta P\right| \geq \epsilon P\right] \leq C\cdot 2^{-b(n)},$$

pour n suffisamment grand. La première partie du théorème 6 est prouvée. La deuxième partie s'obtient facilement avec la relation $P_{[\gamma,\delta]} = P_{<0,\gamma+\delta]} - P_{<0,\gamma]}-$ et la troisième partie est une conséquence directe des relations matricielles. ■

3.5.3 État d'avancement de l'analyse

Nous avons démontré tous les résultats annoncés pour les algorithmes euclidiens sur les entiers. Ces preuves sont toutes basées sur les propriétés analytiques des séries génératrices (propositions 14, 15 et 16). Nous n'avons pas démontré ces propriétés puisqu'elles dépendent de celles des opérateurs. L'analyse des opérateurs étant purement technique et ne relevant pas nécessairement "des grandes étapes usuelles" d'une analyse en moyenne, nous avons décidé de la traiter au prochain chapitre. Nous avons donc terminé l'analyse des algorithmes euclidiens sur les entiers. Au cours de cette analyse, nous avons pointé beaucoup de ressemblances avec l'analyse sur les polynômes. Nous concluons ce chapitre en montrant qu'une analyse dynamique de l'algorithme d'Euclide sur les polynômes est possible.

3.6 Analyse dynamique sur les polynômes

Lors de l'analyse des différents coûts, nous avons mis en évidence des similarités entre les séries génératrices sur les polynômes et les opérateurs de transfert sur les entiers. Ces similarités ne sont pas dues au hasard. Nous allons montrer dans cette partie qu'il existe un système dynamique lié à l'algorithme d'Euclide sur les polynômes et que les opérateurs de transfert associés généralisent les séries génératrices ordinaires. Rappelons que ces deux étapes correspondent aux deux premières de l'analyse dynamique.

3.6.1 Système dynamique des fractions continues

Dans la suite, nous essayons de mettre en parallèle le cas des polynômes et le cas des entiers. Tout d'abord, les entrées appartiennent à l'anneau euclidien $\mathbb{Z}$ dans le cas des entiers alors que sur les polynômes, l'anneau euclidien est $\mathbb{F}_q[X]$.

Dans les deux cas, l'algorithme d'Euclide utilise une suite de divisions euclidiennes de la forme

$$u = m \cdot v + r, \qquad \text{où} \quad \|r\| < \|v\|$$

où la norme correspond valeur absolue correspond à la norme habituelle sur les entiers et à la norme ultramétrique sur les polynômes,

$$\|u\| = 2^{\deg u}.$$

Si (u, v) est le couple de départ, la division euclidienne suivante utilise le couple (v, r) et ainsi de suite. Sur les entiers, il est naturel de regarder les rationnels v/u et r/v qui sont liés par la relation $T(v/u) = r/v$ avec

$$T : [0,1] \to [0,1], \qquad T(x) = \frac{1}{x} - \left\lfloor \frac{1}{x} \right\rfloor = \left\{ \frac{1}{x} \right\}, \qquad T(0) = 0.$$

Sur les polynômes, nous obtenons une formule équivalente. L'ensemble des rationnels $\mathbb{Q}$ est remplacé par l'ensemble des fractions $\mathbb{F}_q(X)$ construit à partir de $\mathbb{F}_q[X]$. La norme ultramétrique sur $\mathbb{F}_q[X]$ s'étend à $\mathbb{F}_q(X)$ par

$$\left| \frac{v}{u} \right| = q^{\deg v - \deg u}.$$

La complétion de $\mathbb{F}_q(X)$ respectivement à la norme ultramétrique est le corps des séries formelles de Laurent $\mathbb{F}_q((1/X))$ de la forme

$$F = \sum_{n \geq n_0} f_n (1/Z)^n, \qquad \text{avec} \quad n_0 \in \mathbb{Z}, f_n \in \mathbb{F}_q.$$

Ces séries sont à rapprocher avec le développement binaire des réels où Z est remplacé par 2. La norme ultramétrique de F est donnée par $|F| = q^{-n_0}$. Sur $\mathbb{F}_q((1/X))$, nous pouvons définir une partie entière et une partie fractionnaire par

$$\lfloor F \rfloor = \sum_{n=n_0}^{0} f_n (1/Z)^n, \qquad \{F\} = \sum_{n>0} f_n (1/Z)^n.$$

Les rationnels v/u et r/v appartiennent tous les deux à la boule unité $\mathcal{X}_q$ sur $\mathbb{F}_q((1/X))$ et sont reliés par $\overline{T}(v/u) = r/v$ où $\overline{T}$ est l'application

$$\overline{T} : \mathcal{X}_q \to \mathcal{X}_q, \qquad \overline{T}(x) = \frac{1}{x} - \left\lfloor \frac{1}{x} \right\rfloor = \left\{ \frac{1}{x} \right\}, \qquad \overline{T}(0) = 0.$$

Le couple $(\overline{T}, \mathcal{X}_q)$ forme un système dynamique qui est l'équivalent du système dynamique $([0,1], T)$ sur les entiers.

3.6.2 Opérateurs de transfert et liens avec les séries

Nous avons défini précédemment le système dynamique $(\overline{T}, \mathcal{X}_q)$. Les branches inverses de $\overline{T}$ admettent la même formule que celles de T et satisfont pour $x \in \mathcal{X}_q$,

$$h_{[m]}(x) = \frac{1}{m+x}, \qquad \text{pour} \quad m \in \mathcal{G}.$$

L'opérateur de transfert pondéré par le coût primitif c est lui aussi donné par la même formule et vérifie

$$\mathbf{G}_{s,w}[f](x) = \sum_{m \in \mathcal{G}} e^{wc(m)} ||h'_{[m]}(x)||^s f \circ h_{[m]}(x).$$

La grande différence entre les deux contextes vient des dérivées des branches inverses. Sur les polynômes, elles sont constantes en norme. La dérivée de $h_{[m]}$ est donnée par

$$h'_{[m]}(x) = \frac{-1}{(m+x)^2} \qquad \text{pour} \quad m \in \mathcal{G}.$$

Comme $||m+x|| = ||m||$ pour $x \in \mathcal{X}_q$ et $m \in \mathcal{G}$, la norme de $h'_{[m]}$ est constante sur $\mathcal{X}_q$ et satisfait

$$||h'_{[m]}(x)|| = ||m||^{-2} = q^{-2 \deg m}.$$

L'opérateur de transfert pondéré se récrit alors

$$\mathbf{G}_{s,w}[f](x) = \sum_{m \in \mathcal{G}} e^{wc(m)} q^{-2s \deg m} f \circ h_{[m]}(x).$$

La fonction constante $\mathbf{1}$ est toujours le vecteur propre dominant de $\mathbf{G}_{s,w}$ associée à la valeur propre dominante $\lambda(s,w)$ où,

$$\lambda(s,w) = \sum_{m \in \mathcal{G}} e^{wc(m)} q^{-2s \deg m}.$$

Dans le cas des polynômes, l'analyse spectrale des opérateurs de transfert est très simplifiée.

La série génératrice $G(z,w)$ est un cas très particulier de l'opérateur de transfert puisque

$$G(z,w) = \sum_{m\in\mathcal{G}} e^{wc(m)} z^{\deg m} = \mathbf{G}_{s,w}[\mathbf{1}](0)$$

avec $q^{-2s} = z$. Il en est de même avec les autres séries génératrices que nous pourrions exprimer en termes d'opérateurs de transfert. En particulier, nous aurions très bien pu faire une analyse dynamique dans le cas des polynômes équivalente à celle présentée sur les entiers. Toutefois, le fait que $\mathbf{1}$ soit le vecteur propre de tous les opérateurs simplifie considérablement les opérations dans le cas des polynômes (puisque les opérateurs sont utilisés avec la fonction $\mathbf{1}$).

3.7 Conclusion

Dans cette partie, nous avons présenté l'analyse complète de tous les paramètres. Nous avons ainsi établi que la complexité binaire étendue admet une loi limite gaussienne (théorème 3). C'est la première fois que la loi limite d'un coût non additif est obtenue. Nous avons également amélioré les résultats connus sur les moments de la complexité binaire classique (théorème 2). La loi limite de la taille du continuant à une fraction de l'exécution a aussi été prouvée (théorème 5). C'est le second paramètre non-additif dont la loi limite est déterminée. L'analyse du continuant à une fraction de l'exécution fait intervenir un nouvel opérateur, le pseudo-quasi-inverse $\mathbb{G}_{s,w}$ dont nous ferons l'analyse au prochain chapitre. Nous montrerons que cet opérateur a des propriétés proches de celles du quasi-inverse. L'analyse des trois paramètres est basée sur un principe de décomposition qui a mis en évidence deux familles de coûts : les coûts additifs à croissance intermédiaire et les coûts terminaux dont nous avons décrit les deux premiers moments (théorème 12). Pour conclure, l'analyse des algorithmes interrompus (théorème 6) s'est réduite à l'étude d'un paramètre M auquel est associé un autre pseudo-quasi-inverse $\mathbb{H}_{s,w}$. Au premier chapitre, nous avons vu qu'il était suffisant d'analyser les algorithmes interrompus pour établir la complexité binaire des algorithmes $\mathcal{KS}_\alpha$ (théorème 8). Là encore, c'est la première fois que la complexité binaire d'un algorithme rapide est analysée.

Nous en avons donc terminé avec l'analyse des algorithmes euclidiens. Reste toutefois à établir les propriétés analytiques des opérateurs, ce que nous faisons au prochain chapitre.

Chapitre 4

Analyse des opérateurs de transfert

Sommaire

Les analyses du chapitre précédent sont toutes basées sur les propositions 14, 15 et 16. Ces propositions résument les propriétés analytiques des séries génératrices de Dirichlet. Comme les séries génératrices s'expriment en fonction des opérateurs de transfert, les propriétés analytiques des séries sont aussi celles des opérateurs et réciproquement. Dans cette partie, nous analysons les propriétés analytiques du quasi-inverse et des pseudo-quasi-inverses puisque ce sont ces opérateurs qui apportent les singularités dominantes. Pour ces opérateurs, nous souhaitons montrer une propriété du type $US(s)$ ou $US(s,w)$ (unique pôle dominant dans une bande avec une borne uniforme, voir section 3.4.1). Ce résultat a déjà été démontré par Dolgopyat [Dol98] pour le quasi-inverse d'un système dynamique avec un nombre fini de branches. Il a été généralisé par Baladi et Vallée [BV05, BV04] au quasi-inverse pondéré d'un système dynamique avec un nombre au plus dénombrable de branches. Dans ce dernier cas, le système dynamique doit vérifier une condition UNI qui est systématique pour les systèmes avec un nombre fini de branches. Si les quasi-inverses sont bien décrits, ce n'est en revanche pas le cas pour les pseudo-quasi-inverses. Jusqu'aux travaux présentés plus bas (et parus dans les articles [LV06b] et [DLMDV06]), des propriétés de type $US(s,w)$ n'étaient pas mises en évidence ces opérateurs. Mais certaines bornes obtenues par Baladi et Vallée s'appliquent à ce nouveau contexte et nous montrons des propriétés de type $US(s,w)$ sur les pseudo-quasi-inverses.

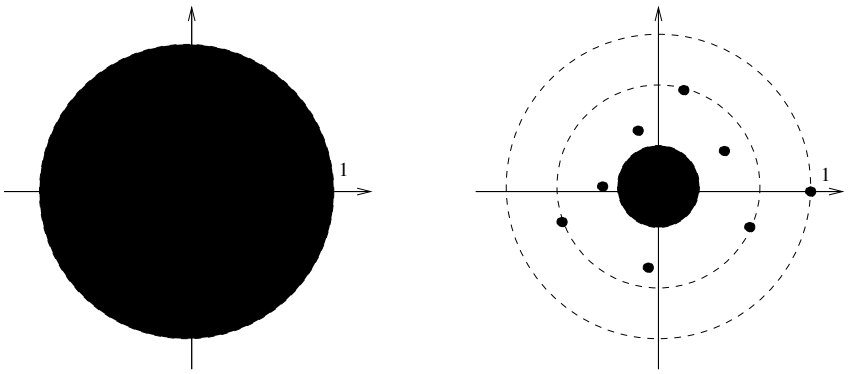
Plan. La première section énumère les propriétés connues des opérateurs de transfert et du quasi-inverse. Les deux autres sections sont consacrées à l'analyse des pseudo- quasi-inverses.

4.1 Propriétés connues des opérateurs de transfert

Dans toute cette section, nous nous contentons d'énumérer des faits déjà connus sur l'opérateur de transfert $\mathbf{G}_s$. Il existe des résultats classiques d'analyse dynamique sur le spectre dominant de $\mathbf{G}_s$ mais aussi des résultats moins classiques sur des propriétés de type US pour ces opérateurs.

4.1.1 Espace fonctionnel et spectre dominant

Les opérateurs agissent sur des espaces de fonctions et selon ces espaces, le spectre diffère pour un même opérateur. Prenons pour exemple le transformateur de densité $\mathbf{G}$. Sur l'espace des fonctions intégrables L^1, le transformateur de densité admet tout le cercle unité (dans $\mathbb{C}$) comme spectre. En revanche, sur l'espace des fonctions C^1 sur l'intervalle $\overline{I} = [0, 1]$ et muni de la norme $||f|| = ||f||_\infty + ||f'||_\infty$, le transformateur de densité admet 1 comme unique valeur propre dominante séparée du reste du spectre par un saut spectral.



spectre sur L^1 spectre sur C^1

En analyse dynamique, un bon espace fonctionnel pour un opérateur est un espace sur lequel il admet une unique valeur propre dominante séparée du reste du spectre par un saut spectral. Chaque algorithme euclidien (algorithme classique, centré, pair, impair, ...) admet un système dynamique associé. Les premiers systèmes étudiés en analyse dynamique comportaient des branches inverses holomorphes et contractantes comme celui des fractions continues. L'espace fonctionnel choisi était alors les fonctions holomorphes sur un voisinage complexe de I ou un produit cartésien de cet espace [Val97a, Val01, Val98a, Val03]. Avec l'analyse des algorithmes α-euclidien [BDV02], des systèmes dynamiques généraux à branches non surjectives ont été étudiés. L'espace fonctionnel adéquat était alors l'espace des fonctions à variation bornée. Plus récemment, Chazal et Maume-Deschamps [CMD04] ont utilisé l'espace des fonctions lipschitziennes par morceaux pour traiter des systèmes dynamiques dits markoviens dont les branches ne sont ni surjectives ni holomorphes. Finalement dans [BV05, BV04], Baladi et Vallée ont préféré l'espace des fonctions $C^1(\overline{I})$ avec une norme particulière pour analyser les opérateurs de transfert pondérés associés aux systèmes dynamiques des algorithmes classique, impair et centré. Cependant, il n'est pas toujours facile de trouver un bon espace fonctionnel. Par exemple, aucun bon espace n'est encore trouvé pour les opérateurs de transfert du système dynamique associé à l'algorithme plus-moins [BK85, Dai04].

Comme nous souhaitons utiliser les résultats de Baladi et Vallée, nous choisissons[5] ici l'espace des fonctions de classe C^1 sur l'intervalle $I = [0,1]$. Sur cet espace, les propriétés spectrales de l'opérateur de transfert $\mathbf{G}_s$ associé au système dynamique des fractions continues sont les suivantes.

Proposition 17 *Soit $\mathbf{G}_s$ l'opérateur de transfert associé au système dynamique des fractions continues et agissant sur l'espace des fonctions de classe C^1 sur $I = [0,1]$.*

(i) Pour s réel et $s > 1/2$, l'opérateur $\mathbf{G}_s$ est quasi-compact. Il admet une unique valeur propre dominante simple $\lambda(s)$, strictement positive, et séparée du reste du spectre par un saut spectral. Si φ_s est une fonction propre associée à $\lambda(s)$, alors φ_s est également strictement positive sur I.

(ii) Pour $s = 1$, $\mathbf{G}_1$ est le transformateur de densité, $\lambda(1) = 1$ et la densité invariante est la densité de Gauss

$$\varphi_1(x) = \frac{1}{\log 2} \cdot \frac{1}{1+x}.$$

De plus, la fonction $s \to \lambda(s)$ est strictement décroissante sur $]1/2, +\infty[$ et si $\rho(s)$ est le rayon spectral de $\mathbf{G}_s$ pour s complexe, alors $\rho(s) < \lambda(\Re(s))$ dès que s n'est pas réel.

(iii) L'application $s \to \mathbf{G}_s$ est analytique en s pour $\Re(s) > 1/2$. Par perturbation analytique, les applications $s \to \lambda(s)$, $s \to \nu(s)$ et $s \to \varphi_s$ sont bien définies dans un voisinage de l'axe réel.

(iv) Pour $s > 1/2$, la fonction de pression $\Lambda(s) = \log \lambda(s)$ est bien définie. La fonction de pression est strictement convexe en s, i.e., $\Lambda''(s) > 0$ et $\Lambda'(1)$ est l'opposé de l'entropie de Kolmogorov du système dynamique des fractions continues,

$$\Lambda'(1) = -\int_I \log|T'(x)|\varphi_1(x)dx = -\frac{\pi^2}{6\log 2} < 0,$$

avec T la fonction de décalage du système dynamique des fractions continues ($T(x) = \{1/x\}$).

4.1.2 Décomposition spectrale et pôle dominant du quasi-inverse

Dans un voisinage de l'axe réel, l'opérateur de transfert admet une unique valeur propre dominante $\lambda(s)$ séparée du reste du spectre par un saut spectral. Cette propriété entraîne une décomposition de l'opérateur en deux parties : une partie dominante liée à la partie dominante du spectre et une partie sous-dominante correspondant au reste du spectre. Cette décomposition, appelée décomposition spectrale, est de la forme

$$\mathbf{G}_s[f] = \lambda(s)\mathbf{P}_s[f] + \mathbf{N}_s[f], \tag{4.1}$$

où $\mathbf{P}_s$ est un projecteur sur l'espace propre dominant et $\mathbf{N}_s$ un opérateur de rayon spectral $\nu(s) < |\lambda(s)|$ (le rayon spectral sous-dominant de $\mathbf{G}_s$). De plus, $\mathbf{P}_s$ commute avec $\mathbf{N}_s$ et pour $s = 1$, sa forme est explicite

$$\mathbf{P}_1[f](x) = \varphi_1(x)\int_I f(t)dt = \frac{1}{\log 2} \cdot \frac{1}{1+x}\int_I f(t)dt.$$

La décomposition spectrale induit aussi une décomposition spectrale des itérés,

$$\mathbf{G}_s^n[f] = \lambda(s)^n\mathbf{P}_s[f] + \mathbf{N}_s^n[f],$$

[5] Nous ferons un autre choix au prochain chapitre

ainsi qu'une décomposition du quasi-inverse,

$$(\mathbf{I}-\mathbf{G}_s)^{-1}[f] = \frac{1}{1-\lambda(s)}\mathbf{P}_s[f] + (\mathbf{I}-\mathbf{N}_s)^{-1}[f].$$

La décomposition du quasi-inverse n'a de sens que si la valeur propre dominante n'est pas égale à 1 ou si 1 n'est pas valeur propre de $\mathbf{N}_s$. Or la proposition 17 (ii) indique que nous sommes dans cette situation avec l'opérateur de transfert $\mathbf{G}_s$ en $s=1$ (i.e. avec le transformateur de densité). Nous en déduisons la première propriété analytique des quasi-inverses $(\mathbf{I}-\mathbf{G}_s)^{-1}$.

Proposition 18 *(i) Dans un voisinage complexe Σ_1 de $s=1$, le quasi-inverse $(\mathbf{I}-\mathbf{G}_s)^{-1}$ admet un unique pôle simple en $s=1$. De plus, il vérifie l'équivalence*

$$(\mathbf{I}-\mathbf{G}_s)^{-1}[f](x) \sim \frac{1}{s-1}\frac{-1}{\lambda'(1)}\varphi_1(x)\int_I f(t)dt, \qquad \varphi_1(x)=\frac{1}{\log 2}\frac{1}{1+x}.$$

(ii) Avec la propriété (ii) de la proposition 17, $s=1$ est l'unique pôle du quasi-inverse sur le demi-plan $\Re(s)\geq 1$, .

4.1.3 Propriété UNI et borne à la Dolgopyat

Dans un voisinage de $s=1$, le quasi-inverse admet un unique pôle simple. Pour démontrer une propriété du type $US(s)$, nous devons montré qu'il existe une bande qui ne contient que ce pôle.

Dolgopyat [Dol98] est le premier à avoir mis en évidence une région sans pôle lorsque s est suffisamment loin de l'axe réel. Ce résultat, limité aux systèmes dynamiques avec un nombre fini de branches a été généralisé par Baladi et Vallée [BV05, BV04] aux systèmes dynamiques avec un nombre dénombrable de branches. Cette généralisation nécessite toutefois une condition supplémentaire appelée condition UNI.

Si $\mathcal{H}$ est l'ensemble des branches inverses d'un système dynamique de l'intervalle, le coefficient de contraction ρ du système dynamique est le plus petit réel ρ tel qu'il existe une constante M vérifiant

$$\sup_I |h'| \leq M\rho^n, \qquad \forall n\geq 1, \forall h\in\mathcal{H}^n.$$

Un système dynamique dont le coefficient de contraction est strictement plus petit que 1 est dit dilatant (car les branches inverses sont contractantes). La condition UNI utilise une notion de distance entre les branches inverses. Pour deux branches inverses h et k de même profondeur, nous posons

$$\Delta(h,k) = \inf_{x\in I}\left|\frac{h''}{h'}(x) - \frac{k''}{k'}(x)\right|.$$

Pour $h\in\mathcal{H}^n$ et $\eta>0$, $J(h,\eta)$ est l'ensemble des intervalles (dit fondamentaux) $I_k=k(I)$ tel $\Delta(h,k)\leq\eta$,

$$J(h,\eta) = \bigcup_{k\in\mathcal{H}^n,\Delta(h,k)\leq\eta} k(I).$$

Définition 7 (Condition UNI) *Un système dynamique (I,T) à branches surjectives et de coefficient de contraction $\rho<1$ satisfait la condition UNI si chaque branche inverse de T se prolonge en une fonction de classe C^3 sur I et si*
(a) pour tout a $(0<a<1)$, $|J(h,\rho^{an})| = O(\rho^{an})$, $\forall n, \forall h\in\mathcal{H}^n$,
(b) il existe $Q<\infty$ tel que $|h'''(x)|\leq Q|h'(x)|$ pour tout $n\geq 1$ et tout $h\in\mathcal{H}^n$.

Bien entendu, le système dynamique des fractions continues vérifie la condition UNI (voir [BV04]). Avec ces conditions, le quasi-inverse admet une borne à la Dolgopyat pour s éloigné de l'axe réel.

Proposition 19 *La norme $||\cdot||_{1,t}$ est la norme définie sur $C^1(I)$ par $||f||_{1,t} = \sup_I |f| + \frac{1}{t}\sup_I |f'|$. Soit $\xi \in [0, 1/5[$ une constante. Il existe un voisinage réel Σ_2 de $s = 1$, des constantes positives M, θ et t_0 avec $\theta \in]0, 1[$ tels que pour tout $s = \sigma + it$ avec $\sigma \in \Sigma_2$ et $|t| > t_0$, les itérés de l'opérateur de transfert satisfont*

$$||\mathbf{G}_s^n||_{1,t} \leq M \cdot |t|^\xi \cdot \theta^n.$$

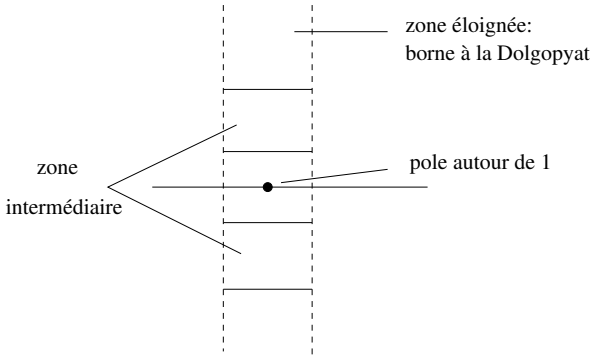
Pour tout $s = \sigma + it$ avec $\sigma \in \Sigma_2$ et $|t| > t_0$, le quasi-inverse n'admet donc pas de pôle et est borné par

$$||(\mathbf{I} - \mathbf{G}_s)^{-1}||_{1,t} \leq \frac{M}{1-\theta} \cdot |t|^\xi.$$

Nous avons énoncé le résultat avec le système dynamique des fractions continues. Toutefois, la proposition reste vrai pour tous les systèmes dynamiques complets (i.e. à branches surjectives), vérifiant la propriété de distorsion des branches (i.e. $|h'(x)/h'(y)|$ uniformément bornée pour tout $x, y \in I$) et la condition UNI. Les systèmes dynamiques associés aux algorithmes centrés et impairs sont aussi des exemples d'applications de la proposition.

4.1.4 Zone intermédiaire

Les deux propositions précédentes ont décrit les propriétés du quasi-inverse dans deux zones distinctes : la première autour de $s = 0$ et la deuxième éloigné de l'axe réel.



Il reste une zone intermédiaire. Avec la condition UNI et la quasi-compacité, il est possible de montrer que sur l'axe $\Re(s) = 1$, le rayon spectral $\rho(s)$ de $\mathbf{G}_s$ est strictement plus petit que 1 dès que $s \neq 1$. Ainsi, comme la zone intermédiaire est compacte, par perturbation analytique, le rayon spectral de $\mathbf{G}_s$ dans un voisinage complexe de le zone intermédiaire est aussi strictement plus petit que 1. Le quasi-inverse est alors bien défini et analytique. Nous venons d'établir la proposition suivante.

Proposition 20 *Soit t_0 et t_1 deux réels tels que $t_1 > t_0 > 0$. Alors, il existe un voisinage réel Σ_3 de $s = 1$ tel que pour tout $s = \sigma + it$ avec $\sigma \in \Sigma_3$ et $|t| \in [t_0, t_1]$, le quasi-inverse est bornée par une constante M_3 positive.*

Une fois de plus, cette proposition ne se limite pas au cadre du système dynamique des fractions continues.

4.1.5 Mise en commun des propositions

En mettant en commun les trois propositions, nous obtenons que le quasi-inverse satisfait une propriété du type $US(s)$.

Théorème DBV (Dolgopyat-Baladi Vallée) *(i) Soit un système dynamique complet (i.e. à branches surjectives), dilatant et satisfaisant la condition UNI. Alors, pour tout $\xi \in]0, 1/5]$, il existe $\alpha > 0$ tel que le quasi-inverse admet un unique pôle en $s = 1$ dans la bande $\Re(s) \in [1-\alpha, 1+\alpha]$.*
(ii) De plus, il existe une constante $M > 0$, un réel positif $\theta < 1$ et $t_0 > 0$ tels que pour tout $s = \sigma + it$ avec $\sigma \in [1-\alpha, 1+\alpha]$ et $|t| > t_0$, les itérés de $\mathbf{G}_s$ ainsi que le quasi-inverse sont majorés en norme par

$$||\mathbf{G}_s^n||_{1,t} \leq M \cdot |t|^\xi \cdot \theta^n, \qquad et \qquad ||(\mathbf{I} - \mathbf{G}_s)^{-1}||_{1,t} \leq \frac{M}{1-\theta} \cdot |t|^\xi.$$

(iii) Finalement, sur la droite verticale $\Re(s) = 1 - \alpha$, le quasi-inverse est majoré par

$$||(\mathbf{I} - \mathbf{G}_s)^{-1}||_{1,\mu(t)} \leq M' \mu(t)^\xi$$

avec $\mu(t) = 1$ si $|t| < t_0$ et $\mu(t) = |t|$ sinon, et M' une constante positive.

Nous aurions pu annoncer directement ce théorème démontré dans sa globalité par Baladi et Vallée. Mais nous avons choisi de présenter la décomposition en trois zones car c'est la méthode que nous utilisons pour l'analyse des pseudo-quasi-inverses. Globalement, pour la zone éloignée de l'axe réel, les bornes sur les itérés de l'opérateur de transfert s'appliquent. Pour la zone intermédiaire, les preuves sont les mêmes. Il n'y a que la zone autour de $s = 1$ (et $w = 0$) où les choses se compliquent. Comme sur les polynômes, il faut faire attention à l'accumulation des pôles à gauches du pôle dominant.

4.2 Analyse du pseudo quasi-inverse $\mathbb{G}_{s,w}$

4.2.1 Analyses des pseudos quasi-inverses

Nous présentons ici les points communs et les différences entre les analyses des deux pseudos quasi-inverses.

Tout d'abord, les deux analyses ont la même structure en trois étapes que l'analyse du quasi-inverse, chacune des étapes étant associée à une zone. Contrairement au quasi-inverse, l'analyse dans la zone éloignée de l'axe réel est facile puisqu'elle s'appuie sur les bornes à la Dolgopyat déjà existantes. En revanche la zone autour de $(0, 1)$ est elle beaucoup plus difficile. En effet, les pseudos quasi-inverses font intervenir dans leur somme deux ou trois puissances d'opérateurs de transfert. Autour de $(0, 1)$, nous appliquerons la décomposition spectrale pour chacun de ces opérateurs ce qui fait apparaître $2^2 = 4$ ou $2^3 = 8$ termes différents. Parmi ces termes, un seul sera dit dominant puisqu'il correspondra au terme qui apporte le pôle dominant. Dans les deux cas, le terme dominant sera de la forme

$$\frac{f(s,w)}{1 - \psi(s,w)^D}$$

où f et ψ sont analytiques, D est le dénominateur du paramètre δ qui intervient dans les opérateurs et $\psi(1, 0) = 1$. La fonction ψ diffère bien entendu entre les deux pseudos quasi-inverses.

C'est à ce moment que l'analyse des pseudos quasi-inverses diffèrent complètement de celle du quasi-inverse. La puissance D entraîne l'existence de D solutions $s = \sigma_k(w)$ $(k = 0 \dots D-1)$ avec

$\sigma_k(0) = e^{2i\pi/D}$ à l'équation $\psi(s,w)^D = 1$ dans un voisinage de $(0,1)$ (pour le quasi-inverse, il n'existait qu'une solution). Face à cette accumulation de pôles, il faut réduire le voisinage autour de $(0,1)$ afin de n'avoir que le pôle dominant (qui sera $s = \sigma_0(w)$). On montre que la taille adéquate du voisinage est de l'ordre de $O(1/D^2)$. Pour le premier pseudo quasi-inverse, ce détail n'est pas important puisque D est constant. En revanche, pour le second pseudo quasi-inverse, D tend vers l'infini et il faut tenir compte du voisinage qui rétrécit pour borner les termes d'erreur.

4.2.2 Résultat principal pour $\mathbb{G}_{s,w}$

Le pseudo quasi-inverse $\mathbb{G}_{s,w}$ intervient dans la série génératrice bivariée du paramètre $\tilde{L}^{[\delta]} = \lg v_{\lfloor \delta p \rfloor}$. Il est définit par

$$\mathbb{G}_{s,w} = \sum_{k \geq 0} \mathbf{G}_{s-w}^{dk} \circ \mathbf{G}_s^{ck},$$

lorsque $\delta = c/(c+d)$. Nous allons montrer que $\mathbb{G}_{s,w}$ satisfait une propriété $US(s,w)$.

Théorème 13 (Propriété US pour $\mathbb{G}_{s,w}$) *(i) Pour tout $\xi \in]0, 1/5]$, il existe un voisinage complexe $\mathcal{W}$ de $w = 0$ et un voisinage réel $\Sigma =]1-\alpha, 1+\alpha[$ de 1 avec $\alpha > 0$ tel que le pseudo quasi-inverse $\mathbb{G}_{s,w}$ admet pour tout $w \in \mathcal{W}$, un unique pôle simple en $s = \sigma(w)$ dans la bande $\Re(s) \in \Sigma$.*

(ii) La fonction $\sigma(w)$ est implicitement définie par $\lambda(\sigma(w) - w)^{1-\delta}\lambda(\sigma(w))^{\delta} = 1$ (où les fonction λ sont bien définies). Les dérivée de la fonction σ en 0 sont données par

$$\sigma'(0) = 2(1-\delta), \qquad \sigma''(0) = 4\delta(1-\delta)\frac{\Lambda''(1)}{|\Lambda'(1)|}.$$

(iii) Il existe $M > 0$ et $t_0 > 0$ tels que pour tout $s = \sigma + it$ avec $\sigma \in [1-\alpha, 1+\alpha]$ et $|t| > t_0$, le pseudo quasi-inverse est majoré en norme par

$$||\mathbb{G}_{s,w}||_{1,t} \leq M \cdot |t|^{\xi}.$$

(iv) Finalement, sur la droite verticale $\Re(s) = 1-\alpha$, le pseudo quasi-inverse est majoré par

$$||\mathbb{G}_{s,w}||_{1,\mu(t)} \leq M'\mu(t)^{\xi}$$

avec $\mu(t) = 1$ si $|t| < t_0$ et $|t|$ sinon, et M' une constante positive.

4.2.3 Zone loin de l'axe réel

Nous suivons les mêmes idées que la preuve de Baladi et Vallée. Le théorème DBV définit un voisinage $\Sigma =]1-\alpha, 1+\alpha[$ sur lequel toutes les majorations du théorème s'appliquent. Nous posons Σ_1 l'intervalle $]1-\alpha/2, 1+\alpha/2[$ et $\mathcal{W}_1$ la boule de centre 0 et de rayon $\alpha/2$. Alors pour tout s dans la bande $\Re(s) \in \Sigma_1$ et pour tout $w \in \mathcal{W}_1$, les complexes s et $s-w$ on une partie réelle dans Σ. En particulier, les bornes du théorème DBV pour $\Im(s) > t_0$ s'appliquent. La norme du pseudo-quasi-inverse est alors bornée par

$$||\mathbb{G}_{s,w}||_{1,t} \leq \sum_{k \geq 0} M^2 \cdot |t|^{2\xi} \cdot \theta^{k(c+d)} = \frac{M^2}{1-\theta^{c+d}}|t|^{2\xi} < \infty.$$

Ceci termine la preuve de (iii) et montre que dans une région éloignée de l'axe réel, le quasi-inverse n'a pas de pôle . La démonstration sera identique pour le pseudo-quasi-inverse $\mathbb{H}_{s,w}$.

4.2.4 Zone autour de $(1,0)$

Autour de $(s,w)=(1,0)$, la solution est plus difficile pour le pseudo-quasi-inverse $\mathbb{G}_{s,w}$ que pour le quasi-inverse classique. Nous décomposons l'opérateur en quatre parties, en utilisant la décomposition spectrale 4.1 de $\mathbf{G}_s$. Soit V_1 un voisinage complexe de $(1,0)$ dont tous les éléments (s,w) vérifient $\Re(s)\in\Sigma_1$, $w\in\mathcal{W}_1$ et où la décomposition spectrale de $\mathbf{G}_s$ et $\mathbf{G}_{s-w}$ s'applique. Pour j et k des entiers, l'opérateur $\mathbf{G}^j_{s-w}\circ\mathbf{G}^k_s$ se décompose en quatre parties

$$\mathbf{G}^j_{s-w}\circ\mathbf{G}^k_s=\lambda(s-w)^j\lambda(s)^k\mathbf{P}_{s-w}\circ\mathbf{P}_s+\lambda(s-w)^j\mathbf{P}_{s-w}\circ\mathbf{N}^k_s+\lambda(s)^k\mathbf{N}^j_{s-w}\circ\mathbf{P}_s+\mathbf{N}^j_{s-w}\circ\mathbf{N}^k_s.$$

Le premier terme est appelé le terme dominant et les trois autres sont appelés des termes de reste.

Termes de reste.

Notons $\nu(t)$ le rayon spectral de l'opérateur $\mathbf{N}_t$ et $R=\log\nu$. Par construction, quitte à rétrécir le voisinage V_1, il existe deux constantes absolue $a<1$ et $M_1>0$ telles que pour tout t dans V_1,

$$||\mathbf{N}^k_t||\leq M_1a^k$$

Alors, la série faisant intervenir uniquement les opérateurs $\mathbf{N}$ est absolument convergente. Considérons maintenant les deux autres séries. Leur norme peut facilement être comparée à un somme géométrique dont le logarithme du terme général est

$$\delta R(s)+(1-\delta)\Re\Lambda(s-w),\qquad \delta\Re\Lambda(s)+(1-\delta)R(s-w).$$

En $s=1$ et $w=0$, $\Lambda(s-w)=\Lambda(s)=0$ et $R(s)=R(s-w)=R(0)\leq a<0$ ce qui montre que les termes sont négatifs. Par perturbation analytique, il en est de même dans un voisinage $V_2\subset V_1$ de $(1,0)$. Pour (s,w) dans ce voisinage V_2, les trois séries relatives aux trois termes de restes sont absolument convergentes et la norme est uniformément bornée par une constante M_2, autrement dit

$$||\sum_{k\geq0}\lambda(s-w)^{kd}\mathbf{P}_{s-w}\circ\mathbf{N}^{kc}_s+\lambda(s)^{kc}\mathbf{N}^{kd}_{s-w}\circ\mathbf{P}_s+\mathbf{N}^{kd}_{s-w}\circ\mathbf{N}^{kc}_s||_{1,1}\leq M_2.$$

A ce stade, l'analyse de $\mathbb{H}_{s,w}$ sera différente car δ pourra tendre vers 0. Il faudra alors déterminer un voisinage qui sera une fonction de δ.

Le terme dominant.

Le terme dominant est donné par

$$\sum_{k\geq0}\lambda(s-w)^{kd}\lambda(s)^{kc}\mathbf{P}_{s-w}\circ\mathbf{P}_s=\frac{1}{1-\psi(s,w)^D}\cdot\mathbf{P}_{s-w}\circ\mathbf{P}_s$$

avec $D=c+d$ le dénominateur de δ et $\psi(s,w)$ la fonction définie sur V_2 par

$$\psi(s,w):=\lambda(s-w)^{1-\delta}\lambda(s)^\delta.$$

Le dénominateur $s\to1-\psi(s,w)^D$ de ψ admet des zéros pour toutes les valeurs de s pour lesquelles

$$\psi(s,w)=\exp[2iL\pi/D]\qquad\text{avec } 0\leq L<D.$$

Cela signifie que la fonction Ψ définie comme $\Psi := \log \psi$ satisfait

$$\Psi(s,w) := (1-\delta)\Lambda(s-w) + \delta\Lambda(s) = \frac{2iL\pi}{D}, \qquad \text{avec} \quad L \in \mathbf{Z}.$$

Pour $w=0$, on a $\Psi(s,w) = \Lambda(s) = 2iL\pi/D$. Le développement de $\Lambda(s)$ près de $s=1$ est de la forme (nous posons $s = \rho + it$)

$$\Lambda(s) = -B(s-1) + A(s)\cdot(s-1)^2 \qquad \text{avec} \quad A(s) := \frac{1}{2}\Lambda''(1+(s-1)\theta), \quad \theta \in [0,1], \tag{4.2}$$

$B := |\Lambda'(1)|$, et $A := A(1) > 0$ entraîne que pour s suffisamment proche de 1,

$$\Re\Lambda(s) \sim -B(\rho-1) - At^2, \qquad \Im\Lambda(s) \sim -Bt.$$

Alors, la courbe $\{s;\ \ \Re\Lambda(s) = 0\}$ est proche de la courbe d'équation $B(\rho-1) + At^2 = 0$ et est contenue dans le demi-plan gauche $\Re(s) \le 1$.
Nous considérons deux parties de cette courbe. La première partie

$$\mathcal{A} := \{s; \quad \Re\Lambda(s) = 0, \quad |\Im\Lambda(s)| > \frac{3\pi}{2D}\}$$

est strictement contenue dans le demi-plan $\{\Re s < 1 - 4\Delta\}$ pour un $\Delta > 0$. Par une petite perturbation, il existe un voisinage $\mathcal{W}_A$ de $w=0$ pour lequel le domaine

$$\mathcal{A}_w := \{s; \quad |\Re\Psi(s,w)| \le \frac{C}{D^2}, \quad |\Im\Psi(s,w)| > \frac{3\pi}{2D}\}$$

est strictement contenu dans le demi-plan $\{\Re s < 1 - 3\Delta\}$, pour tout $w \in \mathcal{W}_A$.
La deuxième partie de la courbe est la portion

$$\mathcal{B} := \{s; \quad \Re\Lambda(s) = 0, \quad |\Im\Lambda(s)| < \frac{\pi}{2D}\},$$

qui est contenue dans la bande $|\Re s - 1| < d$ pour un d fixé. Par une petite perturbation, il existe un voisinage $\mathcal{W}_B$ de $w=0$ pour lequel le domaine

$$\mathcal{B}_w := \{s; \quad |\Re\Psi(s,w)| \le \frac{C}{D^2}, \quad |\Im\Psi(s,w)| < \frac{\pi}{2D}\}$$

est strictement contenu dans la bande $|\Re s - 1| < 2d$, pour tout $w \in \mathcal{W}_{\mathcal{B}}$.

Unique pôle dans un voisinage de $(1,0)$. Le développement 4.2 entraîne que $3\Delta > 2d$. Nous choisissons $\alpha \in]2d, 3\Delta[$ et nous prouvons que la propriété $US(s,w)$ est vérifiée pour la partie dominante (et donc pour tout l'opérateur) dans la bande $|\Re s - 1| < \alpha$ pour $w \in \mathcal{W}_A \cap \mathcal{W}_B$.
Si $\sigma(w)$ satisfait $\Psi(\sigma(w), w) = 2i\pi L/D$, alors en 0, nous avons $\Lambda(\sigma(0)) = 2i\pi L/D$ et $\Re(\Lambda(\sigma(0))) = 0$. Dans ce cas, soit $\sigma(0)$ appartient à $\mathcal{A}$ et il n'appartient pas à la bande $|\Re s - 1| < \alpha$ (car $\mathcal{A}$ est dans le demi-plan $\Re(s) < 1 - 4\Delta$), soit il n'appartient pas à $\mathcal{A}$, et alors $\Im(\Lambda(\sigma(0))) = 2\pi L/D \le \frac{3\pi}{2D}$. Dans ce cas $L = 0$ (car L est entier) et σ est la fonction définie par $\Psi(\sigma(w), w) = 0$. Ceci montre que dans un voisinage de $(1,0)$, $s = \sigma(w)$ est l'unique pôle de la partie dominante.

Borne sur la ligne verticale à gauche.
Il existe seulement deux possibilités sur la ligne verticale $\Re s = 1 - \alpha$,

$$|\Re\Psi(s,w)| > \frac{C}{D^2} \qquad \text{ou} \qquad \frac{\pi}{2D} \le |\Im\Psi(s,w)| \le \frac{3\pi}{2D}.$$

Cela entraîne que le dénominateur $1 - \psi(s,w)^D$ vérifie

$$|\psi(s,w)^D - 1| \ge \exp[C/D] - 1 \ge C/D \qquad \text{ou} \quad |\psi(s,w)^D - 1| \ge 1.$$

Finalement sur la ligne $\Re s = 1 - \alpha$, le terme dominant est d'ordre $O(D)$.

4.2.5 Zone intermédiaire

Nous avons démontré que dans un voisinage de $(1,0)$, il existe un unique pôle simple $s=\sigma(w)$ et que sur la droite verticale à gauche de ce pôle, le terme dominant est borné ainsi que les termes sous-dominants. Dans la zone éloignée de l'axe réel, nous avons obtenu une borne à la Dolgopyat. Reste la zone intermédiaire. La même démarche s'applique puisque en $w=0$, nous retrouvons le vrai quasi-inverse qui est analytique dans la zone intermédiaire. Ensuite, par perturbation analytique, nous obtenons le même résultat pour le pseudo-quasi-inverse. Finalement, en regroupant les les trois zones, nous obtenons le théorème 13. Le calcul des dérivées de σ s'obtiennent facilement en dérivant une ou deux fois par rapport à w la relation $\Psi(\sigma(w),w)=0$.

4.3 Analyse du pseudo quasi-inverse $\mathbb{H}_{s,w}$

L'analyse du pseudo-quasi-inverse $\mathbb{H}_{s,w}$ est quasiment identique à celle du pseudo-quasi-inverse $\mathbb{G}_{s,w}$ exceptée que l'on doit expliciter tous les termes d'erreurs. En effet, l'analyse de $\mathbb{G}_{s,w}$ s'est faite à D constant, où D est le dénominateur de δ. Ici, nous allons considérer des rationnels dont le dénominateur D tend vers l'infini. Par exemple, lors de l'analyse des termes sous-dominants pour $\mathbb{G}_{s,w}$, nous avons dit "Par perturbation analytique, il en est de même dans un voisinage $V_2 \subset V_1$ de $(1,0)$ ...". Pour l'opérateur $\mathbb{H}_{s,w}$, il faudra en plus expliciter la taille du voisinage V_2 en fonction des dénominateurs. De plus, des propriétés pour w réel doivent être démontrée.

Théorème 14 (Propriété US pour $\mathbb{H}_{s,w}$) *Il existe huit constantes $K_0, K_1, K_2, K_3, K_4', K_5', K_6', K_7'$ vérifiant la propriété suivante : pour tout entier $D \geq 1$, notons $\mathcal{S}$ la bande de la forme $|\Re s - 1| \leq K_1/D^2$, et $\mathcal{W}$ le voisinage* réel *de $w=0$ de la forme $|w| \leq K_2/D$. Alors, pour tout rationnel $\delta \in]0,1[$ de dénominateur D, pour tout ϵ avec $1/D \leq |\epsilon| \leq K_3/D$, le pseudo-quasi-inverse $\mathbb{H}_{s,w}$ satisfait les propriétés suivantes :*

(i) Pour tout $w \in \mathcal{W}$, la série $\mathbb{H}_{s,w}$ admet un unique pôle $s=\sigma(w)$ dans la bande $\mathcal{S}$. Ce pôle est d'ordre 1.

(ii) Pour tout $w \in \mathcal{W}$, on a $|\sigma(w)-1| \leq K_4'/(2D^2)$. De plus, il existe $w_0 \in \mathcal{W}$ du même signe que ϵ, pour lequel $\sigma(w_0)-1<0$ et $|\sigma(w_0)-1| \geq \epsilon^2/(4K)$.

(iii) Sur $\mathcal{W}$, le résidu $E(w)$ de $\mathbb{H}_{s,w}[1](0)$ en $s=\sigma(w)$ satisfait $K_5' \leq |E(w)| \leq K_6'$.

(iv) Sur la ligne verticale $\Re s = 1 - K_1/D^2$, et pour tout $w \in \mathcal{W}$, la norme de $\mathbb{H}_{s,w}$ satisfait

$$||\mathbb{H}_{s,w}||_{1,\max(1,t)} \leq K_7' \cdot D^2 \max(1,\Im s)^{\xi}.$$

(v) Il existe $t_0>0$ tel que pour $s \in \mathcal{S}$ avec $|\Re(s)|>t_0$, le quasi-inverse vérifie

$$||\mathbb{H}_{s,w}||_{1,\Im s} \leq K_7' \cdot D^2 (\Im s)^{\xi}.$$

4.3.1 Région éloignée de l'axe réel

L'opérateur $\mathbb{H}_{s,w}$ intervient dans la série bivariée du paramètre M et vérifie

$$\mathbb{H}_{s,w} = \sum_{p \geq 0} \mathbf{G}_{s^-}^{p-\lfloor \gamma p \rfloor - \lfloor \delta p \rfloor} \circ \mathbf{G}_{s^+}^{\lfloor \delta p \rfloor} \circ \mathbf{G}_{s^-}^{\lfloor \gamma p \rfloor}$$

avec $s^- := s-(\delta-\epsilon)w$, $s^+ := s+(1-\delta+\epsilon)w$. Comme pour le pseudo-quasi-inverse $\mathbb{G}_{s,w}$, il existe un voisinage V_0 de $(1,0)$ tel que pour tout $(s,w) \in V_0$, les complexes s^- et s^+ sont dans

le cadre du théorème DBV pour l'opérateur de transfert classique. En utilisant la propriété (ii) du même théorème, la norme de $\mathbb{H}_{s,w}$ est bornée loin de l'axe réel par

$$||\mathbb{H}_{s,w}||_{1,t} \leq \sum_{p\geq 0} M^3 \cdot |t|^{3\xi} \cdot \theta^p = \frac{M^3}{1-\theta}|t|^{3\xi}.$$

Ceci montre que le pseudo-quasi-inverse $\mathbb{H}_{s,w}$ admet une zoné sans pôle pour s loin de l'axe réel et w dans un voisinage fixe de 0.

4.3.2 Décomposition autour de $(1,0)$

La démarche est la même que pour le pseudo-quasi-inverse $\mathbb{G}_{s,w}$. Nous décomposons l'opérateur en huit parties, en utilisant la décomposition spectrale 4.1. Si (s,w) appartiennent à V_0, alors $s^+ = s-(1-\delta+\epsilon)w$ et $s^- := s-\delta-\epsilon w$ appartient à un voisinage dans lequel les décompositions spectrales de $\mathbf{G}_{s^+}, \mathbf{G}_{s^-}$ existent. Comme $\mathbb{H}_{s,w}$ est composée de deux parties contenant $\mathbf{G}_{s^-}$ et d'une partie contenant $\mathbf{G}_{s^+}$, les décompositions spectrales impliquent une décomposition de $\mathbb{H}_{s,w}$ avec 1 terme *dominant* et 7 termes non-dominant.

4.3.2.1 Termes de reste

Chacun des termes de reste est obtenu en remplaçant dans l'opérateur $\mathbb{H}_{s,w}$, pour au moins une valeur de $t = s^+$ ou $t = s^-$, l'itéré $\mathbf{G}_t^k$ par l'itéré de la partie sous-dominante $\mathbf{N}_t^k$, les autres termes étant remplacés par la partie dominante $\lambda(t)^j \cdot \mathbf{P}_t$ (pour un itéré d'ordre j). Nous obtenons sept opérateurs : un qui contient uniquement des opérateur de type $\mathbf{N}_t$, et six opérateurs avec au moins une occurrence de $\mathbf{P}_t$.
Nous posons $\nu(t)$ le rayon spectral de $\mathbf{N}_t$ et $R := \log \nu$. Il existe un voisinage tel que $\nu(s^-)$ et $\nu(s^+)$ sont strictement plus petit que $a < 1$, qui ne dépend pas de D. La série ne contenant que des opérateurs de type $\mathbf{N}_t$ est alors normalement convergence dont la somme est bornée par une constante qui ne dépend pas de D.
Considérons maintenant les six autres séries, dont les normes peuvent se comparer facilement à une somme géométrique où le logarithme du terme général admet la forme générique

$$L(s,w) := \alpha^- R(s^-) + \alpha^+ R(s^+) + \beta^- \Lambda(s^-) + \beta^+ \Lambda(s^+),$$

avec $\alpha^+ + \alpha^- > 1/D, \beta^+ + \beta^- > 1/D$ et $\alpha^+ + \alpha^- + \beta^+ + \beta^- = 1$. Nous prouvons maintenant que la partie réelle de $L(s,w)$ est strictement négative sur un voisinage de la forme. $|1-s| \leq C/D,\ |w| \leq C/D$.

Tout d'abord, il existe un voisinage complexe $\mathcal{V}$ de $\tau = 1$ pour lequel

$$\Re R(\tau) < (1/2)R(1) < 0, \quad |\Lambda'(\tau)| \leq B, \quad \text{pour un } B > 0.$$

Alors on a : $\max(|\Lambda(s^+)|, |\Lambda(s^-)|) \leq (|s-1| + |w|)\, B \leq 2BC/D,$
et, finalement, si $C \leq |R(1)|/(8B)$, la partie réelle de $L(s,w)$ est plus petite que

$$R(1)/(2D) + |R(1)|/(4D) < R(1)/(4D) < 0.$$

Dans un voisinage de $\mathcal{V}$, les six séries sont donc majorées par

$$\frac{1}{1-\exp[R(1)/(4D)]} \leq K_9 \cdot D.$$

4.3.2.2 Terme dominant. Propriétés de la fonction $\psi(s,w)$

Le terme dominant est obtenu en remplaçant chaque occurrence de $\mathbf{G}_t$ par le terme $\lambda(t)\mathbf{P}_t$, et est de la forme $F_M^{[1]}(s,w)\cdot[\mathbf{P}_{s^-}\circ\mathbf{P}_{s^+}\circ\mathbf{P}_{s^-}[1](0)]$ avec

$$F_M^{[1]}(s,w)=\sum_{p=0}^{+\infty}\lambda(s^-)^{p-\lfloor\delta p\rfloor}\cdot\lambda(s^+)^{\lfloor\delta p\rfloor}.$$

Si maintenant $\delta=c/(c+d)$ est rationnel avec un dénominateur $D=c+d$, la série $F_M^{[1]}$ se récrit

$$F_M^{[1]}(s,w)=\left(\sum_{j=0}^{D-1}\lambda^{j-\lfloor\delta j\rfloor}(s^-)\lambda^{\lfloor\delta j\rfloor}(s^+)\right)\left(\sum_{k\geq 0}(\lambda^d(s^-)\lambda^c(s^+))^k\right).$$

C'est une fraction rationnelle par rapport aux deux variables $X=\lambda(s^+)$ et $Y=\lambda(s^-)$ et de la forme

$$F_M^{[1]}(s,w)=\frac{P(\lambda(s^+),\lambda(s^-))}{Q(\lambda(s^+),\lambda(s^-))}$$

où P et Q sont des polynômes de degré total au plus D,

$$P(X,Y):=\sum_{j=0}^{D-1}X^{\lfloor\delta j\rfloor}\cdot Y^{j-\lfloor\delta j\rfloor},\qquad Q(X,Y)=1-X^cY^d.$$

Les singularités de $F_M^{[1]}(s,w)$ sont uniquement due au dénominateur qui peut s'écrire $1-\psi(s,w)^D$ avec

$$\psi(s,w):=\lambda^{1-\delta}(s^-)\lambda^{\delta}(s^+).$$

Le dénominateur $s\mapsto 1-\psi(s,w)^D$ s'annule pour toutes les valeurs de s telles que

$$\psi(s,w)=\exp[2iK\pi/D]\qquad\text{avec } 0\leq K<D.$$

Cela signifie que la fonction $\Psi:=\log\psi$ satisfait

$$\Psi(s,w):=(1-\delta)\Lambda(s^-)+\delta\Lambda(s^+)=\frac{2iL\pi}{D},\qquad\text{avec}\quad L\in\mathbf{Z}.$$

Pour $w=0$, on a $\Psi(s,w)=\Lambda(s)=2iL\pi/D$, et le théorème des fonctions implicites peut être appliqué. Pour tout L, cela définit une courbe sur un petit voisinage de w, de la forme $s=\sigma_L(w)$, qui contient les zéros du dénominateur $1-\psi(s,w)^D$, et les pôles éventuelles de $F_M^{[1]}(s)$. Les zéros les plus proches de 1 sont relatifs aux L tels que $|L|=1$.
Nous décrivons d'abord le comportement de la fonction $\Psi(\sigma+it,w)$ quand $\sigma-1,t,w$ et ϵ sont petits. Avec le développement de $\Lambda(s)$ autour de 1,

$$\Lambda(s)=-B(s-1)+A(s)\cdot(s-1)^2\qquad\text{avec}\quad A(s):=\frac{1}{2}\Lambda''(1+s\theta)>0,\quad\theta\in[0,1],$$

et $B:=|\Lambda'(1)|$, nous déduisons, en utilisant $(1-\delta)(\sigma^--1)+\delta(\sigma^+-1)=(\sigma-1)+\epsilon w$, le développement suivant pour la partie réelle et imaginaire de la fonction $\Psi:=\log\psi$ autour de$(s,w)=(1,0)$,

$$\begin{aligned}\Re\Psi(s,w) &= -[(\sigma-1)-\epsilon w]B+\\ &\quad +A(s)\left[-t^2+(\sigma-1)^2-2(\sigma-1)\epsilon w+\epsilon^2w^2+\delta(1-\delta)w^2\right]\\ \Im\Psi(s,w) &= -Bt+2tA(s)[(\sigma-1)+\epsilon w].\end{aligned}$$

Lorsque t, w, ϵ sont d'ordre $O(1/D)$ et $\sigma - 1$ d'ordre $O(1/D^2)$, les parties réelles et imaginaires de Ψ satisfont

$$\Im\Psi = -Bt + O(\frac{1}{D^3}), \quad \Re\Psi = -[(\sigma - 1) + \epsilon w]B + A(s)[-t^2 + \delta(1-\delta)w^2] + O(\frac{1}{D^3}).$$

Il devient alors facile de prouver le lemme suivant.

Lemme 6 *Soit $\lambda(s)$ la valeur propre dominante de l'opérateur $\mathbf{G}_s$, $\psi(s,w)$ la fonction*

$$\psi(s,w) := \lambda^{1-\delta}(s^-) \cdot \lambda^{\delta}(s^+) = \lambda^{1-\delta}(s - (\delta - \epsilon)w) \cdot \lambda^{\delta}(s + (1 - \delta + \epsilon)w),$$

$\Lambda := \log\lambda$, et $\Psi := \log\psi$. Il existe sept constantes $K_0, K_1, K_2, K_3, K_4, K_5, K_6$ telles que pour tout entier $D \geq 1$, pour tout rationnel $\delta \in]0,1[$ de dénominateur D, les propriétés suivantes sont vérifiées :

(i) Quand (s, w, ϵ) satisfait

$$|t| \leq K_0/D, \quad |\sigma - 1| \leq K_1/D^2, \quad w \text{ réel }, |w| \leq K_2/D, \quad 1/D \leq |\epsilon| \leq K_3/D$$

alors les parties réelle et imaginaire de $\Psi(s,w)$ vérifient

$$|\Im\Psi(s,w)| \leq (3\pi)/(2D), \qquad \Re\Psi(s,w) \leq K_4/D^2,$$

et l'équation $\psi(s,w)^D = 1$ admet une unique racine $s := \sigma(w)$ tel que $\psi(\sigma(w), w) = 1$.

(ii) Quand (s, w, ϵ) satisfait

$$|t| \leq K_0/D, \quad \sigma - 1 = -K_1/D^2, \quad w \text{ réel }, |w| \leq K_2/D, \quad 1/D \leq |\epsilon| \leq K_3/D,$$

alors une des deux conditions suivantes est satisfaite pour la partie réelle et imaginaire de $\Psi(s,w)$,

$$\Re\Psi(s,w) \geq K_5/D^2 \qquad ou \qquad \pi/(2D) \leq \Im\Psi(s,w) \leq (3\pi)/(2D).$$

(iii) Quand (s, w, ϵ) satisfait

$$K_0/D < |t| \leq K_0', \quad |\sigma - 1| \leq K_1/D^2, \quad w \text{ réel }, |w| \leq K_2/D, \quad 1/D \leq |\epsilon| \leq K_3/D,$$

alors $\Re\Psi(s,w) \leq -K_6/D^2$, et $\psi(s,w)$ vérifie $|\psi(s,w)| \leq \exp(-K_6/D^2)$.

4.3.2.3 Terme dominant. Propriétés de la fonction σ.

Il est aussi nécessaire d'étudier la fonction $w \to \sigma(w)$ qui est définie dans un voisinage de $w = 0$ par la relation $\Psi(\sigma(w), w) = 0$.

Lemme 7 *Soit σ la fonction définie dans un voisinage de $w = 0$ par l'équation $\Psi(\sigma(w), w) = 0$. Cette fonction qui dépend des paramètres δ, ϵ, admet une dérivée seconde qui est uniformément bornée en δ, ϵ.*

Preuve. Nous avons :

$$(1-\delta)[\sigma'(w) - \delta - \epsilon]\Lambda'(\sigma^-(w)) + \delta[\sigma'(w) + (1 - \delta + \epsilon)]\Lambda'(\sigma^+(w)) = 0,$$

si bien que la dérivée $\sigma'(w)$ peut s'écrire comme une homographie F de $X(w) = \Lambda'(\sigma^-(w))/\Lambda'(\sigma^+(w))$ sous la forme

$$\sigma'(w) = \frac{(1-\delta)(\delta-\epsilon)X(w) - \delta(1-\delta+\epsilon)}{(1-\delta)X(w)+\delta} = F(X(w)).$$

Il est suffisant de borner les dérivées de F' et X'. Des propriétés bien connues de convexité de Λ implique la borne

$$\exp(-L|w|) \le X(w) := \frac{\Lambda'(\sigma^-(w))}{\Lambda'(\sigma^+(w)} \le \exp(L|w|),$$

ainsi $A_1 < X(w) < A_2$ pour $|w| < 1$, et finalement la dérivée $\sigma'(w)$ est uniformément bornée pour $|w| < 1$. Ensuite, la dérivée $X'(w)$ qui est donnée par

$$\frac{X'(w)}{X(w)} = (\sigma'(w) - \delta - \epsilon)\frac{\Lambda''(\sigma^-(w))}{\Lambda'(\sigma^-(w))} - (\sigma'(w) + (1-\delta+\epsilon))\frac{\Lambda''(\sigma^+(w))}{\Lambda'(\sigma^+(w))}$$

est aussi uniformément bornée pour $|w| \le 1$. D'un autre coté, la dérivée F' satisfait

$$F'(X) = \frac{\delta(1-\delta)}{((1-\delta)X(w)+\delta)^2} \qquad \text{soit } 0 \le F'(X) \le e^L$$

pour $|w| \le 1$, et $|\sigma''|$ est de fait uniformément bornée par une constante K. ■

4.3.2.4 Fin de la preuve du théorème 14

La proposition 16 récapitule toutes les propriétés analytiques de la série génératrice associée au paramètre M. Comme la série est la valeur de $\mathbb{H}_{s,w}$ pour la fonction constante 1, le tout pris en 0, les propriétés analytiques de la séries sont aussi celles du quasi-inverse.

La région proche de l'axe réel donne lieu à deux régions, délimitées par la ligne horizontale $|\Im s| = K_0/D$ du lemme 6.

Région 1. $|t| \le K_0/D, \quad |\sigma - 1| \le K_1/D^2$. Nous utilisons la propriété (i) du lemme 6 qui prouve que l'équation $\psi(s,w)^D = 1$ admet une unique racine $s := \sigma(w)$ qui satisfait $\psi(s,w) = 1$.

Évaluation du résidu en $s = \sigma(w)$. Autour de $s = \sigma(w)$, on a

$$F_M^{[1]}(s,w) \sim \frac{1}{s-\sigma(w)} \; \frac{-1}{\psi'_s(\sigma(w),w)} \frac{1}{D} \sum_{j=0}^{D-1} \lambda^{j-j'}(\sigma(w)^-)\lambda^{j'}(\sigma(w)^+) \tag{4.3}$$

Pour simplifier les notations, nous écrivons σ pour $\sigma(w)$.
Ici, on a $\lambda(\sigma^-) > 1 > \lambda(\sigma^+)$ et $\lambda(\sigma^+) = \lambda^{\frac{\delta-1}{\delta}}(\sigma^-)$. Alors

$$\sum_{j=0}^{D-1} \lambda^{j-\lfloor \delta j \rfloor}(\sigma^-) \cdot \lambda^{\lfloor \delta j \rfloor}(\sigma^+) \ge \sum_{j=0}^{D-1} \lambda^{j-\delta j}(\sigma^-) \cdot \lambda^{\delta j}(\sigma^+) = D.$$

Dans la même veine,

$$\sum_{j=0}^{D-1} \lambda^{j-\lfloor \delta j \rfloor}(\sigma^-) \cdot \lambda^{\lfloor \delta j \rfloor}(\sigma^+) \le \sum_{j=0}^{D-1} \lambda^{j+1-\delta j}(\sigma^-) \cdot \lambda^{\delta j - 1}(\sigma^+) = D\frac{\lambda(\sigma^-)}{\lambda(\sigma^+)}.$$

En combinant ces résultats, le résidu en $s = \sigma(w)$ satisfait

$$\frac{1}{|\psi'_s(\sigma(w),w)|} \le \mathrm{Res}_{s=\sigma} F_M(s) \le \frac{\lambda(\sigma^-)}{\lambda(\sigma^+)} \cdot \frac{1}{|\psi'_s(\sigma(w),w)|},$$

ce qui donne le résultat puisque $\lambda(\sigma^-), \lambda(\sigma^+), \psi'_s(\sigma(w), w)$ admettent des bornes inférieures et supérieures dans un voisinage de $w = 0$ qui ne dépend pas de δ, ϵ.

Majoration de $\Re s = 1 - K_1/D^2, |\Im s| \le K_0/D$. On a

$$|F_M^{[1]}(s)| \le \frac{1}{|\psi(s)^D - 1|} \cdot |B(s)| \qquad \text{avec} \quad B(s) := \sum_{j=0}^{D-1} \lambda(s^-)^{j-\lfloor \delta j \rfloor} \cdot \lambda(s^+)^{\lfloor \delta j \rfloor}.$$

Ici, chaque terme $|\lambda(s^-)|, |\lambda(s^+)|$ est en $\exp[O(1/D)]$ si bien que chaque terme $B(s)$ a un module plus petit que K_{10}, et finalement $|B(s)| \le K_{10} \cdot D$. D'un autre coté, la propriété (ii) du lemme 6 entraîne que

$$|\psi(s)^D - 1| \ge 1 \qquad \text{ou} \quad |\psi(s)^D - 1| \ge \exp[K_5/D] - 1 \ge K_5/D$$

soit

$$|F_M^{[1]}(s)| \le K_{12} \cdot D^2$$

Minoration de $s - \sigma(w)$. Nous prouvons que le pôle $\sigma(w)$ n'est pas trop proche de la ligne verticale $\Re s = 1 - C/D^2$ lorsque w est réel et satisfait $|w| \le C/D$. Nous utilisons le lemme 7 qui montre que $|\sigma'(w) - \sigma'(0)| \le K|w|$. Comme la dérivée de σ' est égale à $-\epsilon$ en $w = 0$, la dérivée $\sigma'(w)$ vérifie

$$-\epsilon w - \frac{1}{2}|\epsilon| < \sigma'(w) < -\epsilon w + \frac{1}{2}|\epsilon| \qquad \text{pour} \quad |w| \le \frac{\epsilon}{2K}.$$

Finalement, nous avons $|\sigma(w) - 1| \le 2|\epsilon| \cdot |w| \le K'_4/(2D^2)$ dès que $|w|$ est majoré $\epsilon/(2K) \le K'_4/D$. Or, si s appartient à la ligne verticale $\Re s = 1 - K_1/D^2$, nous avons $|s - \sigma(w)| \ge K'_4/(2D^2)$ à condition que $K_1 \ge K'_4$.

D'un autre coté, on a

$$\sigma(w_0) - 1 < 0, \qquad |\sigma(w_0) - 1| \ge \frac{1}{2}|\epsilon| \cdot |w| \ge \frac{\epsilon^2}{4K}$$

aussitôt que w_0 est du même signe que ϵ et $|w_0| = \epsilon/2K$.

Région 2. $K_0/D < |t| \le K'_0$, $|\sigma - 1| \le K_1/D^2$, w réel , $|w| \le K_2/D$. Nous rappelons la troisième propriété du lemme 6 : dans cette région, $\psi(s, w)$ est majorée en module par $|\psi(s, w)| \le \exp(-K/D^2) < 1$. Comme $\max(1, |\lambda(s^-)|, |\lambda(s^+)|^{-1}) \le K$, on a

$$|\lambda(s^-)|^{p-\lfloor \delta p \rfloor} \le K|\lambda(s^-)|^{p-\delta p}, \qquad |\lambda(s^+)|^{\lfloor \delta p \rfloor} \le K \cdot \lambda(s^+)^{\delta p}$$

et la série $F_M(s, w)$ satisfait

$$|F_M(s, w)| \le K^2 \sum_{p=0}^{\infty} \psi(s, w)^p \le K^2 \frac{1}{1 - \psi(s, w)} \le K_8 D^2.$$

Ceci termine l'analyse dans la zone autour de $(1, 0)$. Pour la zone intermédiaire, le travail est identique à l'autre pseudo-quasi-inverse. En particulier, nous venons d'établir toutes les propriétés du théorème 14 concernant le pseudo-quasi-inverse $\mathbb{H}_{s,w}$. ■

4.4 Conclusion

Dans cette partie, nous avons adapté les travaux et la démarche de Baladi et Vallée aux deux pseudo-quasi-inverses. L'analyse suit toujours un découpage en trois zones. La première zone est la zone éloignée de l'axe réel où les majorations à la Dolgopyat s'appliquent très simplement. Ainsi, nous montrons que cette zone n'admet pas de pôle. La deuxième zone est celle autour de $(1,0)$. Là encore, les idées restent les mêmes. La décomposition spectrale des opérateurs de transfert entraîne une décomposition des pseudo-quasi-inverses en un terme dominant et plusieurs termes de reste. Les termes de reste sont très vite traités sur des voisinages adéquat. En particulier, ces termes n'apportent pas de pôle dans le voisinage. Le terme dominant fait apparaître une fraction rationnelle dont de dénominateur est de la forme $1-\psi(s,w)^D$. Cette fraction admet un unique pôle dans un voisinage de la forme $\{s : |\Re s - 1| < K/D^2\} \times \mathcal{W}$ où D est le dénominateur des paramètres. Finalement, dans la zone intermédiaire il est facilement démontré que les pseudo-quasi-inverses n'admettent pas de pôle. En collant les trois zones, nous obtenons la propriété $US(s,w)$ dans un voisinage dont la taille est d'ordre $1/D^2$

Toutes ces analyses fonctionnent parce que les pseudo-quasi-inverses correspondent à un découpage en deux ou trois parties régulières de l'exécution. En essayant d'étudier l'équivalent du paramètre N des polynômes (voir théorème 11), l'opérateur suivant apparaît naturellement

$$\mathbf{G}_{s-pw} \circ \mathbf{G}_{s-(p-1)w} \circ \ldots \circ \mathbf{G}_{s-w}.$$

Cette opérateur correspond d'ailleurs à la version opérateur de la série génératrice associée à N. Contrairement aux pseudo-quasi-inverse, aucune grande puissance d'un opérateur de transfert apparaît, ce qui, au voisinage de 0 pose des problèmes pour la décomposition.

Chapitre 5

Calcul de constantes spectrales

Sommaire

5.1 Introduction

Tous les résultats de cette thèse font intervenir des constantes dont certaines admettent une formule close (constantes μ des théorèmes 2 et BV) alors que d'autres n'en admettent pas (constantes ρ des mêmes théorèmes). Pour ces dernières, se pose le problème de les évaluer. Le livre de Finch [Fin03] regroupe plusieurs instances de cette situation avec des constantes provenant de domaines très variés comme l'arithmétique, la théorie des nombres, l'algorithmique, etc. Dans notre contexte, les constantes ρ sont liées à la valeur propre dominante d'opérateurs de

transfert du système dynamique des fractions continues. Nous décrivons une méthode générale, appelée méthode DFV, qui approche ces valeurs propres. Nous en déduisons des algorithmes pour le calcul des constantes ρ mais aussi des constantes de Hensley, de Gauss-Kuz'min-Wirsing et des dimensions de Hausdorff qui sont toutes des grandeurs importantes pour les fractions continues. Ces algorithmes montrent que les constantes associées sont calculables en temps polynomial, c'est-à-dire qu'il existe un réel $r > 0$ tel que les d premiers digits sont calculables en temps $O(d^r)$.

Nous commençons par décrire la méthode DFV avant d'introduire les différentes constantes. Dans la deuxième partie, nous donnons des conditions suffisantes pour pouvoir utiliser la méthode et nous prouvons sa convergence sous ces conditions. L'utilisation de la méthode DFV afin d'obtenir des valeurs prouvées, suppose de savoir calculer certaines constantes. Ceci sera l'objet de la troisième partie. Finalement, nous terminerons avec les algorithmes qui calculent les différentes constantes.

5.1.1 La méthode DFV

La méthode DFV tire son nom de ses trois auteurs Daudé, Flajolet et Vallée qui l'ont introduit dans [DFV97] et qui fut réutilisée plus tard dans [FV00, Val98b]. La méthode DFV approche une partie finie du spectre des opérateurs de transfert mais ses auteurs n'ont pas prouvé cette convergence. Dans la suite, nous en apportons la preuve en exhibant une vitesse explicite.

En dimension finie, il existe plusieurs algorithmes qui calculent les valeurs propres d'un opérateur ou d'une matrice. Malheureusement, les opérateurs de transfert n'agissent en général pas sur des espaces de dimension finie et le calcul exact des valeurs propres est alors presque toujours impossible. Cependant, lorsque les opérateurs de transfert agissent sur l'espace des fonctions analytiques, ils peuvent être vus comme des matrices infinies $\mathcal{M} = (M_{i,j})_{1\leq i,j\leq\infty}$. La méthode DFV considère les matrices tronquées $\mathcal{M}_n = (M_{i,j})_{1\leq i,j\leq n}$ et calcule leurs valeurs propres. En utilisant des résultats de [ALL01], nous montrons que pour toute valeur propre isolée λ de $\mathcal{M}$, il existe une suite λ_n de valeurs propres de $\mathcal{M}_n$ qui converge vers λ.

Nous fixons D un disque de centre x_0 dans le plan complexe et considérons un opérateur $\mathbf{G}$ qui agit sur l'espace $\mathcal{A}_\infty(D)$ des fonctions analytiques sur D et continues sur le bord de D,

$$\mathcal{A}_\infty(D) := \{f : D \to \mathbb{C};\quad f \text{ analytique sur } D \text{ et continue sur } \overline{D}\}.$$

Pour f dans $\mathcal{A}_\infty(D)$, les développements de Taylor de f et $\mathbf{G}[f]$ en x_0 existent et l'opérateur $\mathbf{G}$ peut être vu comme une matrice infinie $\mathcal{M} = (M_{i,j})_{0\leq i,j\leq\infty}$ où le coefficient $M_{i,j}$ est le coefficient de $(z-x_0)^i$ dans $\mathbf{G}[(z-x_0)^j](z)$,

$$M_{i,j} := [(z-x_0)^i]\ \mathbf{G}[(z-x_0)^j](z).$$

La matrice $\mathcal{M}_n = (M_{i,j})_{0\leq i,j\leq n}$ est la matrice d'un opérateur tronqué dont on regarde l'action sur les polynômes $\mathcal{P}_n$ de degrés au plus n. Si π_n est l'opérateur de troncature du développement de Taylor à l'ordre n,

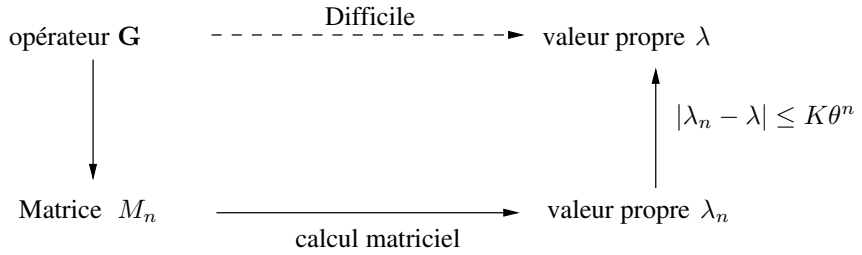
$$\pi_n[f](z) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(z-x_0)^k, \tag{5.1}$$

alors $\mathcal{M}_n$ est la matrice de l'opérateur $\pi_n \circ \mathbf{G}|_{\mathcal{P}_n}$. Il est à noter que l'opérateur $\pi_n \circ \mathbf{G}$ et la matrice $\mathcal{M}_n$ ont le même spectre.

La méthode DFV suit trois grandes étapes :

1. Calculer la matrice tronquée $\mathcal{M}_n$,
2. Calculer le spectre de $\mathcal{M}_n$,
3. Extraire la ou les valeurs propres qui offrent un intérêt.

Dans la suite, nous nous intéressons essentiellement à l'unique valeur propre dominante λ des opérateurs de transfert. Dans [DFV97], les auteurs ont observé que pour n grand, la matrice $\mathcal{M}_n$ admet aussi une unique valeur propre dominante λ_n et que celle-ci semble converger vers λ avec une vitesse exponentielle. Ils ont conjecturé le résultat suivant : *Il existe deux constantes n_0, K et un réel $\theta < 1$ tels que pour tout $n \geq n_0$, $|\lambda_n - \lambda| \leq K\theta^n$.*



Nous allons prouver que ce résultat est vrai pour toute valeur propre isolée des opérateurs de transfert dès que le système dynamique associé admet des branches inverses fortement contractantes (hypothèse $\mathcal{BFC}$). Maintenant si les matrices M_n sont calculables en temps polynomial, la méthode DFV offre un moyen de calcul en temps polynomial des valeurs propres isolées des opérateurs de transfert.

Nous décrivons maintenant plusieurs types de constantes qui jouent un rôle important pour le système dynamique des fractions continues. Ces constantes s'expriment toutes en fonction des valeurs propres des opérateurs de transfert et en s'appuyant sur la méthode DFV, nous proposerons des algorithmes pour les calculer.

5.1.2 Constante de Gauss-Kuz'min-Wirsing

Autour de 1800, Gauss [Gau07] a étudié l'évolution de la distribution initiale sous l'effet des itérés du shift T des fractions continues. Il introduit un opérateur proche du transformateur de densité $\mathbf{G}$ et montre qu'il existe une densité $g(x) = (1/\log 2)(1+x)^{-1}$ invariante sous l'action de T (i.e. $\mathbf{G}[g] = g$). Il conjectura que cette densité est une densité limite c'est-à-dire que quelque soit la densité initiale f, les densités successives après itérations de T convergent vers g, autrement dit $\mathbf{G}^n[f]$ converge vers g. Un siècle plus tard, Kuz'min [Kuz28] et Lévy [Lé29] ont démontré la conjecture. Il était alors important d'obtenir la vitesse de convergence optimale de $\mathbf{G}^n[f]$ vers g. Autour de 1975, Babenko [Bab78] et Wirsing [Wir74] ont complètement résolu le problème et ont montré que la vitesse était exponentielle. Le rapport est donné par l'unique valeur propre sous-dominante du transformateur de densité dont Wirsing a montré qu'elle était réelle et négative. Cette constante, appelée Constante de Gauss-Kuzmin-Wirsing et notée γ_G, ne semble pas liées à d'autres constantes. Elle a été calculée dans [DFV97] avec 30 chiffres après la virgule en utilisant la méthode DFV,

$$\gamma_G \approx -0.30396355092701333\ldots$$

Avec des méthodes similaires, Sebah (non publié) et Briggs [Bri03] ont amélioré la précision à 100 chiffres et 385 chiffres. Cependant, la méthode DFV implique qu'il existe un algorithme polynomial pour calculer la constante γ_G (si l'on admet pour l'instant que les matrices sont calculables en temps polynomial).

5.1.3 Constante de Hensley

Le nombre de divisions que l'algorithme d'Euclide effectue est historiquement le premier paramètre qui a été étudié. Lamé [Lam45] a mis en évidence le pire des cas, Heilbronn [Hei69] et Dixon [Dix70] ont indépendamment effectué l'analyse en moyenne et finalement, Hensley a prouvé la loi limite gaussienne [Hen94]. Le nombre de divisions est un coût à croissance modérée et forme aussi un cas particulier des résultats de Baladi et Vallée. Sur l'ensemble des entrées de taille n, l'espérance et la variance du nombre de divisions P s'expriment en fonction des deux premières dérivées de la valeur propre dominante $\lambda(s)$ de l'opérateur de transfert $\mathbf{G}_s$,

$$\mathrm{E}_n[P] \sim \frac{-2}{\lambda'(1)} n, \qquad \mathrm{V}_n[P] \sim 2\frac{\lambda''(1) - \lambda'(1)^2}{\lambda'(1)^3} n. \tag{5.2}$$

La première dérivée $\lambda'(1)$ est au signe près l'entropie du système dynamique des fractions continues. Puisque la densité invariante (qui est la densité de Gauss) est connue, l'entropie est explicite et $\lambda'(1)$ vaut $-\pi^2/(6\log 2)$. La constante qui apparaît dans le terme dominant de la variance est appelée constante de Hensley et est notée γ_H. Elle s'exprime en fonction de la dérivée seconde $\lambda''(1)$ qui n'est reliée jusqu'à aujourd'hui à aucune autre constante connue. La constante de Hensley a d'abord été calculée dans [FV00] avec la méthode DFV,

$$\gamma_H \approx 0.516024\ldots$$

Dans la suite, nous proposons un algorithme polynomial qui calcule γ_H avec une précision prouvée.

5.1.4 Dimension de Hausdorff de fractions continues contraintes

Soit $\mathcal{E}$ un sous ensemble de $\mathbb{N}^\star$. Dans le contexte des fractions continues, l'ensemble de Cantor de $C_\mathcal{E}$ désigne l'ensemble des réels de l'intervalle $[0,1]$ dont le développement en fraction continue est contraint à $\mathcal{E}$,

$$C_\mathcal{E} := \{x \in [0,1]; x = [m_1, \ldots, m_i, \ldots],\ \forall i,\ m_i \in \mathcal{E}\},$$

où $x = [m_1, \ldots, m_i, \ldots]$ signifie

$$x = \cfrac{1}{m_1 + \cfrac{1}{m_2 + \cfrac{1}{\ddots}}}.$$

Dès que $\mathcal{E}$ est différent de $\mathbb{N}^\star$, l'ensemble $C_\mathcal{E}$ est un ensemble non dénombrable de mesure de Lebesgue nulle. Sa dimension de Hausdorff $s_\mathcal{E}$ est alors adaptée pour en faire une description précise. En particulier, la probabilité qu'un rationnel dont le numérateur et le dénominateur sont plus petits que n appartient à $C_\mathcal{E}$ est $\Theta(n^{2s_\mathcal{E}-2})$. Lorsque $\mathcal{E}$ est fini, les réels de $C_\mathcal{E}$ sont aussi très intéressants puisqu'ils sont tous difficilement estimables par des rationnels [Sha92].
Si l'ensemble $\mathcal{E}$ contient plus de deux éléments, la dimension de Hausdorff de $C_\mathcal{E}$ est un réel $s_\mathcal{E}$ de $]0,1[$. Hensley [Hen92, Hen96] et Vallée [Val98b] ont montré que $s_\mathcal{E}$ est liée à l'unique valeur propre dominante $\lambda_\mathcal{E}(s)$ de l'opérateur de transfert contraint $\mathbf{G}_{s,\mathcal{E}}$:

$$\mathbf{G}_{s,\mathcal{E}}[f](x) = \sum_{m \in \mathcal{E}} \frac{1}{(m+x)^{2s}} f\left(\frac{1}{m+x}\right).$$

La dimension $s_{\mathcal{E}}$ est l'unique s tel que $\lambda_{\mathcal{E}}(s) = 1$.

La dimension de Hausdorff relative à l'ensemble $\mathcal{E} = \{1, 2\}$ a été intensivement étudiée. En 1941, Good [Goo41] a montré que $0.5194 \leq s_{\{1,2\}} \leq 0.5433$ et en 1982, Bumby [Bum82, Bum85] améliore l'estimation et obtient $s_{\{1,2\}} = 0.5313 \pm 10^{-4}$. En 1996, Hensley [Hen96] décrit le premier algorithme en temps polynomial dans la cas d'un ensemble fini $\mathcal{E}$ et obtient la valeur suivante

$$s_{\{1,2\}} \approx 0.5312805062772051416.$$

Finalement en 1999, Jenkinson et Pollicott [JP99] ont développé une algorithme efficace, basé sur les points fixes des branches inverses, et ont calculé les 25 premiers chiffres significatifs de $s_{\{1,2\}}$. Leur algorithme n'est toutefois pas polynomial.

La méthode DFV a déjà été appliquée avec des ensembles généraux $\mathcal{E}$. Elle se généralise aussi très facilement à des contraintes périodiques [Val98b] du type $\mathcal{E}_0 \times \ldots \times \mathcal{E}_{k-1}$ qui signifie que l'entier m_i dans le développement en fraction continue appartient à $\mathcal{E}_{i \mod k}$. Sur ces ensembles, nous montrerons que la méthode converge et dans le cas non-périodique, nous donnerons un algorithme polynomial pour calculer $s_{\mathcal{E}}$.

5.1.5 Constantes ρ

Les constantes $\rho(c)$ et $\rho_{[\delta]}$ sont données par les théorèmes 5 et BV. La première fait intervenir les cinq premières dérivées de la valeur propre dominante $\lambda(s, w)$ de l'opérateur de transfert pondéré relatif à c. La seconde, fait uniquement intervenir les deux premières dérivées de $\lambda(s)$ où $\lambda(s)$ est la valeur propre dominante de l'opérateur de transfert $\mathbf{G}_s$. Le calcul de la constante de Hensley et de $\rho_{[\delta]}$ sont donc similaires et passent par le calcul de $\lambda''(1)$. Comme pour la constante de Hensley, nous proposerons donc un algorithme polynomial pour évaluer $\rho_{[\delta]}$.

La constante $\rho(c)$ fait intervenir un opérateur de transfert différent. Pour l'opérateur pondéré, nous n'avons pas réussi à calculer les constantes qui apparaissent dans la vitesse de convergence (nous verrons pourquoi c'est difficile) et qui sont nécessaires pour contrôler l'erreur sur le calcul des valeurs propres. Nous ne somme donc pas en mesure de donner une valeur prouvée de $\rho(c)$ pour un coût primitif c fixé, mais la méthode DFV converge toujours et une valeur *fortement probable* peut être donnée.

5.2 Hypothèses de convergence pour la méthode DFV

5.2.1 Systèmes à branches fortement contractantes

Comme expliqué dans la section précédente, nous souhaitons prouver la convergence de la méthode DFV. Dans ce contexte, il est naturel de poser la définition suivante.

Définition 8 (Opérateur avec de bonnes troncatures) *Soit D un disque de centre x_0 et $\mathbf{G}$ un opérateur qui agit sur $\mathcal{A}_\infty(D)$. Posons également π_n la projection définie par la formule 5.1 et $\mathbf{G}_n := \pi_n \circ \mathbf{G}$. L'opérateur $\mathbf{G}$ a de bonnes troncatures si la propriété suivante est vérifiée : il existe $\theta < 1$, tel que pour toute valeur propre isolée λ de l'opérateur $\mathbf{G}$, il existe $K > 0$, un entier n_0 et une séquence λ_n de valeurs propres de $\mathbf{G}_n$ pour laquelle*

$$|\lambda_n - \lambda| \leq K\theta^n, \qquad \forall n \geq n_0.$$

Le réel θ est appelé le rapport de troncature.

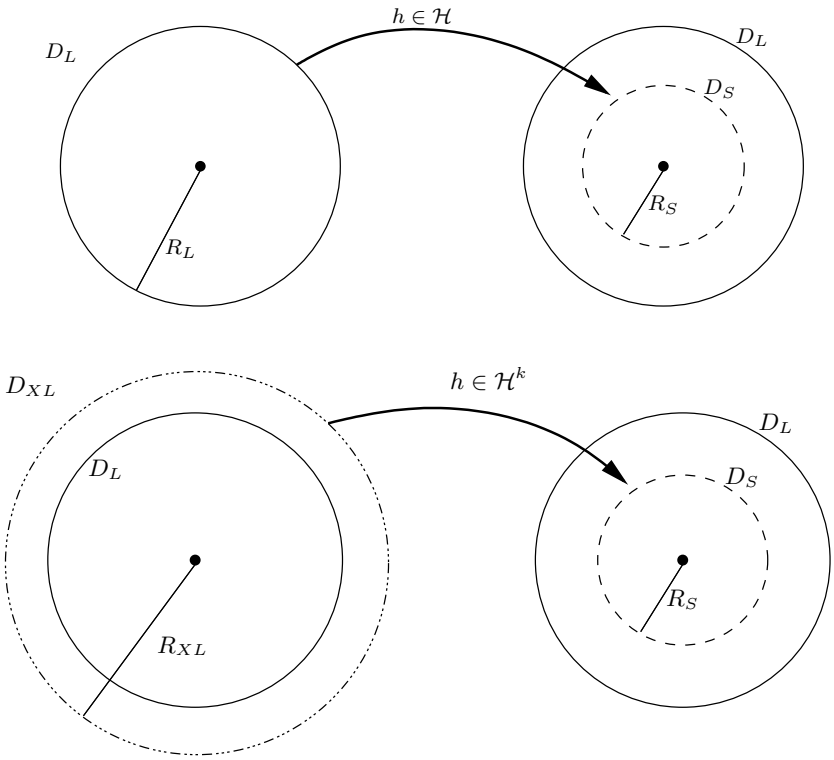


FIG. 5.1 – Branches fortement et très fortement contractantes

Nous avons vu, en introduisant la condition UNI, la propriété de contraction des branches inverses, i.e,

$$\exists n_0 \in \mathbb{N}, \quad \forall n \geq n_0, \quad \forall h \in \mathcal{H}^n, \quad |h'| < 1.$$

Cette propriété de contraction implique de bonnes propriétés sur le système dynamique. Elle entraîne aussi que l'image d'un intervalle par une branche inverse a une mesure strictement plus petite que l'intervalle de départ. Pour démontrer la convergence de la méthode DFV, nous devons supposer un peu plus ce qui motive la définition suivante.

Définition 9 (Branches Fortement Contractantes) *Un système dynamique* holomorphe *est dit à branches fortement contractantes (Propriété BFC) s'il existe deux disques concentriques D_S et D_L, de rayons respectifs $R_S < R_L$ et tels que pour toute branche inverse, l'image du grand disque fermé $\overline{D}_L$ ($L = large$) est contenu dans l'image du petit disque $\overline{D}_S$ ($S = small$).*

Un système dynamique à branches fortement contractantes est dit à branches très fortement contractantes (Propriété BTFC) s'il existe $k > 1$ et un troisième disque D_{XL} ($L = extra-large$) concentriques avec D_S et D_L, de rayon R_{XL} avec $R_{XL} > R_L$ et tel que toute branche inverse h de profondeur k satisfait $h(\overline{D}_{XL}) \subseteq \overline{D}_S$.

La figure 5.1 résume les deux propriétés de contraction. Les opérateurs de transfert classiques, contraints ou pondérés qui interviennent dans les constantes sont tous de la forme

$$\mathbf{G}_{s,\mathcal{A}}[f] = \sum_{h \in \mathcal{A}} \alpha_h^s \; f \circ h,$$

où α_h est soit $|h'|$, soit $e^{wc(h)}|h'|$ et où $\mathcal{A} \subseteq \mathcal{H}$. Pour un disque D sur lequel toutes les fonctions α_h sont définies, nous notons $\delta(s, \mathcal{A}, D)$ la quantité

$$\delta(s, \mathcal{A}, D) := \sum_{h \in \mathcal{A}} \sup_{x \in D} |\alpha_h|^{\Re(s)}.$$

Nous donnons maintenant le premier résultat important pour le calcul des constantes.

Théorème 15 (Les opérateurs de transfert ont de bonnes troncatures) *Pour un système dynamique holomorphe à branches fortement contractantes sur les disques D_S et D_L, si $\delta(s, \mathcal{A}, D_L) < \infty$, alors l'opérateur de transfert $\mathbf{G}_{s,\mathcal{A}} : \mathcal{A}_\infty(D_S) \to \mathcal{A}_\infty(D_L)$ satisfait les points suivants :*

(i) $\mathbf{G}_{s,\mathcal{A}}$ est compact et son spectre est formé de valeurs propres isolées de multiplicité finie, excepté en 0.

(ii) Pour tout réel s tel que $\delta(s, \mathcal{A}, D_L)$ converge, l'opérateur $\mathbf{G}_{s,\mathcal{A}}$ admet une unique valeur propre dominante simple, positive et isolée du reste du spectre par un saut spectral.

(iii) Pour tout réel s tel que $\delta(s, \mathcal{A}, D_L)$ converge, l'opérateur $\mathbf{G}_{s,\mathcal{A}}$ a de bonnes troncatures. Le rapport de troncature θ satisfait $\theta \leq R_S/R_L$ où R_S et R_L sont les rayons des disques D_S et D_L.

(iv) Si le système dynamique est à branche très fortement contractantes sur les disques D_{XL}, D_L et D_S, alors le rapport de troncature θ satisfait $\theta \leq R_S/R_{XL}$ où R_S et R_{XL} sont les rayons des disques D_S et D_{XL}.

Le précédent théorème est fondamental puisqu'il montre que la méthode DFV converge avec les opérateurs de transfert classiques, pondérés et contraints associé à un système dynamique à branches fortement contractantes.

5.2.2 Exemples de systèmes à branches fortement contractantes

Nous considérons ici trois algorithmes euclidiens rapides : l'algorithme d'Euclide standard $\mathcal{S}$, l'algorithme d'Euclide centré $\mathcal{C}$ et l'algorithme d'Euclide impair $\mathcal{I}$. Ces trois algorithmes sont liés à des systèmes dynamiques (I, T), où T est de la forme

$$T(x) := \left| \frac{1}{x} - V\left(\frac{1}{x}\right) \right|,$$

avec $V_{\mathcal{S}}(u)$ la partie entière de u, $V_{\mathcal{C}}(u)$ le plus proche entier de u et $V_{\mathcal{I}}(u)$ le plus proche entier impair de u. Les intervalles $I_{\mathcal{S}}$ et $I_{\mathcal{I}}$ sont égaux à $[0, 1]$ alors que l'intervalle $I_{\mathcal{C}}$ est égal à $[0, 1]$. Les ensembles de branches inverses pour chacun de ces systèmes dynamiques sont donnés par,

$$\mathcal{H}_{\mathcal{S}} = \{x \to \frac{1}{m+x};\ m \geq 1\}, \quad \mathcal{H}_{\mathcal{C}} = \{x \to \frac{1}{m+\epsilon x};\ \epsilon = \pm 1, (m, \epsilon) \geq (2, +1)\}$$

$$\mathcal{H}_{\mathcal{O}} = \{x \to \frac{1}{m+\epsilon x};\ \epsilon = \pm 1, m \text{ impair}, (m, \epsilon) \geq (1, +1)\}$$

où l'ordre $\geq$ est l'ordre lexicographique. Dans les trois cas, les systèmes dynamiques sont à branches fortement contractantes et les paramètres (x_0, R_S, R_L) peuvent être choisit de la manière suivante,

$$\mathcal{S} : (1, 1, 3/2), \qquad \mathcal{C} : (1/4, 5/12, 3/4), \qquad \mathcal{I} : (1, 1, 3/2).$$

Ainsi, les rapport de troncature satisfont $\theta_{\mathcal{S}} = \theta_{\mathcal{I}} = 2/3$ et $\theta_{\mathcal{C}} = 5/9$.

Pour les trois algorithmes d'Euclide, la méthode DFV s'applique et constitue la seule méthode (à notre connaissance) permettant d'évaluer l'équivalent de la constante de Hensley ou des constantes ρ pour les algorithmes centré et impair.

5.3 Preuve de la convergence de la méthode DFV

Nous prouvons maintenant le théorème 15 et explicitons les constantes n_0, K et θ. Les deux premiers résultats sont des résultats classiques qui se démontrent très aisément à partir des travaux de Mayer [May79]. Nous nous concentrons donc sur les troisième et quatrième assertions. La preuve que nous proposons est basée sur deux résultats principaux. Le premier est un résultat classique d'Analyse Fonctionnelle qui dit que deux opérateurs proches en norme ont des spectres également proches. Le second résultat montre que la propriété forte de contraction entraîne la convergence en norme des opérateurs tronqués vers l'opérateur de transfert.

5.3.1 Analyse fonctionnelle

Soit $(\mathcal{B}, ||\ ||)$ un espace de Banach complexe et $\mathbf{G}$ un opérateur qui agit sur $\mathcal{B}$. Le spectre de $\mathbf{G}$ est noté Sp$\mathbf{G}$. Soit λ une valeur propre de $\mathbf{G}$ et $C = C(\lambda, r)$ un cercle de centre λ et de rayon r qui isole λ du reste du spectre. En particulier, r satisfait $r < d(\lambda, \mathrm{Sp}\mathbf{G}\backslash\{\lambda\})$. Les constantes $\alpha_C(\mathbf{G})$ et $\beta_C(\mathbf{G})$ définies par

$$\alpha_C(\mathbf{G}) := \sup_{z \in C} ||(\mathbf{G} - z\mathbf{I})^{-1}||, \tag{5.3}$$

$$\beta_C(\mathbf{G}) := \max\left\{ \frac{1}{2\alpha_C(\mathbf{G})}, \frac{1}{2r\alpha_C^2(\mathbf{G})}, \frac{1}{8r^2\alpha_C^3(\mathbf{G})} \right\}, \tag{5.4}$$

jouent un rôle central dans la suite. Elles interviennent d'abord dans le résultat fondamental suivant.

Lemme 8 *Soit $\mathbf{G}$ et $\widetilde{\mathbf{G}}$ deux opérateurs sur l'espace de Banach $(\mathcal{B}, ||\ ||)$. Supposons que λ est une valeur propre simple et isolée de $\mathbf{G}$ associée au vecteur propre ϕ et posons $C = C(\lambda, r)$ un cercle qui isole λ du reste du spectre. Si $\mathbf{G}$ et $\widetilde{\mathbf{G}}$ satisfont $||\mathbf{G} - \widetilde{\mathbf{G}}|| \leq \beta_C(\mathbf{G})$, alors l'opérateur $\widetilde{\mathbf{G}}$ admet une unique valeur propre simple $\widetilde{\lambda}$ à l'intérieur de C qui satisfait*

$$\begin{aligned} |\widetilde{\lambda} - \lambda| &\leq 2r \cdot \alpha_C(\mathbf{G}) \cdot \frac{||\widetilde{\mathbf{G}}[\phi] - \mathbf{G}[\phi]||}{||\phi||} && (5.5) \\ &\leq 2r \cdot \alpha_C(\mathbf{G}) \cdot ||\widetilde{\mathbf{G}} - \mathbf{G}||. && (5.6) \end{aligned}$$

La condition $||\mathbf{G} - \widetilde{\mathbf{G}}|| \leq \beta_C(\mathbf{G})$ implique, à travers la définition de $\beta_C(\mathbf{G})$, trois conditions ayant un objectif bien précis. La première condition assure que le cercle C ne contient pas de valeur propre de $\widetilde{\mathbf{G}}$. La seconde condition implique que $\widetilde{\lambda}$ est l'unique valeur propre de $\widetilde{\mathbf{G}}$ dans C. Finalement, il est possible de relier les deux espaces propres associés à λ et $\widetilde{\lambda}$ avec la dernière condition.

Le lemme précédent est un résultat classique d'analyse fonctionnelle que l'on peut retrouver dans le livre [ALL01]. Il montre que la convergence en norme d'opérateurs implique la convergence en module des valeurs propres isolées. A la prochaine section, nous montrons que les opérateurs tronqués convergent vers l'opérateur de transfert associé entraînant la preuve du théorème 15. Nous utiliserons aussi le lemme pour le calcul la constante de Hensley (cf. section 5.5.2).

5.3.2 Convergence des opérateurs tronqués pour des systèmes à branches fortement contractantes

Fixons un opérateur $\mathbf{G} : \mathcal{A}_\infty(D_S) \to \mathcal{A}_\infty(D_L)$. Les valeurs propres non nulles de l'opérateur tronqué $\mathbf{G}_n = \pi_n \circ \mathbf{G}$ et celles de la matrice $\mathcal{M}_n$ sont identiques. Selon le lemme précédent, il suffit donc de montrer la convergence en norme de $\pi_n \circ \mathbf{G}_n$ vers $\mathbf{G}$ pour prouver que la méthode DFV converge. C'est l'objectif du lemme suivant qui nécessite l'hypothèse $\mathcal{BFC}$. Une preuve restreinte au cadre des fractions continues peut aussi être trouvée dans [Hen04] avec des espaces fonctionnels légèrement différents.

Lemme 9 *Soit $\mathbf{G} : \mathcal{A}_\infty(D_S) \to \mathcal{A}_\infty(D_L)$ un opérateur de norme $||\mathbf{G}||_{D_S,D_L}$ avec D_S et D_L concentriques et $D_S \subsetneq D_L$. Alors l'opérateur $\mathbf{G}$ a de bonnes troncatures,*

$$||\pi_n \circ \mathbf{G} - \mathbf{G}||_{D_S} \leq ||\mathbf{G}||_{D_S,D_L} \frac{R_L}{R_L - R_S} \left(\frac{R_S}{R_L}\right)^{n+1}.$$

Supposons de plus qu'il existe un très grand disque D_{XL} concentrique avec D_L vérifiant $D_S \subsetneq D_L \subsetneq D_{XL}$ et tel qu'un itéré k de $\mathbf{G}$ satisfait $\mathbf{G}(\mathcal{A}_\infty(D_S)) \subseteq \mathcal{A}_\infty(D_{XL})$. Alors pour tout vecteur propre ϕ de $\mathbf{G}$,

$$\frac{||\pi_n[\phi] - \phi||_{D_S}}{||\phi||_{D_S}} \leq \frac{||\phi||_{D_{XL}}}{||\phi||_{D_S}} \cdot \frac{R_{XL}}{R_{XL} - R_S} \left(\frac{R_S}{R_{XL}}\right)^{n+1}.$$

Preuve. Pour $f \in \mathcal{A}_\infty(D_S)$, le i^e coefficient a_i du développement de Taylor de $g := \mathbf{G}[f]$ en x_0 satisfait, avec la formule de Cauchy, l'inégalité $a_i R_L^i \leq ||g||_{D_L}$. Maintenant, la propriété $\mathcal{BFC}$ entraîne les inégalités suivantes,

$$\begin{aligned} ||\pi_n[g] - g||_{D_S} &\leq ||g||_{D_L} \sum_{i>n} a_i |z|^i \\ &\leq ||g||_{D_L} \sum_{i>n} \left(\frac{R_S}{R_L}\right)^i \\ &= ||g||_{D_L} \frac{R_L}{R_L - R_S} \left(\frac{R_S}{R_L}\right)^{n+1}. \end{aligned}$$

Le premier résultat vient directement de la définition de $||\mathbf{G}||_{D_S,D_L}$. Pour le second résultat, il faut remarquer que toute fonction propre ϕ appartient à $\mathcal{A}_\infty(D_{XL})$ puis appliquer la même démarche avec R_{XL} au lieu de R_L. ■

Les systèmes dynamiques des algorithmes d'Euclide classique, centré et impair satisfont la propriété $\mathcal{BFC}$. Les opérateurs de transfert associés admettent donc tous de bonnes troncatures et la méthode DFV s'applique alors à eux.

5.3.3 Les paramètres θ, K et n_0

Nous retournons maintenant à l'opérateur de transfert $\mathbf{G}_{s,\mathcal{A}}$ et nous considérons une valeur propre simple λ isolée du reste du spectre par un cercle $C = C(\lambda, r)$. La norme $||\mathbf{G}_{s,\mathcal{A}}||_{D_S,D_L}$ est majorée par $\delta(s, \mathcal{A}, D_L)$. Si la propriété $\mathcal{BFC}$ est vérifiée, l'entier n_0 est le plus petit entier n tel que

$$\delta(s, \mathcal{A}, D_L) \frac{R_L}{R_L - R_S} \left(\frac{R_S}{R_L}\right)^{n+1} \leq \beta_C(\mathbf{G}_{s,\mathcal{A}}).$$

Selon le lemme précédent, pour tout $n \geq n_0$, les opérateurs tronqués $\pi_n \circ \mathbf{G}_{s,\mathcal{A}}$ (et les matrices M_n associées) admettent une unique valeur propre λ_n à l'intérieur de C qui satisfait $|\lambda_n - \lambda| \leq K\theta^{n+1}$ avec

$$K = 2r \cdot \alpha_C(\mathbf{G}_{s,\mathcal{A}}) \cdot \delta(s, \mathcal{A}, D_L) \cdot \frac{R_S}{(R_L - R_S)}, \qquad \text{et} \qquad \theta = \frac{R_S}{R_L}.$$

Si la propriété $\mathcal{BTFC}$ est vérifiée, l'entier n_0 reste le même mais les constantes K et θ sont données par

$$K = 2r \cdot \alpha_C(\mathbf{G}_{s,\mathcal{A}}) \cdot \frac{||\phi||_{D_{XL}}}{||\phi||_{D_S}} \cdot \frac{R_S}{(R_L - R_S)}, \qquad \text{et} \qquad \theta = \frac{R_S}{R_{XL}}.$$

5.4 Calcul prouvé de constantes

Dans la section précédente, nous avons montré que la méthode DFV converge et que les valeurs propres isolées des opérateurs de transfert sont approchées exponentiellement rapidement par les valeurs propres des matrices tronquées M_n. Nous désirons maintenant connaître explicitement l'erreur commise afin de prouver un certains nombre de chiffres après la virgule. Pour cela, il est nécessaire de connaître les constantes n_0, K et θ. Ces constantes sont de natures très différentes. Le rapport de troncature θ dépend uniquement de la propriété de contraction forte des branches inverses et principalement des rayons des disques D_S, D_L et D_{XL}. Ces rayons et par suite θ, sont faciles à obtenir dès que les branches inverses sont connues.

Les constantes K et n_0 s'expriment en fonction de $\alpha_C(\mathbf{G})$ qui elle-même dépend du cercle $C = C(\lambda, r)$ isolant λ du reste du spectre. Pour trouver le cercle C, nous avons besoin d'une estimation de λ et d'une borne inférieure du saut spectral autour de λ. Nous allons prouver aux lemmes 10 et 11 que c'est possible dès lors que λ est la valeur propre dominante de l'opérateur de transfert contraint ou classique associé au système dynamique des fractions continues.

Une fois que $C = C(\lambda, r)$ est connu, le calcul de K et n_0 nécessite une borne supérieure de $\beta_C(\mathbf{G})$ et par conséquent de $\alpha_C(\mathbf{G})$. La constante $\alpha_C(\mathbf{G})$ fait intervenir la norme du quasi-inverse $(\mathbf{I} - z\mathbf{G})^{-1}$ et une borne supérieure de cette norme est en général difficile à calculer. Cependant, lorsque l'opérateur $\mathbf{G}$ est normal, il existe une expression exacte de $\alpha_C(\mathbf{G})$ donnée par

$$\alpha_C(\mathbf{G}) = \frac{1}{d(C, \mathbf{G})} \tag{5.7}$$

($\mathbf{G}$ est normal s'il commute avec son dual $\mathbf{G}^\star$). Mais être normal est une propriété rare qu'il est difficile de montrer. Dans le cas du système dynamique des fractions continues, les opérateurs de transfert $\mathbf{G}_{s,\mathcal{A}}$ ne sont pas normaux sur les espaces $\mathcal{A}_\infty(D)$. Cependant, il existe un autre espace fonctionnel, l'espace de Hardy $\mathcal{H}_{s,\mathcal{A}}$, qui dépend de s et de $\mathcal{A}$ où $\mathbf{G}_{s,\mathcal{A}}$ est normal. Il est à remarquer que la normalité ne semble pas être vérifiée pour les systèmes dynamiques des algorithmes d'Euclide centré et impair. Même si l'espace de Hardy et l'espace $\mathcal{A}_\infty(D)$ sont différents, les normes associées peuvent être comparées entraînant une borne supérieure pour $\alpha_C(\mathbf{G}_{s,\mathcal{A}})$. Finalement, nous en arrivons au second résultat important de cette partie.

Théorème 16 *(i) Le système dynamique des fractions continues satisfait la propriété de très forte contraction $\mathcal{BTFC}$.*

(ii) Pour tout sous-ensemble $\mathcal{A} \subset \mathcal{H}$ de branches inverses, les constantes $[K, n_0, \theta]$ sont calculables.

(iii)Dès que la fonction zeta de Hurwitz $\zeta_{\mathcal{A}}$ donnée par

$$\zeta_{\mathcal{A}}(s,x) = \sum_{h_{[m]} \in \mathcal{A}} \frac{1}{(m+x)^s}$$

est calculable en temps polynomial, il existe un algorithme polynomial qui calcule $\lambda_{\mathcal{A}}(s)$ en temps polynomial.

Nous rappelons ici qu'un algorithme est polynomial s'il utilise un nombre polynomial d'opérations arithmétiques. Le reste de la section contient la preuve du théorème.

5.4.1 Disques D_S, D_L et D_{XL} et rapport de troncature θ

Nous rappelons que nous nous limitons au cas de l'algorithme d'Euclide classique et du système dynamique des fractions continues. Nous considérons un sous-ensemble $\mathcal{A}$ de l'ensemble des branches inverses $\mathcal{H}$ et nous notons A l'ensemble des indices de $\mathcal{A}$. Si l'entier $m_{\mathcal{A}}$ est le minimum de A, les disques D_S et D_L peuvent être choisis de la manière suivante,

$$x_0 := \frac{1}{m_{\mathcal{A}}}, \qquad R_S := \frac{1}{m_{\mathcal{A}}}, \qquad R_L := \frac{1}{m_{\mathcal{A}}} + \frac{m_{\mathcal{A}}}{2}, \qquad \frac{R_S}{R_L} = \frac{2}{2+m_{\mathcal{A}}^2}.$$

De plus, l'opérateur $\mathbf{G}_{s,\mathcal{A}}$ transforme $\mathcal{A}_\infty(D_S)$ en un ensemble de fonctions analytiques sur le demi-plan $\{\Re(z) > -m_{\mathcal{A}}\}$. On peut alors choisir le disque D_{XL} tout disque de centre x_0 et de rayon

$$R_{XL} = \frac{1}{m_{\mathcal{A}}} + m_{\mathcal{A}} - \epsilon, \qquad \epsilon \in]0, m_{\mathcal{A}}/2[,$$

si bien que le rapport de troncature θ satisfait

$$\theta = \frac{R_S}{R_{XL}} = \frac{1}{1 + m_{\mathcal{A}}^2 - \epsilon m_{\mathcal{A}}}.$$

Nous venons de montrer que le système dynamique des fractions continues satisfait la propriété $\mathcal{BTFC}$.

5.4.2 Estimation de la valeur propre dominante $\lambda_{\mathcal{A}}(s)$ de $\mathbf{G}_{s,\mathcal{A}}$

Nous utilisons le résultat classique suivant qui a déjà été utilisé dans [DFV97] :

Soit $\mathbf{G}$ un opérateur qui agit sur un espace de fonctions analytiques sur un intervalle $[a,b]$. Supposons de plus que l'opérateur $\mathbf{G}$ est positif (i.e., $\mathbf{G}[f] > 0$ si $f > 0$) et qu'il admet une unique valeur propre dominante λ isolée du reste du spectre par un saut spectral. S'il existe deux constantes c_1 et c_2 et une fonction f analytique sur $[a,b]$, strictement positive et telle que $c_1 f \leq \mathbf{G}[f] \leq c_2 f$, alors la valeur propre dominante λ satisfait $c_1 \leq \lambda \leq c_2$.

Soit $m_{\mathcal{A}}$ le minimum de A et $M_{\mathcal{A}}$ son supremum (éventuellement infini). Par convention, si $M_{\mathcal{A}}$ est infini, nous posons $h_{M_{\mathcal{A}}} = 0$. Chaque branche inverse $h_{M_{\mathcal{A}}} \circ h_{m_{\mathcal{A}}}$ et $h_{m_{\mathcal{A}}} \circ h_{M_{\mathcal{A}}}$ admet comme unique point fixe respectif $a_{\mathcal{A}}$ et $b_{\mathcal{A}}$. Le disque $D_{\mathcal{A}}$ de rayon $[a_{\mathcal{A}}, b_{\mathcal{A}}]$ est le plus petit disque transformé en lui-même par toutes les branches de $\mathcal{A}$. L'application du résultat précédent avec l'opérateur $\mathbf{G}_{s,\mathcal{A}}$ et la fonction $f = 1$ pour la borne supérieure et $f = 1/(1+\beta x)^{2s}$ pour la borne inférieure, conduit à une estimation de $\lambda_{\mathcal{A}}(s)$ faisant intervenir la fonction zeta de Hurwitz $\zeta_{\mathcal{A}}$ restreinte à A,

$$\zeta_{\mathcal{A}}(s,x) := \sum_{m \in A} \frac{1}{(m+x)^s}.$$

Lemme 10 *Soit β le réel $\beta = (-m_{\mathcal{A}} + \sqrt{m_{\mathcal{A}}^2 + 4})/2$. La valeur propre dominante $\lambda_{\mathcal{A}}(s)$ admet l'encadrement suivant,*

$$\zeta_{\mathcal{A}}(2s, \beta) \leq \lambda_{\mathcal{A}}(s) \leq \zeta(2s, 0)$$

Pour obtenir le résultat précédent, nous avons utilisé l'intervalle $[a, b] = [0, 1]$. Comme la fonction $x \to (1 + \beta x)^{2s}\zeta_{\mathcal{A}}(2s, \beta + x)$ est croissante, les estimations précédentes peuvent être améliorées en utilisant le point fixe $a_{\mathcal{A}}$,

$$(1 + \beta a_{\mathcal{A}})^{2s}\zeta_{\mathcal{A}}(2s, \beta + a_{\mathcal{A}}) \leq \lambda_{\mathcal{A}}(s) \leq \zeta(2s, a_{\mathcal{A}}).$$

Preuves. Pour la borne supérieure, nous utilisons la fonctions constantes $\mathbf{1}$. Nous obtenons donc,

$$\frac{\mathbf{G}_{s,\mathcal{A}}[\mathbf{1}](x)}{\mathbf{1}(x)} = \mathbf{G}_{s,\mathcal{A}}[\mathbf{1}](x) = \sum_{m \in A} \frac{1}{(m + x)^{2s}} \leq \zeta(2s, a_{\mathcal{A}}).$$

Avec la fonction $\mathbf{1}$, nous obtenons comme borne inférieure $\zeta(2s, M_{\mathcal{A}}) \geq \zeta(2s, 1)$. Mais celle-ci n'est pas assez précise pour la suite. Nous choisissons plutôt la fonction $(1 + \beta x)^{-2s}$. Nous obtenons alors,

$$(1 + \beta x)^{2s}\mathbf{G}_{s,\mathcal{A}}[\frac{1}{(1 + \beta x)^{2s}}](x) = \sum_{m \in A} \left(\frac{1 + \beta x}{m + x + \beta}\right)^{2s}.$$

La dérivée de la fonction $(1 + \beta x)/(m + x + \beta)$ est du même signe que $\beta m + \beta^2 - 1$ et avec notre choix de β, ce signe est toujours positif. Ainsi, les fonctions $(1 + \beta x)/(m + x + \beta)$ ainsi que la fonction $(1 + \beta x)^{2s}\mathbf{G}_{s,\mathcal{A}}[(1 + \beta x)^{-2s}](x)$ sont croissantes et nous prouvons ainsi la borne inférieure. ■

5.4.3 Estimation du saut spectral

Dans cette seconde étape, nous déterminons une borne inférieure pour le saut spectral $\sigma_{\mathcal{A}}(s)$ entre la valeur propre dominante $\lambda_{\mathcal{A}}(s)$ et le reste du spectre. Pour cela, nous utilisons la trace des opérateurs de transfert. Grothendieck dans [Gro55] a introduit les opérateurs dits nucléaires (d'ordre 0) et a montré qu'ils possèdent une trace qui peut être vue comme une généralisation de la trace des matrices. Les opérateurs de transfert sont nucléaires (d'ordre 0) (voir par exemple [JGU04]) et leur trace est égale à la somme de toutes les valeurs propres. En particulier, $\mathrm{Tr}\mathbf{G}_{s,\mathcal{A}}^2$ est la somme de tous les carrés des valeurs propres de $\mathbf{G}_{s,\mathcal{A}}^2$. Comme l'opérateur est normal et même auto-adjoint pour s réel, les valeurs propres sont réelles et cela entraîne une relation entre $\mathrm{Tr}\mathbf{G}_{s,\mathcal{A}}^2$, la valeur propre dominante $\lambda_{\mathcal{A}}(s)$ et une de ses valeurs propres sous-dominante $\mu_{\mathcal{A}}(s)$,

$$\mu_{\mathcal{A}}^2(s) \leq \mathrm{Tr}\mathbf{G}_{s,\mathcal{A}}^2 - \lambda_{\mathcal{A}}^2(s).$$

L'opérateur $\mathbf{G}_{s,\mathcal{A}}^2$ est la somme d'opérateurs de la forme $\mathbf{L}[f] = |h'|^s \cdot f \circ h$ où h est une branche inverse de profondeur 2 de $\mathcal{A}^2$. Ces opérateurs sont connus sous le nom d'opérateurs de composition et leur spectre a été largement étudié par Shapiro dans [Sha93]. En appliquant ses résultats, si $h = h_i \circ h_j$ avec $(i, j) \in A^2$, le spectre de $\mathbf{L}$ est exactement une suite géométrique de la forme $\{\tau_{i,j}^{-2s-2n} : n \geq 0\}$ avec

$$\tau_{i,j} = \frac{1}{2}(ij + (i^2j^2 + 4ij)^{1/2} + 2).$$

Finalement, avec la propriété d'additivité de la trace, la trace de $\mathbf{G}_{s,\mathcal{A}}^2$ satisfait

$$\mathrm{Tr}\mathbf{G}_{s,\mathcal{A}}^2 = \sum_{i,j\in A} \frac{\tau_{i,j}^{-2s}}{1-\tau_{i,j}^{-2}}.$$

En écrivant $\mathrm{Tr}\mathbf{G}_{s,\mathcal{A}}^2 - 2\zeta_{\mathcal{A}}(2s,\beta)^2$ sous la forme suivante,

$$\mathrm{Tr}\mathbf{G}_{s,\mathcal{A}}^2 - 2\zeta_{\mathcal{A}}(2s,\beta)^2 = \sum_{i,j\in\mathcal{A}} \frac{\tau_{i,j}^{-2s}}{1-\tau_{i,j}^{-2}} - \frac{2}{(i+\beta)^{2s}(j+\beta)^{2s}},$$

nous montrons par de simples calculs que chaque terme est négatif, autrement dit que la relation suivante est vérifiée,

$$\mathrm{Tr}\mathbf{G}_{s,\mathcal{A}}^2 - \zeta_{\mathcal{A}}(2s,\beta)^2 < \zeta_{\mathcal{A}}(2s,\beta)^2.$$

La relation entre la trace, la valeur propre dominante et une des valeurs propres sous-dominante combinée avec l'encadrement de la valeur propre $\lambda_{\mathcal{A}}(s)$ entraîne le résultat suivant.

Lemme 11 *Le saut spectral pour les opérateurs de transfert est plus grand que* $2r_{\mathcal{A}}(s)$ *avec*

$$2r_{\mathcal{A}}(s) := \zeta_{\mathcal{A}}(2s,\beta) - \left(Tr\mathbf{G}_{s,\mathcal{A}}^2 - \zeta_{\mathcal{A}}(2s,\beta)^2\right)^{1/2}$$

et $\beta = (\frac{1}{2})(m_{\mathcal{A}} - (m_{\mathcal{A}}^2+4)^{1/2})$.

L'estimation du saut spectral peut être amélioré si l'encadrement amélioré de $\lambda_{\mathcal{A}}(s)$ est utilisé. Wirsing [Wir74] a montré que la constante γ_G satisfait $0.3020 \leq |\gamma_G| \leq 3043$. Comme la valeur propre dominante de $\mathbf{G}_1$ est 1, en utilisant la trace, nous obtenons $|\gamma_G| - |\mu| \geq 0.18959$ où μ est une des valeurs propres sous-sous-dominante de $\mathbf{G}_1$. Nous améliorons ainsi le précédent résultat connu pour le saut spectral autour de γ_G qui était $|\gamma_G| - |\mu| \geq 0.031$.

5.4.4 Normalité sur des espaces de Hardy

Nous avons déjà dit que la constante $\alpha_C(\mathbf{G}_{s,\mathcal{A}})$ a une forme close 5.7 dès que l'opérateur $\mathbf{G}_{s,\mathcal{A}}$ est normal. Les opérateurs de transfert ne sont pas normaux sur les espaces $\mathcal{A}_\infty(D)$ mais ils le sont sur un autre espace appelé espace de Hardy [JGU04] noté $\mathcal{H}_{s,\mathcal{A}}$. Pour $x \in \mathbb{R}$, P_x est le demi-plan $P_x = \{z : \Re(z) > x\}$. L'espace de Hardy $\mathcal{H}_{s,\mathcal{A}}$ est composé des fonctions analytiques sur $P_{-m_{\mathcal{A}}/2}$, bornées sur tous les demi-plans P_x avec $x > -m_{\mathcal{A}}/2$ et qui admettent la représentation intégrale suivante,

$$f(z) = \int_0^{+\infty} t^{s-\frac{1}{2}} e^{-tz} \phi(t) d\nu_{\mathcal{A}}(t),$$

avec

$$d\nu_{\mathcal{A}}(t) = \sum_{n\in\mathcal{A}} e^{-nt} dt \qquad \text{et} \qquad \phi \in L^2(\nu_{\mathcal{A}}).$$

Muni de la norme

$$||f||_{<s,\mathcal{A}>}^2 = \int_0^{+\infty} |\phi(t)|^2 d\nu_{\mathcal{A}}(t),$$

l'espace $\mathcal{H}_{s,\mathcal{A}}$ est un espace de Banach.

Il existe des relations étroites entre les espaces $\mathcal{H}_{s,\mathcal{A}}$ et $\mathcal{A}_\infty(D_S)$. Pour $\mathcal{A} = \mathcal{H}$, Babenko [Bab78] et Mayer [May91] ont prouvé que le comportement de $\mathbf{G}_s$ est comparable sur $\mathcal{H}_{s,\mathcal{A}}$ et sur $\mathcal{A}_\infty(D_s)$. Cependant, leurs méthodes ne se généralise pas facilement au cas où $\mathcal{A} \neq \mathcal{H}$. Nous donnons ici une méthode différente qui utilise les polynômes de Laguerre généralisés.

Lemme 12 *Pour tout complexe s tel que $\delta(s, \mathcal{A}, D_L)$ converge,*

(i) l'opérateur de transfert $\mathbf{G}_{s,\mathcal{A}} : \mathcal{H}_{s,\mathcal{A}} \to \mathcal{H}_{s,\mathcal{A}}$ est conjugué à un opérateur intégral ; il est normal et auto-adjoint pour des valeurs réels de s. En particulier, pour s réel, le spectre de $\mathbf{G}_{s,\mathcal{A}}$ est réel.

(ii) Le spectre de $\mathbf{G}_{s,\mathcal{A}}$ sur $\mathcal{H}_{s,\mathcal{A}}$ et sur $\mathcal{A}_\infty(D_S)$ sont les mêmes.

(iii) Soit D un disque intermédiaire de centre x_0 et de rayon R avec $R_S < R < R_L$ et f une fonction de $\mathcal{A}_\infty(D)$. Pour tout sous-ensemble $\mathcal{A} \subset \mathcal{H}$, la fonction $\mathbf{G}_{s,\mathcal{A}}[f]$ appartient à $\mathcal{H}_{s,\mathcal{A}}$.

(iv) Nous définissons pour tout R avec $R_S < R < R_L$, les trois constantes κ_1, κ_2 et κ_3 par

$$\kappa_1 = \zeta_{\mathcal{A}}(2s, x_0 - R), \qquad \kappa_2 = \Gamma(2s) \cdot \zeta_{\mathcal{A}}(2s, 2(x_0 - R)), \tag{5.8}$$

$$\kappa_3 = \sum_{j \geq 0} \left(\frac{R_S}{R}\right)^j \left(\frac{j!}{\Gamma(2s+j)} \frac{(\gamma_j R_S)^{2s}}{\gamma_j^{2s} - 1} + \frac{\zeta_{\mathcal{A}}(2s, 0)}{\Gamma(2s)^2} \right)^{1/2} \tag{5.9}$$

avec $\gamma_j = e^{x_0/(j+1)}$. Alors les inégalités suivantes sont vérifiées,

$$||\mathbf{G}_{s,\mathcal{A}}[f]||_D \leq \kappa_1 \cdot ||f||_{D_S}, \qquad \textit{pour } f \in \mathcal{A}_\infty(D) \tag{5.10}$$

$$||f||_D \leq \kappa_2 \cdot ||f||_{<s,\mathcal{A}>}, \qquad \textit{pour } f \in \mathcal{H}_{s,\mathcal{A}} \tag{5.11}$$

$$||\mathbf{G}_{s,\mathcal{A}}[f]||_{<s,\mathcal{A}>} \leq \kappa_3 ||f||_D, \qquad \textit{pour } f \in \mathcal{A}_\infty(D). \tag{5.12}$$

Avant de prouver le lemme, nous expliquons comment il conduit à une estimation de $\alpha_C(\mathbf{G}_{s,\mathcal{A}})$. Considérons le cercle C de centre $\lambda_{\mathcal{A}}(s)$ et de rayon $r_{\mathcal{A}}(s)$ donné au lemme 11 et fixons $z \in C$. Les deux inclusions

$$\mathbf{G}_{s,\mathcal{A}}[\mathcal{A}_\infty(D_S)] \subset \mathcal{A}_\infty(D), \quad \text{et} \quad \mathbf{G}_{s,\mathcal{A}}[\mathcal{A}_\infty(D)] \subset \mathcal{H}_{s,\mathcal{A}},$$

combinées avec la relation

$$z(\mathbf{G}_{s,\mathcal{A}} - z\mathbf{I})^{-1} = (\mathbf{G}_{s,\mathcal{A}} - z\mathbf{I})^{-1}\mathbf{G}_{s,\mathcal{A}} - \mathbf{I}$$

entraînent les deux inégalités

$$|z| \; ||(\mathbf{G}_{s,\mathcal{A}} - z\mathbf{I})^{-1}||_{D_S} \leq \kappa_1 \cdot ||(\mathbf{G}_{s,\mathcal{A}} - z\mathbf{I})^{-1}||_D + 1, \tag{5.13}$$

$$|z| \; ||(\mathbf{G}_{s,\mathcal{A}} - z\mathbf{I})^{-1}||_D \leq \kappa_2\kappa_3 1 \cdot ||(\mathbf{G}_{s,\mathcal{A}} - z\mathbf{I})^{-1}||_{<s,\mathcal{A}>} + 1. \tag{5.14}$$

Maintenant, l'opérateur $\mathbf{G}_{s,\mathcal{A}}$ est normal sur $\mathcal{H}_{s,\mathcal{A}}$ si bien que

$$||(\mathbf{G}_{s,\mathcal{A}} - z\mathbf{I})^{-1}||_{<s,\mathcal{A}>} = \frac{1}{d(z, \mathrm{Sp}\mathbf{G}_{s,\mathcal{A}})} = \frac{1}{r_{\mathcal{A}}(s)}.$$

Finalement, avec les formules 5.13 et 5.14, l'inégalité

$$||(\mathbf{G}_{s,\mathcal{A}} - z\mathbf{I})^{-1}||_{\mathcal{A}_\infty(D_S)} \leq \frac{1}{|z|^2}\left(\frac{\kappa_1 \cdot \kappa_2 \cdot \kappa_3}{r_{\mathcal{A}}(s)} + 1\right) + \frac{1}{|z|}$$

est vérifiée. Maintenant, l'estimation de $\lambda_{\mathcal{A}}(s)$ donnée au lemme 10 montre que tout z du cercle C satisfait

$$|z| \geq \zeta_{\mathcal{A}}(2s, b) - r_{\mathcal{A}}(s) > 0$$

et une majoration de $\alpha_C(\mathbf{G}_{s,\mathcal{A}})$ s'en déduit. Finalement, si nous résumons ces étapes nous obtenons le lemme suivant.

Lemme 13 *Si $r_{\mathcal{A}}(s)$ est le rayon défini au lemme 11, pour tout rayon intermédiaire R avec $R_S < R < R_L$ et tout s tel que $\delta(s, \mathcal{A}, D_L) < \infty$, il existe des constantes κ_i définies au lemme 12 pour lesquelles*

$$\alpha_C(\mathbf{G}_{s,\mathcal{A}}) \leq \frac{\kappa_1\kappa_2\kappa_3 + r_{\mathcal{A}}(s)[1 - r_{\mathcal{A}}(s) + \zeta_{\mathcal{A}}(2s, b_{\mathcal{A}})]}{r_{\mathcal{A}}(s)[\zeta_{\mathcal{A}}(2s, b_{\mathcal{A}}) - r_{\mathcal{A}}(s)]^2}.$$

Preuve du lemme 12. Les deux premiers points sont démontrés dans l'article [JGU04] de Jenkinson, Gonzalez et Urbański. L'inégalité 5.10 est une conséquence directe de la propriété forte de contraction.

Toute fonction f de $\mathcal{H}_{s,\mathcal{A}}$ admet une expression intégrale. En utilisant l'inégalité de Cauchy-Schwartz avec la relation

$$\Gamma(s)\zeta_{\mathcal{A}}(s, z) = \int_0^\infty t^{s-1}e^{-zt}d\nu_{\mathcal{A}}(t),$$

l'inégalité 5.11 s'obtient très facilement. La preuve de l'inégalité 5.12 est plus difficile. Tout d'abord, Hensley [Hen04] a montré que pour tout $j \geq 0$, la fonction $\mathbf{G}_{s,\mathcal{A}}$ est un élément de $\mathcal{H}_{s,\mathcal{A}}$ dont la représentation intégrale est étroitement liée aux polynômes de Laguerre généralisés $L_j^{(2s-2)}$. Les polynômes de Laguerre $(L_j^{(p)})$ forment une base orthogonale pour le poids $t^p e^{-t}$ sur $]0, \infty[$ et ils satisfont la formule

$$L_j^{(p)}(x) = \frac{\Gamma(p+1+j)}{j!}\sum_{k=0}^{j} j(-1)^k \binom{j}{k} \frac{x^k}{\Gamma(p+1+k)}.$$

La fonction $\mathbf{G}_{s,\mathcal{A}}[(X - x_0)^j]$ satisfait alors

$$\mathbf{G}_{s,\mathcal{A}}[(X - x_0)^j](z) = \int_0^\infty t^{s-1/2}e^{-tz}\left[\frac{(-x_0)^j j!}{\Gamma(2s+j)} t^{s-1/2} L_j^{(2s-1)}\left(\frac{t}{x_0}\right)\right] d\nu_{\mathcal{A}}(t).$$

Nous en déduisons que la norme est donnée par

$$||\mathbf{G}_{s,\mathcal{A}}[f]||^2_{<s,\mathcal{A}>} = \left[\frac{(-x_0)^j j!}{\Gamma(2s+j)}\right]^2 \int_0^\infty t^{2s-1} L_j^{(2s-1)}(t) d\nu_{\mathcal{A}}(x_0 t).$$

Mais les polynômes de Laguerre sont positifs et décroissants sur $[0, 2s/(j+1)]$. En utilisant cette propriété avec les relations d'orthogonalité et en coupant l'intégrale $\int_0^\infty$ en $\int_0^{2s/(j+1)} + \int_{2s/(j+1)}^\infty$, nous montrons que

$$\forall j \geq 1,\ ||\mathbf{G}_{s,\mathcal{A}}[f]||_{<s,\mathcal{A}>} \leq K_j \qquad \text{avec} \qquad \frac{K_{j+1}}{K_j} \to R_S.$$

Maintenant, le i^e coefficient c_i du développement de Taylor de $f \in \mathcal{A}_\infty(D)$ en x_0 satisfait $R^j|c_j| \leq ||f||_D$ et finalement,

$$||\mathbf{G}_{s,\mathcal{A}}[f]||_{<s,\mathcal{A}>} \leq \kappa_3 \cdot ||f||_D \qquad \text{avec} \qquad \kappa_3 = \sum_{j\geq 0} \frac{K_j}{R^j}.$$

Il est à noter que la série précédente converge exponentiellement rapidement. ■

Finalement, les constantes K et n_0 font intervenir le cercle C de centre $\lambda_{\mathcal{A}}(s)$ et de rayon $r_{\mathcal{A}}(s)$ ainsi que la constante $\alpha_C(\mathbf{G}_{s,\mathcal{A}})$. Comme toutes ces constantes sont calculables, nous avons prouvé le théorème 16.

5.4.5 Calculabilité de la matrice

Nous n'avons jusqu'à pas présent montré que la matrice M_n est calculable en temps polynomial.

Lemme 14 *Pour $\mathcal{A} \subset \mathcal{H}$, si les fonctions zeta $\zeta_{\mathcal{A}}$ de Hurwitz sont calculables en temps polynomial, alors la matrice tronquée M_n est calculable en temps polynomial.*

Preuve. Nous allons donner la formule exacte des coefficients qui sont tous une somme finie faisant intervenir les fonctions $\zeta_{\mathcal{A}}$. Nous avons,

$$\mathbf{G}_{s,\mathcal{A}}[(X-x_0)^j](x) = \sum_{i\in A} \frac{1}{(m+x)^{2s}} (\frac{1}{m+x} - x_0)^j.$$

En développant selon la formule du binôme, une seconde somme apparaît et en échangeant les deux signes sommes, nous obtenons

$$\mathbf{G}_{s,\mathcal{A}}[(X-x_0)^j](x) = \sum_{u=0}^{j} \binom{j}{u} (-x_0)^{j-u} \zeta_{\mathcal{A}}(2s+u,x).$$

Maintenant, le développement de Taylor de la fonction $\zeta_{\mathcal{A}}(2s+u,x)$ au point x_0 satisfait

$$\zeta_{\mathcal{A}}(2s+u,x) = \sum_{i=0}^{\infty} (-1)^i \binom{2s+u+i-1}{i} \zeta_{\mathcal{A}}(2s+u+i,x_0)(x-x_0)^i.$$

Le coefficient $M_{n,i,j}$ de la matrice M_n est alors donné par

$$M_{n,i,j} = [(x-x_0)^i]\mathbf{G}_{s,\mathcal{A}}[(X-x_0)^j](x) = (-1)^i \sum_{u=0}^{j} \binom{j}{u} \binom{2s+u+i-1}{i} (-x_0)^{j-u} \zeta_{\mathcal{A}}(2s+u+i,x_0).$$

La dernière égalité montre que les coefficients de la matrice sont calculables en temps polynomial dès que les fonctions $\zeta_{\mathcal{A}}$ le sont. ■

Pour tous les ensembles classiques $\mathcal{A} \subset \mathcal{H}$ ou de manière équivalente $A \subset \mathbb{N}^\star$, les fonctions $\zeta_{\mathcal{A}}$ sont calculables en temps polynomial. Nous pouvons par exemple considérer les ensembles A finis, $A = \mathbb{N}^\star$ ou encore A l'ensemble des nombres pairs, etc. Dans la preuve, nous n'avons considérer que les opérateurs de transfert non pondérés. Pour les opérateurs de transfert pondérés, l'expression reste exactement la même sauf que la fonction zeta de Hurwitz est remplacée par la fonction zeta $\zeta_{\mathcal{A},[c]}$ pondérée

$$\zeta_{\mathcal{A},[c]}(s,x) = \sum_{m\in A} \frac{e^{wc(m)}}{(m+x)^s}.$$

5.5 Algorithmes polynomiaux

Cette section applique les résultats précédents et donne des valeurs numériques prouvées pour les constantes de Gauss-Kuz'min-Wirsing et Hensley et les dimensions de Hausdorff des ensembles de Cantor $C_{\mathcal{A}}$ une valeur *non prouvée* pour la constante $\rho(\ell)$ qui intervient dans le théorème BV.

Pour tout sous-ensemble A de $\mathbb{N}^\star$, construire la matrice M_n nécessite le calcul de $2n+1$ fonctions zeta et $O(n^3)$ opérations arithmétiques. Ensuite, le calcul des valeurs propres avec une méthode naïve utilise $O(n^4)$ opérations arithmétiques. Maintenant, pour une précision de d chiffres après la virgule, la taille n est linéaire en d ($n = -(d\log 10+\log K)/\log\theta$) soit la méthode DFV calcule avec une précision de d chiffres en utilisant $O(d^4)$ opérations arithmétiques et $O(d)$ calculs de fonctions zeta.

Théorème 17 (Algorithmes polynomiaux) *Il existe un algorithme qui calcule en temps polynomial la constante de Hensley et la constante de Gauss-Kuz'min-Wirsing. Si $A \subset \mathbb{N}^\star$ est tel que la fonction ζ_A soit calculable en temps polynomial, alors il existe un algorithme polynomial qui calcule la dimension de Hausdorff de l'espace de Cantor C_A.*

5.5.1 Algorithme pour la constante de Gauss-Kuz'min-Wirsing

La constante de Gauss-Kuz'min-Wirsing γ_G est l'unique valeur propre sous-dominante de $\mathbf{G}_1$. Elle est réelle et négative. Wirsing a donné un encadrement prouvé de γ_G et nous avons vu qu'il était possible de minorer le saut spectral autour de γ_G. Le cercle C et la constante $\alpha_C(\mathbf{G}_1)$ sont alors calculables et la méthode DFV s'applique. Pour estimer γ_G, une seule matrice est à calculer et les fonctions zeta sont les fonctions zeta classiques. Nous avons implémenté un algorithme qui étant donné le nombre de chiffres exacts attendus retourne une estimation de γ_G. Nous avons résumé les valeurs obtenues dans le tableau suivant.

chiffres	temps	valeur obtenue
10	11s	-0.3036630028
20	1m46	-0.30366300289873265859
30	9m54	-0.303663002898732658597448121901
40	34m	-0.3036630028987326585974481219015562331108
50	1h41	-0.303663002898732658597448121901556233110877352253 65

5.5.2 Algorithme pour la constante de Hensley et $\rho_{[\delta]}$

La constante de Hensley (voir 5.2) ainsi que la constante $\rho_{[\delta]}$ (voir Théorème 5) s'expriment avec les deux premières dérivées de $\lambda(s)$ en $s = 1$. La première dérivée $\lambda'(1)$ admet une formule close $\lambda'(1) = -\pi^2/(6\log 2)$. Il reste la deuxième dérivée $\lambda''(1)$ à calculer. Considérons un intervalle I_h de la forme $I_h := [1-h, 1+h]$ et supposons qu'il existe des estimations $\widetilde{\lambda}$ de λ aux points $1 \pm h$ vérifiant

$$\min(|\lambda(1+h) - \widetilde{\lambda}(1+h)|, |\lambda(1-h) - \widetilde{\lambda}(1-h)|) \leq \frac{h^2\epsilon}{3}.$$

La formule de Taylor entraîne la majoration suivante,

$$\left|\lambda''(1) - \frac{\widetilde{\lambda}(1+h) + \widetilde{\lambda}(1-h) - 2}{h^2}\right| \leq \frac{2\epsilon}{3} + \frac{h^2}{24}\sup_{I_h}|\lambda^{(4)}|.$$

Il est alors suffisant de connaître une majoration de la quatrième dérivée sur l'intervalle I_h. L'application $s \to \mathbf{G}_s$ est analytique et sa dérivée $\mathbf{G}'_s$ satisfait $||\mathbf{G}'_s||_{D_S} \leq 8$ pour $s \geq 0.9$. Alors, $||\mathbf{G}_s - \mathbf{G}_1||_{D_S} \leq 8|s-1|$. Nous appliquons le lemme 10 : le cercle C de centre 1 et de rayon $r_1 = (1-\gamma_G)/2$ est un cercle qui isole $\lambda(1) = 1$. Maintenant, si s satisfait $|s-1| < r_2$ avec $r_2 = \beta_C(\mathbf{G}_1)/8$, l'opérateur $\mathbf{G}_s$ admet une unique valeur propre dominante $\lambda(s)$ dans C qui satisfait $|\lambda(s) - 1| \leq r$ avec $r := 16 r_1 r_2 \alpha_C(\mathbf{G}_1)$. Comme l'application $s \to \mathbf{G}_s$ est analytique, la

fonction $s \to \lambda(s)$ est aussi analytique. La formule de Cauchy appliquée au cercle de centre 1 et de rayon r_1 conduit à la majoration de la quatrième dérivée,

$$\sup_{I_h} |\lambda^{(4)}| \leq 4! \frac{1+r}{(r_1-h)^4}.$$

On en déduit que la dérivée seconde $\lambda''(1)$ et par suite que γ_H et $\rho_{[\delta]}$ sont calculables dès que deux estimations de $\lambda(1 \pm h)$ sont connues, ce qui est le cas avec la méthode DFV. Il faut remarquer que les deux estimations nécessitent (asymptotiquement) deux fois plus de précision que la précision finale attendue ainsi que deux calculs de matrices d'autant plus grande. Le tout reste toutefois polynomial mais le temps de calcul est considérablement augmenté. Le tableau suivant résume les résultats numériques pour la constante de Hensley.

chiffres	temps	valeur obtenue pour γ_H
5	2m30s	0.51606
10	7m30s	0.5160624088
15	41m	0.516062408899991
20	2h33	0.51606240889999180681

En utilisant la dérivée seconde déjà calculée pour la constante de Hensley, nous obtenons la valeur suivante pour la constante $\rho_{[\delta]}$,

$$\rho_{[\delta]} = \delta(1-\delta)\frac{\lambda''(1)-\lambda'(1)^2}{|\lambda'(1)|} \approx 1.4531 \cdot \delta(1-\delta)$$

5.5.3 Algorithme pour les dimensions de Hausdorff

La dimension de Hausdorff de l'espace de Cantor C_A est l'unique réel s_A tel que $\lambda_A(s_A) = 1$. L'algorithme utilise un principe de dichotomie et calcule une suite d'intervalles de longueur 2^{-k} qui contient la dimension de Hausdorff s_A. Considérons l'intervalle $[u_{k-1}, v_{k-1}]$ obtenus après $(k-1)$ étapes. Nous posons w_k le milieu de $[u_{k-1}, v_{k-1}]$ et considérons $\widetilde{\lambda}$ une estimation à $2^{-(k+1)}$ près de $\lambda(w_k)$ obtenue avec la méthode DFV. Il y a trois cas possibles :

(i) si $\widetilde{\lambda} - 1 - 2^{-(k+1)} \geq 0$ alors $s_A \geq w_k$ et $[u_k, v_k] = [w_k, v_{k-1}]$.

(ii) si $\widetilde{\lambda} - 1 - 2^{-(k+1)} \leq 0$ alors $s_A \leq w_k$ et $[u_k, v_k] = [u_{k-1}, w_k]$.

(iii) sinon $[u_k, v_k] = [w_k - 2^{-(k+1)}, w_k + 2^{-(k+1)}]$.

La preuve que s_A appartient à l'intervalle $[u_k, v_k]$ est basé sur la stricte décroissance de la fonction $s \to \lambda(s)$ avec l'inégalité $|\lambda_A(s+h) - \lambda_A(s)| \geq h$ (voir [Hen96]). Pour une précision de d chiffres, un nombre linéaire en d de dichotomies sont nécessaires et chaque dichotomie utilise un nombre polynomial d'opérations arithmétiques soit l'algorithme est polynomial. Nous avons résumé les valeurs numériques obtenues pour l'ensemble $A = \{1, 2\}$ dans le tableau suivant.

chiffres	temps	valeur obtenue pour $C_{\{1,2\}}$
5	2min	0.53128
10	8min	0.5312805062
15	25min	0.531280506277205
20	1h	0.53128050627720514162
30	4h26	0.531280506277205141624468647368
40	14h11	0.5312805062772051416244686473684717854930
45	23h10	0.531280506277205141624468647368471785493059109

Bien entendu, l'algorithme s'applique avec d'autres ensembles A possiblement infinis. De tels calculs ont été effectués dans [Val98b]. Dans le même article, la méthode DFV a aussi été utilisée pour des contraintes périodiques. En suivant le même schéma de preuve, nous montrons que la méthode DFV converge également dans ce contexte mais nous ne savons alors pas calculer les constantes K et n_0.

5.5.4 Calcul de $\rho(\ell)$

La constante $\rho(\ell)$ intervient dans le terme dominant de la variance de la complexité binaire étendue (cf. théorème 3). Elle s'exprime en fonction des cinq premières dérivées de $\lambda(s,w)$ où $\lambda(s,w)$ est la valeur propre dominante de l'opérateur de transfert pondéré $\mathbf{G}_{s,w,[\ell]}$. Nous ne savons pas si cet opérateur est normal sur un espace de Hardy et nous ne pouvons donc pas calculer les constantes n_0 et K. Toutefois, nous savons que la méthode DFV converge et en appliquant les mêmes idées que pour la constante de Hensley, nous pouvons estimer les cinq premières dérivées de $\lambda(s,w)$ par rapport à s et w. Cela nous a conduit à l'estimation suivante

$$\rho(\ell) \approx 0.09152$$

5.6 Conclusion

Nous avons prouvé que la méthode DFV est une méthode *efficace* pour calculer les valeurs propres des opérateurs de transfert à conditions que ceux-ci admettent de bonnes troncatures et que les matrices tronquées soit facilement calculables.
Cependant, si l'on est intéressé par des valeurs numériques prouvées, il est nécessaire de connaître les constantes K et n_0 qui interviennent dans la description de la vitesse de convergence. Ces paramètres sont en général difficiles à calculer mais nous avons résolu ce problème pour les opérateurs de transfert non pondérés associés au système dynamique des fractions continues.

La méthode DFV peut aussi être utilisée pour calculer la dimension de Hausdorff d'ensembles de Cantor à contraintes simples ou périodiques. Dans le premier cas, nous obtenons un algorithme polynomial alors que dans le second cas, nous avons uniquement la convergence de la méthode DFV.

Finalement, les auteurs de [DFV97] et Sebah ont utilisé la méthode DFV avec le point $x_0 = 1/2$. Ce choix particulier n'entre pas dans notre cadre puisque aucun disque de centre $1/2$ est strictement envoyé dans lui-même. Nous pouvons utiliser tout disque D_L de centre $1/2+\delta$ et de rayon $1/2+2\delta$ avec $\delta \leq 1/2$. Le rapport de troncature R_S/R_{XL} tend alors vers $1/3-\epsilon$ lorsque δ tend vers 0 ce qui est la vitesse de convergence observée par les auteurs de [DFV97].

Conclusion de la Partie A

Résumé des résultats. Dans cette partie, nous obtenons une description probabiliste fine des paramètres fondamentaux de l'algorithme d'Euclide classique : nous exhibons CINQ lois–limite gaussiennes, sur les SIX lois escomptées, pour les paramètres suivants : complexité binaire étendue (polynômes et entiers), complexité binaire standard (polynômes seulement), taille du reste à une fraction (fixe) de l'exécution (polynômes et entiers). Seule manque à l'appel la complexité binaire standard de l'algorithme d'Euclide sur les entiers. Dans ce dernier cas, même si nous ne sommes pas parvenus au résultat final escompté, nous obtenons un résultat qui précise le comportement de la variance. Nous avons également exhibé un phénomène de régularité de l'algorithme classique, qui a son intérêt propre mais qui permet aussi d'analyser en moyenne la complexité binaire de l'algorithme de Knuth-Schönhage. Au cours des preuves de ces résultats, nous avons aussi rencontré d'autres paramètres que nous avons également analysés : coûts additifs à croissance modérée ou intermédiaire, coût du type longueur de cheminement.

Dans le cadre des polynômes, nous avons utilisé les méthodes classiques de combinatoire analytique, qui s'appliquent bien et conduisent assez rapidement aux résultats. L'analyse dans le cas des entiers est fondée, elle, sur les méthodes d'analyse dynamique. Le domaine est moins bien balisé, et l'analyse dynamique en distribution en est encore à ses tous débuts. Tout en reprenant la méthodologie générale définie par Baladi et Vallée, nous avons été conduits à l'étendre et à la généraliser, afin de pouvoir obtenir, pour des perturbations du quasi-inverse [appelés pseudo-quasi-inverses] les mêmes résultats que ce qui déjà était connu pour le vrai quasi-inverse. L'analyse du pseudo-quasi-inverse a permis de comprendre le rôle du paramètre δ qui représente la fraction de l'exécution à laquelle on s'intéresse, et de montrer que les deux cas – δ rationnel ou δ irrationnel– sont sensiblement différents. L'analyse du pseudo-quasi-inverse a permis aussi de montrer que l'algorithme est bien régulier, quand la longueur des phases n'est pas trop petite –supérieure à $\sqrt{n}$ pour des données de taille n–.

Cette partie se termine avec le calcul de constantes spectrales, liées aux objets spectraux des opérateurs de transfert du système dynamique euclidien. Nous prouvons les propriétés d'un algorithme dû à Daudé Flajolet et Vallée, et nous montrons que trois constantes – constante de Gauss-Kuz'min-Wirsing, la constante de Hensley, dimensions de Hausdorff d'ensembles de Cantor– sont calculables en temps polynomial.

Perspectives. Il reste finalement peu à faire sur l'analyse de l'algorithme classique étendu. Par contre, même si nous sommes convaincus que l'algorithme classique standard a, sur les entiers, un comportement limite gaussien, nous n'avons pas obtenu de preuve de ce fait. Notre forte conviction se fonde sur l'obtention de ce résultat dans les cas des polynômes, et sur l'heuristique générale suivante : "Ce qui est vrai dans $\mathbb{F}_q[X]$ l'est sur $\mathbb{Z}$". Mais il y a une deuxième heuristique qui annonce "Mais c'est beaucoup plus difficile à démontrer". Nous sommes donc incertains de la difficulté de ce résultat. Une piste pour obtenir cette loi limite gaussienne consisterait à utiliser les résultats récents que Eda Cesaratto [Ces06] a obtenus sur les propiétés spectrales des opérateurs

de transfert qui agissent sur des fonctions de deux variables.

En ce qui concerne l'analyse des algorithmes rapides, très liée aux plénomènes de régularité de l'algorithme d'Euclide, nous avons échoué à obtenir l'analyse en moyenne de ces algorithmes, quand la multiplication est trop rapide. C'est très lié à la régularité de l'algorithme, que nous ne savons pas prouver quand les phases sont de longueur trop petite, inférieure à $\sqrt{n}$. Mais est-ce une limite réelle ? Le phénomène n'est-il plus vrai pour des phases trop courtes ? ou est-ce seulement notre méthode qui échoue ?

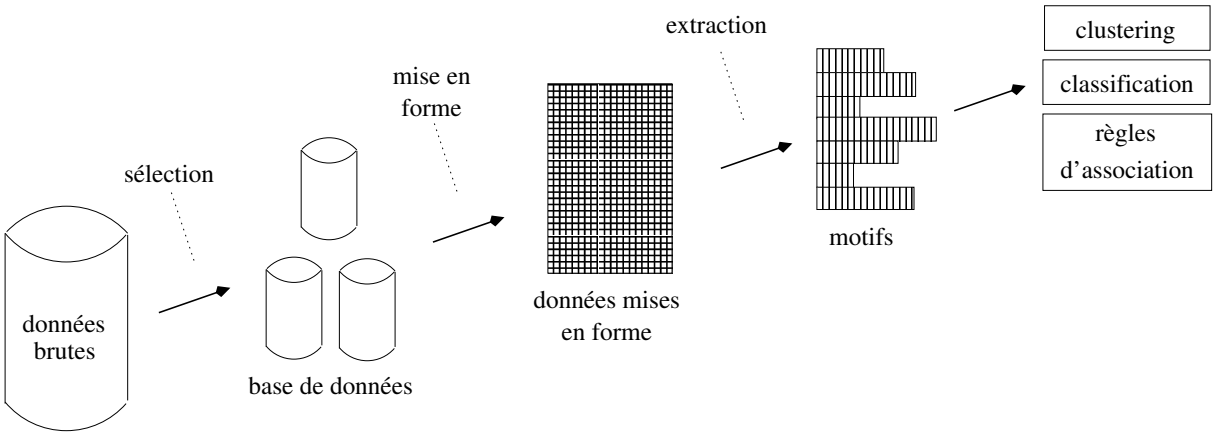
Il y a aussi beaucoup d'autres algorithmes de type Euclide, comme l'algorithme binaire, l'algorithme Plus-Moins, l'algorithme à bits de poids faible, tous les algorithmes qui sont appelés lents, etc... Pour tous ces algorithmes, SAUF l'algorithme Plus-Moins, le comportement moyen des complexités binaires est connu. Pour Plus-Moins, on ne sait rien. Pour les autres algorithmes, la distribution de cette complexité moyenne n'est pas connue, car la démarche générale de Baladi et Vallée échoue dans ces cas. On doit s'attendre d'ailleurs, dans le cas des algorithmes lents, à des lois limite non gaussiennes. Ce sont des questions très ouvertes, et sans doute difficiles.

Partie B :
Nombre moyen de motifs fréquents et fermés dans une base de données

Introduction de la Partie B

La croissance des volumes d'informations disponibles, avec leur côté multiforme, créent de nouveaux enjeux pour les systèmes d'information, et tout particulièrement pour les bases de données. L'internet, avec le Web et le commerce électronique, a été l'un des facteurs principaux qui a déclenché la globalisation de l'information, de ses sources et de ses usages. Mais, de fait, toutes les sciences [notamment, la physique, la biologie, la médecine, l'environnement mais également l'ingénierie, les sciences humaines, etc] sont des domaines fortement consommateurs et producteurs d'informations. Plus largement, c'est la société toute entière, dans les domaines privés et publics, économiques, culturels, qui produit et cherche de l'information. La croissance explosive des données produites appelle de nouveaux processus de traitement, avec des capacités d'analyse sophistiquées et puissantes, pour produire des connaissances à partir de ces données.

Le domaine de la fouille de données. L'Extraction de Connaissances dans les Bases de Données (ECBD), également appelée Fouille de Données, est une des composantes du traitement des masses de données, qui s'attache à extraire des informations dans une base de données. La fouille de données se compose généralement de deux grandes phases.



La phase de pré-traitement transforme les données brutes en données pré-traitées, qui admettent une mise en forme avec un format bien défini (tables relationnelles, tableaux multidimensionnels, ...). Le prétraitement effectue aussi d'autres opérations : les données en double sont effacées, les erreurs corrigées, les valeurs manquantes comblées [FPSS96, Dyr97], etc. Chacune de ces opérations fait l'objet de nombreux travaux, non abordés ici. Une fois le prétraitement achevé, l'extraction de (nouvelles) connaissances peut commencer, et c'est l'objet de la deuxième phase, celle qui nous intéresse ici.

La deuxième phase est composée, elle aussi, de plusieurs étapes : segmentation ou clustering, classification, recherche de règles d'associations. La segmentation regroupe entre eux les objets qui se ressemblent. La pertinence des groupes ainsi formés, appelés clusters ou classes, est ensuite

validée par les experts. Puis, lors de la classification, on attribue une classe à chaque objet. Enfin, il faut trouver des règles d'association dans les bases. Ces règles, qui sont du type $pain, jambon \Rightarrow beurre$ (80%) [cela signifie que 80% des personnes qui achètent du pain et du jambon achètent aussi du beurre], présentent un grand intérêt, et sont centrales dans les sciences : bio-informatique, médecine, chimie, etc.

Pour trouver des règles d'associations dans une base de données, on doit rechercher des corrélations entre les éléments de la base. Cette recherche de corrélations est liée à la découverte de motifs dits fréquents. Définissons cette notion quand la base de données est une matrice binaire où les lignes (en nombre n) représentent les objets observés (personnes, chromosomes, fleurs, protéines,...) alors que les colonnes (en nombre m) représentent les attributs (grand, petit, jaune, présent, ...). Dans ce cas, un motif est un ensemble d'attributs, et un motif est dit fréquent, s'il apparaît plusieurs fois dans la base. La présence d'un motif fréquent indique que beaucoup d'objets partagent le même ensemble d'attributs, ce qui indique une forte corrélation entre les attributs qui composent le motif. Un motif est dit γ-fréquent s'il est partagé par au moins γ-objets, et l'entier γ est appelé le seuil de fréquence. Ce seuil est généralement défini par l'utilisateur. Le nombre de motifs γ–fréquents est potentiellement grand, puisque, dans le pire des cas, et pour toute valeur du seuil γ, ce nombre est exponentiel en le nombre total m d'attributs de la base. Pourtant, il apparaît expérimentalement, qu'avec des seuils raisonnables, le nombre de motifs fréquents explicitement présents est en général raisonnable, si bien que les algorithmes qui recherchent ces motifs fonctionnent bien. En revanche, quand le seuil γ est trop petit, le nombre de motifs γ-fréquents explose, et les algorithmes ne sont plus efficaces.

Le cadre de l'étude et les résultats. C'est ainsi une situation typique où le pire des cas n'est pas du tout représentatif de la situation réelle, et où l'analyse en moyenne trouve sa place naturelle. Dans cette thèse, nous cherchons à expliciter la manière dont le nombre moyen de motifs γ-fréquents évolue en fonction du seuil γ considéré. Nous avons aussi cherché à comparer (toujours en moyenne) le nombre de motifs fréquents et le nombre de motifs fermés [représentation condensée des motifs fréquents].

Toute analyse en moyenne débute par une étape de modélisation des structures de données. Ici, il faut trouver un "bon modèle" probabiliste pour les bases de données. Pour une telle structure de données, utilisée dans des contextes si divers, le modèle choisi est forcément réducteur, et assez peu réaliste. Nous avons cherché à ce qu'il ne soit pas trop simpliste, qu'il prenne en compte des corrélations potentielles [qui existent, puisque tout le but est de les trouver !], tout en restant accessible à des techniques fines d'analyse. Notre culture et notre environnement scientifique nous a ainsi incités à choisir un modèle de bases de données, qui fait explicitement référence au contexte de la théorie de l'information. Nous utilisons le concept de source, qui est, par définition, un mécanisme qui produit des symboles. Une base de données "informationnelle" , avec m lignes-objets, et m colonnes–attributs, est alors créée par une unique source qui produit, de manière indépendante, n mots–lignes–objets, formés de la succession des m valeurs binaires de ses attributs.

C'est notre modèle général, où, bien sûr, l'hypothèse d'indépendance entre les lignes peut paraître très restrictive. Dans la fin de cette partie, nous étudierons plus particulièrement le cas d'une base de données "dynamique", où la source émettrice est une source dynamique, source créée par un système dynamique,

Dans le modèle général, la recherche de motifs est un problème de Pattern–Matching généralisé. L'analyse du Pattern Matching classique vise à déterminer le nombre moyen de motifs d'un type donné apparaissant dans une seule séquence : les motifs peuvent être très généraux, formés de symboles contigüs ou non, et il existe déjà des méthodes d'analyse dans ce domaine, quand la

source est dynamique. Nous voulons ici évaluer le nombre moyen de motifs communs à plusieurs séquences, et nous devons ainsi généraliser la notion et l'étude de la coincidence, qui est limitée à l'étude de préfixes communs à plusieurs mots produits par la même source. De même, l'étude des motifs généralisés (non contigus) a aussi déjà été faite, mais dans le cas d'un seul mot.

Ici, le paramètre de référence est le seuil γ et nous obtenons trois résultats correspondant à trois types de seuils de fréquence. Nous considérons d'abord une base de données "informationnelle" générale (Chapitre 7), et nous exhibons, selon le seuil étudié, des hypothèses sur la source, sous lesquelles nous pouvons obtenir nos résultats. Nous montrons ensuite dans le Chapitre 8 que ces hypothèses sont vérifiées par certaines sources dynamiques, ayant de bonnes propriétés. Le premier type de seuil étudié est un seuil linéaire en le nombre d'objets n (hypothèse 6). Si la source émettrice vérifie l'hypothèse de décroissance exponentielle des probabilités (hypothèse 8), nous montrons que le nombre de motifs γ-fréquents est polynomial en le nombre m d'attributs (théorème 18). C'est un résultat intéressant puisqu'il explique pourquoi les algorithmes fonctionnent bien en pratique pour des seuils raisonnables alors que le pire des cas indique toujours un comportement exponentiel.

Le second type de seuil étudié est un seuil dit logarithmique, défini par une fonction faiblement croissante du nombre d'objets (hypothèse 7). Avec une hypothèse plus forte de décroissance exponentielle des probabilités (hypothèse 9), nous prouvons que le nombre moyen de motifs γ-fréquents est équivalent au nombre moyen de motifs γ-fermés (théorème 19). Ce résultat est déjà connu pour des bases de données faiblement corrélées du type panier de la ménagère. Il n'est plus vrai avec des bases fortement corrélées.

Le dernier type de seuil abordé est le seuil fixe (hypothèse 5), qui correspond en pratique à un seuil petit vis-à-vis du nombre n d'objets, qui, lui, a tendance à être grand. Sous une hypothèse plus fine sur le nombre moyen de motifs présents dans exactement γ objets (hypothèse 10), nous montrons que le nombre moyen de motifs fréquents est exponentiel en le nombre n d'objets (n est aussi le nombre de mots) et polynomial en le nombre m d'attributs (m est aussi la longueur des mots) (théorème 20).

Organisation de la Partie B. Cette partie s'organise en trois chapitres.

Chapitre 6. Ce chapitre introduit toutes les notions sur la fouille de données mais aussi les objectifs à plus long terme que nous souhaitons atteindre. Ce qui y est présenté dépasse donc très largement ce qui est analysé dans la suite....

Chapitre 7. Ce chapitre décrit le modèle de base de données informationnelle (i.e. construites à partir d'une source) adopté, présente les hypothèses sur la source émettrice et explicite les cas des seuils étudiés. Les résultats sont énoncés et prouvés dans ce modèle général de source, satisfaisant aux hypothèses demandées. Nous terminons en montrant des exemples de sources émettrices qui vérifient les hypothèses envisagées.

Chapitre 8. Ce chapitre introduit le modèle de base de *données dynamique*[6], qui est donc une base de données informationnelle, où la source émettrice est une *source dynamique*. Après quelques rappels sur le modèle des sources dynamiques, [qui contient à la fois des sources simples comme les sources sans mémoire, des sources plus évoluées comme les chaînes de Markov, mais aussi des sources complexes à mémoire non bornée], nous montrons que les sources dynamiques vérifient les conditions 8, 9 et 10 des théorèmes 18 et 20. Nous avons ainsi explicité un modèle de bases de données, assez vaste, où les résultats du précédent chapitre sont valides.

[6]Ici, "dynamique" ne signifie pas que les bases évoluent avec le temps. Au contraire, une fois générées à partir d'une source, elles restent inchangées

Chapitre 6

Fouille de données et motifs

Sommaire

6.1 Introduction

Nous avons décrit au chapitre précédent le cadre très général de la fouille de données. Notre étude porte toutefois sur un cadre bien précis. Il existe différents types de données : les données constituées d'attributs continus comme la taille, la distance entre deux points, l'heure, ..., les données constituées d'attributs discrets comme les réponses à des questionnaires, la couleur, ..., et les données qui évoluent avec le temps comme les flots de données [GGR02], les bases relationnelles qui mettent en correspondance plusieurs bases, ... Nous nous plaçons dans le cadre de données discrètes ou de manière équivalente de données binaires. Une base de données binaire est un tableau de 0 et de 1. Les colonnes représentent les attributs et les lignes sont les objets observés. Chaque objet est décrit à l'aide d'attributs pouvant être soit présents (un 1 dans le tableau), soit absents (un 0). Il existe un grand nombre de méthodes dont nous ne discuterons pas ici pour l'obtention de bases binaires à partir d'attributs continus [SA96, ZRRF99, GB02]. La Figure 6.1 forme un exemple d'une base binaire avec 7 attributs et 8 objets.

On définit un motif comme un ensemble d'attributs. Un motif qui apparaît un nombre significatif de fois est d'un grand intérêt puisqu'il indique une forte corrélation entre les attributs qui le compose. L'extraction de ces motifs dits *fréquents* est une première étape possible pour la création de règles d'association [AS94], mais aussi pour la classification (recherche de motifs qui caractérisent une classe) et le clustering. Mais les motifs fréquents sont souvent trop nombreux et des représentations condensées comme les motifs fermés ou libres (voir ci-dessous), ou des motifs contraints [BAG99, SC05] sont préférées.

Depuis une dizaine d'années, les chercheurs proposent des algorithmes pour la recherche de motifs. Aujourd'hui, l'extraction de motifs dispose d'algorithmes de référence comme APRIORI [AS94], ECLAT [ZPOL97, Zak00] ou FP-GROWTH [HPY00] accompagnés de leurs nombreuses améliorations. Il est très difficile de comparer ces algorithmes tout d'abord parce que tous les chercheurs ne mettent pas le code source à disposition. Ensuite, deux implémentations d'un même algorithme avec des structures de données différentes peuvent avoir des temps d'exécution très variables. Finalement, le temps de calcul est très dépendant des propriétés des bases. Compte tenu de tous ces éléments, certaines affirmations sur les temps d'exécution se trouvent parfois contredites par la suite [Goe03].

Afin de comparer les implémentations, des sites comme le FIMI[7] regroupent les codes sources ainsi que des bases de données tests. Tout utilisateur est ainsi capable d'évaluer les performances, d'analyser l'impact des bases sur le temps d'exécution et de proposer des améliorations ou des cadres d'applications des algorithmes. Cette démarche met en avant l'expérimentation. Nous proposons une nouvelle approche basée cette fois-ci sur l'analyse. Les bases réelles sont remplacées par des modèles de bases qui mettent en avant les propriétés à observer. Ensuite, le comportement des algorithmes vis-à-vis de ces bases est analysé (en moyenne) à travers certains paramètres comme la complexité en temps ou en espace. Finalement, soit les résultats théoriques confirment les hypothèses déduites de l'expérimentation, soit ils apportent un nouvel éclairage qui guidera de nouvelles expérimentations ou hypothèses.

Le point de vue théorique possède de nombreux avantages. Tout d'abord, on peut tenir compte d'un très grand nombre de bases de données dans les analyses alors que les expérimentations ne peuvent se limiter qu'aux quelques bases à disposition. Ensuite, il est possible de privilégier certaines propriétés dans la modélisation des données et d'en étudier les effets sur les algorithmes. En pratique, le nombre de bases satisfaisant une propriété donnée est faible et on est en droit de se demander si les exécutions d'un algorithme sur quelques bases particulières sont suffisamment représentatives pour en déduire un comportement globale. Finalement, les résultats théoriques peuvent être utilisés pour régler certains paramètres des algorithmes, pour valider ou invalider des hypothèses, pour optimiser certains algorithmes, etc.

Toutefois, la théorie a ses limites. Les modèles ne peuvent appréhender toute la complexité du monde réel. De plus avec des modèles trop complexes, les analyses deviennent difficiles. Finalement, l'analyse en moyenne fournit des résultats asymptotiques qui sont vérifiables uniquement avec de très grandes bases.

Notre positionnement est ambitieux car il vise à expliquer de manière théorique des phénomènes réels avec des analyses en moyenne et cela en Fouille de Données. L'objectif à long terme est de calculer la complexité en temps et en espace des algorithmes classiques comme APRIORI, ECLAT ou FP-GROWTH et de déterminer leur meilleur cadre d'application. De plus, nous désirons comprendre quelle est la réelle complexité de l'extraction de motifs. Par exemple, il est connu que le nombre de motifs fréquents est au pire exponentiel en le nombre d'attributs mais en pratique et en réglant certains paramètres, les algorithmes sont très efficaces. La complexité réelle

[7] adresse : http ://fimi.cs.helsinki.fi/

semble donc différente de la complexité au pire. Notre approche vise donc à mieux comprendre ces comportements.

Objectifs de ce chapitre et plan. Ce chapitre a un rôle particulier dans cette thèse. Une thèse rassemble des résultats en les situant dans leur contexte général et celle-ci n'échappe pas à la règle (du moins, je l'espère). Mais dans le prochain chapitre, nous appliquerons pour la première fois les techniques d'analyse en moyenne à la Fouille de Données. Nous pourrions nous limiter au strict cadre de notre analyse mais comme nous l'avons précisé dans la précédente partie, nos objectifs sont à plus long terme. C'est pourquoi nous faisons une description plus large que nos besoins afin de mettre en évidence différents axes de recherches.

Plan. Dans la première section, nous présenterons le cadre de Mannila et Toivonen pour l'extraction de motifs. Dans la seconde section, nous introduirons les principes généraux de trois algorithmes qui servent de référence pour l'extraction des motifs. Finalement, dans une troisième partie et avant de conclure, nous présenterons quelques axes de recherche.

6.2 Cadre de Mannila et Toivonen

Dans [MT97], Mannila et Toivonen introduisent un cadre formel pour l'extraction de motifs sous contrainte anti-monotone dans les bases de données mais de nos jours, il existe de nombreuses approches pour des contraintes plus générales [DJLM02, BAG99, DJLM02, SC05]. Nous ne traiterons pas ces contraintes générales car la prochaine partie se focalise uniquement sur la contrainte anti-monotone la plus utilisée : la contrainte de fréquence.

Dans cette section, nous commencerons par définir précisément les objets formels que nous manipulerons : les bases, les motifs, les contraintes, etc. Ensuite, nous aborderons la représentation sous forme de treillis des motifs. Finalement, nous introduirons la notion de bordure qui est un point clé dans la complexité des algorithmes.

6.2.1 Base de données binaire et représentations

Tout d'abord, définissons formellement ce qu'est une base binaire.

Définition 10 (Base de données binaire) *Une base de données binaire $\mathcal{B}$ est un triplet $(\mathcal{A}, \mathcal{O}, f_{\mathcal{B}})$ où $\mathcal{A} = \{a_1, \ldots, a_m\}$ est l'ensemble d'attributs, $\mathcal{O} = \{o_1, \ldots, o_n\}$ est l'ensemble des objets et $f_{\mathcal{B}}$ est une application de $\mathcal{A} \times \mathcal{O}$ dans $\{0, 1\}$.*
On dira que l'objet o contient l'attribut a si $f_{\mathcal{B}}(a, o) = 1$.

La table 6.1 présente un exemple de base de données binaire comportant 7 attributs et 8 objets. Une croix est mise entre un objet o et un attribut a dès que l'on a $f_{\mathcal{B}}(a, o) = 1$. En pratique, les bases sont codées avec le modèle transactionnel c'est à dire que chaque ligne représente un objet et énumère le numéro des attributs contenus dans l'objet correspondant. La figure 6.2 montre le codage transactionnel pour l'exemple de la table 6.1. Il existe d'autres codages comme le codage vertical où chaque colonne représente un attribut et énumère tous les objets le contenant. Le codage matriciel est souvent utilisé pour mettre en mémoire la base (lorsque c'est possible). Les lignes (resp. colonnes) représentent les objets (resp. attributs) et le coefficient (i, j) vaut 1 dès que o_i et a_j satisfont $f_{\mathcal{B}}(a_j, o_i) = 1$. Il est aussi possible de représenter une base de données binaire sous la forme d'un graphe bipartite dont l'ensemble des sommets est $\mathcal{A} \cup \mathcal{O}$ et il existe une arête entre a et o si $f_{\mathcal{B}}(a, o) = 1$. Finalement, en prenant le complémentaire du graphe bipartite, on obtient un graphe co-bipartite composé de deux cliques et d'arêtes entre ces deux cliques. Les sommets des deux cliques contiennent respectivement les attributs et les

	attributs						
objets	a_1	a_2	a_3	a_4	a_5	a_6	a_7
o_1	×		×		×		
o_2		×	×		×		
o_3	×		×		×		
o_4	×			×		×	
o_5		×	×			×	
o_6		×	×			×	
o_7	×			×			×
o_8		×		×			×

FIG. 6.1 – Exemple d'une base de données binaire.

objets et il existe une arête entre a et o si l'on a $f_{\mathcal{B}}(a,0)=0$. Les représentations par graphes ont par exemple été utilisées par les auteurs de [LLSW05, Sig02, BS04]. Finalement, il est possible de définir deux hypergraphes $\mathcal{H}_{\mathcal{B}}$ et $\mathcal{H}_{\overline{\mathcal{B}}}$ pour représenter la base $\mathcal{B}$. Les sommets sont les attributs et les hyper-arêtes de $\mathcal{H}_{\mathcal{B}}$ (resp. $\mathcal{H}_{\overline{\mathcal{B}}}$) sont les objets (resp. les complémentaires des objets).

Toutes ces représentations sont équivalentes. Le modèle transactionnel n'est autre que la représentation par liste de successeurs des sommets correspondant aux objets du graphe bipartite. De manière similaire, le codage vertical est la représentation par liste de successeurs des attributs alors que le codage matriciel est la matrice d'adjacence du graphe bipartite. Le modèle transactionnel est aussi le codage naturel de l'hypergraphe $\mathcal{H}_{\mathcal{B}}$.

6.2.2 Motifs et contraintes anti-monotone

Notre problématique est de trouver des associations d'attributs qui ont un intérêt. Une association d'attributs n'est autre qu'un motif et l'intérêt d'un attribut sera représenté par une contrainte que l'attribut vérifiera ou non.

Définition 11 (Motif) Un motif *(d'attributs) est un sous ensemble de* $\mathcal{A}$.

Il existe aussi des motifs d'objets mais nous ne les utiliserons pas. Pour alléger l'écriture, un motif sera noté sous la forme d'une chaîne plutôt que sous forme ensembliste (i.e. a_1a_2 au lieu de $\{a_1,a_2\}$). De plus, un objet $o\in\mathcal{O}$ pourra également être considéré comme un motif : $o=\{a\in\mathcal{A}\mid f_{\mathcal{B}}(a,o)=1\}$. Nous abordons maintenant la notion de support et de fréquence.

Définition 12 *Un objet* $o\in\mathcal{O}$ *supporte un motif* $X\subset\mathcal{A}$ *(ou* X *est présent dans* o*) si* $X\subset o$ *(ou* $\forall a\in X, f_{\mathcal{B}}(a,o)=1$*).*
Le support *d'un motif* X *dans la base* $\mathcal{B}$*, noté* $Supp_{\mathcal{B}}(X)$ *est l'ensemble des objets qui supportent* X*.*
La fréquence de X *dans la base* $\mathcal{B}$*, notée* $Freq_{\mathcal{B}}(X)$*, est le cardinal du support de* X*.*

Dans l'exemple de la table 6.1, le motif a_1a_5 admet comme support $Supp_{\mathcal{B}}(a_1a_5)=\{o_1,o_3\}$ et comme fréquence 2 alors que le motif a_1 a comme support $Supp_{\mathcal{B}}(a_1)=\{o_1,o_3,o_4,o_7\}$ et une fréquence 4.

Les premiers motifs auxquels les chercheurs se sont intéressés sont les motifs qui apparaissent fréquemment dans la base de données. Ces motifs indiquent qu'une corrélation forte existe entre les attributs qui les composent.

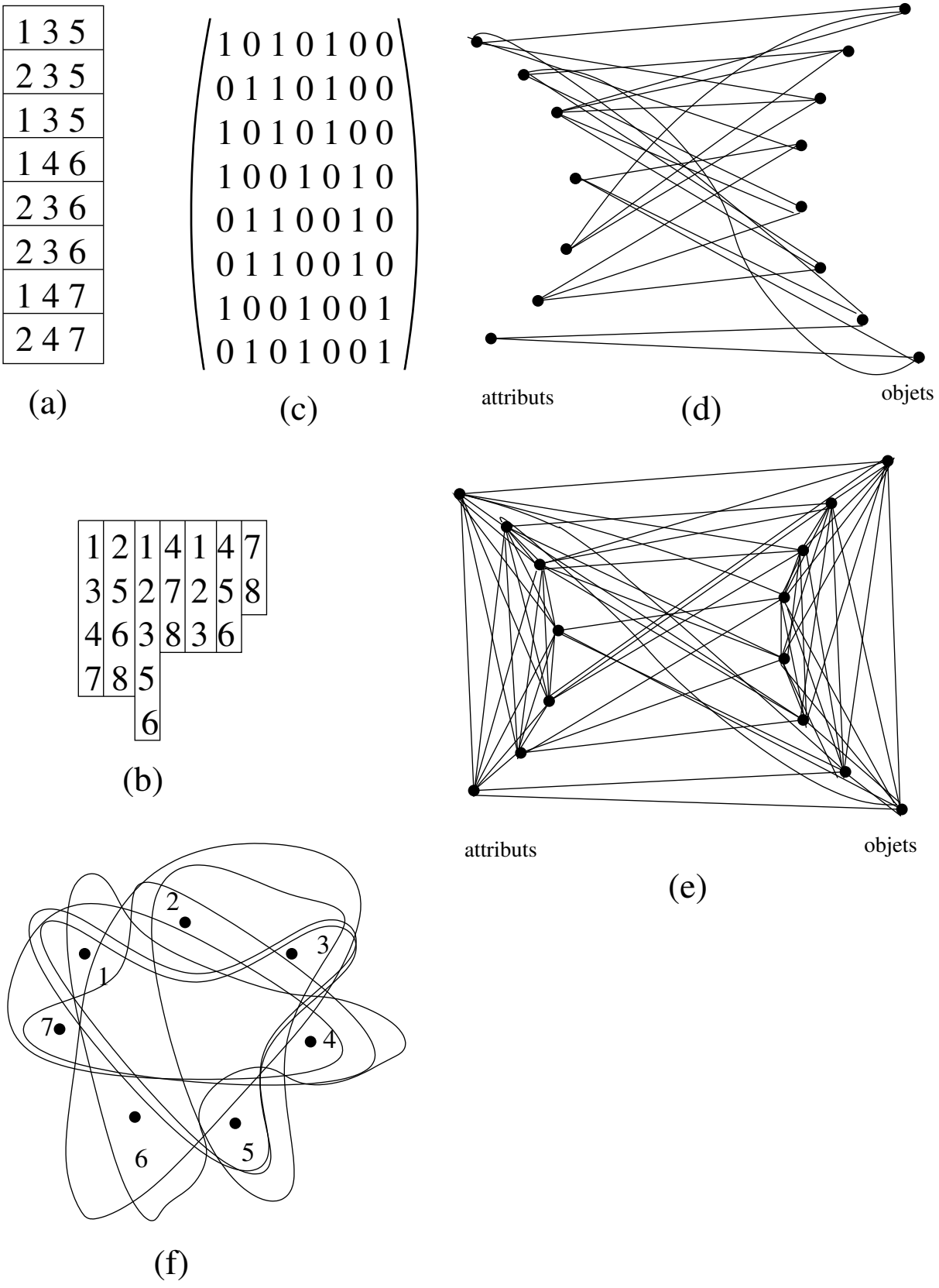


FIG. 6.2 – Codages de l'exemple sous forme (a) transactionnel (b) verticale (c) matricielle (d) de graphe bipartite (e) de graphe co-bipartite (f) hypergraphe $\mathcal{H}_{\mathcal{B}}$.

Définition 13 (Motif fréquent) *Pour un entier positif γ, un motif X est dit γ-fréquent (dans la base $\mathcal{B}$) si sa fréquence est plus grande que γ,*

$$Freq_{\mathcal{B}}(X) \geq \gamma.$$

Dans l'exemple de la table 6.1, le motif a_1a_5 est 1 ou 2-fréquent mais pas 3-fréquent. L'extraction de motifs fréquents est à la base de la recherche de règles d'association [AS94], mais les motifs fréquents ne forment qu'un cas particulier de motif plus généraux satisfaisant une contrainte anti-monotone.

Définition 14 (Contrainte) *Une* contrainte binaire *q est une application qui à une base $\mathcal{B}$ et un motif X renvoie une valeur dans $\{0,1\}$. On dira qu'un motif X de la base $\mathcal{B}$ est* valide *(resp.* non valide*) sous la contrainte q si*

$$q(X,\mathcal{B}) = 1 \qquad (resp.\ q(X,\mathcal{B}) = 0).$$

On dira aussi que X satisfait la contrainte q (resp. ne satisfait pas la contrainte).

La contrainte de γ-fréquence se définit de la manière suivante :

$$q_\gamma(X,\mathcal{B}) = \begin{cases} 1 & \text{si } Freq_{\mathcal{B}}(X) \geq \gamma \\ 0 & \text{sinon} \end{cases}$$

L'ensemble des motifs γ-fréquents dans $\mathcal{B}$ est l'ensemble des motifs qui satisfont la contrainte de γ-fréquence dans la base $\mathcal{B}$. De manière similaire, un motif X qui satisfait la contrainte

$$\widetilde{q}_\ell(X,\mathcal{B}) = \begin{cases} 1 & \text{si } |X| \geq \ell \\ 0 & \text{sinon} \end{cases}$$

pour un entier $\ell \geq 1$ est un motif qui comporte au moins ℓ attributs. On parle de la contrainte de longueur (sur les motifs).

La contrainte de longueur vérifie une propriété de monotonie à savoir que si X est un motif de longueur au moins ℓ, alors tout sur-motif de X (motif contenant X) est aussi de longueur au moins ℓ. Il existe d'autres contraintes satisfaisant cette propriété que l'on regroupe sous le terme de contrainte monotone.

Définition 15 (contrainte monotone) *Une contrainte q est dite* monotone *si pour un motif X satisfaisant la contrainte, tout sur-motif de X satisfait aussi la contrainte,*

$$\forall X, Y \subset \mathcal{A},\ X \subset Y \Rightarrow (q(X,\mathcal{B}) = 1 \Rightarrow q(Y,\mathcal{B}) = 1).$$

Par opposition, les contraintes de fréquences satisfont une propriété d'anti-monotonie. En effet, si X est γ-fréquent, tout les sous-motifs de X (motifs inclus dans X) sont aussi γ-fréquents.

Définition 16 (contrainte anti-monotone) *Une contrainte q est dite* anti-monotone *si pour un motif X satisfaisant la contrainte, tout sous-motif de X satisfait aussi la contrainte,*

$$\forall X, Y \subset \mathcal{A},\ Y \subset X \Rightarrow (q(X,\mathcal{B}) = 1 \Rightarrow q(Y,\mathcal{B}) = 1).$$

Il existe une multitude de contraintes monotones ou anti-monotones portant par exemple sur le coût total d'un motif lorsque les attributs sont valués, l'inclusion de motifs dans d'autres ou qui sont des conjonctions de contraintes monotones et anti-monotones [DJLM02].

Dorénavant et jusqu'à la fin de ce chapitre, on se fixe une contrainte q anti-monotone sur une base de données binaire $\mathcal{B} = (\mathcal{A}, \mathcal{O}, f_{\mathcal{B}})$.

6.2.3 Treillis des motifs contraints

L'ensemble des motifs s'organisent naturellement dans un treillis d'inclusion (cf Figure 6.3). La première ligne contient le motif vide, la deuxième contient les singletons, la troisième les couples et ainsi de suite jusqu'au motif comportant tous les attributs. Le treillis s'élargit rapidement jusqu'aux motifs comportant la moitié des attributs et rétrécit ensuite. C'est la raison pour laquelle les treillis sont souvent représentés par des losanges.

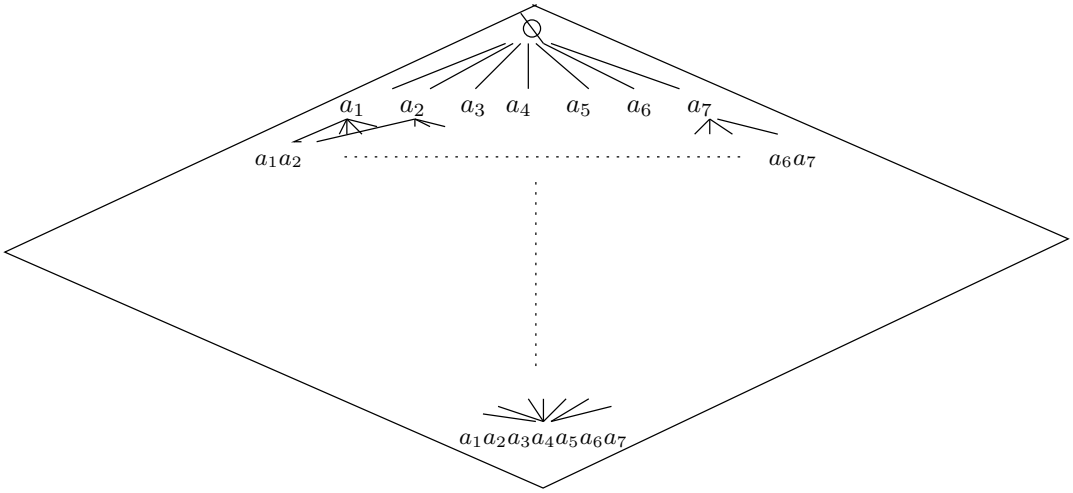


FIG. 6.3 – Treillis d'inclusion des motifs.

Les motifs satisfaisant une contrainte anti-monotone se représentent aussi sous la forme d'un sous-treillis du treillis des motifs (cf Figure 6.4). Le treillis des motifs contraints admet alors une forme similaire. Au premier niveau on trouve le motif vide, au second niveau les singletons satisfaisant la contrainte et au niveau $k+1$ les motifs de longueur k satisfaisant la contrainte. Par habitude, on ajoute aussi la fréquence de chaque motif.

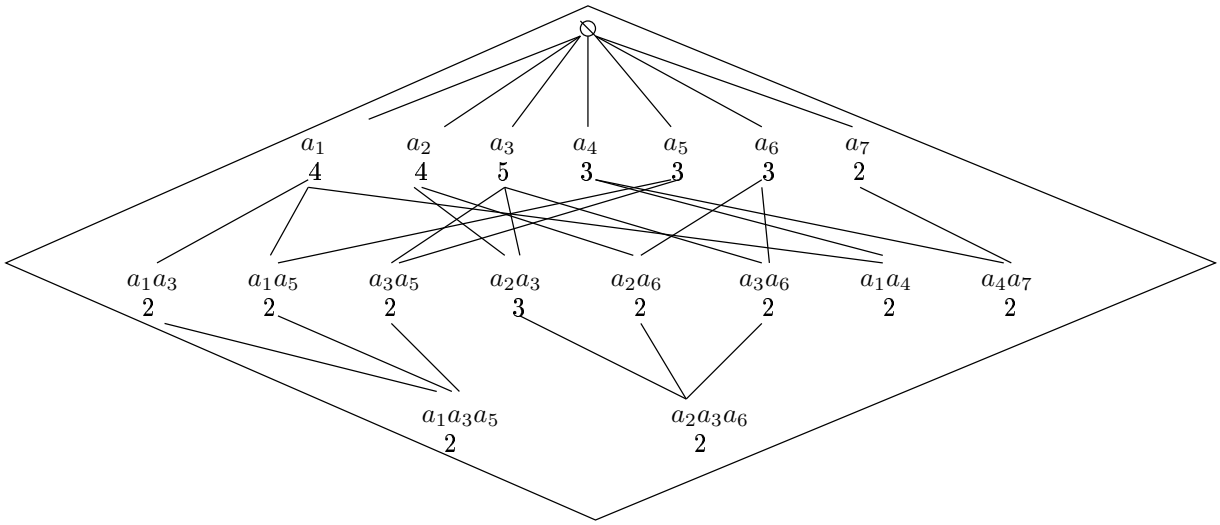


FIG. 6.4 – Treillis d'inclusion des motifs 2-fréquents de l'exemple avec les fréquences.

Toute la difficulté est d'énumérer tous les motifs qui satisfont une contrainte sans parcourir complètement le treillis des motifs. En effet, le treillis des motifs contient 2^m éléments (pour m attributs) et il est préférable d'avoir une stratégie qui fait gagner du temps. Nous verrons à la section 6.3 que l'anti-monotonie se révèle importante.

6.2.4 Représentations condensées

Parmi l'ensemble des motifs contraints, il existe des sous ensembles de motifs qui permettent de retrouver les autres motifs avec leur fréquence. On parle alors de représentation condensée. Parmi ces motifs, il y a les motifs libres et les motifs fermés qui se définissent à partir des classes d'équivalences.

Définition 17 (Classe d'équivalence) *Deux motifs contraints X et Y sont équivalents si et seulement si ils ont le même support. On écrira alors $X \sim Y$. La relation $\sim$ est une relation d'équivalence et partitionne de fait l'ensemble des motifs contraints en classes d'équivalences.*

La figure 6.5 représente l'ensemble des classes d'équivalences sur le treillis des motifs 2-fréquents correspondant à la table 6.1.

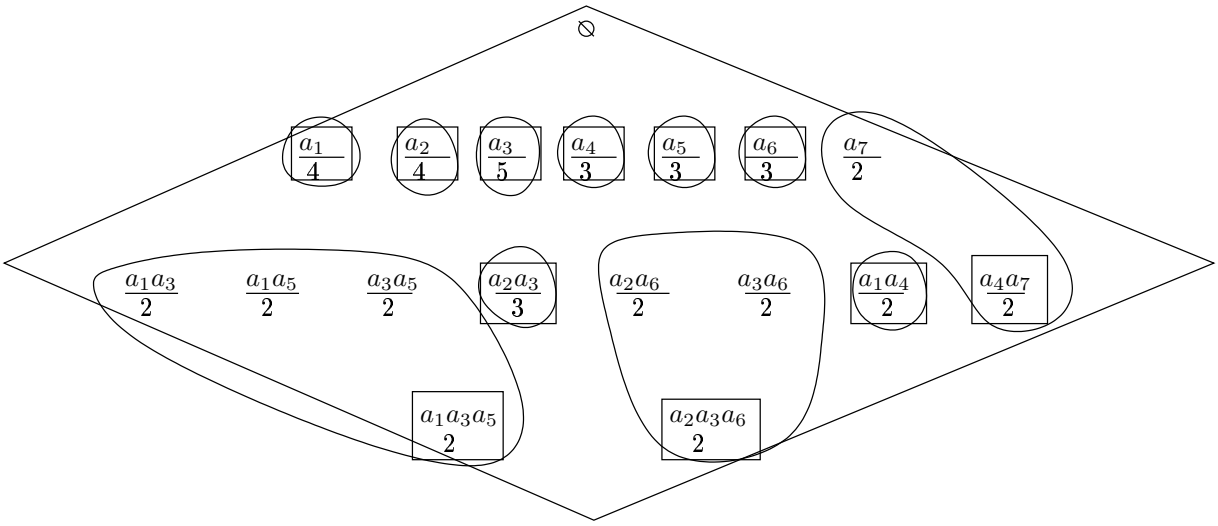


FIG. 6.5 – Classes d'équivalences des motifs 2-fréquents. Motifs fermés et motifs libres.

Pasquier et al. ont proposé dans [PBTL99a, PBTL99b] une alternative à l'extraction de tous les motifs contraints. L'idée est d'extraire des motifs contraints dits fermés (avec leur fréquence) qui constituent un système de générateurs minimal grâce auquel il est possible de retrouver tous les motifs contraints avec leur fréquence.

Définition 18 *Chaque classe d'équivalence admet un unique motif de plus grande taille. Cet unique motif est dit fermé.*

Intuitivement, un motif fermé est un motif qui n'admet pas de sur-motifs contraints ayant la même fréquence. Plus qu'une représentation condensée, les motifs fermés admettent pour la contrainte de 1-fréquence une propriété de transposition qui est très utilisée pour des bases avec un grand nombre d'attributs et peu d'objets [JR05]. Les motifs fermés sont aussi très liés aux

séparateurs minimaux dans le graphe co-bipartite codant la base de données. Cette relation a permis aux auteurs de [Sig02, BS04] d'importer des méthodes de graphes en Fouille de Données. Nous avons représenté les motifs fermés encadrés d'un rectangle dans la figure 6.5.

Nous venons de voir que les motifs fermés forment une représentation condensée optimale des motifs contraints. Nous présentons maintenant une autre représentation condensée, très utile mais non optimale, basée sur les motifs libres.

Définition 19 (Motif libre) *Les motifs libres sont les éléments minimaux au sens de l'inclusion des classes d'équivalences.*

Intuitivement, un motif libre ne possède pas de sous motif ayant le même support. Les motifs libres furent introduits par Bastide et al. [BTP+00, BTP+02] sous la notion de motifs clés et par Boulicaut et al. dans [BBR00]. Ces derniers utilisèrent les motifs libres pour générer des règles d'association de confiance 100% et à prémisses minimales.
Les motifs libres sont soulignés dans la figure 6.5.

6.2.5 Bordures

Les bordures (positive et négative) sont deux autres représentations condensées de tous les motifs valides. A l'inverse des motifs libres ou fermés, il n'est cependant pas possible de recalculer les fréquences des motifs. Les bordures sont les limites entre les motifs qui satisfont une contrainte (bordure positive) et ceux qui ne la satisfont pas (bordure négative).

Définition 20 (Bordure positive) *La bordure positive est l'ensemble des motifs maximaux au sens de l'inclusion satisfaisant la contrainte.*

De manière équivalente, tout sur-motif d'un motif de la bordure positive ne satisfait pas la contrainte. La bordure positive associé à la table 6.1 et à la contrainte de 2-fréquence est donnée par $\{a_1a_3a_5, a_2a_3a_6, a_1a_4, a_4a_7\}$.

Définition 21 (Bordure négative) *La bordure négative est l'ensemble des motifs minimaux au sens de l'inclusion qui ne satisfont pas la contrainte.*

Par opposition à la bordure positive, tout sous-motif d'un motif de la bordure négative satisfait la contrainte. Pour la contrainte de 2-fréquence, la bordure négative de la table 6.1 est constituée de tous les motifs non-fréquents à deux attributs (13 éléments).

La notion de bordure négative est étroitement liée à la génération de candidats pour les algorithmes par niveaux [MT97]. En particulier, Mannila et Toivonen ont montré que ces algorithmes testent la contrainte au moins autant de fois qu'il y a de motifs dans les bordures positive et négative. Finalement, le calcul des bordures négatives est étroitement lié au calcul de traverses minimales dans des hypergraphes [MR86, DT95].

6.3 Algorithmes de recherche de motifs

Depuis une dizaine d'années, la fouille de données dispose d'un grand nombre d'algorithmes pour extraire des motifs sous contrainte anti-monotone. Nous avons choisi de n'en présenter que deux : Apriori et Eclat qui sont les algorithmes fondateurs de deux approches possibles pour l'extraction de motifs. Toutefois nous aborderons succinctement d'autres algorithmes.

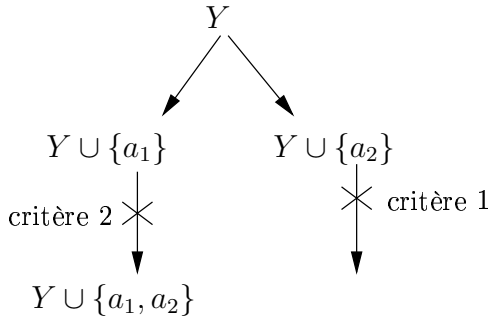


FIG. 6.6 – Critères d'élagage.

Cette section commence par exposer les propriétés fondamentales qui sont à la base des deux algorithmes. Ensuite, l'algorithme APRIORI qui parcourt le treillis des motifs en largeur est présenté. La troisième sous-section se concentre sur l'algorithme en profondeur ECLAT. Finalement, nous présentons succinctement d'autres techniques pour l'extraction de motifs.

6.3.1 Propriétés fondamentales des algorithmes

Une méthode simple pour extraire tous les motifs est de parcourir le treillis des motifs au complet et de ne garder que les motifs contraints. Cette méthode est à proscrire puisqu'elle nécessite le parcours de 2^m motifs s'il y a m attributs. Il faut donc pouvoir élaguer le treillis des motifs très rapidement.

Les contraintes anti-monotones offrent deux critères d'élagage pour optimiser la recherche.

Critère 2 *Si un motif X ne satisfait pas une contrainte anti-monotone, tous les sur-motifs de X ne la vérifient pas non plus.*

Critère 3 *Si un motif X satisfait une contrainte anti-monotone, tous les sous-motifs de X vérifient aussi la contrainte.*

La figure 6.6 montre comment les deux critères d'élagage peuvent intervenir. On se place dans la situation ou Y est un motif valide dont le sur-motif $Y \cup \{a_1\}$ est aussi valide mais pas le sur-motif $Y \cup \{a_2\}$. L'algorithme peut tout d'abord considérer le motif $Y \cup \{a_1, a_2\}$ comme un sur-motif de $Y \cup \{a_1\}$ mais après vérification du critère 2, il s'aperçoit que $Y \cup \{a_2\}$ est un sous-motif non valide du motif $Y \cup \{a_1, a_2\}$. Ce dernier n'est donc pas valide et peut donc être ignoré. Ensuite comme $Y \cup \{a_2\}$ est non valide, le critère 1 indique à l'algorithme qu'il n'est pas nécessaire de générer tous les sur-motifs de $Y \cup \{a_2\}$.

Maintenant, si $Y \cup \{a_1\}$ et $Y \cup \{a_2\}$ sont des motifs valides, alors aucun des deux critères ne peut être utilisé pour affirmer (ou infirmer) que $Y \cup \{a_1, a_2\}$ satisfait la contrainte. Il faut alors vérifier dans la base directement (ou disposer d'autres informations). Un tel motif est communément appelé un *motif candidat*.

La génération des candidats est un point clé pour l'efficacité des algorithmes. Il est bien évident qu'il est très facile de ne pas créer un sur-motif à partir d'un motif non valide. Il est cependant plus difficile de vérifier que tous les sous-motifs d'un motif sont valides (critère 2). La vérification de ce dernier critère nécessite une structure adéquate pour gérer les motifs valides mais la génération de candidats doit rester simple. Han et al. [HPY00] considèrent cette étape comme un goulet d'étranglement et Bayardo et al. [BAG99] proposent même de ne pas vérifier le

second critère (Algorithme DENSE-MINER). Beaucoup plus de motifs (appelés aussi candidats) sont générés et vérifiés dans la base mais le coût de vérification du second critère est nul. Les expérimentations montrent que Dense-Miner reste comparable aux autres algorithmes ce qui montre à la fois l'efficacité du second critère d'élagage mais aussi son coût. Le point faible de ces algorithmes est la mise en mémoire de tous les motifs candidats comme nous allons le voir avec l'algorithme APRIORI.

6.3.2 APRIORI : algorithme de recherche en largeur d'abord

Les algorithmes de recherche (ou d'extraction) en largeur d'abord, ou algorithmes par niveaux, ont tous en commun qu'ils génèrent des motifs candidats de longueur $k+1$ à partir de motifs valides de longueur k. Il existe de nombreuses instances d'algorithmes par niveaux dont le premier, APRIORI, fut inventé par Agrawal et al. en 1993 [AIS93, AS94]. L'algorithme APRIORI est présenté à la figure 6.7.

Nous notons $\mathcal{F}_k$ l'ensemble des motifs valides de longueur k et C_k l'ensemble des motifs candidats de longueur k. APRIORI commence par lire la base de données $\mathcal{B}$ pour construire l'ensemble $\mathcal{F}_1$ des motifs valides de longueur 1 (ligne 1). L'algorithme alterne ensuite les phases de génération de candidats (fonction **apriori-gen**) et de vérification des candidats dans la base (lignes 4-7).

La génération de candidats de longueur $k+1$ à partir de l'ensemble F_k s'effectue de la manière suivante. APRIORI commence par fusionner 2 à 2 les motifs de $\mathcal{F}_k$ ayant le même préfixe de longueur $k-1$ (ligne 2 de **apriori-gen**) pour obtenir des motifs de longueur $k+1$. Ensuite, le second critère est testé pour chaque nouveau motif (fonction **verifie-critere2**). Si le critère est vérifié, alors le motif est un motif candidat et est ajouté à C_k, sinon il est ignoré.

Une fois que tous les candidats sont calculés, la phase de vérification parcourt la base pour ne garder que les motifs valides (lignes 4 à 7). Une exécution de APRIORI sur l'exemple 6.1 est donnée à la figure 6.8 pour la contrainte de 2-fréquence.

APRIORI parcourt autant de fois la base que la longueur du plus grand motif valide. Ce parcours est très coûteux surtout si la base ne peut être mise en mémoire principale. Une astuce pour réduire ce coût est de considérer le codage vertical[8] plutôt que le codage transactionnel, voire une approche hybride [HGN00].
L'inconvénient majeur de APRIORI est d'utiliser énormément de mémoire. En effet à chaque étape, l'algorithme garde en mémoire tous les motifs valides de longueur $k-1$ et tous les motifs candidats de longueur k. Pour de grandes bases de données denses, c'est à dire avec beaucoup de motifs valides, APRIORI conduit quasi-systématiquement à une saturation de la mémoire ce qui a motivé l'élaboration d'algorithmes en profondeur d'abord (cf. prochaine section).
La génération de tous les candidats ainsi que le calcul de leur support (pendant la vérification) nécessite une structure de données adaptée à cette tâche. Dans la littérature, plusieurs exemples sont proposés comme les tables de hachage, les hash-tree [AMS+96], les tries ou prefix-tries [BK02, BMUT97],etc. Dans son état de l'art [Goe03], Goethals a constaté que APRIORI n'avait pas du tout la même performance selon la structure choisie[9].
La génération, le stockage et la vérification sont donc primordiaux pour améliorer la complexité de APRIORI et il existe plusieurs optimisations possibles à ces étapes .

[8] Ainsi, le calcul du support d'un motif se résume à l'intersection des colonnes qui correspondent aux attributs qui composent le motif. Les autres colonnes sont ignorées. Avec le codage transactionnel, la base doit être parcourue entièrement.

[9] ils ont choisi le trie sans justifier ce choix

Algorithme Apriori

Input : une base $\mathcal{B}$ et une contrainte q
Output : l'ensemble des motifs valides $\mathcal{F}$

0 : $\mathcal{F} := \{\}$
1 : $C_1 := \{\{a\} | a \in \mathcal{A}\}$
2 : $k := 1$
3 : **Tant que** $C_k \neq \{\}$ **faire**
4 : // vérification des candidats
5 : **Pour tout** objet o de $\mathcal{B}$ **faire**
6 : mettre à jour les variables pour la vérification des candidats
7 : **fin pour**
8 : //Extraction des motifs valides
9 : $\mathcal{F}_k := \{X |\ q(X, \mathcal{B}) = 1\}$
10 : $\mathcal{F} := \mathcal{F} \cup \mathcal{F}_k$
11 : //Génération des candidats
12 : C_{k+1} :=**apriori-gen**$(\mathcal{F}_k)$
13 : $k++$
14 : **fin tant que**
15 : **retourner**$\mathcal{F}$

Fonction apriori-gen :génère les candidats

Input : ensemble $\mathcal{F}_k$ de motifs valides de longueur k
Output : l'ensemble C_{k+1} des motifs candidats de longueur $k+1$

0 : $C_{k+1} = \{\}$
1 : **Pour tout** motifs $X, Y \in \mathcal{F}_k$ avec $X[1..k-1] = Y[1..k-1]$ et $X[k] \neq Y[k]$ **faire**
2 : $Z := X \cup \{Y[k]\}$
3 : **Si verifie-critere2**$(Z, \mathcal{F}_k)$ **alors**
4 : $C_{k+1} = C_{k+1} \cup \{Z\}$
5 : **Fin si**
6 : **Fin pour**
7 : **retourner** C_{k+1}

Fonction vérifie-critere2

Input : ensemble $\mathcal{F}_k$ de motifs valides de longueur k et Z un motif généré
Output : vrai si Z vérifie le critère 2, faux sinon

0 : **Pour** i de 1 à $k+1$ **faire**
1 : **Si** $Z \backslash Z[i] \notin \mathcal{F}_k$ **alors**
2 : **retourner** faux
3 : **Fin si**
4 : **Fin pour**
5 : **retourner** vrai

FIG. 6.7 – Algorithme APRIORI.

C_1	$\mathcal{F}_1$	C_2	$\mathcal{F}_2$	C_3	$\mathcal{F}_3$	C_4
a_1	a_1 (4)	toutes	a_1a_3 (2)	$a_1a_3a_5$	$a_1a_3a_5$ (2)	
a_2	a_2 (4)	les	a_1a_4 (2)	$a_2a_3a_6$	$a_2a_3a_6$ (2)	
a_3	a_3 (5)	paires	a_1a_5 (2)			
a_4	a_4 (3)	de	a_2a_3 (3)			
a_5	a_5 (3)	a_1a_2	a_2a_6 (2)			
a_6	a_6 (3)	à	a_3a_5 (3)			
a_7	a_7 (2)	a_6a_7	a_3a_6 (2)			
			a_4a_7 (2)			

FIG. 6.8 – Exécution de APRIORI .

6.3.3 ECLAT : algorithme de recherche en profondeur

D'une manière générale, les algorithmes par niveaux sont très gourmands en ressource mémoire. Pour pallier à cette faiblesse, Zaki [ZPOL97, Zak00] proposa une nouvelle approche basée sur le parcours en profondeur d'abord du treillis et donna naissance à l'algorithme ECLAT. Depuis, d'autres algorithmes en profondeur sont apparus [AAP00, AAP01] dont un des plus connus est FP-GROWTH [HPY00] proposé par Han et al.

Dorénavant, nous supposons que les attributs qui composent les motifs sont rangés dans l'ordre lexicographique. Ainsi, le motif $a_5a_3a_8$ s'écrira $a_3a_5a_8$. Dans toute la suite, pour un motif $X = x_1 \dots x_k$ avec $x_1 <_{lex} x_2 <_{lex} \dots <_{lex} x_k$ et un attribut a, la notation Xa supposera toujours que $x_k <_{lex} a$. Pour un motif X de longueur $k-1$, l'ensemble $F[X]$ désigne les motifs valides de longueur au moins k ayant comme préfixe X (les attributs sont ordonnés). Tous les algorithmes en profondeur sont basés sur la propriété récursive suivante des ensembles F_k,

$$F[X] = \bigcup_{a \in \mathcal{A} \setminus X,\ Xa \text{ fréquent}} F[Xa]. \tag{6.1}$$

Toute la difficulté est de calculer le plus rapidement possible les ensembles $F[Xa]$. Pour certaines contraintes comme la contrainte de fréquence, la base est maintenue à jour lors des appels récursifs pour n'avoir à considérer que les informations nécessaires aux calculs des $F[Xa]$. Ainsi, plus le motif X est grand, plus la vérification des motifs candidats se fera rapidement car la taille de la base diminuera. Par exemple, la formule 6.1 pour une contrainte de fréquence q se simplifie en

$$F_{\mathcal{B}}[X] = \bigcup_{a \in \mathcal{A} \setminus X,\ Xa \text{ fréquent}} Xa \cdot F_{\mathcal{B}^{<Xa>}}[\emptyset], \tag{6.2}$$

où $\mathcal{B}^{<Y>}$ est la base initiale privée des attributs de Y et qui ne contient que les transactions qui supportent Y.

Une version de ECLAT pour la contrainte de γ-fréquence est donnée à la figure 6.9. Par hypothèse l'algorithme utilise le codage vertical des bases et suppose que tous les attributs dans la base passée en paramètre sont γ-fréquents. On constate que la construction des bases $\mathcal{B}^a$ forme le coût principal à chaque étape. Tous les motifs (candidats) de la forme $\{a, b\}$ sont testés (lignes 6-7) et seuls les attributs b avec les objets supportant a sont ajoutés dans la base $\mathcal{B}^a$ (ligne 8).

ECLAT teste la contrainte sur beaucoup plus de motifs candidats que APRIORI. L'approche en profondeur fait qu'ECLAT ne dispose pas de tous les motifs fréquents de longueur $k-1$ pour

Algorithme Eclat pour la γ-fréquence
Input : une base $\mathcal{B}$, un seuil de fréquence γ et un motif X
Output : $\{X \cup Y \mid Y$ motif γ-fréquent dans $\mathcal{B}\}$

```
0 :  F := {}
1 :  pour tout a ∈ A faire
2 :      F := F ∪ {X ∪ {a}}
3 :      //Création de la base B^a
4 :      B^a := {}
5 :      pour tout b ∈ A avec b > a faire
6 :          C := support({a}) ∩ support({b})
7 :          si |C| ≥ σ alors
8 :              B^a := B^a ∪ {(b, C)}
9 :          fin si
10 :     fin pour
11 :     //appel récursif
12 :     F_a := Eclat(B^a, γ, X ∪ {a})
13 :     F := F ∪ F_a
14 : fin pour
15 : retourner F
```

FIG. 6.9 – Implémentation de ECLAT pour la contrainte de fréquence.

vérifier le second critère d'élagage sur les motifs candidats de longueur k (essentiellement toutes les fusions de deux motifs valides ayant le même préfixe sont testées ce qui correspond uniquement à l'étape de fusion de APRIORI sans la vérification du second critère d'élagage). Toutefois, les vérifications sont plus efficaces comparées à celles de APRIORI et l'espace mémoire requis est sans comparaison moindre. De plus, ECLAT est facilement parallélisable car les ensembles $F[Xa]$ pour des attributs avec $a \notin X$ sont tous disjoints et peuvent être calculés séparément [LM04].

Comme APRIORI , ECLAT connaît des améliorations. Il est possible d'ordonner les attributs à chaque étape pour diminuer le nombre de motifs candidats. La fusion de deux motifs peut se révéler inutile modulo des informations complémentaires sur les *diffsets*[10] [ZG01, ZH02]. Finalement, une approche hybride combinant APRIORI pour les motifs de petites tailles et ECLAT pour les motifs plus longs a été proposée par Hipp et al. dans [HGN00].

6.3.4 Autres algorithmes

Jusqu'à présent, nous avons décrit deux algorithmes qui furent respectivement les pionniers pour la recherche en largeur et en profondeur de motifs. L'inconvénient des algorithmes en profondeur est qu'ils nécessitent de mettre en mémoire la base complète. Cela n'est pas toujours possible. Pour contourner ce problème, les auteurs de [BMUT97] et [SON95] ont respectivement présenté deux algorithmes DIC et PARTITION qui découpent le base initiale en petits morceaux qui tiennent en mémoire, puis calculent les motifs fréquents locaux à chaque base et vérifient si les motifs locaux sont fréquents dans la base initiale. Ces approches génèrent toutefois beaucoup plus de motifs candidats que APRIORIpuisqu'un motif peut être fréquent dans un des morceaux

[10] différence entre le support d'un motif et le support de tous ses sous-motifs immédiats

sans l'être dans la base complète[11].

Une alternative au partitionnement est l'échantillonnage proposé par Toivonen [Toi96]. Un échantillon d'objets est prélevé au hasard dans la base initiale et les motifs fréquents y sont ensuite extraits. Le choix du seuil de fréquence pour l'extraction dans l'échantillon est délicat car il doit assurer que tous les motifs fréquents de la base initiale sont aussi des motifs fréquents de l'échantillon. Pour s'en assurer, le seuil est abaissé ce qui a pour effet de générer beaucoup plus de candidats.

Nous terminons cette section consacrée aux algorithmes par un nouvel algorithme en profondeur FP-Growth. L'objectif de cet algorithme est de tirer avantage à la fois des codages transactionnel et vertical. Pour cela, une nouvelle structure de données appelée FP-tree (Frequent Pattern Tree) est créée. Toutes les informations sur cette structure sont fournies dans [HPY00]. Cependant, Goethals [Goe03] semble indiquer que malgré ses nombreux avantages, la gestion du FP-tree à chaque étape est très coûteuse ce qui fait de FP-Growth un algorithme comparable à Apriori ou Eclat.

Cette section a présenté une liste d'algorithmes de natures très différentes. Cette liste n'est bien entendu pas exhaustive. Nous n'avons pas abordé par exemple les algorithmes d'extraction de motifs fermés dont les plus connus sont Close [PBTL99a], Charm [ZH02] et Closet [PHM00, WHP03]. La prochaine section traite des apports possibles de l'analyse en moyenne à la compréhension des algorithmes d'extraction de motifs.

6.4 Que peut apporter l'analyse en moyenne à la fouille de données ?

La première section a présenté le problème général de l'extraction de motifs satisfaisant une contrainte anti-monotone. La seconde section a décrit plusieurs algorithmes pour résoudre ce problème. Dans l'introduction, nous avons expliqué notre positionnement. A savoir expliquer les phénomènes observés en pratique par des analyses en moyenne, mieux comprendre la difficulté de l'extraction de motifs et si possible, utiliser les résultats théoriques pour optimiser les algorithmes. Atteindre ces objectifs passe nécessairement par l'étude de certains paramètres.

Dans cette section, nous présentons quelques paramètres qui ont un grand intérêt pour l'extraction. Nous tenterons aussi de donner des pistes pour aborder ces problèmes avant d'analyser dans le prochain chapitre, le nombre de motifs fréquents et fermés présents dans une base de données prise au hasard.

6.4.1 Nombre de motifs valides, fermés et libres

Il est bien connu que le nombre maximum de motifs valides, fermés ou libres est d'au plus 2^m pour m attributs et que cette borne est atteinte pour des instances très particulières. Cependant, les expériences montrent que ces motifs sont beaucoup moins nombreux, et jusqu'à aujourd'hui, nous ne disposons pas d'ordres de grandeur plus fidèles à la réalité.

Nous nous proposons dans un premier temps (cf. prochaine partie) d'étudier le nombre de ces motifs dans une base de données aléatoire (pour la contrainte de fréquence). Les intérêts sont multiples. Tout d'abord, le nombre de motifs valides est le travail minimum auquel ne peut

[11] En revanche, si un motif est 10% fréquent dans la base initiale, il existe au moins un morceau dans lequel il est 10% fréquent.

échapper tout algorithme d'extraction. Ce paramètre constitue donc une *borne inférieure* de la complexité des algorithmes d'extraction. Ensuite, il permet de mieux appréhender la réelle difficulté du problème. Ainsi, existe-t-il un nombre exponentiel de motifs valides ou plutôt un nombre polynomial ? D'un autre coté, différents modèles de bases peuvent mettre en évidence l'influence des bases (et de leurs propriétés) sur le nombre de motifs valides. Finalement, comme les algorithmes de comptage probabiliste, nous espérons à plus long terme pouvoir compter rapidement le nombre de motifs valides avec un minimum d'information sur la base. Ceci permettrait aux experts d'extraire rapidement la quantité d'information voulue sans avoir à répéter le processus d'extraction plusieurs fois.

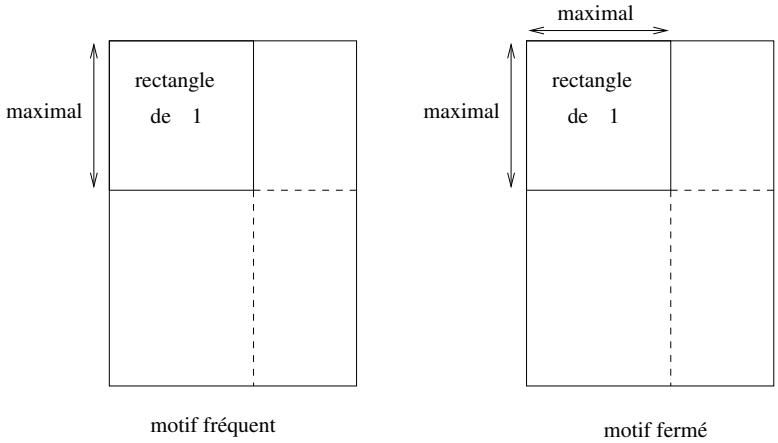
En outre, les motifs γ-fréquents ou γ-fermés ont un sens particulier dans les représentations matricielles et par graphes (cf. figure 6.10). Dans la représentation matricielle, un motif γ-fréquent est à une permutation près des lignes et des colonnes, un rectangle de 1 maximal en hauteur d'au moins γ lignes. Un motif fermé est quant à lui un rectangle de 1 maximal en hauteur et en largeur d'au moins γ lignes. Dans la représentation en graphe bipartite, un motif 1-fermé avec son support est un sous-graphe bipartite complet et maximal [LLSW05] alors que dans le graphe co-bipartite, les 1-fermés avec leurs supports sont en correspondance (bijective) avec les séparateurs minimaux [BS04, Sig02] (ensemble de sommets qui lorsqu'ils sont supprimés déconnectent le graphe). Énumérer dans un modèle aléatoire précis le nombre de motifs fréquents ou fermés revient donc à énumérer les objets combinatoires précédemment cités.

6.4.2 Taille de la bordure négative

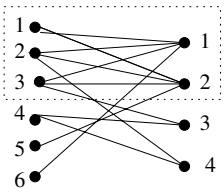
Un motif de la bordure négative est un motif non valide dont tous les sous-motifs sont valides. Mannila et Toivonen [MT97] ont montré que les algorithmes par niveaux effectuent au moins une vérification pour chaque motif des bordures négative et positive. Comme la bordure positive est incluse dans l'ensemble des motifs valides, le nombre total de vérifications correspond à la taille de la bordure négative plus le nombre de motifs valides.

Un second paramètre à analyser est donc le nombre de motifs dans la bordure négative. Associé à l'analyse du nombre de motifs valides, on en déduira le nombre total de vérifications exécutées par les algorithmes en largeur.

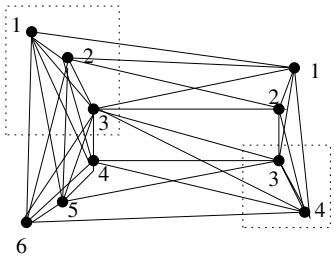
Mais ce paramètre a un autre intérêt dans la théorie des hypergraphes. Comme nous l'avons vu, une base de données $\mathcal{B}$ se modélise naturellement sous la forme d'un hypergraphe $\mathcal{H}_{\overline{B}}$. Une traverse minimale d'un hypergraphe est un ensemble minimal (au sens de l'inclusion) de sommets qui intersecte toutes les hyper-arêtes. Les auteurs de [MT97] ont montré que les traverses minimales de l'hypergraphe $\mathcal{H}_{\overline{B}}$ forment les éléments de la bordure négatives. Déterminer la taille moyenne de la bordure négative d'une base revient donc à calculer le nombre moyen de traverses minimales que possède un hypergraphe.



(a) Représentation matricielle

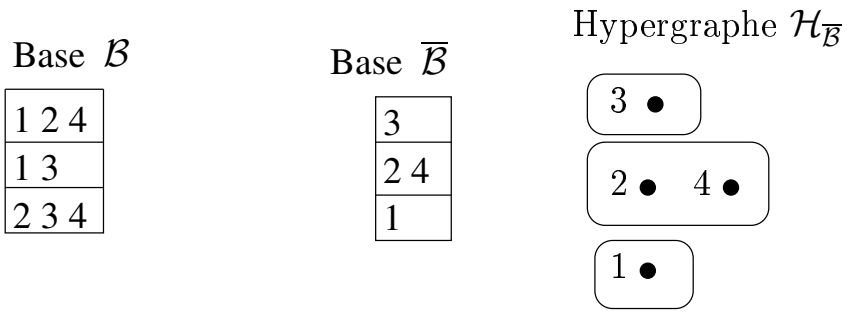


(b) motif fermé {1,2,3} de support {1,2}=
sous−graphe bipartite maximal complet {1,2,3}×{1,2}



(c) motif fermé {4,5,6} de support {1,2}=
séparateur minimal {1,2,3}×{3,4}

FIG. 6.10 – Correspondances entre les motifs fréquents ou fermés et les rectangles maximaux dans la matrice, les sous-graphes bipartites complets maximaux et les séparateurs minimaux.



traverses minimales $\{1, 2, 3\}$ et $\{1, 3, 4\}$
bordure négative=$\{\{1, 2, 3\}, \{1, 3, 4\}\}$

6.4.3 Taille du plus long motif

A chaque niveau du treillis des motifs valides, APRIORI génère des candidats puis parcourt la base de données pour savoir s'ils satisfont la contrainte. La taille du plus long motif est donc le nombre de passes sur la base que APRIORI utilise pour toutes ses vérifications. Dans la littérature, le parcours complet de la base est souvent considéré comme très coûteux surtout si la base est en mémoire externe. C'est pourquoi on essaie souvent d'optimiser ce paramètre.

Les expériences avec la contrainte de fréquence montrent qu'il existe très peu de longs motifs valides. Pour des bases aléatoires où chaque attribut appartient à une transaction avec une probabilité p fixée, les auteurs de [AMS+96] ont montré qu'il y avait peut de motifs de longueur plus grande que $\log r/logp$ où le seuil de fréquence est donné par $\gamma = r \times n$ avec $r \in]0, 1[$ et n est le nombre de transactions. Cependant, la distribution exacte de la taille du plus long motif n'est pas connue.

6.4.4 Complexité des algorithmes de recherche en largeur

La complexité en temps et la complexité en espace sont deux paramètres fondamentaux pour les algorithmes de recherche en largeur.

Les algorithmes par niveaux alternent les phase de génération de candidats et les phases de vérification. Comme nous l'avons vu précédemment, le coût total de toutes les phases de vérifications est étroitement lié à la taille de la bordure négative plus le nombre de motifs valides. Mais plus que le nombre de motifs, il faudrait aussi tenir compte du coût de la vérification d'un motif qui est lui très dépendant de la structure utilisée (trie, Hashtree). Le coût total des phases de génération comprend le coût total des phases de fusion et le coût total des vérifications du second critère d'élagage. Ces coûts dépendent bien entendu de la structure de données utilisée mais par exemple le nombre de couples fusionnés permettrait d'avoir une bonne compréhension de l'importance de la phase de fusion.

La complexité en mémoire est aussi un point clé des algorithmes par niveau. L'espace mémoire maximum utilisé par APRIORI intervient au moment où le nombre de motifs candidats de longueur k et le nombre de motifs valides de longueur $k-1$ est maximum. Actuellement, tout ce que l'on sait sur la complexité en mémoire est du même ordre que ce que l'on sait sur le nombre de motifs valides à savoir que dans le pire des cas, cela explose de manière exponentielle. Mais

en pratique et pour des contraintes raisonnables, APRIORI reste efficace ce qui semble montrer un comportement différent du pire des cas.

6.4.5 Complexité des algorithmes en profondeur

Contrairement aux algorithmes en largeur, la complexité en mémoire est peu importante pour les algorithmes en profondeur car ils ont été conçus pour ne pas utiliser trop de cette ressource. En revanche, la complexité en temps reste fondamentale.

La complexité de l'algorithme ECLAT (pour la contrainte de fréquence) est essentiellement constituée du coût total de construction des bases $\mathcal{B}^a$. Contrairement aux algorithmes par niveaux, la complexité des algorithmes en profondeur me semble plus accessible. Il est traditionnel en algorithmique de représenter les exécutions d'algorithmes récursifs sous la forme d'arbres. Ensuite, l'analyse de certains paramètres de l'algorithme se fait en attribuant un coût aux feuilles et un coût aux noeuds internes des exécutions. Pour calculer la complexité de ECLAT, il me semble suffisant de considérer un coût aux feuilles valant 1 et un coût au noeud équivalent au coût total pour générer toutes les bases $\mathcal{B}^a$. Un bémol cependant, les arbres considérés sont très déséquilibrés ce qui est inhabituel en analyse d'algorithmes.

6.4.6 Combinaison de contraintes

Nous terminons cette section avec ce que nous souhaiterions faire à long terme en Fouille de Données. Pendant très longtemps, la communauté scientifique s'est concentrée uniquement sur les contraintes monotones ou anti-monotones. Depuis quelques années, la recherche de motifs sous contrainte s'est généralisée. Aujourd'hui, nous disposons d'algorithmes capables d'extraire des motifs qui satisfont une contrainte générale issue de grammaires de contraintes [SC05].

Notre objectif, peut-être ambitieux ou utopique est de créer une grammaire d'objets combinatoires, à l'image des dictionnaires pour les séries génératrices, grâce à laquelle il serait possible d'analyser systématiquement certains des paramètres précédemment rencontrés.

6.5 Conclusion

Dans ce chapitre, nous avons présenté les notions essentielles sur l'extraction de motifs sous contrainte anti-monotone. Nous avons aussi introduit deux algorithmes importants dans l'histoire de l'extraction de motifs. Le premier, APRIORI , fut créé pour extraire les motifs avec une stratégie de recherche en largeur d'abord. ECLAT est quant à lui le premier algorithme utilisant la recherche en profondeur d'abord.

Nous nous proposons d'apporter un nouvel éclairage sur l'extraction de motif, celui de l'analyse en moyenne. Dans un domaine où les analyses dans le pire des cas sont nombreuses mais non représentatives de la complexité réelle, l'analyse en moyenne trouve naturellement sa place. Nos objectifs à court et long terme sont multiples. Tout d'abord, nous désirons vérifier par la théorie des constatations pratiques. Ensuite nous espérons apporter de nouveaux éléments concernant la réelle complexité de l'extraction de motifs. Finalement à long terme, nous espérons pouvoir analyser concrètement la complexité des algorithmes de références. Le prochain chapitre constitue le premier pas vers tous ces objectifs et présente la première analyse en moyenne du nombre de motifs fréquents et fermés dans une base de données aléatoire.

Chapitre 7

Analyse des motifs fréquents et fermés dans une base de données

Sommaire

7.1 Introduction

La partie précédente résume les aspects essentiels de l'extraction de motifs sous une contrainte anti-monotone. Nous nous concentrons maintenant sur le nombre moyen de motifs valides pour la contrainte la plus utilisée : la contrainte de fréquence définie pour un entier positif γ, un motif X et une base $\mathcal{B}$ par,

$$q_\gamma(X, \mathcal{B}) = \begin{cases} 1 & \text{si } Freq_{\mathcal{B}}(X) \geq \gamma \\ 0 & \text{sinon.} \end{cases}$$

L'extraction des motifs fréquents est la première étape mais aussi la plus coûteuse pour construire les règles d'associations. Comme les règles d'associations s'appliquent à des domaines très différents (Biologie, Chimie, Géographie, Commerce, etc.), il est nécessaire de bien comprendre quelle est la réelle difficulté de l'étape d'extraction.

Complexité théorique. D'un point de vue de la théorie de la complexité, l'extraction est une tâche très difficile. En utilisant le problème de sous-graphe bipartite complet et équilibré, les auteurs de [GMS97, PGG04] ont montré que l'existence d'un motif fréquent (maximal) de longueur donnée est un problème NP-complet. Le problème de comptage associé est quant à lui #P-difficile puisque le calcul du nombre d'affectations satisfaisant une formule 2-CNF s'y réduit [GMKT97].

Nombre de motifs. Nous disposons (à notre connaissance) de très peu d'informations sur le nombre de motifs qu'ils soient fréquents, libres, fermés ou même candidats. Dans [GGV01], Geerts, Goethals et Van den Bussche donnent des bornes supérieures du nombre de motifs candidats générés à partir de motifs fréquents en se basant sur des décompositions d'entiers en somme de coefficients binomiaux. Leur résultat donne une borne supérieure optimale mais cette borne n'admet pas de formule close. Il n'est alors pas facile d'avoir une idée exacte du nombre de candidats. De leur coté, Dexters et Calders [DC04] ont étudié le lien entre la longueur maximale d'un motif libre et le nombre de motifs libres. Ils ont prouvé que si ℓ est cette longueur maximale, alors il y a 2^ℓ motifs libres.

Pires et meilleurs cas. Si tous les attributs sont présents dans tous les objets (ou de manière équivalente, si le codage matriciel ne contient que des 1, si le graphe bipartite est *complet*, si le graphe co-bipartite n'est pas connexe), alors tous les motifs sont fréquents quel que soit le seuil γ avec $1 \leq \gamma \leq n$ (où n est le nombre d'objets). Ainsi, pour m attributs, il y a 2^m motifs fréquents. Maintenant, si aucun attribut n'est contenu dans les objets (ou de manière équivalente, si le codage matriciel ne contient que des 0, si le graphe bipartite n'a que des sommets isolés, si le graphe co-bipartite est complet), alors seul le motif vide est fréquent quel que soit le seuil γ. Le nombre de motifs γ-fréquents est donc au pire exponentiel, au mieux constant. Il en est de même avec les motifs libres et fermés. Pour le codage matriciel ne contenant que des 1, tous les motifs ont le même support, il n'y a donc qu'une classe d'équivalence avec un unique motif fermé (avec tous les attributs) et un unique motif libre (le motif vide). Considérons maintenant la matrice composée uniquement de blocs de un attribut et de γ objets. Supposons de plus que tous les blocs contiennent des 1 sauf les blocs diagonaux qui contiennent uniquement des 0. Ainsi, les $2^m - 1$ motifs non complets sont fermés et libres. La figure 7.1 résume ces deux situations.

Analyses en moyennes existantes. Il existe un saut théorique important entre le pire et le meilleur des cas pour le nombre de motifs fréquents, fermés et libres mais à notre connaissance, il n'existe pas d'autres ordres de grandeurs plus précis sur la difficulté réelle de la tâche. C'est pour cette raison que nous avons effectué une analyse en moyenne. Nous avons trouvé peu d'analyses probabilistes en extraction de motifs. Lorsque Agrawal et al. introduisent pour la première fois l'algorithme Apriori [AMS+96], ils montrent que le nombre moyen de motifs fréquents de longueur ℓ dans une base de données aléatoires (sous le modèle de Bernoulli, cf. section 7.4.1) est d'au plus $m^\ell e^{-2(\gamma - np^\ell)^2/n}$ avec p la probabilité qu'un attribut appartiennent à un objet et $\gamma > np^\ell$. Cette borne, obtenue avec une inégalité de Chernoff, montre qu'il y a (en moyenne) peu de motifs fréquents de longueur plus grande que $\log(\gamma/n)/\log p$ (modulo certaines conditions sur m, *gamma* et n).
Toivonen a aussi utilisé des résultats probabilistes pour justifier sa méthode d'échantillonnage [Toi96]. Il montre par exemple qu'un motif fréquent dans la base de départ est presque sûrement fréquent dans l'échantillon (à condition d'adapter le seuil γ).

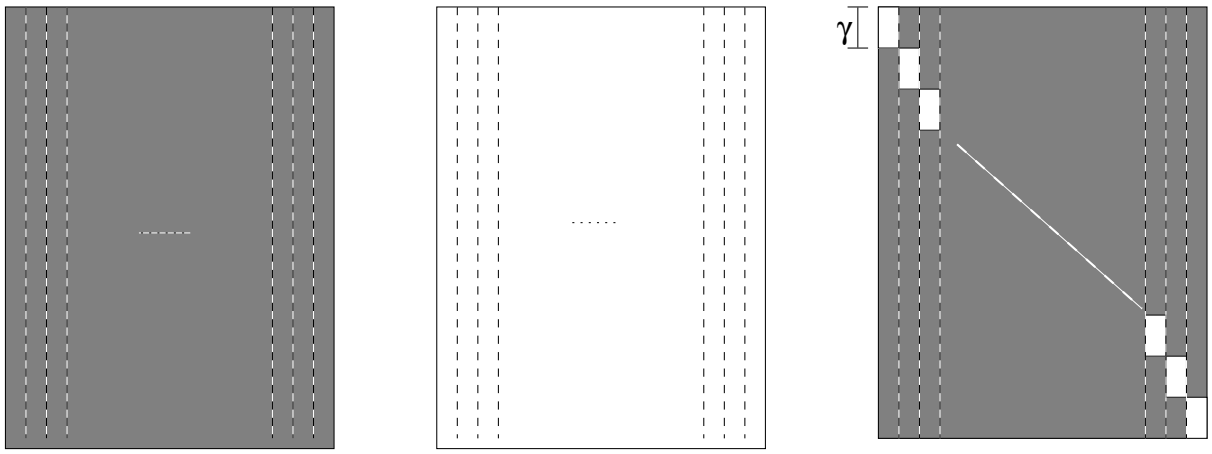


FIG. 7.1 – Pires et meilleurs des cas pour les motifs fréquents, fermés et libres : (a) nombre constant de motifs fréquents, fermés et libres (b) nombre exponentiel de motifs fréquents, nombre constant de motifs fermés et libres (c) nombre exponentiel de motifs fréquents, fermés et libres.

Plus récemment, Purdom et al. [PGG04] ont analysé sous le même modèle de bases aléatoires que [AMS+96], le taux d'échec ou de succès de APRIORI. A chaque niveau, APRIORI génère des candidats et le taux d'échec (resp. de succès) est la proportion de candidats non valides (resp. valides). En utilisant une fois de plus les bornes de Chernoff, les auteurs ont mis en évidence quatre régions où les taux de succès ou d'échec sont proches de 0 ou 1.
Des analyses probabilistes ont également été réalisées par Christodoulakis [CHR83b, CHR83a, CHR84] pour montrer l'influence des hypothèses d'indépendance et d'uniformité sur l'estimation de paramètres liés à l'évaluation du coût d'une requête dans une base de données relationnelle. En particulier, il montre que l'uniformité et l'indépendance conduisent souvent à une majoration du coût réel, entraînant un surcroît de travail pour obtenir les réponses aux requêtes.

Plan. Ce chapitre est organisé de la manière suivante. La première section présente différents points de vue sur le problème de recherche de motifs fréquents. La section suivante introduit les hypothèses et les nouveaux résultats en moyenne. La troisième section présente des modèles pour lesquels les hypothèses s'appliquent. Nous terminons finalement avec les preuves.

7.2 Points de vue sur le problème

Au chapitre précédent, nous avons décrit plusieurs représentations possibles des bases de données. Dans chacune de ces représentations, les motifs γ-fréquents ou γ-fermés ont une signification particulière que nous décrivons maintenant.

7.2.1 Point de vue matriciel

Les bases de données considérées s'expriment sous la forme de matrices binaires. Si $\mathcal{M} = (m_{i,j})_{i\in\{1,\ldots,n\},j\in\{1,\ldots,m\}}$ est une matrice binaire, un motif $X \subset \{1,\ldots,m\}$ (où l'attribut a_i est identifié avec son indice i) est un motif γ-fréquent si son support $S \subset \{1,\ldots,n\}$ (où l'objet o_i est identifié avec son indice i) est de cardinal au moins γ. Quitte à permuter les lignes et les colonnes,

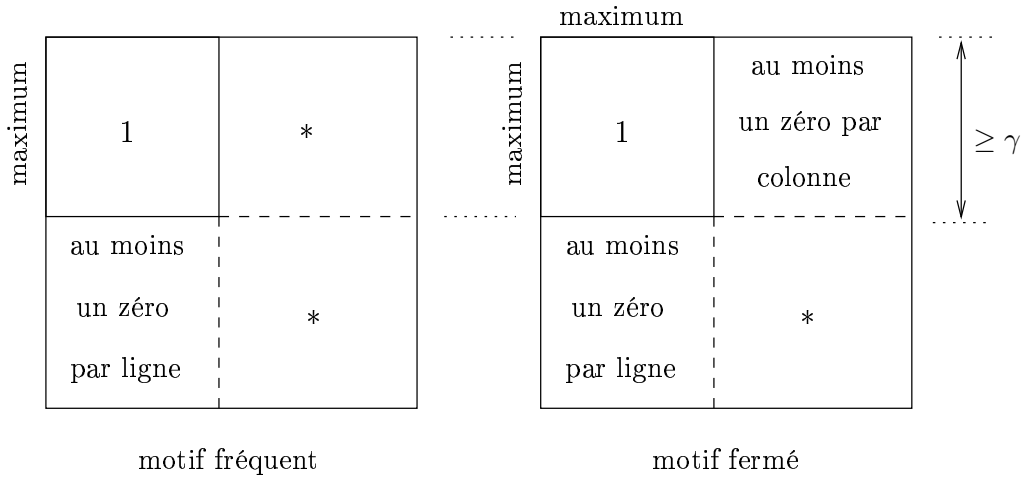


FIG. 7.2 – Propriétés matricielles des motifs fréquents et fermés

nous pouvons supposer que X et S sont respectivement de la forme $\{1, \ldots, j\}$ et $\{1, \ldots, i\}$. La matrice $\mathcal{M}$ est alors constituée d'un rectangle de 1 en haut à gauche qui ne peut être agrandi vers le bas et qui est de hauteur au moins γ. Nous en déduisons la propriété suivante :

Les motifs γ-fréquents sont à une permutation des lignes et des colonnes près, des rectangles de 1 maximaux en hauteur et de hauteur au moins γ.

La figure 7.2 résume cette situation. Le motif $X = \{1, \ldots, j\}$ est en plus γ-fermé si l'ajout d'un attribut fait diminuer strictement le support. A une permutation des colonnes près, cela signifie que le rectangle de 1 ne peut être agrandi en largeur (en gardant la même hauteur). Cela nous amène à la propriété suivante :

Les motifs γ-fermés sont à une permutation des lignes et des colonnes près, des rectangles de 1 maximaux en hauteur et en largeur, de hauteur au moins γ.

La figure 7.2 résume aussi cette situation. Nous venons de donner deux caractérisations matricielles des motifs fréquents et fermés, et leur dénombrement se ramène donc au dénombrement de rectangles de 1 particuliers dans une matrice binaire.

7.2.2 Point de vue graphe bipartite

Une base de données (resp. matrice) binaire s'interprète aussi sous la forme d'un graphe bipartite non-orienté (S_1, S_2, E) où S_1 est l'ensemble des sommets $S_1 = \{a_1, \ldots, a_m\}$ correspondant aux attributs (resp. colonnes) et $S_2 = \{o_1, \ldots o_n\}$ est l'ensemble des sommets correspondant aux objets (resp. lignes). L'ensemble des arêtes E est un sous-ensemble de $S_1 \times S_2$ tel que (a, o) est une arête si et seulement si l'objet o contient l'attribut a (resp. il y a un 1 dans la colonne associée à a et la ligne associée à o). Par identification, un motif est un sous-ensemble S_1' de S_1 et son support est un sous ensemble S_2' de S_2. En particulier, chaque élément de S_1' est relié par une arête à chaque élément de S_2'. Le sous-graphe ainsi formé est un sous-graphe bipartite complet du graphe initial. Par définition, le motif S_1' est γ-fréquent si son support S_2' est de cardinal au moins γ. De plus, le support est le plus grand ensemble d'objets qui contient le motif. S_2' est alors le plus grand sous-ensemble de S_2 dont les éléments sont tous reliés aux éléments de S_1'. Nous en déduisons la propriété suivante :

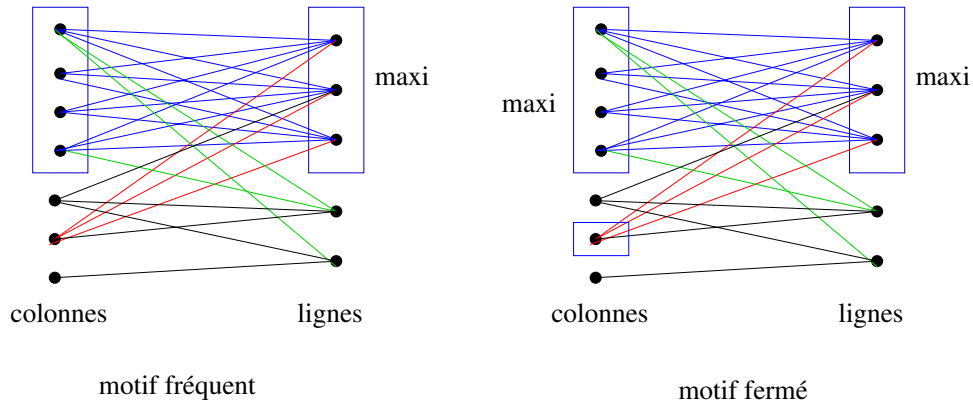


FIG. 7.3 – Liens entre les motifs fréquents ou fermés avec les sous-graphes bipartites complets.

Un motif γ-fréquent est un sous-graphe bipartite complet $(S'_1, S'_2, E|_{S'_1 \times S'_2})$ de (S_1, S_2, E) tel que S'_2 est de cardinal au moins γ et tel que S'_2 est le plus grand sous ensemble de S_2 vérifiant $(S'_1, S'_2, E|_{S'_1 \times S'_2})$ est complet.

Le motif S'_1 est en plus γ-fermé si l'ajout d'un élément entraîne que le support diminue strictement. Dans ce cas, S'_1 est le plus grand sous-ensemble de S_1 tel que tout élément de S'_1 est relié à tout élément de S'_2. Par définition, le couple (S'_1, S'_2) forme un sous-graphe bipartite complet maximum c'est-à-dire, un sous graphe-bipartite complet dont tout sur-ensemble de sommets ne forme pas un graphe bipartite complet.

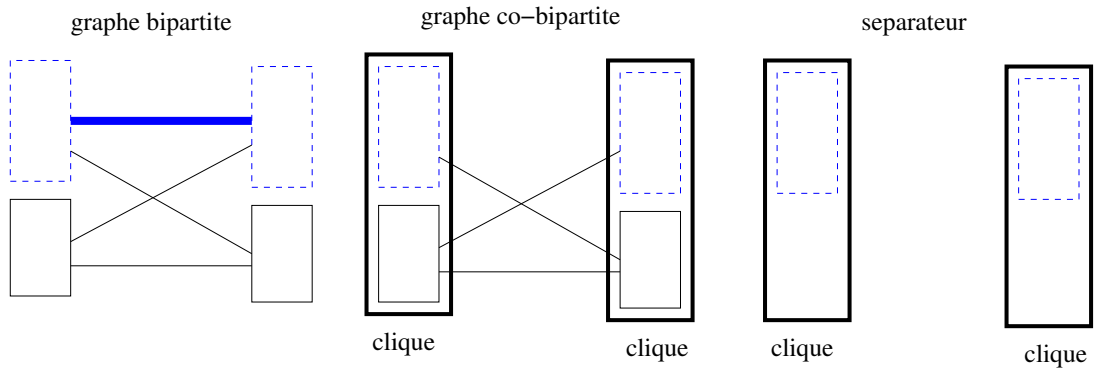
Un motif γ-fermé est un sous-graphe bipartite $(S'_1, S'_2, E|_{S'_1 \times S'_2})$ complet maximum (au sens de l'inclusion) tel que S'_2 est de cardinal au moins γ.

La figure 7.3 résume les propriétés précédentes. Avec ces relations, le dénombrement de motifs fréquents ou fermés se ramène au dénombrement de sous-graphes bipartites complets maximums ou maximums d'un coté. Ces relations ont déjà été utilisées par les auteurs de [LLSW05] pour extraire les motifs fermés avec des techniques de graphes.

7.2.3 Point de vue graphe co-bipartite

Le graphe co-bipartite $(S_1, S_2, \overline{E})$ associé à une base de données binaire se construit de la manière suivante. L'ensemble S_1 (resp. S_2) est l'ensemble des attributs (resp. objets). Tous les éléments de S_1 (resp. S_2) sont reliés entre eux et forment ainsi une clique. Ensuite, il y a une arête entre un attribut a de S_1 et un objet o de S_2 si l'objet o ne contient pas l'attribut a. Le graphe ainsi construit est donc composé de deux cliques reliées entre elles par des arêtes. Ce graphe est exactement le co-graphe du graphe bipartite décrit dans la partie précédente. Pour obtenir un co-graphe, le principe est le suivant : là où il y a des arêtes, on les enlève et là où il n'y en a pas, on en met. Dans la section précédente, les fréquents ou les fermés étaient des sous-graphes bipartites (S'_1, S'_2) complets. Dans le graphe co-bipartite, les ensembles S'_1 et S'_2 n'ont pas d'arête commune. Si les ensembles $S_1 \backslash S'_1$ et $S_2 \backslash S'_2$ sont supprimés ainsi que les arêtes qui sont issues de ces ensembles, alors il reste les deux sous-cliques associés à S'_1 et S'_2 qui sont déconnectées. La suppression de $S_1 \backslash S'_1$ et $S_2 \backslash S'_2$ rend donc le graphe non connexe. Or un ensemble de sommets tel que s'il est supprimé, le graphe devient non connexe est appelé un séparateur. La figure suivante montre le passage d'un sous-graphe bipartite complet à un séparateur. Une ligne en gras signifie

qu'il n'y a que des arêtes entre les deux ensembles de sommets alors que les autres lignes signifient qu'il peut exister des arêtes.



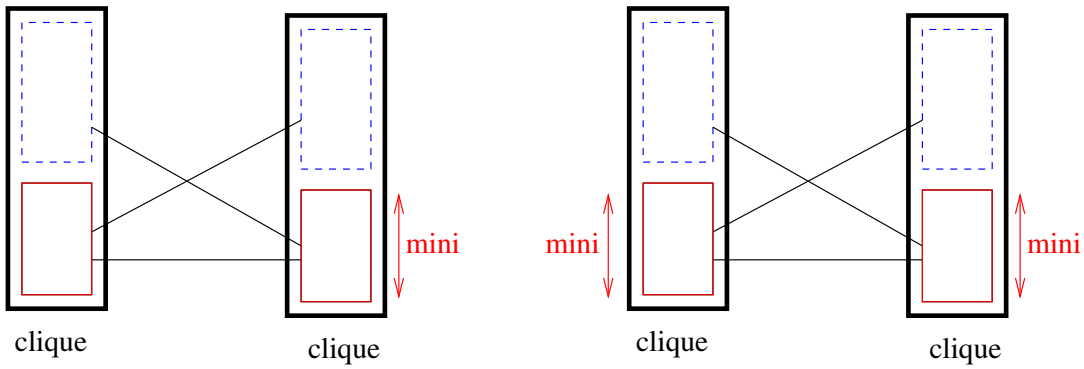
Un motif fréquent S'_1 de support S'_2 correspond donc dans le graphe co-bipartite au séparateur $(S_1 \backslash S'_1, S_2 \backslash S'_2)$. Pour un motif fréquent, S_2 est maximal au sens de l'inclusion et a un cardinal d'au moins γ. L'ensemble $S_2 \backslash S'_2$ est donc minimal et a pour cardinal au plus $n - \gamma$ (où n est le nombre d'objets. On en déduit le point suivant :

L'ensemble des motifs γ-fréquents est en bijection avec l'ensemble des séparateurs (S'_1, S'_2) du graphe co-bipartite avec $|S'_2| \leq n - \gamma$ et tels que S'_2 est le plus petit ensemble S vérifiant (S'_1, S) est un séparateur.

De même, pour un motif fermé, S_1 est aussi maximal au sens de l'inclusion et l'ensemble $S_1 \backslash S'_1$ est alors minimal. Les motifs fermés vérifient alors la propriété suivante :

L'ensemble des motifs γ-fermés est en bijection avec l'ensemble des séparateurs (S'_1, S'_2) minimums (au sens de l'inclusion) du graphe co-bipartite et tels que $|S'_2| \leq n - \gamma$.

La relation entre les séparateurs minimums et les motifs fermés a déjà été utilisée par les auteurs de [Sig02, BS04] pour importer les techniques de graphes en fouille de données. La figure suivante résume les deux types de séparateurs correspondants aux motifs fréquents et fermés.



7.2.4 Point de vue Pattern Matching

Une source est un procédé aléatoire qui émet des symboles à chaque top d'une horloge. Le pattern matching est le domaine de l'informatique qui s'intéresse aux propriétés des motifs

(ou mots) produits par une source aléatoire. Les questions posées sont très diverses : un motif appartient-il à un texte produit par la source ? Si oui, combien de fois ? Combien de temps faut-il attendre pour voir apparaître le motif ? etc ... Au départ, les motifs considérés étaient des lettres consécutives alors que de nos jours, des contraintes plus générales sont utilisées.

Pour introduire le pattern matching en fouille de données, il nous faut un texte. Nous considérons chaque ligne de la matrice binaire, ou de manière équivalente chaque objet, comme un mot produit par une source aléatoire sur l'alphabet $\{0,1\}$. Une base de données à m attributs et n objets est alors une suite de n mots $(L_1, \ldots, L_n)$ produit par la source dont chacun est de longueur m. Nous appelons *motif ligne* X, un sous-ensemble de $\{1, \ldots, m\}$.

Définition 22 *Nous dirons qu'un motif ligne $X \subset \{1, \ldots, m\}$ est contenu dans une ligne ou dans un mot L si $L = \ell_1 \ldots \ell_m$ et si pour tout $i \in X$, $\ell_i = 1$. Nous écrirons alors $X \subset L$.*

Il revient au même de dire que la ligne admet des 1 aux endroits pointés par le motif ligne et qu'ailleurs, il peut y avoir soit 0 soit 1. Si $\mathcal{A} = \{a_1, \ldots, a_m\}$, à chaque motif ligne $X \subset \{1, \ldots, m\}$ correspond naturellement un motif (d'attributs) $X_{\mathcal{A}}$. Le problème de recherche de motifs γ-fréquents se récrit comme suit.

Etant donné n mots de longueur m produits par une source binaire $\mathcal{S}$, trouver l'ensemble des motifs lignes de $\{1, \ldots, m\}$ qui sont contenus dans au moins γ mots.

A notre connaissance, ce problème de pattern matching n'a jamais été traité. Il est très souvent recherché des motifs présents dans tous les mots ou bien des préfixes présents dans au moins deux mots (voir par exemple les tries [CFV01]).
Pour les motifs fermés, il faut tenir compte du support qui doit strictement décroître lorsqu'un attribut est ajouté. Voici donc le problème équivalent au problème de recherche de motifs fermés.

Etant donné n mots de longueur m produits par une source binaire $\mathcal{S}$, trouver l'ensemble des motifs lignes de $\{1, \ldots, m\}$ présents dans au moins γ mots et tel que tout sur-motif ligne Y, $X \subsetneq Y$, est présent dans strictement moins de mots.

Dans la suite de ce chapitre, nous allons considérer les bases de données comme une liste de mots et nous nous plaçons dans le cadre du pattern matching.

7.3 Modélisation des bases et résultats

Pour la première fois en fouille de données, des résultats précis concernant le nombre moyen de motifs fréquents et fermés sont donnés. Ces résultats sont obtenus dans un modèle de bases aléatoires construites à partir de sources et que nous décrivons dans la première section. Plus le seuil de fréquence est grand, moins il y a de motifs fréquents. Pour décrire cette évolution, nous considérons dans la deuxième section trois types de seuils : fixe, logarithmique et linéaire. Finalement, il est peu probable d'obtenir des résultats généraux valables pour toutes les sources. A partir d'une hypothèse faite pour chaque type de seuil, nous énonçons dans la dernière section les trois résultats principaux sur les motifs fréquents et fermés.

7.3.1 Modélisation des bases de données

Dans toute la suite, le nombre d'attributs (de colonnes) est noté m et n désigne le nombre d'objets (de lignes). Il est clair que le nombre de motifs fréquents dépend fortement de la taille de

la base. On sait par exemple qu'il y a beaucoup plus de motifs fréquents dans les bases génétiques (comportant un grand nombre d'attributs par rapport au nombre d'objets) que dans les bases plus rectangulaires. Dans nos analyses, nous nous plaçons dans un cadre large de "matrices rectangulaires" formalisé par l'hypothèse suivante.

Hypothèse 2 (Base rectangulaire) *Le nombre d'attributs m est au plus polynomial en le nombre d'objets n et réciproquement,*

$$\exists c > 0, \qquad \log n \sim c \log m$$

Les objets sont en pratique des observations qui sont décrites avec les attributs. A priori, ces observations peuvent ne pas être indépendantes mais nous ferons l'hypothèse simplificatrice suivante.

Hypothèse 3 (Indépendance des objets) *L'ensemble des objets forme une famille indépendante de variables aléatoires.*

Chaque ligne (ou objet) est indépendante des autres mais les colonnes (attributs) peuvent être corrélées. Les corrélations sont modélisées avec une source générale. Nous adoptons donc le point de vue du pattern matching.

Hypothèse 4 (Source) *Les objets, vus comme des mots sur l'alphabet $\{0, 1\}$, sont les préfixes d'une* même *source $\mathcal{S}$.*

Cette hypothèse n'est pas restrictive puisque dès que des symboles sont émis, nous pouvons les considérer comme le résultat d'une source inconnue. Ce sont les hypothèses sur la source $\mathcal{S}$ qui fixeront les limites du modèle.

Toutes les bases de données considérées ici satisfont ces trois hypothèses. Mise à part la seconde hypothèse, nous obtenons un modèle très large de bases de données dont les limites sont celles imposées par la source $\mathcal{S}$.

7.3.2 Trois types de seuils

Lorsque le seuil de fréquence γ passe de 1 à n (n étant le nombre d'objets), les expériences indiquent que le nombre de motifs fréquents est tout d'abord exponentiel puis décroît pour devenir constant. Pour démontrer ces résultats, nous considérons trois types de seuils : les seuils constants (hypothèse 5), les seuils logarithmiques (hypothèse 7) et les seuils proportionnels au nombre d'attributs (hypothèse 6).

Hypothèse 5 (γ constant) *Le seuil de fréquence γ est constant par rapport au nombre n d'objets.*

En pratique, un seuil constant signifie un seuil très petit vis-à-vis du nombre d'objets (lignes). D'un point de vue plus théorique, un seuil de fréquence constant entraîne que la contrainte *s'affaiblit* lorsque la taille de la base s'agrandit. Etre 20-fréquent dans une base de 50 objets est beaucoup moins probable (et donc beaucoup plus fort) qu'être 20-fréquent dans une base de 10000 objets.
Posons maintenant la deuxième hypothèse.

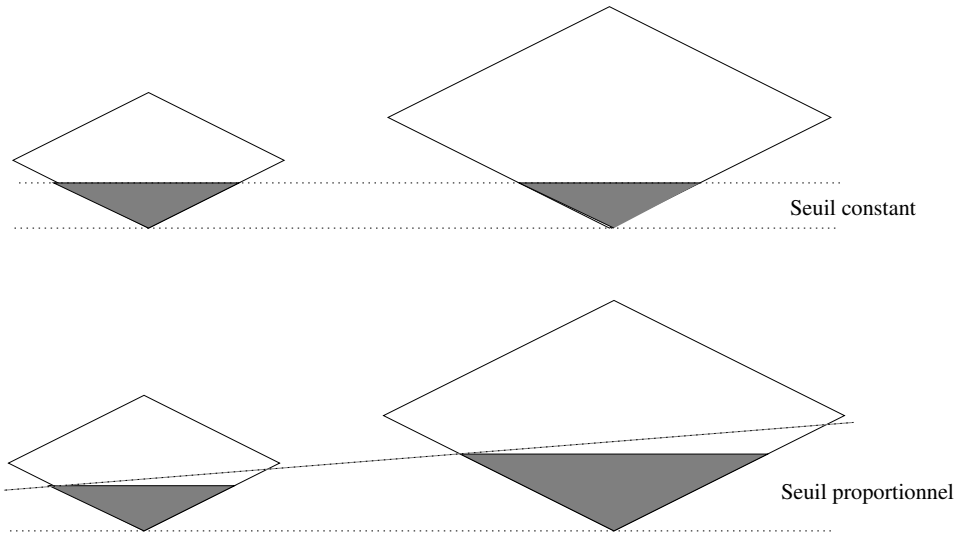


FIG. 7.4 – Différence entre le seuil linéaire et le seuil proportionnel. La partie grisée est la partie élaguée par la contrainte de fréquence.

Hypothèse 6 (γ proportionnel) *Le seuil de fréquence γ est proportionnel au nombre n d'objets et satisfait*

$$\gamma = r \cdot n, \qquad \textit{avec} \qquad r \in]0,1[.$$

Un seuil proportionnel (ou linéaire) signifie en pratique un seuil non négligeable (voir même assez élevé) comparé au nombre d'objets. D'un point de vue théorique, un seuil de fréquence proportionnel entraîne que la contrainte *ne s'affaiblit pas* lorsque la taille de la base s'agrandit. Si un motif est 5%-fréquent dans une base de 1000 objets, il sera certainement 5%-fréquent dans une base de 100000 objets et réciproquement. C'est sur ce principe qu'est fondée la méthode d'échantillonnage de Toïvonen [Toi96]. La figure 7.4 montre aussi que choisir un seuil proportionnel élague beaucoup plus le treillis des motifs qu'un seuil constant lorsque la base de données s'agrandit.
Nous terminons avec un nouveau type de seuil grâce auquel nous avons aussi mis en évidence un phénomène intéressant.

Hypothèse 7 (γ logarithmique) *Le seuil de fréquence γ est d'un ordre plus grand que le logarithme du nombre n d'objets,*

$$\log n = o(\gamma).$$

Sous l'hypothèse 2 de la base rectangulaire, on a $\log n \sim c \log m$ et l'hypothèse 7 revient à $\log m = o(\gamma)$.

7.3.3 Trois seuils, trois hypothèses, trois résultats

A chacun des seuils décrits précédemment, nous associons une hypothèse afin de mettre en évidence un résultat sur les motifs fréquents ou fermés. Pour un motif ligne $X \subset \{1, \ldots, m\}$, nous notons p_X la probabilité qu'il soit présent dans une ligne (ou dans un mot). Les hypothèses font intervenir ces probabilités.

7.3.3.1 Cas du seuil proportionnel

Afin d'exprimer le nombre moyen de motifs fréquents, nous posons $k(\mathcal{S}, r)$ la plus petite longueur de motif pour laquelle tous les motifs de longueur supérieure ont une probabilité strictement plus petite que r,

$$k(\mathcal{S}, r) = \max\{k : \ \forall X \subset \mathcal{A}, \ |X| > k, \ p_X < r\}.$$

L'hypothèse suivante implique une borne constante (en m et n) sur $k(\mathcal{S}, r)$.

Hypothèse 8 (Décroissance exponentielle des probabilités) *Les motifs lignes ont une probabilité exponentiellement décroissante avec leur taille. Plus précisément, il existe $K > 0$ et $\theta \in]0, 1[$ tels que pour tout motif ligne $X \subset \{1, \ldots, m\}$ de cardinal $|X|$, on a*

$$p_X \leq K \cdot \theta^{|X|}.$$

L'hypothèse de décroissance exponentielle des probabilités entraînent que les motifs longs sont très peu probables. Ils ont ainsi une faible probabilité d'être présents simultanément dans beaucoup d'objets. Pour un seuil de fréquence proportionnel, le théorème suivant confirme cette intuition.

Théorème 18 *Dans une base de données aléatoire rectangulaire (hypothèse 2) dont les objets sont générés indépendamment (hypothèse 3) à partir d'une source $\mathcal{S}$ (hypothèse 4) et dont les motifs lignes ont une probabilité exponentiellement décroissante avec leur taille (hypothèse 8) le nombre moyen de motifs γ-fréquents, noté Fr_γ, pour un seuil de fréquence proportionnel de type $\gamma = r \cdot n$ (hypothèse 6) est au plus polynomial en le nombre d'attributs et vérifie*

$$Fr_\gamma = O(m^{k(\mathcal{S},r)}).$$

Pour la première fois en extraction de motifs, un comportement polynomial du nombre de motifs fréquents est mis en évidence.
Nous pouvons relâcher l'hypothèse en considérant une condition de la forme

$$p_X \leq K\theta^{|X|}, \qquad \text{où} \qquad (1-\theta) \cdot \min(m, n) \to \infty.$$

Sous cette condition, l'asymptotique n'est plus nécessairement polynomiale mais elle est de la forme

$$Fr_\gamma = O(\sum_{j=1}^{k(\mathcal{S},r)} \binom{m}{j}).$$

Cette hypothèse plus faible réduit considérablement la contrainte mais les exemples classiques de sources entrent tous dans le cadre du théorème.

7.3.3.2 Cas du seuil logarithmique

Le seuil est dit logarithmique s'il est de la forme $\log n = o(\gamma)$. Nous proposons une hypothèse plus forte que la précédente pour traiter ce type de seuil.

Hypothèse 9 *Le rapport de probabilité entre un motif ligne et ses sur-motifs lignes est strictement plus petit que* 1,

$$\exists \theta \in]0, 1[, \quad \forall X \subset \{1, \ldots, m\}, \forall Y \subset \{1, \ldots, m\} \qquad X \subsetneq Y \Rightarrow \frac{p_Y}{p_X} \leq \theta,$$

Remarquons que si $p_Y/p_X < \theta$, alors les probabilités p_X sont exponentiellement décroissantes avec le cardinal de X et la source $\mathcal{S}$ satisfait aussi la condition 8 du théorème 18. En particulier sous la condition 9 et pour un seuil proportionnel, le nombre de motifs fermés est aussi polynomial en le nombre d'attributs (de colonnes). Pour un seuil logarithmique, les motifs fréquents sont aussi fermés à une proportion négligeable près.

Théorème 19 *Dans une base de données aléatoire rectangulaire (hypothèse 2) dont les objets sont générés indépendamment (hypothèse 3) à partir d'une source $\mathcal{S}$ (hypothèse 4) et dont les motifs lignes satisfont l'hypothèse 9, le nombre moyen Fr_γ de motifs γ-fréquents et le nombre moyen $Ferm_\gamma$ de motifs γ-fermés sont équivalents dès que le seuil de fréquence est logarithmique (hypothèse 7),*

$$Fr_\gamma \sim Ferm_\gamma, \qquad \textit{dès que} \qquad \log n = o(\gamma).$$

Comme précédemment, il est possible d'affaiblir l'hypothèse mais il faut alors augmenter le seuil logarithmique. De manière générale, si le seuil de fréquence vérifie

$$m\theta^\gamma \to 0, \qquad \text{lorsque} \qquad m, n \to \infty,$$

alors le nombre moyen de motifs γ-fréquents est équivalent au nombre moyen de motifs γ-fermés.

7.3.3.3 Cas du seuil fixe

L'hypothèse dans le cas du seuil fixe peut s'exprimer de deux manières complètement équivalentes. La première utilise les probabilités p_X des motifs lignes alors que la seconde se base sur une source construite à partir de la source initiale et d'une série génératrice associée.

L'hypothèse pour le seuil fixe fait intervenir les sommes $S_{\gamma,m}$ définies pour γ entier par

$$S_{\gamma,m} = \sum_{X \subset \{1,\dots,m\}} p_X^\gamma.$$

L'hypothèse suppose que la somme $S_{\gamma,m}$ est exponentiellement plus grande que la somme $S_{\gamma+1,m}$.

Hypothèse 10 (version 1) *Pour tout entier γ, il existe trois constantes κ_γ, λ_γ et θ_γ avec κ_γ positif, $\lambda_\gamma > 1$ et $\theta_\gamma \in]0,1[$ telles que*

$$S_{\gamma,m} = \kappa_\gamma \cdot \lambda_\gamma^m (1 + O(\theta_\gamma^m)) \quad \textit{et} \quad \lambda_\gamma > \lambda_{\gamma+1}.$$

L'inconvénient avec ce type d'hypothèse est qu'avant de pouvoir l'utiliser, il faut calculer les sommes $S_{\gamma,m}$. L'autre version de l'hypothèse part d'une formulation différente des sommes en fonction des probabilités des mots.
Dans toute la suite, le réel p_M désigne la probabilité que le mot M soit généré par la source $\mathcal{S}$. En particulier, nous avons la relation simple

$$p_X = \sum_{M : X \subset M} p_M.$$

En injectant l'égalité précédente dans les sommes S_γ puis en échangeant les signes sommes, les S_γ satisfont

$$\begin{aligned} S_{\gamma,m} &= \sum_{X\subset\{1,\dots,m\}} \sum_{\substack{M_1,\dots,M_\gamma \\ M_i\in\{0,1\}^m,\ X\subset M_i}} p_{M_1}\cdots p_{M_\gamma} \\ &= \sum_{M_1,\dots,M_\gamma} 2^{c_\gamma(M_1,\dots,M_\gamma)} p_{M_1}\cdots p_{M_\gamma} \end{aligned} \tag{7.1}$$

où $c_\gamma(M_1,\dots,M_\gamma)$ est le nombre de positions (de colonnes) où tous les M_i ont un 1. Par exemple $c_2(0110001, 1001001) = 1$ car la seule position commune où tous les mots ont un 1 est la dernière position.

$$\begin{array}{cccccc|c} 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{array}$$

Pour une base de données fixée, l'égalité 7.1 montre qu'il est suffisant de considérer tous les γ-uples d'objets. Pour γ fixé, le nombre de γ-uples est polynomial et d'ordre $O(n^\gamma)$. D'un point de vue probabiliste, l'égalité 7.1 est aussi l'espérance sur l'ensemble des γ-uples d'objets de la variable aléatoire 2^{c_γ}. En tirant au hasard un grand nombre de fois γ mots puis en calculant 2^{c_γ} pour ces mots, la loi des grands nombres donne un procédé simple pour estimer $S_{\gamma,m}$.

Nous ramenons maintenant l'étude de la variable c_γ à un problème de pattern matching connu : le nombre d'occurrences d'un motif dans un texte. A partir de γ copies $(\mathcal{S}_1,\dots,\mathcal{S}_\gamma)$ de la source $\mathcal{S}$, nous construisons une source $\underline{\mathcal{S}}_\gamma$ dont l'alphabet est $\{\underline{0},\underline{1}\}$ et qui suit le procédé de génération suivant : si toutes les sources $\mathcal{S}_1$ à $\mathcal{S}_\gamma$ émettent en même temps un 1, alors la source $\underline{\mathcal{S}}_\gamma$ émet un $\underline{1}$, sinon elle émet un $\underline{0}$. L'exemple ci-dessous donne le mot produit par $\underline{\mathcal{S}}_3$ à partir de $\gamma = 3$ copies de $\mathcal{S}$.

$$\begin{array}{lccccc} \mathcal{S}_1 : & 0 & 1 & 0 & 1 & 1 \quad \dots \\ \mathcal{S}_2 : & 1 & 1 & 0 & 0 & 1 \quad \dots \\ \mathcal{S}_3 : & 0 & 1 & 0 & 0 & 1 \quad \dots \\ & & & & & \\ \underline{\mathcal{S}}_3 : & \underline{0} & \underline{1} & \underline{0} & \underline{0} & \underline{1} \quad \dots \end{array}$$

On vérifie facilement que c_γ est exactement le nombre d'occurrences de $\underline{1}$ produit par la source $\underline{\mathcal{S}}_\gamma$. L'analyse en moyenne du nombre d'occurrences d'un motif fixé est un problème résolu pour toutes les sources simples connues [BV02, BV06]. En particulier, ce nombre suit une loi limite gaussienne d'espérance et de variance linéaires en la taille du texte. Pour démontrer ce résultat, l'outil principal est la série génératrice bivariée

$$S(z,w) = \sum_{M\in\{\underline{0},\underline{1}\}^\star} e^{wc_\gamma(M)} z^{|M|},$$

où $|M|$ est la longueur du mot M et $c_\gamma(M)$ est le nombre d'occurrences de $\underline{1}$ dans M. Notre intérêt se porte plutôt sur la variable 2^{c_γ} puisque $S_{\gamma,m}$ est l'espérance de 2^{c_γ} sur tous les mots de longueur m. Nous introduisons donc $S_\gamma(z)$ la série génératrice du premier moment de c_γ définie par

$$S_\gamma(z) = \sum_{M\in\{\underline{0},\underline{1}\}^\star} 2^{c_\gamma(M)} z^{|M|}. \tag{7.2}$$

Par construction, le coefficient de z^m dans $S_\gamma(z)$ est exactement $S_{\gamma,m}$. Il est bien connu que la nature et la position des singularités dominantes d'une série génératrice déterminent la croissance des coefficients. L'équivalent de l'hypothèse 10 porte sur les pôles de la série $S_\gamma(z)$.

Hypothèse 10 (version 2) *Pour γ un entier fixé, les séries $S_\gamma(z)$ et $S_{\gamma+1}(z)$ admettent un unique pôle simple dominant respectivement en $z = z_\gamma$ et $z = z_{\gamma+1}$ avec $z_\gamma < z_{\gamma+1} < 1$.*

Nous pouvons maintenant énoncer notre résultat pour un seuil de fréquence constant.

Théorème 20 (Seuil constant) *Dans une base de données aléatoire rectangulaire (hypothèse 2) dont les objets sont générés indépendamment (hypothèse 3) à partir d'une source $\mathcal{S}$ (hypothèse 4) et dont les sommes $S_{\gamma,m}$ vérifient l'hypothèse 10, le nombre moyen de motifs γ-fréquents, pour un seuil de fréquence γ constant (hypothèse 5), est exponentiel en le nombre d'attributs (colonnes) et polynomial en le nombre d'objets (lignes),*

$$Fr_\gamma = \kappa_\gamma \binom{n}{\gamma} \lambda_\gamma^m (1 + O(n\theta_\gamma^m))$$

où κ_γ est le résidu de $S_\gamma(z)$ en z_γ, $\theta_\gamma = z_\gamma / z_{\gamma+1} \in]0,1[$ et $\lambda_\gamma = 1/z_\gamma$.

Le théorème 20 présente un résultat asymptotique mais en fait, la preuve conduit à l'encadrement suivant pour Fr_γ :

$$\binom{n}{\gamma} S_{\gamma,m} - (n-\gamma)\binom{n}{\gamma} S_{\gamma+1,m} \leq Fr_\gamma \leq \binom{n}{\gamma} S_{\gamma,m}. \tag{7.3}$$

Les deux versions de l'hypothèse 10 entraînent que la somme $S_{\gamma,m}$ est exponentiellement plus grande que $S_{\gamma+1,m}$. Avec l'hypothèse de base rectangulaire, l'asymptotique se déduit facilement. Puisqu'il suffit que $S_{\gamma,m}$ soit exponentiellement plus grande que $S_{\gamma+1,m}$, il est possible d'affaiblir les hypothèses en supposant par exemple que la série $S_\gamma(z)$ a un disque de convergence de rayon r_γ avec $r_\gamma < r_{\gamma+1}$. On obtient alors l'équivalence

$$Fr_\gamma \sim \binom{n}{\gamma} S_{\gamma,m} \sim \kappa_{\gamma,m} (1/r_\gamma)^m$$

où $\limsup_m \kappa_{\gamma,m}^{1/m} = 1$.

7.4 Modèles applicables

Dans tous les théorèmes, nous n'avons pas décrit la source utilisée. Dans cette partie, nous donnons plusieurs types de sources pour lesquelles les théorèmes s'appliquent.

7.4.1 Modèle de Bernoulli

Le modèle de base de données de Bernoulli est un modèle de complète indépendance. Les objets ainsi que les attributs sont indépendants deux à deux alors que les bases réelles sont corrélées. Cependant, toutes les analyses en moyenne dont nous avons parlé dans l'introduction utilisent ce modèle. La raison en est certainement sa simplicité qui en fait un cadre idéal pour les analyses probabilistes.

Nous proposons plusieurs modèles de Bernoulli : un modèle *simple*, un modèle *groupé* qui tient compte des attributs continus qui sont binarisés, et un modèle *évolué* qui tient compte de la répartition de chaque attribut. Une source est dite de Bernoulli si elle émet un 1 avec une probabilité p et un 0 avec une probabilité $1-p$. Le modèle simple suppose que la source $\mathcal{S}$ est une source de Bernoulli.

Le modèle simple de Bernoulli suppose que chaque attribut appartient à un objet (une ligne) avec une probabilité $p \in]0,1[$ fixée et identique pour tous les attributs

En d'autres mots, si $\chi_{i,j}$ est la variable aléatoire à valeurs dans $\{0,1\}$ qui vaut 1 si et seulement si a_j est un attribut présent dans o_i, alors $\chi_{i,j}$ suit une loi de Bernoulli de paramètre p.
Pour le codage sous forme de graphe bipartite (resp. co-bipartite), une arête entre les deux ensembles de sommets existe avec une probabilité p (resp. $1-p$). C'est le modèle probabiliste classique de graphe $G(n,p)$ (resp. $G(n,1-p)$) adapté aux graphes bipartites (resp. co-bipartites).

Le modèle simple de Bernoulli satisfait les conditions des trois théorèmes. En effet, pour un motif ligne X, sa probabilité est la probabilité d'avoir des 1 aux positions pointées par X soit $p_X = p^{|X|}$. La décroissance exponentielle des probabilités se vérifie et le théorème 18 s'applique. Dans [LRS05a], nous obtenons même un équivalent du nombre de motifs fréquents qui est

$$Fr_\gamma \sim \frac{m^{k(\mathcal{S},r)}}{k(\mathcal{S},r)!} \qquad \text{si} \quad \gamma = r \cdot n.$$

De même, si Y est un sur-motif de X, alors $p_Y/p_X \leq p$ et le théorème 19 s'applique aussi. Finalement pour γ sources, la probabilité d'avoir une colonne de 1 (ou que la source $\underline{\mathcal{S}}_\gamma$ émette un $\underline{1}$) est p^γ. Nous posons $m(z,u) = z((1-p^\gamma) + up^\gamma)$ la série associée à l'ensemble $\{\underline{0},\underline{1}\}$ telle que z marque les symboles et u marque les $\underline{1}$. Maintenant, un mot est un élément de $\{\underline{0},\underline{1}\}^\star$ et en utilisant les dictionnaires sur les séries génératrices, nous avons

$$S(z,u) := \sum_{M \in \{\underline{0},\underline{1}\}^\star} z^{|M|} u^{c_\gamma(M)} = \frac{1}{1-m(z,u)}.$$

Pour $u=2$, on retrouve la série $S_\gamma(z)$ qui satisfait

$$S_\gamma(z) = \frac{1}{1-z(1+p^\gamma)}.$$

S_γ admet un unique pôle simple en $z = 1/(1+p^\gamma)$ et le théorème 20 montre que le nombre moyen de motif γ-fréquents pour γ fixe est

$$Fr_\gamma = \binom{n}{\gamma}(1+p^\gamma)^m(1+O(\theta^m)).$$

La proposition suivante résume les hypothèses satisfaites par les sources de Bernoulli.

Proposition 21 *Les sources de Bernoulli satisfont les hypothèses 8, 9 et 10 liées respectivement aux théorèmes 18, 19 et 20.*

7.4.2 Modèle de Bernoulli par groupes

Le modèle simple de Bernoulli ne tient pas compte des exclusions provenant d'attributs continus qui ont été découpés en plusieurs attributs binaires. Par exemple la taille, qui peut s'exprimer en centimètres, se décompose en trois attributs : grand, petit et moyen. Une personne ou un animal ne peut être à la fois petit et moyen et il ne peut être aucun des trois. C'est pour modéliser ce type de corrélations que nous avons introduit le modèle de Bernoulli par groupes. Ce modèle suppose que tous les attributs réels sont découpés en K attributs binaires ($K \geq 2$) équiprobables.

Soit K un entier, $K \geq 2$ et supposons que le nombre d'attributs m est un multiple de K, $m = K \times m_1$. On note $\chi_{i,j}$ la variable aléatoire à valeurs dans $\{0,1\}$ qui vaut 1 si et seulement si a_j est un attribut présent dans o_i. Le modèle de Bernoulli par groupes suppose que les groupes de K variables aléatoires de la forme $(\chi_{i,(j-1)K+1}, \ldots, \chi_{i,jK})$ n'admettent qu'un seul 1 positionné aléatoirement et uniformément.

Avec le modèle simple de Bernoulli, une plante pouvait être grande et petite à la fois ou ni grande, ni petite, ni moyenne. Avec le modèle par groupes, ces situations ne sont plus possibles. Une plante est nécessairement soit grande, soit moyenne, soit petite et de manière équiprobable. Il n'existe pas de traduction naturelle de ce modèle probabiliste pour les codages matriciel, par graphes ou hypergraphes. Notons toutefois que c'est la première fois qu'un modèle probabiliste de bases de données tenait compte d'exclusions locales (aussi simples soient elles) lorsque nous avons introduit ce modèle [LRS05a].

Dans ce modèle, la probabilité qu'un attribut appartiennent à un objet est $1/K$. Par conséquent, $p_X \leq (1/K)^{|X|}$ et si p_X est non nulle, pour tout sur-motif Y de X nous avons $p_Y/p_X \leq (1/K)$. Les deux conditions des deux premiers théorèmes sont vérifiées. Un équivalent est aussi donné dans [LRS05a].
Pour le dernier théorème, les attributs, ou de manière équivalente les colonnes, sont groupées par K. Considérons un de ces groupes. Pour γ copies de la source $\mathcal{S}$, la probabilité qu'il y ait une colonne de 1 dans le groupe est $(1/K)^{\gamma-1}$ et pour chaque groupe, la source $\underline{\mathcal{S}}_\gamma$ émet un $\underline{1}$ dans ce groupe avec la même probabilité. Nous posons $m(z,u) = z^K((1-(1/K)^{\gamma-1}) + u(1/K)^{\gamma-1})$ la série associée à l'ensemble des K-uples de $\{\underline{0},\underline{1}\}^K$ telle que z marque les ensembles de K symboles (tous les groupes possibles) et u marque les ensembles de K symboles contenant un $\underline{1}$. Un texte produit par $\underline{\mathcal{S}}_\gamma$ est de la forme $(\{\underline{0},\underline{1}\}^K)^\star$ et en utilisant les dictionnaires sur les séries génératrices, nous avons

$$S(z,u) := \sum_{M \in (\{\underline{0},\underline{1}\}^K)^\star} z^{|M|} u^{c_\gamma(M)} = \frac{1}{1-m(z,u)}.$$

Pour $u = 2$, on retrouve la série $S_\gamma(z)$ qui satisfait

$$S_\gamma(z) = \frac{1}{1 - z^K(1 + (1/K)^{\gamma-1})}$$

et $S_{\gamma,K\cdot m_1} = S_{\gamma,m}$ est de la forme $(1 + (1/K)^{\gamma-1})^{m/K}$. Le nombre moyen de motifs fréquents pour un seuil de fréquence fixe est donc

$$Fr_\gamma \sim \binom{n}{\gamma}(1 + (1/K)^{\gamma-1})^{m/K}$$

La proposition suivante résume les hypothèses satisfaites par les sources de Bernoulli par groupes.

Proposition 22 *Les sources de Bernoulli par groupes satisfont les hypothèses 8, 9 et 10 liées respectivement aux théorèmes 18, 19 et 20.*

7.4.3 Un modèle de Bernoulli évolué

Pour les deux modèles de Bernoulli précédents, les attributs ou groupes d'attributs suivaient la même distribution. Nous considérons maintenant un modèle plus évolué qui tient compte de la probabilité de présence de chaque attribut. La symétrie entre les attributs est ainsi supprimée.

Soit $(p_j)_{j\geq 1}$ une suite de réels de l'intervalle $]\theta, 1-\theta[$ pour un θ positif. Dans le modèle de Bernoulli évolué, l'attribut a_j appartient à un objet avec la probabilité p_j.

Avec ce modèle, nous retournons dans le cadre d'attributs indépendants mais dont la probabilité de présence est respectée. Comme les probabilités sont toutes à distance au moins θ de 1, les hypothèses 8 et 9 se vérifient facilement.
En revanche, la troisième hypothèse n'est pas toujours vérifiée (cela dépend de la suite $(p_j)_j$). Mais la somme $S_{\gamma+1,m}$ reste exponentiellement plus petite que $S_{\gamma+1,m}$. En effet, la probabilité que la colonne j soit une colonne de 1 est p_j^γ et en utilisant la définition de $S_{\gamma,m}$, nous obtenons

$$S_{\gamma,m} = \sum_{k=1}^{m} \sum_{\{i_1,\ldots,i_k\}\subset\mathcal{A}} p_{i_1}^\gamma \ldots p_{i_k}^\gamma = \prod_{j=1}^{m}(1+p_j^\gamma) - 1.$$

Comme les probabilités $p_j \in]\theta, 1-\theta[$ sont à la fois distantes de 0 et de 1, les sommes $S_{\gamma,m}$ satisfont (asymptotiquement)

$$\frac{S_{\gamma+1,m}}{S_{\gamma,m}} \leq \left(1 - \frac{\theta(1-\theta)^\gamma}{(1+\theta)^\gamma}\right)^m$$

et en considérant l'encadrement 7.3, nous obtenons l'asymptotique suivante,

$$Fr_\gamma \sim \binom{n}{\gamma} S_{\gamma,m} \sim \binom{n}{\gamma} \prod_{j=1}^{m}(1+p_j^\gamma).$$

Proposition 23 *Les sources de Bernoulli par groupes satisfont les hypothèses 8 et 9 liées respectivement aux théorèmes 18 et 19. Pour un seuil fixe, l'hypothèse 10 n'est pas vérifiée pour toute suite $(p_j)_{j\geq 1}$. Mais, un équivalent du nombre de motifs fréquents est donné par*

$$Fr_\gamma \sim \binom{n}{\gamma} \prod_{j=1}^{m}(1+p_j^\gamma).$$

Les expériences (cf. section 7.5) montrent que le modèle de Bernoulli évolué conduit à une *bonne* estimation du nombre de motifs fréquents. Nous supposons que ce modèle s'étend aussi à un modèle par groupes dont chaque groupe suivrait un modèle de Bernoulli adapté (mais nous n'avons pas [encore] effectué les vérifications).

7.4.4 Chaîne de Markov

Avec les modèles de Bernoulli, les attributs (ou groupes d'attributs) sont indépendants les uns des autres. Avec les chaînes de Markov, il est possible d'ajouter des corrélations entre les attributs proches. Notre modèle de Markov introduit dans [LRS05b] suppose que les objets forment une famille indépendante de chaînes de Markov.

Définition 23 (Chaîne de Markov) *Une famille $(\chi_i)_{i\geq 1}$ de variables aléatoires est une chaîne de Markov d'ordre $K \in \mathbb{N}^\star$, si pour tout $i \geq 1$,*

$$\begin{aligned}&\mathbb{P}[\chi_{i+K} = x_{i+K} \mid \chi_{i+K-1} = x_{i+K-1}, \ldots, \chi_1 = x_1]\\ &\quad = \mathbb{P}[\chi_{i+K} = x_{i+K} \mid \chi_{i+K-1} = x_{i+K-1}, \ldots, \chi_i = x_i]\\ &\quad = \mathbb{P}[\chi_{K+1} = x_{i+K} \mid \chi_K = x_{i+K-1}, \ldots, \chi_1 = x_i]\end{aligned}$$

Les formules précédentes montre que pour une chaîne de Markov d'ordre K, la valeur d'une variable aléatoire ne dépend que de la valeur des K précédentes variables aléatoires et que cette dépendance est invariante avec le temps.
Une chaîne de Markov est complètement décrite par la distribution des K premières variables aléatoires et les probabilités de transition. Nous noterons toujours $f = (f_w)_{w\in\{0,1\}^K}$ la distribution initiale des K premières variables aléatoires et $p_{x|w}$ la probabilité que le nouveau symbole soit $x \in \{0,1\}$ sachant que les K précédents sont $w \in \{0,1\}^K$. La paire constituée de la distribution initiale et des probabilités de transitions forment une source également appelée chaîne de Markov. Le modèle de Markov suppose que la source $\mathcal{S}$ est une chaîne de Markov.

Modèle de Markov d'ordre K : soit $K \geq 1$, $f = (f_w)_{w\in\{0,1\}^K}$ la distribution initiale et $(p_{x|w})_{w\in\{0,1\}^K, x\in\{0,1\}}$ les probabilités de transitions toutes strictement positives d'une chaîne de Markov d'ordre K sur $\{0,1\}$. La source $\mathcal{S}$ est la chaîne de Markov associée à la paire $((p_{x|w}), f)$ qui construit les objets de longueur m avec le procédé suivant : pour un objet $o = (\chi_1, \ldots, \chi_m)$, la valeur des K premières variables est calculée suivant la distribution initiale f. Puis les valeurs de $\chi_{K+1}, \ldots, \chi_m$ sont séquentiellement obtenues à l'aide des K précédentes valeurs en utilisant les probabilités de transition $(p_{x|w})_{w\in\{0,1\}^K, x\in\{0,1\}}$.

Une chaîne de Markov d'ordre K sur $\{0,1\}$ est équivalent à une chaîne de Markov d'ordre 1 sur $\{0,1\}^K$. La densité initiale reste la même mais les probabilités de transition $p_{w_2|w_1}$ se font entre deux mots de longueur K, $w_1 = a_1 \ldots a_K$ et $w_2 = a_{K+1} \ldots a_{2K}$, et satisfont,

$$p_{w_2|w_1} = \prod_{i=K+1}^{2K} p_{a_i|a_{i-1}\ldots a_{i-K}}.$$

De ce point de vue, la source $\mathcal{S}$ émet des groupes de K symboles. La condition *toutes strictement positives* assure que la chaîne de Markov est irréductible et apériodique. Ces deux conditions impliquent un bon mélange des symboles et limitent les corrélations. Toutefois, nous supposons que les conditions usuelles d'irréductibilité et d'apériodicité sont suffisantes pour obtenir les mêmes résultats.

Les chaînes de Markov sont très utilisées pour approcher des phénomènes réels. Par exemple, des expériences menées sur l'anglais ont montré qu'avec un ordre K relativement faible, un processus de Markov génère des phrases parfois grammaticalement correctes [Sha01]. L'anglais possède toutefois une structure que n'ont pas nécessairement les bases de données.

Nous avons montré [LRS05b] que les chaînes de Markov satisfont les deux conditions 8 et 10 des théorèmes 18 et 20.

Les preuves dans le modèle de Markov ne s'expriment malheureusement pas aussi simplement que dans les modèles de Bernoulli. Nous ne donnons donc pas ici ces preuves. Cependant, les chaînes de Markov sont des cas particuliers de sources dynamiques que nous présentons en détail au prochain chapitre. Les sources dynamiques sont basées sur des systèmes dynamiques et nous montrerons qu'elles satisfont les hypothèses 8, 9 et 10. En particulier, pour les chaînes de Markov (avec une matrice de transition et une densité initiale strictement positives), nous obtenons la proposition suivante.

Proposition 24 *Les chaînes de Markov ayant une matrice de transition et une densité initiale strictement positives satisfont les hypothèses 8, 9 et 10 liées respectivement aux théorèmes 18, 19 et 20.*

Une matrice de transition strictement positive implique une propriété d'irréductibilité forte. Si l'on suppose la chaîne simplement irréductible (i.e. il existe un entier k tel que la matrice élevée à la puissance k est strictement positive), alors la chaîne vérifie les hypothèses 8 et 10. Dans ce cas, nous ne savons pas démontrer la condition 9.

7.5 Expériences

Cette section présente les expériences que nous avons réalisé avec les bases de données classiques issues du workshop FIMI (Frequent Itemset Mining Implementations, http ://fimi.cs.helsinki.fi/).

Les graphiques de la figure 7.5 représentent le nombre de motifs fréquents en fonction du seuil de fréquence. Le trait plein représente le nombre de motifs réellement fréquents dans la base initiale. Le trait pointillé le plus proche du trait plein correspond au calcul (exact et non asymptotique) obtenu avec le modèle de Bernoulli évolué. Enfin, le trait discontinu en forme d'escalier correspond au modèle de Bernoulli simple. Les tests ont été effectués sur une base de données synthétique T10I4D100K et deux bases réelles Chess et Mushroom. Dans tous les cas, le modèle de Bernoulli évolué conduit à une *bonne* estimation du nombre de motifs fréquents (attention à l'échelle logarithmique) alors que le modèle simple de Bernoulli est très rapidement éloigné de la réalité. La raison en est que tous les attributs n'ont pas du tout la même fréquence alors que le modèle de Bernoulli (simple) implique qu'elles sont proches. S'il n'est pas étonnant que dans les bases synthétiques, le modèle évolué donne de bons résultats (il est bien connu que les attributs sont peu corrélés entre eux), il l'est beaucoup plus pour les bases réelles. La forme en escalier du modèle simple s'explique aussi très facilement. Dans [LRS05a], nous avons montré que pour un seuil de fréquence proportionnel $\gamma = r \cdot n$, le nombre de motif γ-fréquents est équivalent à $\binom{m}{j}$ si $p^{j+1} < r < p^j$. Donc si r est dans l'intervalle $]p^{j+1}, p^j[$, il y a asymptotiquement un nombre constant de motifs fréquents.

Il est très difficile d'évaluer la croissance du nombre de motifs fréquents dans les bases réelles puisque l'on ne dispose pas de procédé pour les générer. En revanche, il est possible de tester l'équivalence entre les motifs fréquents et les motifs fermés et affirmé par le théorème 19 page 167 (sous certaines conditions). Cette fois, nous avons utilisé deux bases synthétiques, T10I4D100K et T40I10D100K, ainsi que trois bases réelles Chess, Mushroom et Connect. Les résultats sont présentés à la figure 7.6. Pour les bases synthétiques, l'équivalence entre le nombre de motifs fréquents et fermés se constate pour un seuil de fréquence très bas (20 sur 100000 pour T10I4D100K et 400/100000 pour T40I10D100K). En revanche, l'équivalence n'a jamais lieu pour les bases

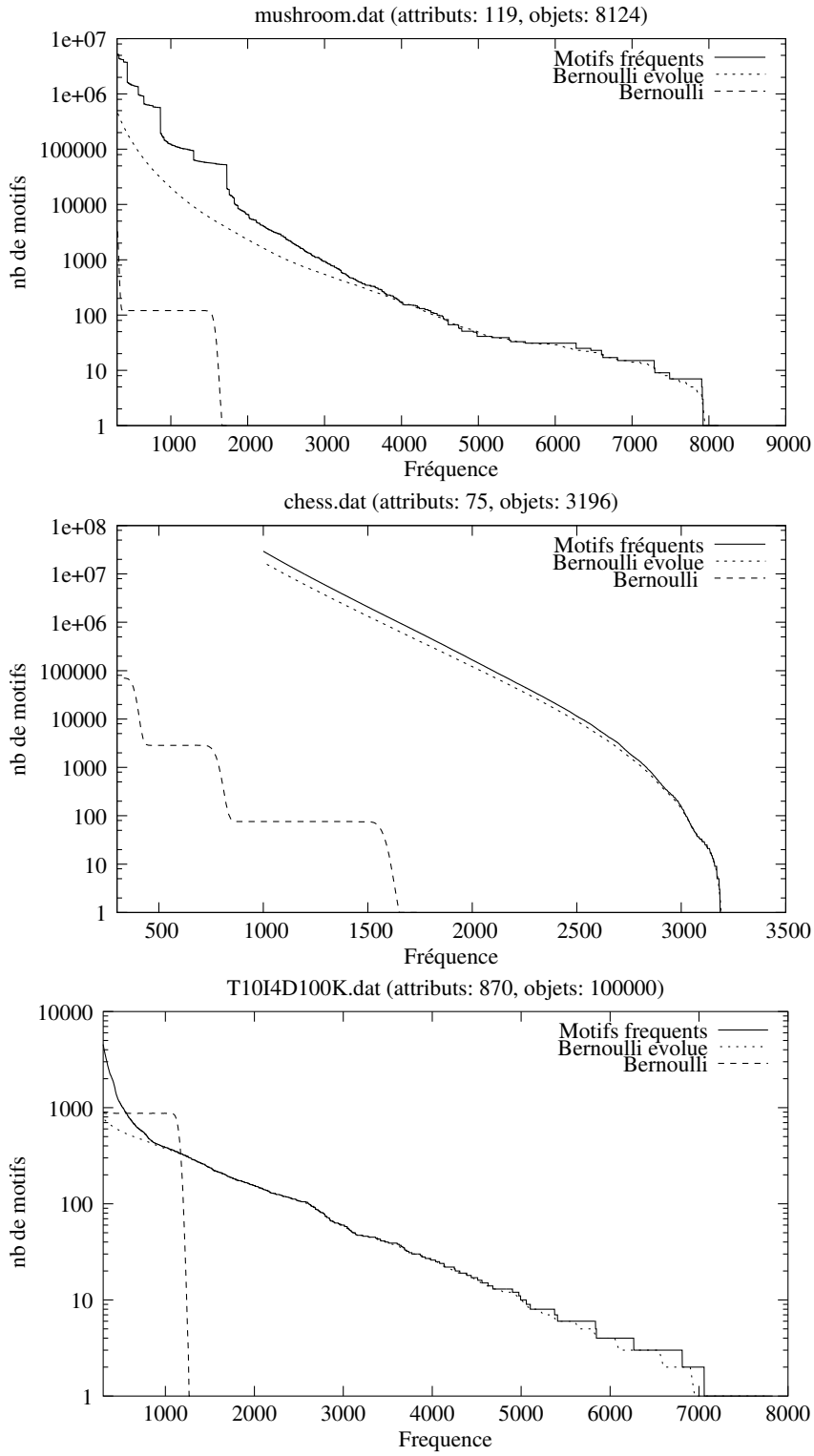


FIG. 7.5 – Comparaison entre le nombre de motifs fréquents présents réellement, le nombre de motifs fréquents théoriques issu de modèle de Bernoulli et le nombre de motifs fréquents théoriques issu de modèle de Bernoulli évolué.

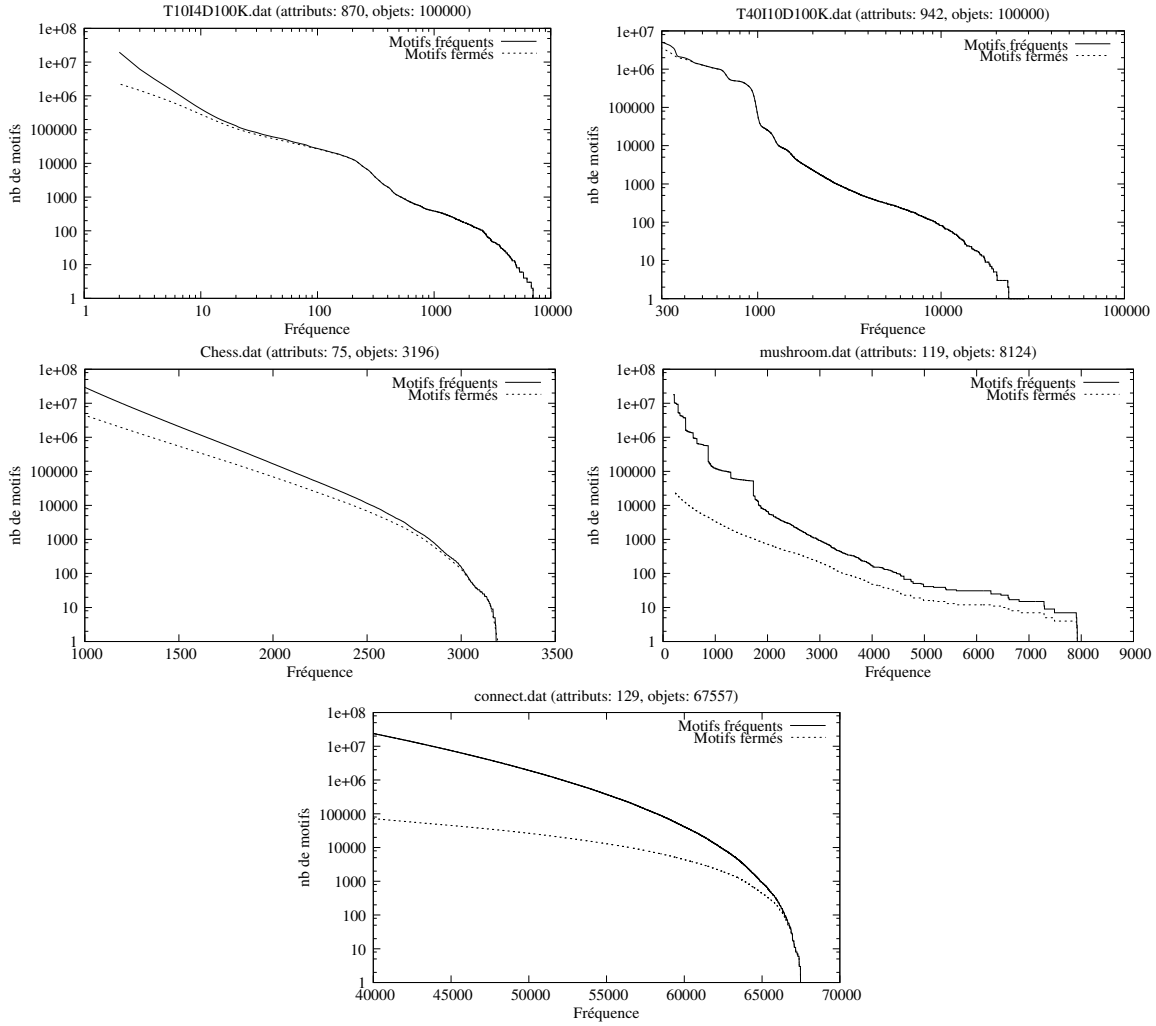


FIG. 7.6 – Equivalence entre le nombre de motifs fréquents et le nombre de motifs fermés.

réelles. La raison en est que certains attributs ont une probabilité de présence très proche de 1 pour ces bases ce qui contredit l'hypothèse du théorème 19.

7.6 Preuves des trois théorèmes

7.6.1 Démarche générale

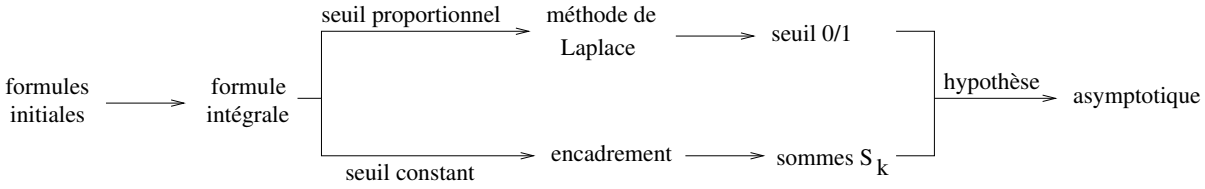
Notre démarche pour démontrer les trois résultats suit quatre étapes. La première étape consiste à trouver deux formules générales pour le nombre moyen de motifs fréquents et le nombre moyen de motifs fermés et qui sont valables pour toute base de données satisfaisant les trois hypothèses de base rectangulaire (hypothèse 2) dont les objets sont indépendants (hypothèse 3) et générés à partir d'une même source $\mathcal{S}$ (hypothèse 4). A partir des deux formules, nous prouvons le théorème 19 sur l'équivalence entre le nombre de motifs fréquents et le nombre de motifs fermés.

L'objectif de la deuxième étape est de transformer la formule initiale pour les motifs fréquents

en une formule intégrale. C'est une étape simple mais la nouvelle expression intégrale est l'élément essentiel pour les démonstrations.

Une valeur exacte de l'intégrale est certainement très difficile à obtenir sauf dans des cas simples. L'objectif de la troisième étape est d'approcher au mieux cette intégrale qui dépend notamment du seuil de fréquence γ. La manière d'approcher ne sera pas la même si le seuil est constant ou proportionnel. Pour un seuil proportionnel, nous utilisons une méthode de Laplace qui donne une asymptotique de l'intégrale. Pour un seuil constant, nous encadrons la fonction intégrée et trouvons un encadrement de l'intégrale.

La dernière étape consiste à utiliser les estimations de l'intégrale pour trouver les asymptotiques du nombre moyen de motifs fréquents. Pour un seuil proportionnel, la méthode de Laplace montre que l'intégrale passe brusquement de 1 à 0 et combinée avec l'hypothèse du théorème 18, nous obtenons quasi-immédiatement le résultat. Pour un seuil constant, l'encadrement de l'intégrale conduit à estimer les sommes $S_{\gamma,m}$ et $S_{\gamma+1,m}$ puis à les comparer. C'est ici qu'intervient l'hypothèse du théorème 20. La figure ci-dessous résume les étapes successives.



7.6.2 Formules de départ

Dans toute la suite, nous nous plaçons dans le cadre d'une base de données binaire rectangulaire (hypothèse 2) dont les objets sont indépendants (hypothèse 3) et générés à partir d'une même source $\mathcal{S}$ (hypothèse 4). Nous supposons de plus que la base admet m attributs $\mathcal{A} = \{a_1, \ldots, a_m\}$ et n objets $\mathcal{O} = \{o_1, \ldots, o_n\}$ et nous fixons γ un seuil de fréquence.

Nous rappelons qu'un motif ligne X est un sous-ensemble de l'ensemble $E_m = \{1, \ldots, m\}$ et qu'il naturellement associé à un motif d'attributs dont les attributs sont ceux d'indices présents dans X. Par abus de notation, nous noterons aussi X ce motif d'attributs. La probabilité du motif X est noté p_X et pour une ligne L productible par la source, nous posons p_L sa probabilité.

Le point de départ de toutes nos preuves est le lemme simple suivant.

Lemme 15 *Le nombre moyen Fr_γ de motifs γ-fréquents dans une base de données aléatoire à m objets et n attributs, dons les objets sont générés indépendamment (hypothèse 3) par une même source (hypothèse 4) vérifie*

$$Fr_\gamma = \sum_{X \subset \mathcal{A}} \sum_{i=\gamma}^{n} \binom{n}{i} p_X^i (1 - p_X)^{n-i} \tag{7.4}$$

Sous les mêmes hypothèses, le nombre moyen $Ferm_\gamma$ de motifs γ-fermés satisfait

$$Ferm_\gamma = \sum_{X \subset \mathcal{A}} \sum_{i=\gamma}^{n} \binom{n}{i} (1 - p_X)^{n-i} \sum_{X \subseteq Y \subseteq \mathcal{A}} (-1)^{|Y \setminus X|} p_Y^i \tag{7.5}$$

Preuve du lemme. Soit X un motif γ-fréquent de fréquence i. Alors $i \geq \gamma$ et le nombre de supports possibles pour X est $\binom{n}{i}$. Maintenant étant donné i objets, la probabilité que X appartiennent à ces i objets est p_X^i et la probabilité qu'il n'appartient pas aux $n-i$ objets restants est $(1-p_X)^{n-i}$.

En sommant sur tous les motifs et tous les i on retrouve la formule 7.4.
Soit X est un motif γ-fermé de fréquence i. Alors $i \geq \gamma$ et le nombre de supports possibles est $\binom{n}{i}$. Étant donné i objets, la probabilité que X n'appartient pas aux $n-i$ objets restants est toujours $(1-p_X)^{n-i}$. Maintenant comme X est fermé, X appartient aux i objets formant son support et aucun sur-motif Y n'a le même support. Finalement, si $o_1, \ldots, o_i$ sont i objets, la probabilité que X soit exactement dans l'intersection des i objets est donnée par

$$\mathbb{P}\left(X = \bigcap_{k=1}^{i} o_k\right) = \mathbb{P}\left(\left[X \subset \bigcap_{k=1}^{i} o_k\right] \setminus \bigcup_{a \in \mathcal{A}\setminus X} \left[Xa \subset \bigcap_{k=1}^{i} o_k\right]\right).$$

De plus, nous avons clairement l'inclusion

$$\bigcup_{a \in \mathcal{A}\setminus X} \left[Xa \subset \bigcap_{k=1}^{i} o_k\right] \subseteq \left[X \subset \bigcap_{k=1}^{i} o_k\right]$$

ce qui entraîne la deuxième relation

$$\mathbb{P}\left(X = \bigcap_{k=1}^{i} o_k\right) = \mathbb{P}\left(\left[X \subset \bigcap_{k=1}^{i} o_k\right]\right) - \mathbb{P}\left(\bigcup_{a \in \mathcal{A}\setminus X} \left[Xa \subset \bigcap_{k=1}^{i} o_k\right]\right).$$

La première probabilité dans le membre de droite est donnée par p_X^i. Le principe d'inclusion-exclusion et l'indépendance des objets conduisent également à l'égalité

$$\mathbb{P}\left(\bigcup_{a \in \mathcal{A}\setminus X} \left[Xa \subset \bigcap_{k=1}^{i} o_k\right]\right) = \sum_{X \subsetneq Y \subseteq \mathcal{A}} (-1)^{|Y\setminus X|+1} p_Y^i.$$

En combinant les deux dernières identités, nous trouvons

$$\mathbb{P}\left(X = \bigcap_{k=1}^{i} o_k\right) = \sum_{X \subseteq Y \subseteq \mathcal{A}} (-1)^{|Y\setminus X|} p_Y^i.$$

La formule 7.5 découle immédiatement de ces relations. ■

Nous pouvons dès à présent démontrer le théorème 19 page 167 concernant le seuil logarithmique. Le théorème est basé sur l'hypothèse 9 qui suppose que la probabilité d'un sur-motif Y du motif X satisfait $p_Y/p_X < \theta$ avec $\theta < 1$. Sous cette hypothèse, le théorème annonce que le nombre de motifs γ-fréquents est équivalent à celui des γ-fermés.
preuve du théorème 19. Pour démontrer l'équivalence, il suffit de montrer que pour tout $i \geq \gamma$, la dernière somme est équivalente à p_X^i. Dans la succession d'égalités présentes dans la preuve de la formule 7.5, nous avons en particulier utilisé l'égalité

$$\sum_{X \subseteq Y \subseteq \mathcal{A}} (-1)^{|Y\setminus X|} p_Y^i = p_X^i - \mathbb{P}\left(\bigcup_{a \in \mathcal{A}\setminus X} \left[Xa \subset \bigcap_{k=1}^{i} o_k\right]\right). \tag{7.6}$$

Nous utilisons maintenant l'hypothèse 9 pour borner la probabilité à droite,

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{a \in \mathcal{A}\setminus X} \left[Xa \subset \bigcap_{k=1}^{i} o_k\right]\right) &\leq \sum_{a \in \mathcal{A}\setminus X} \mathbb{P}\left(Xa \subset \bigcap_{k=1}^{i} o_k\right) \\
&= \sum_{a \in \mathcal{A}\setminus X} p_{Xa}^i \qquad \text{(indépendance des objets)} \\
&= m\theta^i p_X^i \qquad \text{(hypothèse 9)}.
\end{aligned}$$

L'hypothèse 7 du seuil logarithmique ($\gamma = o(\log n)$) et l'hypothèse 2 d'une base rectangulaire ($\log n \sim c \log m$) impliquent que $\log m = o(\gamma)$, soit

$$m\theta^i \leq m\theta^\gamma = \exp(\log m + \gamma \log \theta) \to 0.$$

L'égalité 7.6 s'écrit alors

$$\sum_{X \subseteq Y \subseteq \mathcal{A}} (-1)^{|Y \setminus X|} p_Y^i = p_X^i (1 + O(m\theta^\gamma))$$

ce qui montre l'équivalence attendue.■

7.6.3 Formule intégrale

A ce stade, nous avons démontré le seul résultat concernant les motif fermés. Nous nous concentrons maintenant sur les motifs fréquents. Dans cette section, nous montrons que le nombre moyen de motifs fréquents satisfait une formule alternative faisant intervenir une intégrale. Le point de départ est le lemme suivant.

Lemme 16 *L'égalité suivante est satisfaite,*

$$\sum_{i=\gamma}^{n} \binom{n}{i} x^i (1-x)^{n-i} = \gamma \binom{n}{\gamma} \int_0^x t^{\gamma-1} (1-t)^{n-\gamma} dt.$$

En admettant le lemme, le nombre de motifs γ-fréquents dans une base de données dont les objets sont indépendants (hypothèse 3) et générés par une même source (hypothèse 4) vérifie

$$Fr_\gamma = \gamma \binom{n}{\gamma} \sum_{X \subset \mathcal{A}} \int_0^{p_X} t^{\gamma-1} (1-t)^{n-\gamma} dt. \tag{7.7}$$

Preuve du lemme 16.

Commençons par développer $(1-x)^{n-i}$. Nous obtenons,

$$\sum_{i=\gamma}^{n} \binom{n}{i} x^i (1-x)^{n-i} = \sum_{i=\gamma}^{n} \binom{n}{i} \sum_{u=0}^{n-i} \binom{n-i}{u} (-1)^u x^{i+u}.$$

Maintenant, le changement de variable $v = u + i$ et l'inversion des deux sommes conduisent à la nouvelle égalité,

$$\sum_{i=\gamma}^{n} \binom{n}{i} x^i (1-x)^{n-i} = \sum_{v=\gamma} \binom{n}{v} x^v \sum_{i=\gamma}^{v} \binom{v}{i} (-1)^{v-i}. \tag{7.8}$$

Un raisonnement par récurrence montre que la seconde somme se simplifie en

$$\sum_{i=\gamma}^{v} \binom{v}{i} (-1)^{v-i} = \binom{v}{i} (-1)^{v-i} = (-1)^{v-\gamma} \binom{v-1}{\gamma-1},$$

et l'égalité 7.8 vérifie alors l'identité

$$\sum_{i=\gamma}^{n} \binom{n}{i} x^i (1-x)^{n-i} = \sum_{v=\gamma} \binom{n}{v} x^v (-1)^{v-\gamma} \binom{v-1}{\gamma-1}.$$

Avec la relation $\binom{n}{v}\binom{v-1}{\gamma-1} = \frac{\gamma}{v}\binom{n}{\gamma}\binom{n-\gamma}{v-\gamma}$ et le changement de variable $w = v-\gamma$, l'égalité précédente devient

$$\sum_{i=\gamma}^{n} \binom{n}{i} x^i (1-x)^{n-i} = \gamma \binom{n}{\gamma} \sum_{w=0}^{n-\gamma} \binom{n-\gamma}{w} (-1)^w \frac{x^{w+\gamma}}{w+\gamma}.$$

Pour $x = 0$, la seconde somme vaut zéro et la dérivée par rapport à x est exactement $x^{\gamma-1}(1-x)^{n-\gamma}$. Il suffit donc d'intégrer $x^{\gamma-1}(1-x)^{n-\gamma}$ à partir de 0 pour retrouver la somme initiale. ■

Remarquons au passage que l'intégrale de 0 à 1 du lemme 16 est reliée par la relation suivante à la fonction Γ,

$$\int_0^1 t^{x-1}(1-t)^{y-1}dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}. \tag{7.9}$$

En particulier, pour $x = \gamma$ et $y = n - \gamma + 1$, nous obtenons

$$\int_0^1 t^{\gamma-1}(1-t)^{n-\gamma}dt = \frac{1}{\gamma\binom{n}{\gamma}}.$$

7.6.4 Cas du seuil proportionnel $\gamma = r \cdot n$

Nous abordons maintenant la troisième étape c'est-à-dire estimer l'intégrale de l'équation 7.7. Dans le cas du seuil proportionnel, nous appliquons les mêmes idées que la méthode de Laplace à savoir que nous identifions le maximum de la fonction intégrée et nous montrons que toute intégrale dont le domaine d'intégration fait intervenir ce maximum est équivalente à l'intégrale sur le domaine complet. Le lemme suivant résume ce que nous obtenons.

Lemme 17 (Méthode de Laplace) *Soit $x \in [0,1]$ et ϵ tel que $\epsilon < \min\{r, 1-r\}$. Alors*

$$\gamma\binom{n}{\gamma}\int_0^x t^{\gamma-1}(1-t)^{n-\gamma}dt = \begin{cases} 1 & si\ x = 1 \\ O(w_{r,n}(x)) & si\ x < r-\epsilon \\ 1 - O(w_{r,n}(x)) & si\ x > r+\epsilon \end{cases}$$

où le terme d'erreur dans le O est uniforme en x et la fonction $w_{r,n}$ satisfait

$$w_{r,n}(x) = \frac{r\sqrt{n}}{\sqrt{2\pi r(1-r)}}\widetilde{w}_r^n(x) \quad et \quad \widetilde{w}_r(x) = \left(\frac{x}{r}\right)^r \left(\frac{1-x}{1-r}\right)^{1-r}.$$

La fonction $\widetilde{w}_r$ est croissante sur $]0,r[$ puis décroissante sur $]r,1[$ et admet 1 comme maximum en r. En particulier, si $|x-r| > \epsilon$ alors

$$w_{r,n}(x) = O(\theta^n) \quad avec \quad \theta = \max\{\widetilde{w}_r(r-\epsilon/2), \widetilde{w}_r(r+\epsilon/2)\} < 1.$$

Le lemme montre que l'intégrale admet un effet de seuil. Si la probabilité du motif X est plus grande qu'un certain $r+\epsilon$, alors l'intégrale est proche de 1 et le motif X aura un poids proche de 1 dans la valeur de Fr_γ. Réciproquement, Si la probabilité du motif X est plus petite qu'un certain $r-\epsilon$, alors l'intégrale est proche de 0 et le motif X aura un poids proche de 0 dans la valeur de Fr_γ. Cependant, il n'est pas inconcevable que la somme de tous les motifs ayant un poids faible soit non négligeable, mais la décroissance exponentielle des probabilités (hypothèse 8 page 166) empêche ce phénomène.
En admettant le lemme 17, nous pouvons démontrer le théorème 18.
Preuve théorème 18. Selon l'hypothèse 8 de décroissance exponentielle des probabilités, il existe

$j_0 \geq 1$ tel que pour tout motif X de longueur plus grande que j_0, on a $p_X \leq M\theta^{|X|} < r - \epsilon$ (pour un epsilon positif). En décomposant la somme 7.7 entre les éléments de longueur plus grande que j_0 et les éléments de longueur plus petite que j_0, puis en majorant l'intégrale par 1 pour les motifs de longueur au plus j_0, nous obtenons la majoration

$$Fr_\gamma \leq \sum_{j=1}^{j_0} \binom{m}{j} + \gamma \binom{n}{\gamma} \sum_{X \subset \mathcal{A}: |X| > j_0} \int_0^{p_X} t^{\gamma-1}(1-t)^{n-\gamma} dt.$$

Cette fois en appliquant le lemme 17, la seconde somme est bornée par

$$\begin{aligned} \gamma \binom{n}{\gamma} \sum_{X \subset \mathcal{A}: |X| > j_0} \int_0^{p_X} t^{\gamma-1}(1-t)^{n-\gamma} dt &\leq \gamma \binom{n}{\gamma} \sum_{j=j_0+1}^{m} \binom{m}{j} \int_0^{M\theta^j} t^{\gamma-1}(1-t)^{n-\gamma} dt \\ &= O\Big(\sum_{j=j_0+1}^{m} \binom{m}{j} w_{r,n}(M\theta^j) \Big) \end{aligned}$$

(puisque le terme d'erreur dans le O du lemme 17 est uniforme). On pose $v_j = \binom{m}{j} w_{r,n}(Mp^j)$. Le rapport v_{j+1}/v_j satisfait

$$\frac{v_{j+1}}{v_j} = \frac{m-j}{j+1}\theta^\gamma \left(1 + M\theta^j \frac{1-\theta}{1-M\theta^j}\right)^{n-\gamma} := \frac{m-j}{j+1} \exp(n \cdot f_n)$$

où f_n est définie par

$$f_n = r \log \theta + (1-r) \log(1 + M\theta^j \frac{1-\theta}{1-M\theta^j}).$$

Maintenant pour tout $x > -1$, la fonction $\log(1+x)$ est majorée par x et en utilisant le fait que $M\theta^j < r - \epsilon$, nous obtenons la majoration suivante pour f_n,

$$f_n \leq (1-\theta)(-r + (1-r)\frac{r-\epsilon}{1-r+\epsilon}) < 0.$$

Avec l'hypothèse sur les bases rectangulaires (hypothèse 2), ceci montre que v_{j+1}/v_j tend vers 0 pour tout $j > j_0$ lorsque m et n tendent vers l'infini. La somme sur les éléments plus grands que j_0 suit alors l'équivalence,

$$\sum_{j=j_0+1}^{m} \binom{m}{j} w_{r,n}(Mp^j) \sim \binom{m}{j_0+1} w_{r,n}(Mp^{j_0+1}).$$

Le nombre moyen de motifs γ-fréquents vérifie alors

$$\begin{aligned} Fr_\gamma &= O\left(\sum_{j=1}^{j_0} \binom{m}{j} + \binom{m}{j_0+1} w_{r,n}(Mp^{j_0+1}) \right) \\ &= O\left(\binom{m}{j_0} (1 + m w_{r,n}(Mp^{j_0+1})) \right) \\ &= O(m^{j_0}(1 + m\theta^n)) = O(m^{j_0}), \end{aligned}$$

car $\binom{m}{j_0} \sim m^{j_0}/j_0!$ et avec l'hypothèse 2 des bases rectangulaires, $m^{j_0+1}\theta^n = o(1)$. Ceci termine la preuve du théorème 20. ■

Il reste maintenant à démontrer le lemme 17 pour compléter l'analyse dans le cas du seuil de fréquence proportionnel.

Preuve lemme 17. Rappelons que le réel $r = \gamma/n$ est fixé et que l'on fait tendre n vers l'infini. La première égalité est issue directement du lien avec la fonction Γ (cf. formule 7.9 page 180). Soit la fonction f_n définie sur $]0,1[$ par

$$f_n(t) = \frac{(\gamma-1)}{n}\log t + \frac{(n-\gamma)}{n}\log(1-t).$$

La fonction f_n est croissante sur $]0, \frac{\gamma-1}{n-1}[$, décroissante sur $]\frac{\gamma-1}{n-1}, 1[$ et sa dérivée seconde est négative. Pour $x < r - \epsilon$, l'intégrale est alors majorée par

$$\int_0^x \exp(nf_n(t))dt \leq x^\gamma (1-x)^{n-\gamma}.$$

En utilisant la formule de Stirling avec le coefficient binomial, nous obtenons

$$\gamma\binom{n}{\gamma}\int_0^x \exp(nf_n(t))dt = O(w_{r,n}(x))$$

ce qui montre le second résultat du lemme. La preuve du troisième résultat est identique sauf que le point de départ est la relation

$$\int_0^x \exp(nf_n(t))dt = \int_0^1 \exp(nf_n(t))dt - \int_x^1 \exp(nf_n(t))dt = 1 - \int_x^1 \exp(nf_n(t))dt.$$

Ensuite, les mêmes calculs s'appliquent à l'intégrale $\int_x^1$ sauf que l'on utilise la décroissance de f_n sur $[r+\epsilon, 1]$. La dérivée de $\log \widetilde{w}_r(x)$ est donnée par

$$(\log \widetilde{w}_r(x))' = \frac{r-x}{x(1-x)}.$$

En particulier, si $x < r$ (resp. $x > r$), $\widetilde{w}_r$ est strictement croissante (resp. décroissante) et $\widetilde{w}_r$ atteint son maximum en $x = r$ et satisfait $\widetilde{w}_r(r) = 1$. ■

Nous en avons terminé avec la preuve du théorème 18.

7.6.5 Cas du seuil constant

Le nombre de motifs fréquents est donné par *l'expression intégrale* 7.7 page 7.7. Dans le cas d'un seuil de fréquence constant, nous utilisons l'encadrement suivant de la fonction $(1-t)^{n-\gamma}$,

$$1 - (n-\gamma)t \leq (1-t)^{n-\gamma} \leq 1.$$

Le nombre moyen de motifs γ-fréquents vérifie alors l'encadrement

$$\gamma\binom{n}{\gamma}\left(\frac{S_{\gamma,m}}{\gamma} - (n-\gamma)\frac{S_{\gamma+1,m}}{\gamma+1}\right) \leq Fr_\gamma \leq \binom{n}{\gamma}S_{\gamma,m}, \tag{7.10}$$

où $S_{\gamma,m}$ est la somme

$$S_{\gamma,m} = \sum_{X\subset\mathcal{A}} p_X^\gamma. \tag{7.11}$$

L'encadrement 7.10 est valable quel que soit le seuil de fréquence. Sous l'hypothèse 10 page 167, la série $S_\gamma(z)$ (dont le coefficient de z^m est exactement $S_{\gamma,m}$) admet un unique pôle simple en $z = z_\gamma$ avec $z_\gamma < z_{\gamma+1} < 1$, l'extraction du coefficient z^m dans $S_\gamma(z)$ par la formule de Cauchy montre que l'asymptotique des sommes $S_{\gamma,m}$ satisfait

$$S_{\gamma,m} = [z^m]S_\gamma(z) = \frac{\kappa_\gamma}{z_\gamma^m}(1 + O(\theta_\gamma^m)) \qquad \text{avec} \qquad 0 < \theta_\gamma < 1,$$

et κ_γ le résidu de $S_\gamma(z)$ en z_γ. Par suite, le nombre moyen de motifs γ-fréquents vérifie

$$Fr_\gamma = \binom{n}{\gamma}\frac{\kappa_\gamma}{z_\gamma^m}(1 + O(\theta^m)),$$

où $\theta = \max(\theta_\gamma, \theta_{\gamma+1}, z_{\gamma+1}/z_\gamma)$. Ceci termine la preuve du théorème 20.■

7.7 Conclusion

Dans cette partie, nous avons présenté un modèle aléatoire de base de données basé sur les sources. Dans ce modèle, nous avons obtenu trois résultats dont chacun dépend d'une hypothèse faite sur la source. Pour toutes les sources classiques, ces hypothèses sont vérifiées mais les sources classiques conduisent systématiquement à des bases de données faiblement corrélées. Un des travaux futurs sera d'améliorer à la fois les modèles et les hypothèses afin d'élargir le plus possible le cadre d'application.

Les trois résultats concernent le nombre moyen de motifs fréquents pour trois seuils de fréquence différents. Pour un seuil de fréquence constant, nous montrons que le nombre moyen de motifs fréquents est polynomial en le nombre d'objets et exponentiel en le nombre d'attributs. Pour un seuil de fréquence au moins logarithmique, le nombre de motifs fermés est équivalent au nombre de motifs fréquents. Finalement, pour un seuil proportionnel au nombre d'objets, le nombre moyen de motifs fréquents est au plus polynomial. Le premier résultat est conforme à l'intuition générale et le second résultat avait déjà été constaté pour des bases faiblement corrélées du type panier de la ménagère. Toutefois, le troisième résultat est complètement nouveau car il met pour la première fois en évidence un comportement polynomial du nombre de motifs fréquents. A posteriori, ce n'est pas étonnant puisque pour un seuil fixe, ce nombre est exponentiel alors que pour un seuil maximal de n, il n'y a que le motif vide qui est fréquent. La transition n'étant pas brutale, il existe bien un seuil au-delà duquel le comportement est polynomial.
Nous terminons cette partie quelques points que nous n'avons pas résolu ou abordé :

1. Quel est le nombre moyen de motifs γ-fermés dans le cadre des sources de Bernoulli pour un seuil constant ? Nous rappelons que les 1-fermés sont en correspondance avec les séparateurs minimaux dans le graphe co-bipartite ou avec les sous-graphes bipartites maximaux du graphe bipartite. Il est donc très intéressant d'étudier ce paramètre.
2. Mêmes questions pour les motifs libres et candidats. Les motifs libres sont essentiels pour construire les règles d'association à prémisse minimale alors que les motifs candidats sont tous les motifs dont la fréquence est calculée.
3. Quelle est la taille t pour laquelle le nombre de motifs fréquents (resp. candidats) est maximum ? Et quel est ce maximum ? Ces paramètres correspondent à la taille mémoire nécessaire à l'exécution de Apriori.
4. Quelle est la taille du plus long motif ? La taille du plus long motif correspond au nombre de passes qu'effectue Apriori sur la base de données.

Chapitre 8

Point de vue dynamique sur les bases de données

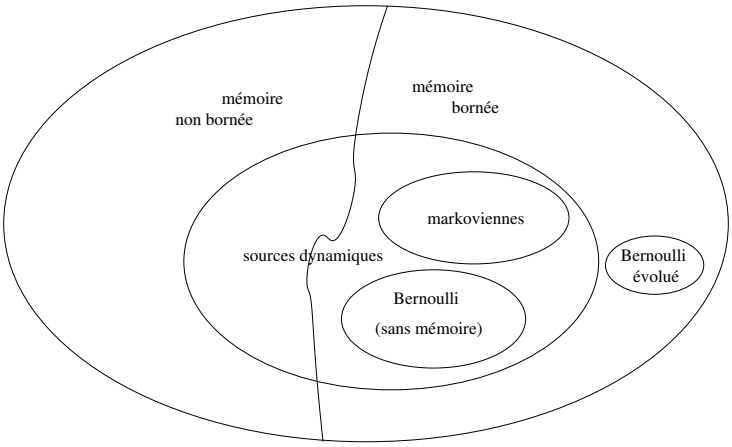
Sommaire

Dans la partie précédente, nous avons montré comment les sources pouvaient intervenir pour la modélisation des bases de données. Une source est un procédé (aléatoire) qui émet des symboles à chaque top d'une horloge. On distingue généralement trois catégories de sources : les sources sans mémoire, à mémoire bornée et à mémoire non-bornée. Les sources sans mémoire n'utilisent pas les symboles précédents pour émettre le nouveau symbole. C'est le cas par exemple des sources de Bernoulli ou de la source issue du modèle de Bernoulli évolué. Les sources à mémoire bornée utilisent une partie bornée du passé pour émettre le nouveau symbole. Les chaînes de Markov appartiennent à cette catégorie. Finalement, les sources à mémoire non-bornée utilisent tout le passé pour émettre le nouveau symbole. La source des fractions continues qui à un réel x tiré au hasard, associe le mot formé des quotients successifs du développement en fraction continue est une source à mémoire non-bornée. Les sources sans mémoire et à mémoire bornée sont dites simples par opposition aux sources à mémoire non bornée qui sont dites complexes.

Les sources sans mémoire amènent à un contexte de complète indépendance ou de non corrélations pour l'analyse. Les probabilistes savent très bien que c'est un contexte idéal pour toute étude mathématique mais la réalité est souvent plus complexe. Par exemple, les bases de données ne sont jamais complètement non corrélées. Les chaînes de Markov offrent plus de corrélations entre les symboles mais ces corrélations restent locales ou bornées. Une fois encore, les chaînes de

Markov sont trop simples pour modéliser la complexité de la réalité même si cela constitue une étape supplémentaire. Par exemple, les bases de données réelles ne contiennent pas uniquement des corrélations locales.

Dans cette partie, nous utilisons essentiellement les sources dynamiques introduites par Brigitte Vallée dans [Val01] afin de construire les bases de données. Certaines sources dynamiques sont à mémoire non bornée (comme la source des fractions continues) mais les corrélations sont décroissantes exponentiellement avec le temps. Les sources dynamiques sont basées sur des systèmes dynamiques. Tous les outils des systèmes dynamiques dont les opérateurs de transfert sont donc disponibles pour l'analyse de ces sources. De plus, les sources dynamiques englobent à la fois des sources simples comme les sources de Bernoulli simples (sans mémoire) et les chaînes de Markov (à mémoire bornée), mais aussi des sources complexes.



Les sources dynamiques ont été utilisées à la fois en pattern matching et pour l'analyse des arbres digitaux qui codent les dictionnaires (aussi appelés tries). Les problèmes de pattern matching concernaient essentiellement le nombre moyen d'occurrences d'un motif (généralisé) dans un texte produit par une source dynamique (voir par exemple [BV02, BV06]). Pour les tries, les mots du dictionnaire étaient produits par la source dynamique et les analyses portaient sur la taille, la hauteur, la longueur de cheminement des arbres digitaux [CFV01, Bou01, BNV01]...

Plan. Notre objectif dans cette partie est dans un premier temps de présenter les sources dynamiques. Cette description sera très précise afin de mettre en évidence tous les aspects des sources. Ensuite, nous allons montrer que les sources dynamiques (complètes ou markoviennes) vérifient (sous certaines hypothèses) les conditions 8 , 9 et 10 (pages 166, 166, 167) des théorèmes 18, 19 et 20 (pages 166, 167, 169). Ainsi, nous aurons montré que les résultats annoncés au précédent chapitre s'appliquent à un grand nombre de modèles de sources.

8.1 Sources dynamiques

Les sources dynamiques se construisent à partir de systèmes dynamiques. L'étude des systèmes dynamiques est une branche à part entière des mathématiques et nous n'abordons ici que les principaux aspects des systèmes dynamiques de l'intervalle. Pour une introduction plus précise, nous conseillons le livre [LM94].

8.1.1 Sources dynamiques et mots produits

Un système dynamique de l'intervalle est défini par quatre objets :

1. un alphabet $\mathcal{E}$ fini,
2. une partition topologique de $I =]0,1[$ en intervalles ouverts disjoints I_p pour $p \in \mathcal{P}$, i.e. $\overline{I} = \cup_{p\in\mathcal{P}} \overline{I}_p$,
3. une application de codage $\sigma : I \to \mathcal{E}$ constante intervalle I_p de la partition,
4. une fonction de décalage (ou shift) $T : I \to I$ injective et C^2 sur chaque I_p. On pose $J_p = T(I_p)$ l'image par T de I_p, $h_p : J_p \to I_p$ l'inverse local (ou branche inverse) de T sur I_p et $\mathcal{H}$ l'ensemble des branches inverses $\mathcal{H} = \{h_p : p \in \mathcal{P}\}$.
5. De plus, si $\sigma(I_p) = \sigma_{I_q}$, alors $J_p \cap J_q = \emptyset$.

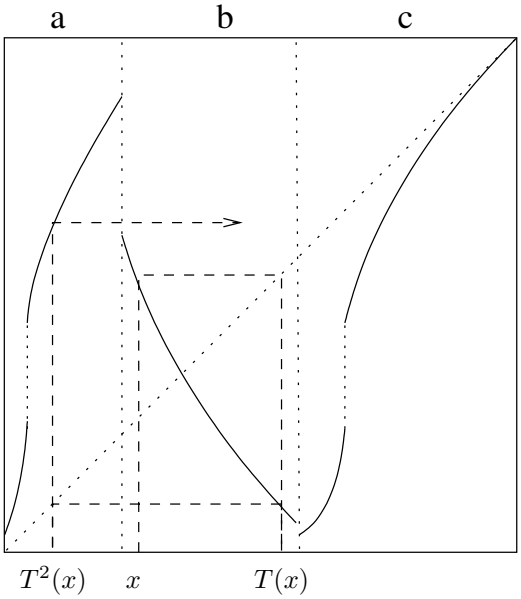
Si $h = h_{p_1} \circ \ldots \circ h_{p_n}$ est la composition de p branches inverses avec h définie sur un intervalle J_h et à valeur dans un intervalle I_h, alors h est appelée branche inverse de profondeur n. On note $\mathcal{H}^n$ l'ensemble des branches inverses de profondeur n. Les branches inverses de profondeur n sont les branches inverses du n^e itéré de T, T^n.

La trajectoire associée à un point x est la suite des itérés de x sous l'action du shift T,

$$\mathcal{T}(x) = (x, T(x), T^2(x), \ldots, T^k(x), \ldots).$$

En appliquant la fonction de codage à chaque composante de $\mathcal{T}(x)$, on définit un mot $\mathcal{M}(x)$ associé à x,

$$\mathcal{M}(x) = (\sigma(x), \sigma(T(x)), \sigma(T^2(x)), \ldots, \sigma(T^k(x)), \ldots).$$



Pour un élément m de l'alphabet $\mathcal{E}$, l'intervalle fondamental $I_{[m]}$, est l'ensemble des intervalles de la partition codant la lettre m,

$$I_{[m]} = \cup_{p\in\mathcal{P}:\ \sigma(I_p)=m} I_p.$$

L'image par T de $I_{[m]}$ définit l'intervalle $J_{[m]}$ avec

$$J_{[m]} = \cup_{p\in\mathcal{P}:\ \sigma(I_p)=m} J_p.$$

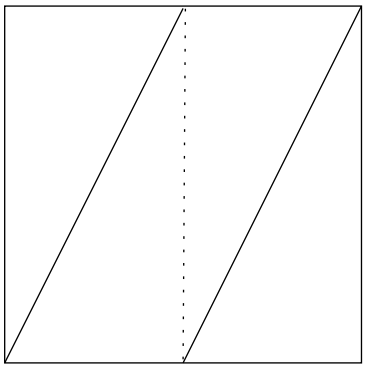
Nous considérons maintenant la branche inverse associée à m et définie par

$$h_{[m]}(x) = \sum_{p\in\mathcal{P}:\ \sigma(I_p)=m} h_p(x)\mathbf{1}_{J_p}(x).$$

Pour chaque mot $w = m_1 m_2 \dots m_k$ sur l'alphabet $\mathcal{E}$, il existe un intervalle fondamental $I_{[w]}$ contenant tout les $x \in I$ dont les mots associés $\mathcal{M}(x)$ commence par w. L'image par T de $I_{[w]}$ définit l'ensemble $J_{[w]}$. La branche inverse $h_{[w]}$ associée à w est quant à elle définie par

$$h_{[w]} : J_{[w]} \to I_{[w]} \quad \text{et} \quad h_{[w]} = h_{[m_1]} \circ h_{[m_2]} \circ \dots \circ h_{[m_k]}.$$

Prenons pour exemple le système dynamique du développement binaire composé de deux branches affines identiques dont la première (resp. la seconde) correspond au symbole 0 (resp. 1). Le shift T est donné par $T(x) = 2x[1]$.



Pour $x = 5/8$, on a

$$\sigma(\frac{5}{8}) = 1, \quad \sigma(T(\frac{5}{8})) = \sigma(\frac{1}{4}) = 0, \quad \sigma(T^2(\frac{5}{8})) = \sigma(\frac{1}{2}) = 1, \quad \sigma(T^k(\frac{5}{8})) = \sigma(0) = 0, \quad \text{pour } k \geq 3.$$

Le mot associé à $5/8$ est donc $\mathcal{M}(5/8) = (1, 0, 1, 0, 0, \dots)$. On constate que c'est le développement binaire de $5/8$. Cette propriété se vérifie pour tout x de l'intervalle. De manière similaire, le shift $T(x) = b \cdot x[1]$ pour un entier positif b définit un système dynamique pour la numération en base b.
Dans la partie précédente, le système dynamique utilisé était le système dynamique des fractions continues. Ce système admet une infinité de branches dont aucune n'est affine. Pour $x \in]0, 1[$, le mot $\mathcal{M}(x)$ est composé des entiers successifs qui apparaissent dans le développement en fractions continues de x, c'est à dire,

$$\mathcal{M}(x) = (m_1, m_2, m_3, \dots) \quad \text{équivaut à} \quad x = \cfrac{1}{m_1 + \cfrac{1}{m_2 + \cfrac{1}{m_3 + \cfrac{1}{\ddots}}}}$$

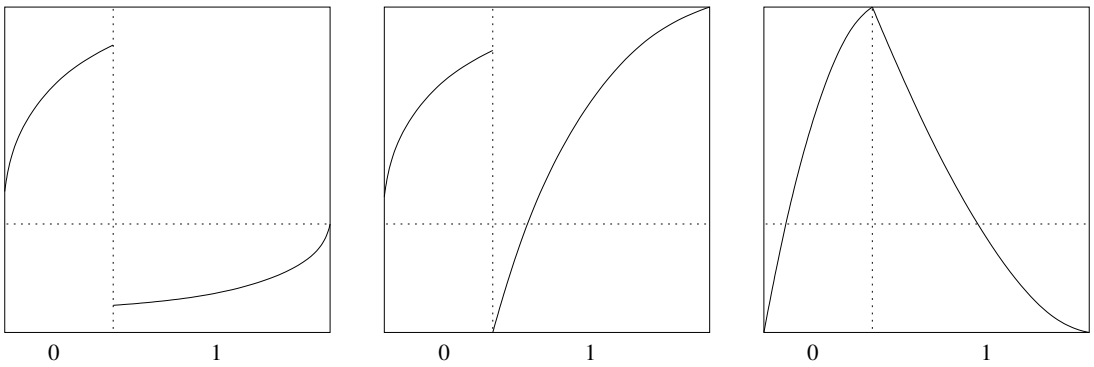
Nous donnons maintenant la définition d'une source dynamique en adaptant à notre contexte celles introduites par Brigitte Vallée [Val01].

Définition 24 (Source Dynamique) *Une source dynamique est la donnée d'un système dynamique $\mathcal{D} = (I, T, \sigma, \mathcal{E})$ et d'une densité initiale f sur I. L'initialisation de la source se fait en choisissant x suivant la densité f. Ensuite, les lettres successivement émises sont celles qui composent le mot $\mathcal{M}(x)$.*

Une source dynamique ressemble à un générateur pseudo-aléatoire. Une graine est utilisée pour initialiser le processus et ensuite, toutes les étapes sont entièrement déterministes.

8.1.2 Géométrie des branches

La géométrie des branches a une influence énorme sur le comportement d'une source dynamique. Pour nous convaincre, considérons les trois sources suivantes.



Avec le premier système, les zéros et les uns alternent. Avec le second système, un zéro est nécessairement suivi d'un un mais après un un, on peut trouver les deux symboles. Finalement, avec le dernier système, tous les mots sur $\{0, 1\}$ sont possibles.
Suivant la géométrie des branches, les systèmes (ou sources) dynamiques sont classé(e)s en trois catégories : les systèmes complets, les systèmes markoviens et les systèmes généraux.

8.1.2.1 Systèmes complets

Un système dynamique est dit complet si $\mathcal{P} = \mathcal{E}$ et si pour tout intervalle I_m, T est bijective sur I_m, i.e., $T(\overline{I}_m) = \overline{I}$. Une source dynamique associée à un système dynamique complet est dite complète. La source des fractions continues, les sources de Bernoulli ou la source du développement binaire sont des sources complètes. Après chaque symbole, tous les symboles sont possibles. Intuitivement, c'est le type de source (ou système) qui offre le moins de corrélations entre les symboles.

Les sources de Bernoulli de paramètre p est une source dynamique à branches complètes $(I, T, \sigma, \mathcal{E}, f)$ sur l'alphabet à deux symboles $\mathcal{E} = \{0, 1\}$. La fonction T se compose de deux branches affines correspondant à chaque lettre de l'alphabet et le tout est définit de la manière suivante :

$$I = [0,1], \qquad f(x) = 1, \qquad \begin{cases} T(x) = x/p, & \sigma(x) = 1, \quad \text{pour } x \in [0, p[\\ T(x) = (x-p)/(1-p), & \sigma(x) = 0, \quad \text{pour } x \in [p, 1] \end{cases}$$

Nous avons aussi considérer un modèle groupé de Bernoulli sur des groupes d'attributs de taille t. La source dynamique associée est basée sur une source dynamique $(I, T, \sigma, \mathcal{E}, f)$ sur l'alphabet des t-uples ne contenant qu'un seul 1,

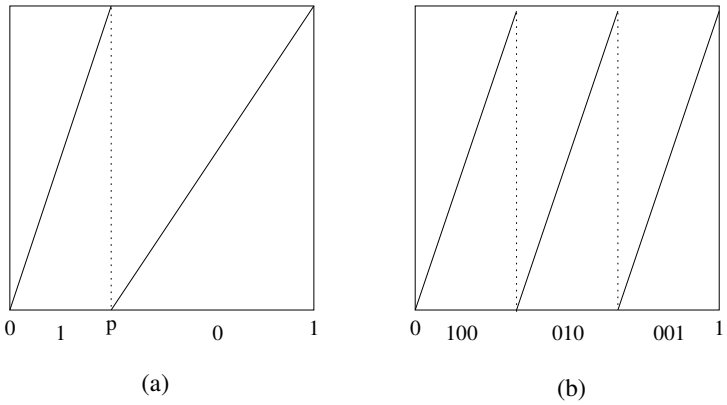
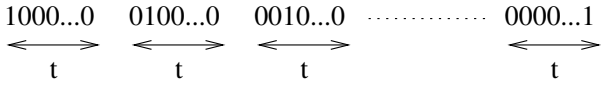
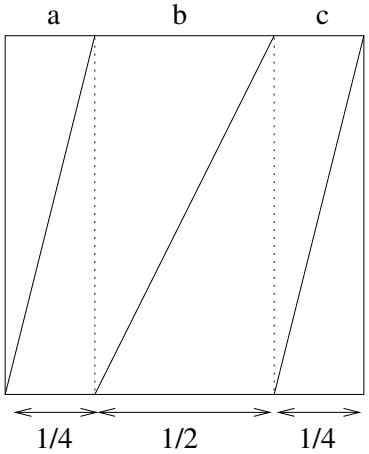


FIG. 8.1 – Sources dynamiques correspondant (a) au modèle simple de Bernoulli de paramètre p et (b) au modèle groupé de Bernoulli de paramètre $t = 3$.



La fonction T admet t branches affines de même pente t et la distribution initiale reste la fonction constante $f = 1$. Nous avons représenté les sources dynamiques correspondant aux deux modèles de Bernoulli à la figure 8.1.
Dans les deux cas, ce ne sont pas les seules manières de faire. Pour le modèle simple de Bernoulli, il est par exemple possible d'échanger les deux branches ou d'utiliser des branches décroissantes. Il en est de même avec le modèle groupé.

Plus généralement, les sources sans-mémoire se modélisent naturellement avec des sources dynamiques complètes à branches affines. Si p_ω est la probabilité d'émettre la lettre ω, alors il suffit de considérer le système dynamique complet à branches affines de densité initiale $f = 1$ et de partition topologique I_ω avec $|I_\omega| = p_\omega$ (une telle partition existe et n'est pas unique). La figure suivante correspond au système dynamique d'une source sans mémoire sur l'alphabet $\{a, b, c\}$ avec $p_a = p_c = 1/4$ et $p_b = 1/2$. Notons qu'il n'est pas possible de représenter le modèle de Bernoulli évolué avec un système dynamique (a fortiori complet).



8.1.2.2 Systèmes markoviens

Nous nous limitons au cas des alphabets finis bien qu'il existe des définitions pour des alphabets infinis. Un système dynamique est dit markovien si

1. $\mathcal{P} \subseteq \mathcal{E}^2$. Pour $(m_1, m_1) \in \mathcal{P}$, l'intervalle associé dans la partition est alors noté $I_{m_2|m_1}$.
2. pour tout $(m_1, m_2) \in \mathcal{P}$, $T(I_{m_2|m_1}) = I_{[m_2]}$ et $\sigma(I_{m_2|m_1}) = m_1$.

Les intervalles $I_{[m]}$ sont l'équivalent des états dans les chaînes de Markov. Lorsque $x \in I_{m_2|m_1}$, le système dynamique est dans l'état m_1 et l'état suivant, donné par $\sigma(T(x))$, est l'état m_2. La matrice binaire sous-jacente à un système markovien décrit les passages possibles entre les états et est définie par $P = (P_{m_2|m_1})$ avec

$$P_{m_2|m_1} = 1 \qquad \text{ssi} \qquad (m_1, m_2) \in \mathcal{P}.$$

Autrement dit, $P_{m_2|m_1} = 1$ si et seulement si le système (resp. la source) peut passer de l'état m_1 à l'état m_2 (resp. émettre m_1 puis m_2).

Définition 25 (système irréductible) *Un système markovien est dit irréductible (resp. fortement irréductible) si une puissance de P est strictement positive (resp. si P est strictement positive).*

Une source dynamique est dite markovienne (resp. fortement markovienne) si elle est associée à un système dynamique irréductible (resp fortement irréductible).

L'irréductibilité implique que tous les symboles peuvent être émis après un certains nombre d'étapes à partir de n'importe quel symbole. Une étape suffit avec l'irréductibilité forte.

Les chaînes de Markov se modélisent naturellement avec les sources dynamiques markoviennes. Fixons par exemple $(f_w)_{w\in\{0,1\}^K}$ une densité initiale et $(p_{w_1|w_2})_{w_1,w_2\in\{0,1\}^K}$ les probabilités de transition. Pour $w \in \{0,1\}^K$, nous posons a_w l'entier dont w est la représentation binaire et $I_{[w]}$ l'intervalle $I_{[w]} = [a_w, a_w + 1]$. L'intervalle I est donné par la réunion des intervalles $I_{[w]}$ soit $I = [0, 2^K]$. La source dynamique markovienne que nous construisons sera définie sur l'intervalle I.
L'alphabet $\mathcal{E}$ est $\{0,1\}^K$. La fonction de codage est constante et égale à w sur chaque $I_{[w]}$ ainsi que la densité initiale qui est constante et égale à f_w sur chaque I_w. Il reste le shift T à construire et la partition $(I_p)_{p\in\mathcal{P}}$. Chaque intervalle $I_{[w_1]}$ admet une partition topologique $(I_{w_2|w_1})_{w_2\in\mathcal{E}}$ où $I_{w_2|w_1}$ est de longueur $p_{w_2|w_1}$. La partition est donnée par l'ensemble des intervalles $I_{w_2|w_1}$. Pour terminer la construction, le shift T est affine par morceaux et tel que ses restrictions $T : I_{w_2|w_1} \to I_{[w_2]}$ sont des bijections.

La figure 8.2 montre le système dynamique markovien correspondant à une chaîne de Markov d'ordre 1 sur l'alphabet $\{00, 01, 10, 11\}$ ayant pour densité initiale $f = (f_{00}, f_{10}, f_{01}, f_{11})$ avec

$$f_{00} = 1/4, \quad f_{01} = 1/4, \quad f_{10} = 1/6, \quad f_{11} = 1/3,$$

et pour probabilités de transition,

	00	01	10	11
00	1/4	1/4	1/4	1/4
01	1/3	1/3	1/6	1/6
10	1/2	1/4	1/8	1/8
11	1/5	1/5	2/5	1/5

Notons que les sources dynamiques complètes sont des sources markoviennes fortement irréductibles. Il suffit de considérer la partition naturelle de T^2.

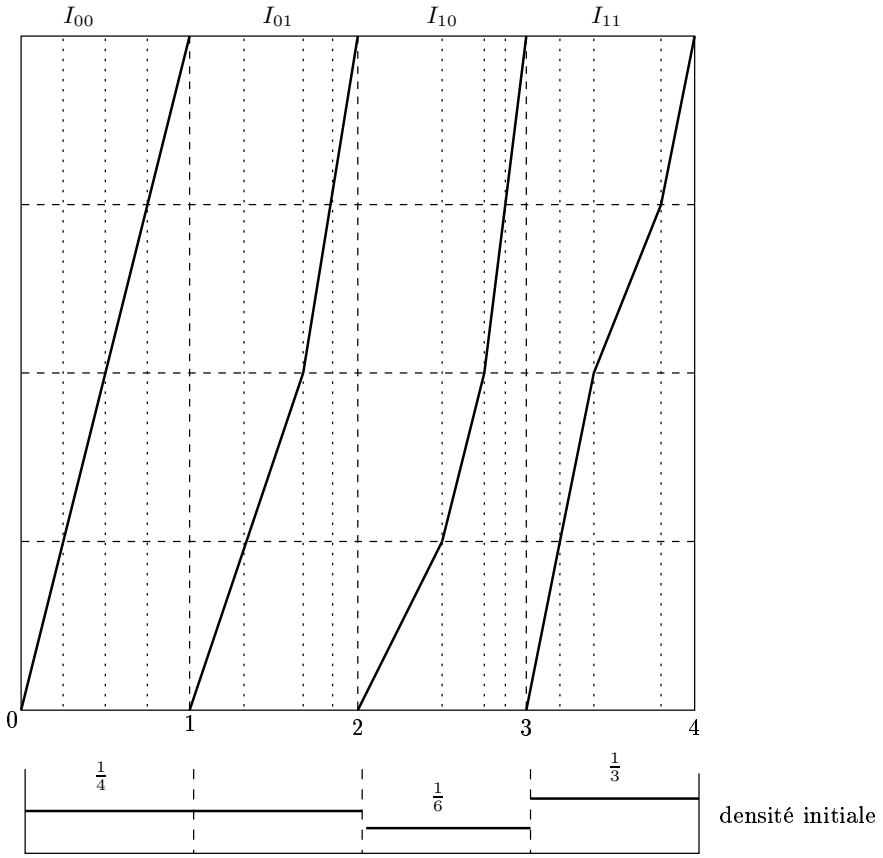


FIG. 8.2 – Source dynamique markovienne associée à une chaîne de Markov.

8.1.2.3 Systèmes généraux

On appellera source dynamique générale un système dynamique qui n'est ni complet, ni markovien. De tels systèmes apparaissent par exemple lors de l'étude des algorithmes α-euclidiens [BDV02]. Ces systèmes sont beaucoup plus complexes à étudier. Les preuves que nous utilisons profitent des propriétés géométriques des sources dynamiques complètes ou markoviennes et ne s'étendent donc pas aux sources dynamiques générales.

8.1.3 Régularités des branches

Nous venons de voir les aspects géométriques des branches d'un système dynamique. Nous considérons maintenant la régularité des branches. Cette régularité a aussi une influence sur le comportement d'une source dynamique.

8.1.3.1 Système dilatant ou branches inverses contractantes

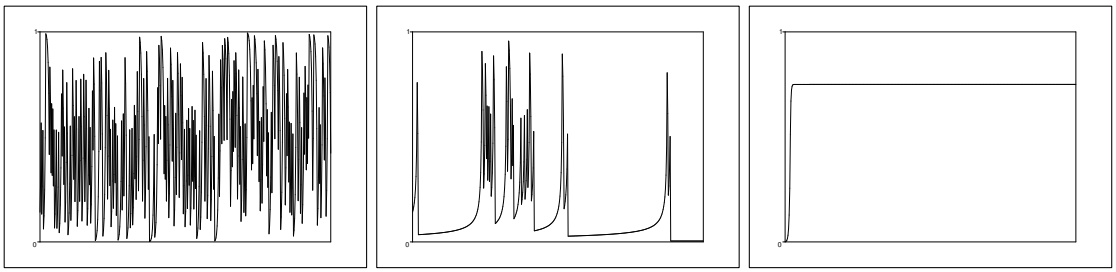
Un système dynamique est dit dilatant si la dérivée du shift vérifie

$$|T'| \geq \rho^{-1} > 1.$$

Il revient au même de dire que les branches inverses sont contractantes, i.e.

$$\forall h \in \mathcal{H}, \quad |h'| \leq \rho < 1$$

Le plus petit réel ρ satisfaisant cette propriété est appelé le coefficient de contraction du système dynamique. Un système dynamique dilatant n'admet pas de points fixes attractifs ($T(x) = x$ et $|T'(x)| < 1$). Cela évite que les mots soient tous des suites stationnaires. Il n'admet pas non plus de points indifférents ($T(x) = x$ et $|T'(x)| = 1$). Un système dynamique qui admet un point indifférent met un certain temps avant de s'en éloigner. La source a donc tendance à émettre plusieurs fois la même lettre (celle autour du point fixe) avant d'en émettre une nouvelle. Ceci implique un mauvais mélange des lettres. En terme de trajectoire, on voit très clairement la différence entre un système dilatant, un système qui admet un point indifférent et un système qui admet un point fixe (de gauche à droite).



Les trajectoires d'un système dilatant sont très chaotiques. Si le système admet un point indifférent, on constate des zones de calme, appelées phénomènes d'intermittences, qui correspondent au temps qu'il faut pour s'éloigner du point indifférent. Finalement, avec un point attractif, les trajectoires convergent vers le point attractif.
Avoir un point indifférent a aussi des conséquences algorithmiques. Les auteurs de [BDV02, Val03] ont classé les algorithmes euclidiens en deux classes : les algorithmes lents et rapides. Les algorithmes lents sont caractérisés par des systèmes dynamiques ayant un point indifférent contrairement aux algorithmes rapides.

8.1.3.2 Distorsion

Un système dynamique satisfait la propriété de distorsion s'il existe un réel $A > 0$ tel que pour toute branche inverse $h \in \mathcal{H}$, et tout $x \in J_h$

$$\frac{|h''(x)|}{|h'(x)|} \leq A.$$

Plus la distorsion est proche de 0, plus les branches ressemblent à des branches affines. Or, les systèmes à branches affines sont les moins corrélés. La constante de distorsion est donc très liées aux corrélations de la source.

Dans le cas d'un alphabet fini, la condition de distorsion est équivalente à la condition $|h''| > 0$ pour toute branche inverse h.

Le lemme suivant montre que la propriété de distorsion se vérifie aussi pour toutes les branches inverses de n'importe quel ordre.

Lemme 18 *Considérons un système dynamique à branches contractantes, de coefficient de contraction ρ. Supposons de plus que le système vérifie la propriété de distorsion avec la constante A. Alors, pour toute branche inverse $h \in \mathcal{H}^\star$, la branche h vérifie la distorsion*

$$\forall x \in J_h, \quad \frac{|h''(x)|}{|h'(x)|} \leq \frac{A}{1-\rho}.$$

En particulier, si $D = e^{-A/(1-\rho)}$, alors

$$\forall h \in \mathcal{H}^\star, \quad \forall x, y \in J_h, \qquad D \leq \frac{|h'(x)|}{|h'(y)|} \leq D^{-1}.$$

Preuve : Si $h = g \circ k$ est une branche inverse d'ordre $p+1$ avec $g \in \mathcal{H}$ et $k \in \mathcal{H}^p$, alors

$$\frac{(g \circ k)''}{(g \circ k)'} = k' \frac{g'' \circ k}{g' \circ k} + \frac{k''}{k'}.$$

Comme k est une branche inverse d'ordre p, sa dérivée vérifie $|k'| \leq \rho^p$. Il est alors très facile de montrer par récurrence que, pour toute branche inverse ℓ d'ordre p,

$$\forall x \in J_\ell, \quad \frac{|\ell''(x)|}{|\ell'(x)|} \leq A(\sum_{i=0}^{p-1} \rho^i).$$

Ceci démontre la première partie du lemme. La deuxième partie s'obtient très facilement à partir de la relation intégrale

$$\int_y^x \frac{|h''(t)|}{|h'(t)|} dt = \log \frac{|h'(x)|}{|h'(y)|}.$$

Ceci termine la preuve du lemme.■

8.1.4 Intervalles fondamentaux et probabilités fondamentales

Dorénavant, nous fixons une source dynamique $\mathcal{S}$ ayant pour système dynamique $(I, T, \sigma, \mathcal{E})$ et densité initiale f. Pour un mot ω, p_ω désigne la probabilité fondamentale qu'un mot de $\mathcal{S}$ commence par ω et $I_{[\omega]}$ l'*intervalle fondamental* de ω, c'est-à-dire l'ensemble des x tel que $\mathcal{M}(x)$ commence par ω. L'intervalle fondamental est un véritable intervalle pour les sources complètes alors que c'est une réunion d'intervalles dans le cas général.

Si $\omega = m$ avec m une lettre de l'alphabet, alors l'intervalle fondamental de ω est $I_{[m]} = h_{[m]}(I)$. La probabilité p_ω est donnée par

$$p_\omega = \int_{I_{[m]}} f(x)dx = \int_I |h'_{[m]}(x)| f \circ h_{[m]}(x) \mathbf{1}_{J_{[m]}}(x)dx.$$

Maintenant si $\omega = m_1 m_2$ est un mot (productible) de deux lettres et si $\mathcal{M}(x)$ commence par ω, alors $x \in I_{[m_1]}$ et $T(x) \in I_{[m_2]}$ soit $x \in h_{[m_1]} \circ h_{[m_2]}(I) = I_{h_{[m_1]} \circ h_{[m_2]}}$. Ce résultat se généralise à tout mot de longueur n quelconque. Il est alors clair qu'un mot ω de longueur n est associée à une unique branche inverse de profondeur n notée $h_{[\omega]}$, définie sur un intervalle $J_{[\omega]}$ et à valeurs dans $I_{[\omega]}$. Si $\omega = m_1 \dots m_n$, l'intervalle fondamental de ω est

$$I_{[\omega]} = I_{h_{[\omega]}} = h_{[\omega]}(I) \qquad \text{où} \qquad h_{[\omega]} = h_{[m_1]} \circ h_{[m_2]} \circ \dots \circ h_{[m_n]}.$$

Avec ces notations, la probabilité fondamentale de ω satisfait aussi la formule intégrale

$$p_\omega = \int_{I_{[\omega]}} f(x)dx = \int_I |h'_{[\omega]}(x)| f \circ h_{[\omega]}(x) \mathbf{1}_{J_{[\omega]}}(x)dx.$$

Pour un système markovien, la partition $\mathcal{P}$ vérifie $\mathcal{P} \subset \mathcal{E}^2$ et les intervalles de la partition sont notés $I_{m_2|m_1}$ ($(m_1, m_2) \in \mathcal{P}$). On désigne par $h_{m_2|m_1}$ la bijection inverse de T sur $I_{m_2|m_1}$.Alors, pour tout mot $w = m_1 m_2 \dots m_k$ de longueur au moins 2, nous avons

$$h_{[w]} = h_{m_2|m_1} \circ h_{m_3|m_2} \circ \dots \circ h_{m_k|m_{k-1}} \quad \text{et} \quad p_w = \int_{I_{[m_k]}} |h'_{[w]}(t)| f \circ h_{[w]}(t)dt.$$

8.1.5 Opérateurs générateurs

Lors de l'analyse de la complexité binaire des algorithmes euclidiens, nous avons déjà utilisé des opérateurs pour générer les séries génératrices. Les opérateurs de transfert pondérés ou non sont des généralisations du transformateur de densité que nous décrivons à la section 8.1.5.1. Dans le cadre de la fouille de données, nous utilisons également deux autres généralisations que sont les opérateurs contraints (section 8.1.5.2) et les opérateurs de transfert multidimensionnels (section 8.1.5.3).

8.1.5.1 Transformateur de densité

La probabilité qu'un mot commence par la lettre m est donnée par

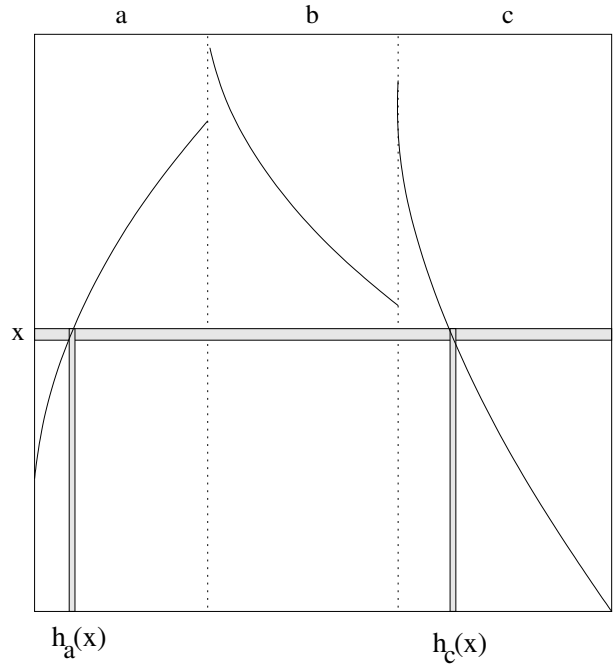
$$\int_{I_{[m]}} f(x)dx = \int_I |h'_{[m]}(x)| f \circ h_{[m]}(x) \mathbf{1}_{J_{[m]}}(x)dx,$$

où f est la densité initiale. Une question naturelle est quelle est la probabilité que la deuxième lettre soit un m ? Même chose pour la troisième ? Pour répondre à ces questions, il faut connaître

la densité après une ou deux itérations du shift T. Si f_n désigne la densité après n itérations du shift, alors il existe un opérateur $\mathbf{G}$, appelé *transformateur de densité*, tel que $f_n = \mathbf{G}^n[f]$. Comme son nom l'indique, le transformateur de densité transforme une densité en une autre et vérifie la formule,

$$\mathbf{G}[f](x) = \sum_{p \in \mathcal{P}} |h'_p(x)| f \circ h_p(x) \mathbf{1}_{J_p}(x) = \sum_{m \in \mathcal{E}} |h'_{[m]}(x)| f \circ h_{[m]}(x) \mathbf{1}_{J_{[m]}}(x).$$

Intuitivement, la densité en un point x après une itération est la somme de tous les morceaux de densités apportés par les antécédents de x qui sont les $h_p(x)$. La pondération par les dérivées $|h'_p|$ est due à la formule du changement de variable.



Avec la propriété de multiplicativité de la dérivée d'une fonction composée, le n^e itéré de $\mathbf{G}$ s'exprime en fonction des branches inverses de profondeur n,

$$\mathbf{G}^n[f](x) = \sum_{h \in \mathcal{H}^n} |h'(x)| f \circ h(x) \mathbf{1}_{J_h}(x).$$

Comme les branches inverses de profondeur n sont associées à des mots de longueur n, le n^e itéré de $\mathbf{G}$ s'écrit aussi,

$$\mathbf{G}^n[f](x) = \sum_{\omega \in \mathcal{E}^{<n>}} |h'_{[\omega]}(x)| f \circ h_{[\omega]}(x) \mathbf{1}_{J_{[\omega]}}(x),$$

où $\mathcal{E}^{<n>}$ est l'ensemble des préfixes de longueur n productibles par la source. Nous posons $\mathbf{G}_{[\omega]}$ l'opérateur composante

$$\mathbf{G}_{[\omega]}[f](x) = |h'_{[\omega]}(x)| f \circ h_{[\omega]}(x) \mathbf{1}_{J_{[\omega]}}(x).$$

Cas des systèmes markoviens. Si $w = m_1 m_2 \dots m_k$ est un mot de longueur au moins deux, alors la branche inverse $h_{[w]}$ est donnée par $h_{[w]} = h_{m_2|m_1} \circ \dots \circ h_{m_p|m_{p-1}}$. Nous définissons l'opérateur composante $\mathbf{G}_{m_2|m_1}$ par

$$\mathbf{G}_{m_2|m_1}[f] = |h'_{m_2|m_1}\ f \circ h_{m_2|m_1} \mathbf{1}_{I_{[m_2]}}.$$

On vérifie facilement que l'opérateur composante associé à w s'écrit

$$\mathbf{G}_{[w]} = \mathbf{G}_{m_p|m_{p-1}} \circ \ldots \circ \mathbf{G}_{m_3|m_2} \circ \mathbf{G}_{m_2|m_1}.$$

8.1.5.2 Opérateur de transfert contraint

Dans la partie consacrée aux calculs de constantes (Partie 5), nous avons déjà introduit les opérateurs de transfert contraints dont la valeur propre dominante était liée à la dimension de Hausdorff d'un espace de Cantor. Nous nous étions alors limités au cas des fractions continues mais de tels opérateurs existent bien entendu dans un cadre plus général.

Nous fixons $\mathcal{E}'$ un sous-alphabet de l'alphabet initial $\mathcal{E}$. La probabilité qu'un mot de longueur n commence par n lettres de $\mathcal{E}'$ est

$$\mathbb{P}\left[\omega \in (\mathcal{E}')^n\right] = \sum_{\omega \in (\mathcal{E}')^n} \int_I |h'_{[\omega]}(x)| f \circ h_{[\omega]}(x) \mathbf{1}_{J_{[\omega]}}(x) dx.$$

En notant $\mathbf{G}_{[\mathcal{E}']}$ l'opérateur de transfert contraint à $\mathcal{E}'$ et défini par,

$$\mathbf{G}_{[\mathcal{E}']}[f] = \sum_{m \in \mathcal{E}'} \mathbf{G}_{[m]}[f] = \sum_{m \in \mathcal{E}'} |h'_{[m]}| f \circ h_{[m]} \mathbf{1}_{J_{[m]}},$$

la probabilité précédente s'écrit

$$\mathbb{P}\left[\omega \in (\mathcal{E}')^n\right] = \int_I \mathbf{G}^n_{[\mathcal{E}']}[f](x) dx.$$

Remarque : Par perturbation de l'opérateur de transfert contraint,

$$\mathbf{G}_{s,[\mathcal{E}']}[f] = \sum_{m \in \mathcal{E}'} |h'_{[m]}|^s f \circ h_{[m]} \mathbf{1}_{J_{[m]}}.$$

nous retrouvons l'opérateur que nous avons utilisé dans le cadre des fractions continues lors du calcul des constantes.

8.1.5.3 Opérateur de transfert multidimensionnel

Dans cette section, nous fixons γ un entier. Jusqu'à présent, nous avons considéré l'évolution d'une unique source dynamique $\mathcal{S} = (I, T, \sigma, \mathcal{E}, f)$. Considérons maintenant, en parallèle et de manière indépendante, l'évolution de γ copies de $\mathcal{S}$. A chaque étape, un γ-uplet de lettres est produit. Cela définit une source dynamique multidimensionnelle $\overline{\mathcal{S}}_\gamma = (\overline{I}, \overline{T}, \overline{\sigma}, \overline{\mathcal{E}}, \overline{f})$ avec

$$\overline{I} = I^\gamma, \qquad \overline{\mathcal{E}} = \mathcal{E}^\gamma, \qquad \overline{T}(x_1, \ldots, x_\gamma) = (T(x_1), \ldots, T(x_\gamma)),$$

$$\overline{\sigma}(x_1, \ldots, x_\gamma) = (\sigma(x_1), \ldots, \sigma(x_\gamma)), \qquad \overline{f}(x_1, \ldots, x_\gamma) = f(x_1) \times \ldots \times f(x_\gamma).$$

Pour $\overline{x} \in \overline{I}$, la trajectoire et le mot associés à $\overline{x}$ sont alors respectivement définis par,

$$\overline{T}(\overline{x}) = (\overline{x}, \overline{T}(\overline{x}), \overline{T}^2(\overline{x}), \ldots), \qquad \overline{\mathcal{M}}(\overline{x}) = (\overline{\sigma}(\overline{x}), \overline{\sigma}(\overline{T}(\overline{x})), \overline{\sigma}(\overline{T}^2(\overline{x})), \ldots).$$

Pour $\overline{m} = (m_1, \ldots, m_\gamma) \in \overline{\mathcal{E}}$, le shift $\overline{T}$ est par construction C^2 par morceaux et injectif sur chaque ensemble $I_{[\overline{m}]} = I_{[m_1]} \times \ldots \times I_{[m_\gamma]}$. L'image de $I_{[\overline{m}]}$ par le shift $\overline{T}$ est donné par l'ensemble

$J_{[\overline{m}]} = J_{[m_1]} \times \ldots \times J_{[m_\gamma]}$. La branche inverse multidimensionnelle associée à $\overline{m}$ et notée $h_{[\overline{m}]}$, est la bijection réciproque de $\overline{T} : I_{[\overline{m}]} \to J_{[\overline{m}]}$ et est définie par

$$h_{[\overline{m}]}(x_1, \ldots, x_\gamma) = (h_{[m_1]}(x_1), \ldots, h_{[m_\gamma]}(x_\gamma)).$$

En particulier, son jacobien satisfait

$$\mathrm{Jac}(h_{[\overline{m}]})(x_1, \ldots, x_\gamma) = \prod_{i=1}^{\gamma} |h'_{[m_i]}(x_i)|.$$

Comme pour les sources dynamiques classiques, à chaque préfixe $\overline{[\omega]} = \overline{m}_1 \ldots \overline{m}_p$ productible par la source $\overline{\mathcal{S}}$, il existe une branche inverse multidimensionnelle $h_{[\overline{\omega}]} = h_{[\overline{m}_1]} \circ \ldots \circ h_{[\overline{m}_p]}$ de profondeur p qui lui est associée. Cette branche est définie sur un intervalle $J_{h_{[\overline{\omega}]}} = J_{[\overline{\omega}]}$ et est à valeurs dans $I_{h_{[\overline{\omega}]}} = I_{[\overline{\omega}]}$.

Tous les opérateurs pour les sources dynamiques classiques s'étendent à la source multidimensionnelle $\overline{\mathcal{S}}_\gamma$. L'opérateur composante associé à un préfixe $\overline{\omega}$ est défini par

$$\mathbb{G}_{<\gamma>,[\overline{\omega}]}[F] := \mathrm{Jac}(h_{[\overline{\omega}]})F \circ h_{[\overline{\omega}]}\mathbf{1}_{J_{[\overline{\omega}]}}.$$

où F est une fonction complexe définie sur $\overline{I}$. La probabilité du mot $\overline{\omega}$ est alors donnée par

$$p_{\overline{\omega}} = \int_{\overline{I}} \mathbb{G}_{<\gamma>,[\overline{\omega}]}[\overline{f}](x)dx \tag{8.1}$$

où $\overline{f}$ est la densité initiale. De son coté, le transformateur de densité satisfait

$$\mathbb{G}_{<\gamma>}[F] := \sum_{\overline{m} \in \overline{\mathcal{E}}} \mathbb{G}_{<\gamma>,[\overline{m}]}[F] = \sum_{\overline{m} \in \overline{\mathcal{E}}} \mathrm{Jac}(h_{[\overline{m}]})F \circ h_{[\overline{m}]}\mathbf{1}_{[J_{\overline{m}}]}$$

et l'opérateur contraint à un sous-alphabet $\mathcal{E}'$ est défini par

$$\mathbb{G}_{<\gamma>,[\mathcal{E}']}[F] := \sum_{\overline{m} \in \overline{\mathcal{E}'}} \mathbb{G}_{<\gamma>,[\overline{m}]}[F] = \sum_{\overline{m} \in \overline{\mathcal{E}'}} \mathrm{Jac}(h_{[\overline{m}]})F \circ h_{[\overline{m}]}\mathbf{1}_{J_{[\overline{m}]}}$$

Nous introduisons maintenant un nouvel opérateur : l'opérateur de transfert pondéré multidimensionnel. Il admet une expression similaire aux opérateurs de transfert pondérés que nous avons rencontré lors de l'analyse des coûts additifs sur les entiers, mais il ne dépend que d'un paramètre complexe w et d'un coût élémentaire c sur l'alphabet $\overline{\mathcal{E}}$. Il est défini par

$$\mathbb{G}_{<\gamma>,w,[c]} = \sum_{\overline{m} \in \overline{\mathcal{E}}} \mathbb{G}_{<\gamma>,w,[c],[\overline{m}]} \quad \text{avec} \quad \mathbb{G}_{<\gamma>,w,[c],[\overline{m}]}[F] = e^{wc(\overline{m})}\mathrm{Jac}(h_{[\overline{m}]})F \circ h_{[\overline{m}]}\mathbf{1}_{J_{[\overline{m}]}},$$

et si c est étendu de façon additive sur les branches inverses de $\overline{T}$ ou sur les mots, i.e.,

$$c(\overline{m}_1\overline{m}_2 \ldots \overline{m}_p) = c(h_{\overline{m}_1} \circ h_{\overline{m}_2} \circ \ldots \circ h_{\overline{m}_p}) = c(\overline{m}_1) + c(\overline{m}_2) + \ldots + c(\overline{m}_p),$$

les itérés de l'opérateur de transfert pondéré multidimensionnel satisfont

$$\mathbb{G}^n_{<\gamma>,w,[c]} = \sum_{\overline{\omega} \in \overline{\mathcal{E}}^n} \mathbb{G}_{<\gamma>,w,[c],[\omega]} \quad \text{avec} \quad \mathbb{G}_{<\gamma>,w,[c],[\omega]}[F] = e^{wc(\overline{\omega})}\mathrm{Jac}(h_{[\overline{\omega}]})F \circ h_{[\overline{\omega}]}\mathbf{1}_{J_{[\overline{\omega}]}}.$$

Il est à noter que si la source $\mathcal{S}$ est markovienne, elle est par définition irréductible. Cette irréductibilité se transmet aussi à la source multidimensionnelle.
Nous retrouverons naturellement l'opérateur de transfert pondéré multidimensionnel pour estimer les sommes $S_{\gamma,m}$.

8.1.6 Lien entre les sources

La source $\underline{\mathcal{S}}_\gamma$ est définie à partir de γ copies de la source $\mathcal{S}$ et émet un $\underline{1}$ (resp. un $\underline{0}$) si les γ copies de $\mathcal{S}$ émettent (resp. n'émettent pas) en même temps un 1. La source $\overline{\mathcal{S}}_\gamma$ est définie à partir de γ copies de la source $\mathcal{S}$ et émet des symboles qui sont les γ-uplets des symboles 0 ou 1 émis par les γ sources $\mathcal{S}$. Le lien entre les deux sources est maintenant claire. Si $\widetilde{\sigma}$ est la fonction définie sur les γ-uplets de la manière suivante,

$$\widetilde{\sigma} : \{0,1\}^\gamma \to \{\underline{0},\underline{1}\}, \qquad \text{et} \qquad \widetilde{\sigma}(x) = \underline{1} \quad \text{ssi} \quad x = (1,\ldots,1),$$

alors nous pouvons écrire $\widetilde{\sigma}(\overline{\mathcal{S}}_\gamma) = \underline{\mathcal{S}}_\gamma$ où $\widetilde{\sigma}(\overline{\mathcal{S}}_\gamma)$ signifie que chaque γ-uplet émis par $\overline{\mathcal{S}}_\gamma$ est transformé par $\widetilde{\sigma}$ en un élément de $\{\underline{0},\underline{1}\}$.
Raisonner sur la source multidimensionnelle $\overline{\mathcal{S}}$ revient donc exactement à raisonner sur la source $\underline{\mathcal{S}}$.

8.1.7 Opérateur associé à un langage et dictionnaire

Soit $\mathcal{E}$ un alphabet. Dans cette section, l'opérateur $\mathbf{H}_{[\omega]}$ désigne l'un des opérateurs composantes $\mathbf{G}_{[\omega]}$, $\mathbb{G}_{<\gamma>,[\omega]}$, $\mathbb{G}_{<\gamma>,w,[c],[\omega]}$ (c est additif). Dans tous les cas, l'opérateur associé au mot $\omega = \omega_1\omega_2\ldots\omega_m$ de longueur m est donné par

$$\mathbf{H}_{[\omega]} = \mathbf{H}_{[\omega_m]} \circ \mathbf{H}_{[\omega_{m-1}]} \circ \ldots \circ \mathbf{H}_{[\omega_1]}.$$

Définition 26 (Opérateur associé à un langage) *Etant donné un langage L de $\mathcal{E}^\star$ (i.e. $L \subset \mathcal{E}^\star$), l'opérateur associé au langage L, noté $\mathbf{H}_{[L]}$, est défini par*

$$\mathbf{H}_{[L]} = \sum_{\omega \in L} \mathbf{H}_{[\omega]}.$$

Comme pour les séries génératrices ordinaires, un dictionnaire existe pour les opérateurs associés à des langages. Si L_1 et L_2 sont deux langages, alors les opérateurs associés à la concaténation $L_1 \cdot L_2$ ou à l'union disjointe $L_1 \oplus L_2$ sont donnés par

$$\mathbf{H}_{[L_1\cdot L_2]} = \mathbf{H}_{[L_2]} \circ \mathbf{H}_{[L_1]}, \quad \text{et} \quad \mathbf{H}_{[L_1\oplus L_2]} = \mathbf{H}_{[L_1]} + \mathbf{H}_{[L_2]}.$$

En particulier, l'opérateur associé au langage $L_1^\star$ vérifie

$$\mathbf{H}_{[L_1^\star]} = \sum_{m\geq 0} \mathbf{H}_{[L_1]}^m = (\mathbf{I} - \mathbf{H}_{[L_1]})^{-1}.$$

Dans la suite, nous utiliserons exclusivement les langages associés à un motif ligne, c'est-à-dire l'ensemble des mots qui contiennent le motif ligne. Nous verrons que ces langages sont des concaténations d'autres langages et nous exprimerons les opérateurs associés à l'aide du dictionnaire précédent.

8.2 Sources dynamiques et fouille de données

La partie précédente a décrit tous les aspects importants des sources dynamiques et a introduit plusieurs opérateurs. Pour nos problèmes de fouille de données, nous allons nous concentrer sur les sources dynamiques fortement markoviennes qui comprennent à la fois les les sources de Bernoulli simples et par groupes, les sources dynamiques complètes, ainsi que les chaînes de Markov à matrice de transition strictement positive. Pour ce type de sources et avec un alphabet adéquat, nous montrons que les conditions 8 page 166, 9 page 166 et 10 page 167 respectivement associées aux théorèmes 18 page 166, 19 page 167 et 20 page 169 sont vérifiées.

8.2.1 Alphabets des sources dynamiques

Les bases de données que nous considérons sont binaires. A priori, l'alphabet adéquat est $\{0,1\}$. Mais le modèle de Bernoulli par groupe peut être considéré comme une source de Bernoulli sur l'alphabet $\{100\ldots0, 010\ldots0, \ldots, 000\ldots1\}$. De même, les chaînes de Markov d'ordre K sur $\{0,1\}$ sont des chaînes de Markov d'ordre 1 sur $\{0,1\}^K$. Nous sommes naturellement amenés à considérer tous les alphabets $\mathcal{E}$ de la forme

$$\mathcal{E} \subset \{0,1\}^K \qquad \text{avec} \qquad K \in \mathbb{N}^\star.$$

Avec ce type d'alphabet, les mots produits sont vus comme une concaténation de morceaux de taille K. Il faut toutefois éviter certains alphabets. Par la suite, nous allons "coller les morceaux" pour construire des mots qui contiennent des motifs lignes. Si X et Y sont deux motifs lignes avec $X \subset Y$, nous voulons que le nombre de mots qui contiennent X soit strictement plus grand que celui de Y. L'hypothèse suivante nous assure cette situation.

Hypothèse 11 (Alphabet) *Soit $w \in \{0,1\}^K$ pour $K \geq 2$. Si $w = w_1 \ldots w_K$, on note $I(w) = \{i \ : \ w_i = 1\}$ l'ensemble des indices où w admet un 1, et pour $X \subset \{1,\ldots,K\}$, l'ensemble $\mathcal{E}_X$ désigne les mots w de $\mathcal{E}$ tels que $X \subset I(w)$. Si $X \subsetneq Y \subset \{1,\ldots,K\}$, alors $\mathcal{E}_Y \subsetneq \mathcal{E}_X$.*

Les principaux alphabets qui nous intéressent, à savoir les alphabets des sources de Bernoulli par groupe ou l'alphabet $\{0,1\}^K$ tout entier vérifient l'hypothèse précédente.
L'hypothèse 8 suppose la décroissance exponentielle des probabilités des motifs, i.e., il existe une constante M et un réel $\theta < 1$ tels que pour tout motif ligne X, $p_X \leq \theta^{|X|}$. L'hypothèse 9 est plus forte que l'hypothèse 8 et suppose qu'il existe $\theta < 1$ tel que pour tout couple de motifs lignes X, Y avec $X \subsetneq Y$, on a $p_Y/p_X < \theta$. Finalement, l'hypothèse 10suppose que pour tout entier γ, il existe trois constantes κ_γ, λ_γ et θ_γ avec $\kappa_\gamma > 0$, $\lambda_\gamma > 1$ et $\theta_\gamma \in]0,1[$ telles que

$$S_{\gamma,m} = \kappa_\gamma \cdot \lambda_\gamma^m (1 + O(\theta_\gamma^m)) \quad \text{et} \quad \lambda_\gamma > \lambda_{\gamma+1}.$$

Le théorème suivante montre que les sources dynamiques fortement markoviennes vérifient les trois hypothèses précédentes.

Théorème 21 *Soit une source dynamique $\mathcal{S}$ fortement markovienne, dilatante, dont l'alphabet vérifie l'hypothèse 11.*
Alors, $\mathcal{S}$ vérifie les trois hypothèses 8, 9 et 10 et donc les trois théorèmes principaux du chapitre précédent, à savoir les théorèmes 18, 19 et 20.

8.2.2 Génération des probabilités des motifs lignes

Les probabilités des motifs lignes sont essentielles puisqu'elles interviennent dans les conditions de tous les théorèmes. Nous fixons donc $X = \{v_1, \ldots, v_j\}$ un motif ligne de $\{1,\ldots,m\}$. Nous rappelons que le motif ligne est contenu dans un objet o (ou dans une ligne) si o contient les attributs $a_{v_1}, \ldots, a_{v_j}$. D'un point de vue texte, cela signifie que le mot associé à o admet des 1 aux positions pointées par X. Dans ce cas, le motif ligne est dit contenu dans le mot. La difficulté dans notre contexte est que la source émet des paquets de K symboles 0 ou 1 à travers l'alphabet $\mathcal{E}$. Avoir un 1 dans une position donnée implique donc des contraintes autour de cette position. De plus, deux 1 peuvent appartenir au même paquet.

Dans toute la suite, nous fixons K un entier positif et $\mathcal{E}$ un alphabet inclus dans $\{0,1\}^K$. Nous supposons également que la longueur m des mots est de la forme $m = K \cdot m_1$ avec m_1 un entier positif.

Définition 27 *Soit X un motif ligne de $\{1,\ldots,m\}$. Le langage $L(X)$ associé au motif ligne X est l'ensemble des mots de longueur m_1 sur l'alphabet $\mathcal{E}$ contenant le motif ligne X,*

$$L(X) = \{\omega = \omega_1 \ldots \omega_{m_1} = l_1 \ldots l_m \mid \forall i,\ \omega_i \in \mathcal{E} \quad et \quad \forall j \in X,\ l_j = 1\}. \tag{8.2}$$

La probabilité p_X qu'un motif ligne soit présent dans un mot s'exprime en fonction de l'opérateur associé au langage $L(X)$,

$$p_X = \sum_{\omega \in L(X)} p_\omega = \sum_{\omega \in L(X)} \int_0^1 \mathbf{G}_{[\omega]}[f](t)dt = \int_0^1 \mathbf{G}_{[L(X)]}[f](t)dt.$$

Nous allons maintenant décomposer l'ensemble $L(X)$ et appliquer le dictionnaire sur les opérateurs associés à un langage afin de trouver une nouvelle expression des p_X.

Nous appelons contrainte élémentaire (sur les mots de l'alphabet) un sous-ensemble de $\{1,\ldots,K\}$. A un motif ligne X, nous associons un m_1-uplet de contraintes élémentaires $(X_1,\ldots,X_{m_1})$ où

$$X_i = \{j \in \{1,\ldots,K\} \quad \text{tel que} \quad u = K\cdot(i-1)+j \quad \text{appartient à } X\}.$$

Par exemple, pour $m_1 = 4$, $K = 3$ et $X = \{2,4,6,7,11\}$, on a $(X_1,X_2,X_3,X_4) = (\{2\},\{1,3\},\{1\},\{2\})$.

indice dans le mot	1	2	3	4	5	6	7	8	9	10	11	12
indice dans le morceau	1	2	3	1	2	3	1	2	3	1	2	3
motif ligne X		2			4	6	7				11	
contraintes élémentaires		{2}			{1,3}			{1}			{2}	

Les contraintes élémentaires nous indiquent les positions où il doit y avoir un 1 dans les mots de l'alphabet pour satisfaire le motif ligne en un morceau donné.

Le langage $L(Y)$ associé à une contrainte élémentaire $Y \subset \{1,\ldots,K\}$ est l'ensemble $\mathcal{E}_Y$ des mots de l'alphabet ayant un 1 aux positions indicées par Y,

$$\mathcal{E}_Y = \{\omega = m_1 \ldots m_K \in \mathcal{E} \mid \forall i \in Y, m_i = 1\}.$$

Avec l'exemple précédent, $\mathcal{E}_{X_2} = \{101, 111\}$ si l'alphabet est $\mathcal{E} = \{0,1\}^K$ et $\mathcal{E}_{X_2} = \emptyset$ si l'alphabet est $\{001, 010, 100\}$.

Il est maintenant clair que le langage associé à un motif ligne X se décompose en fonction des langages des contraintes élémentaires,

$$L(X) = L(X_1)\cdot L(X_2)\cdot \ldots \cdot L(X_{m_1}).$$

En appliquant le dictionnaire sur les opérateurs, l'opérateur $\mathbf{G}_{[L(X)]}$ s'écrit également sous la forme

$$\mathbf{G}_{[L(X)]} = \mathbf{G}_{[L(X_{m_1})]} \circ \ldots \circ \mathbf{G}_{[L(X_1)]}$$

et la probabilité p_X qu'un mot contient le motif ligne X satisfait la formule alternative

$$p_X = \int_0^1 \mathbf{G}_{[L(X_{m_1})]} \circ \ldots \circ \mathbf{G}_{[L(X_1)]}[f](t)dt. \tag{8.3}$$

8.2.3 Sources dynamiques markoviennes : première et deuxième conditions

Nous souhaitons montrer que pour toute source dynamique fortement markovienne dilatante, il existe une constante $\theta < 1$ telle que pour tout motifs X, Y avec $X \subsetneq Y$, on a $p_Y/p_X < \theta$. Ceci établira l'hypothèse 9 pour ces sources, qui elle-même entraîne l'hypothèse 8.

Considérons deux motifs X et Y avec $|Y| = |X| + 1$. Les langages associés à X et Y sont alors de la forme

$$L(X) = L_1 \cdot \mathcal{E}_A \cdot \mathcal{E}_X \cdot \mathcal{E}_B \cdot L_2 \quad \text{et} \quad L(Y) = L_1 \cdot \mathcal{E}_A \cdot \mathcal{E}_X \cdot \mathcal{E}_B \cdot L_2,$$

avec L_1 et L_2 des langages, et $\mathcal{E}_A$, $\mathcal{E}_B$, $\mathcal{E}_X$ et $\mathcal{E}_Y$ des langages associés à des contraintes élémentaires. L'hypothèse sur l'alphabet induit en plus l'inclusion stricte $\mathcal{E}_Y \subsetneq \mathcal{E}_X$. Nous notons L le langage $L(X)/L(Y)$. Alors, nous avons clairement

$$\begin{aligned} p_X &= p_Y + p_L = p_Y + \sum_{w \in L} p_w \geq p_Y + |L| \min_{w \in L} p_w \\ &\geq p_Y (1 + \frac{|L|}{|L(Y)|} \times \frac{\min_{w \in L} p_w}{\max_{w \in L(Y)} p_w}). \end{aligned}$$

Comme la source est fortement irréductible, toutes les probabilités p_w sont strictement positives et la taille de $L(Y)$ et L vérifient $|L|/|L(Y)| = (|\mathcal{E}_X| - |\mathcal{E}_Y|)/|\mathcal{E}_Y| \geq 1/2^K$. Si γ est défini par

$$\gamma = \frac{\min_{w \in L} p_w}{\max_{w \in L(Y)} p_w},$$

alors $p_X \geq (1 + \gamma/2^K) \cdot p_Y$, et, dans le cas où γ n'est pas nul, la condition 9 est démontrée avec $\theta = 1/(1 + \gamma/2^K)$.

Lemme 19 *Considérons γ le réel défini par*

$$\gamma = \frac{\min_{w \in L} p_w}{\max_{w \in L(Y)} p_w}.$$

Alors, le réel γ est minoré par une constante ne dépendant que de la source et de la densité initiale, et non des langages L et $L(Y)$. Plus précisément, nous avons

$$\gamma \geq B \cdot D^2 \qquad \text{où} \qquad B = \frac{\min f}{\max f}$$

et D est la constante de distorsion du lemme 18.

Preuve. Soit w un mot de L. Alors w s'écrit sous la forme

$$w = \alpha \cdot a \cdot x \cdot b \cdot \beta \qquad \text{avec} \qquad (\alpha, a, x, b, \beta) \in L_1 \times \mathcal{E}_A \times (\mathcal{E}_X/\mathcal{E}_Y) \times \mathcal{E}_B \times L_2.$$

Comme la source est fortement markovienne, la probabilité du mot w peut s'écrire sous la forme

$$p_w = \int_0^1 |h'_{[w]}(t)| f \circ h_{[w]}(t) \mathbf{1}_{J_{[w]}}(t) dt \quad \text{avec} \quad h_{[w]} = h_{[\alpha]} \circ h_{x|a} \circ h_{b|x} \circ h_{[\beta]},$$

et f la densité initiale. En minorant f par son minimum (f est strictement positive) puis en effectuant le changement de variable $z = h_{[\beta]}(t)$, la probabilité p_w est minorée par

$$p_w \geq (\min_I f) \int_{I_{[\beta]}} |(h_{[\alpha]} \circ h_{x|a} \circ h_{b|x})'(z)| dz.$$

La propriété de distorsion donnée par le lemme 18 implique qu'il existe une constante D telle que pour toute branche inverse h de $\mathcal{H}^\star$, on ait $\inf_{J_h} |h'|/\sup_{J_h} |h'| \geq D > 0$. En particulier, pour $y \in \mathcal{E}_Y$, nous obtenons

$$p_w \geq (\min_I f) D \int_{I_{[\beta]}} |(h_{[\alpha]} \circ h_{y|a} \circ h_{b|y})'(z)| \frac{|(h_{x|a} \circ h_{b|x})'(z)|}{|(h_{y|a} \circ h_{b|y})'(z)|} dz.$$

En appliquant une seconde fois la propriété de distorsion, nous avons aussi

$$\min_{(a,b,x,y)\in\mathcal{E}^4} \frac{\min_{I_{[b]}} |(h_{y|a} \circ h_{b|y})'|}{\max_{I_{[b]}} |(h_{x|a} \circ h_{b|x})'|} \geq D,$$

et la probabilité de w satisfait alors

$$p_w \geq (\min_I f) \cdot D^2 \cdot \int_{I_{[\beta]}} |(h_{[\alpha]} \circ h_{y|a} \circ h_{b|y})'(z) dz.$$

En réintroduisant f dans l'intégrale avec la minoration $f/(\max_I f) \leq 1$, on trouve finalement

$$p_w \geq B \cdot D^2 \cdot p_{\overline{w}}$$

avec $\overline{w} \in L(Y)$ et $B = \min f / \max f$.
Ceci termine la preuve du lemme. ■
Le lemme montre que la source vérifie la condition 9 avec un θ explicite,

$$\theta = \frac{1}{1 + B \cdot D^2 \cdot 2^{-K}}.$$

Ceci montre deux des trois assertions du théorème 21.■

8.2.4 Sommes $S_{\gamma,m}$ et opérateur multidimensionnel

Les sommes $S_{\gamma,m}$ sont les objets principaux dans l'énoncé des hypothèses du théorème 20. Elles sont définies par

$$S_{\gamma,m} = \sum_{\underline{\omega}\in\{\underline{0},\underline{1}\}^m} 2^{c_\gamma(\underline{\omega})} p_{\underline{\omega}}.$$

Le lien entre la source $\underline{\mathcal{S}}$ (qui produit des mots sur $\{\underline{0},\underline{1}\}$) et la source multidimensionnelle $\overline{\mathcal{S}}$ (qui produit des γ-uplets de 0 et 1) donne aussi la formule alternative suivante,

$$S_{\gamma,m} = \sum_{\overline{\omega}\in\overline{\mathcal{E}}^{m/K}} 2^{\overline{c}_\gamma(\overline{\omega})} p_{\overline{\omega}}$$

où $\overline{c}_\gamma(\overline{w})$ est le nombre de γ-uplets de 1 dans le mot $\overline{w}$, K est la taille des éléments de l'alphabet et $\underline{c}_\gamma(\underline{\omega})$ est le nombre de $\underline{1}$ dans $\underline{\omega}$. Si ω_1 et ω_2 sont deux mots sur $\overline{\mathcal{E}}$, alors le nombre de colonnes de 1 dans le mot $\omega_1\omega_2$ est bien le nombre de colonnes de 1 dans ω_1 plus le nombre de colonnes de 1 dans ω_2. Le coût élémentaire $\overline{c}_\gamma$ est donc additif et satisfait

$$\overline{c}_\gamma(\omega_1 \cdot \omega_2) = \overline{c}_\gamma(\omega_1) + \overline{c}_\gamma(\omega_2).$$

Considérons maintenant l'opérateur de transfert multidimensionnel $\mathbb{G}_{<\gamma>,w,[\overline{c}_\gamma]}$ associé au coût additif $\overline{c}_\gamma$. Les itérés de $\mathbb{G}_{<\gamma>,w,[\overline{c}_\gamma]}$ sont donnés par

$$\mathbb{G}^p_{<\gamma>,w,[\overline{c}_\gamma]} = \sum_{\omega\in\overline{\mathcal{E}}^p} e^{w\overline{c}_\gamma(\omega)} \mathbb{G}_{<\gamma>,[\omega]}.$$

Avec la formule 8.1 qui exprime la probabilité d'un mot ω en fonction de l'opérateur associé, nous obtenons aussi l'égalité

$$\sum_{\omega \in \overline{\mathcal{E}}^{m/K}} e^{w\overline{c}_\gamma(\omega)} p_\omega = \int_{\overline{I}} \mathbb{G}^p_{<\gamma>,w,[\overline{c}_\gamma]}[\overline{f}](t)dt$$

où nous rappelons que la fonction $\overline{f}$ est la densité initiale sur $\overline{I}$ induite par la densité initiale de la source $\mathcal{S}$. En posant $w = \log 2$, la somme $S_{\gamma,m}$ satisfait

$$S_{\gamma,m} = \int_{\overline{I}} \mathbb{G}^{m/K}_{<\gamma>,w,[\overline{c}_\gamma]}[\overline{f}](t)dt. \tag{8.4}$$

En particulier, la série génératrice $S_\gamma(z,u)$ définie par

$$S_\gamma(z,u) = \sum_{\overline{\omega} \in \overline{\mathcal{E}}^\star} z^{|\overline{\omega}|} u^{\overline{c}_\gamma(\overline{\omega})} p_{\overline{\omega}}$$

vérifie

$$S_\gamma(z,u) = \int_{\overline{I}} (\mathbf{I} - z^K \mathbb{G}_{<\gamma>,\log u,[\overline{c}_\gamma]})^{-1}[\overline{f}](t)dt.$$

8.2.5 Sources dynamiques markoviennes : troisième condition

Nous montrons dans cette section que les sources dynamiques fortement markoviennes vérifient la troisième hypothèse, à savoir, pour tout entier γ, il existe trois constantes $\kappa_\gamma > 0$, $\lambda_\gamma > 1$ et $\theta_\gamma \in]0,1[$ telles que

$$S_\gamma = \kappa_\gamma \cdot \lambda_\gamma^m (1 + O(\theta_\gamma^m)) \quad \text{et} \quad \lambda_\gamma > \lambda_{\gamma+1}.$$

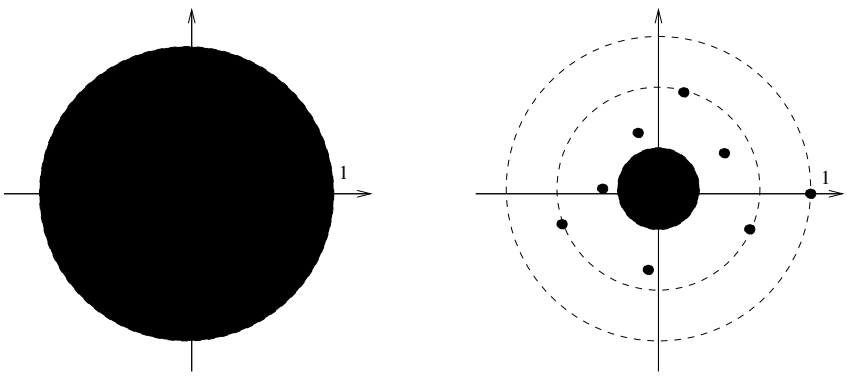
Les sommes S_γ sont définies par

$$S_{\gamma,m} = \sum_{X \in \{1,\ldots,m\}} p_X^\gamma$$

Les propriétés spectrales de l'opérateur multidimensionnel sont ici essentielles.

8.2.5.1 Spectre de l'opérateur multidimensionnel

Espaces fonctionnels. Selon l'espace de fonctions, le spectre des opérateurs changent. Par exemple, sur l'espace des fonctions intégrables L^1, le transformateur de densité admet tout le cercle unité (dans $\mathbb{C}$) comme spectre. En revanche, sur l'espace des fonctions C^1 sur l'intervalle $\overline{I} = [0,1]$ et muni de la norme $||f|| = ||f||_\infty + ||f'||_\infty$, le transformateur de densité admet 1 comme unique valeur propre dominante séparée du reste du spectre par un saut spectral.



spectre sur L^1 spectre sur C^1

Un bon espace fonctionnel pour nos opérateurs est un espace sur lequel ils admettent une unique valeur propre dominante séparée du reste du spectre par un saut spectral. Selon les propriétés des branches du système dynamique, il est possible de choisir des espaces de fonctions analytiques [Val97a, Val01, Val98a, Val03] pour des systèmes à branches holomorphes contractantes. Pour certains systèmes généraux [BDV02], l'espace adéquat est l'ensemble des fonctions à variation bornée. Plus récemment, Chazal et Maume-Deschamps [CMD04] ont utilisé l'espace des fonctions lipschitziennes par morceaux pour traiter des sources dynamiques markoviennes dont les branches ne sont pas holomorphes. Finalement dans [BV05, BV04], Baladi et Vallée ont préféré l'espace des fonctions $C^1(\overline{I})$ avec une norme particulière pour analyser les opérateurs de transfert pondérés.

Nous disposons donc d'un certain nombre d'espaces pour nos opérateurs. Dans cette partie, seuls des systèmes dynamiques fortement markoviens sont abordés. Nous n'utiliserons donc pas les fonctions à variation bornée. Dans le chapitre "calcul de constantes" (Chapitre 5), l'espace fonctionnel était un espace de fonctions analytiques sur un disque alors que pendant l'analyse des algorithmes euclidiens (cf. chapitre 3), nous utilisions l'espace des fonctions C^1.

Dans toute la suite, nous choisissons l'espace fonctionnel $C^1(I)$ pour les sources dynamiques complètes, l'espace des fonctions C^1 sur chaque élément de la partition $(I_{[m]})_{m\in\mathcal{E}}$ pour les sources dynamiques markoviennes, l'espace $C^1(\overline{I})$ pour la source dynamique multidimensionnelle $\overline{\mathcal{S}}_\gamma$ associée à un source complète et l'espace des fonctions C^1 sur chaque élément de la partition $(I_{[\overline{m}]})_{\overline{m}\in\overline{\mathcal{E}}}$ pour la source dynamique multidimensionnelle $\overline{\mathcal{S}}_\gamma$ associée à un source markovienne.

Avec les espaces précédemment cités, les propriétés du spectre de l'opérateur multidimensionnel sont résumées par la proposition suivante.

Proposition 25 *Pour w réel, l'opérateur de transfert pondéré multidimensionnel $\mathbb{G}_{\gamma,w,[c]}$ admet une unique valeur propre dominante $\lambda(\gamma, w, [c])$ positive, isolée du reste du spectre par un saut spectral et dont le vecteur propre associé est strictement positif. De plus, si w est strictement positif, la valeur propre dominante est strictement plus grande que 1 (dès que le coût est non-nul).*

Décomposition spectrale. Les propriétés spectrales dominantes de l'opérateur multidimensionnel fait qu'il peut être séparé en deux parties : une partie dominante liée à la partie dominante du spectre et une partie sous-dominante correspondant au reste du spectre. Cette décomposition, appelée décomposition spectrale, est de la forme

$$\mathbb{G}_{<\gamma>}[f] = \lambda_{<\gamma>}\mathbb{P}_{<\gamma>}[f] + \mathbb{N}_{<\gamma>}[f],$$

où $\mathbb{P}_{<\gamma>}$ est un projecteur sur l'espace propre dominant et $\mathbb{N}_{<\gamma>}$ un opérateur de rayon spectral strictement plus petit que $|\lambda|$. De plus, $\mathbb{P}_{<\gamma>}$ et $\mathbb{N}_{<\gamma>}$ commutent.
La décomposition spectrale induit aussi une décomposition spectrale des itérés,

$$\mathbb{G}^n_{<\gamma>}[f] = \lambda^n_{<\gamma>}\mathbb{P}_{<\gamma>}[f] + \mathbb{N}^n_{<\gamma>}[f],$$

ainsi qu'une décomposition du quasi-inverse,

$$(\mathbf{I} - \mathbb{G}_{<\gamma>})^{-1}[f] = \frac{1}{1-\lambda_{<\gamma>}}\mathbb{P}_{<\gamma>}[f] + (\mathbf{I} - \mathbb{N}_{<\gamma>})^{-1}[f].$$

8.2.5.2 Troisième hypothèse

Nous montrons maintenant que les sources dynamiques fortement markoviennes vérifient la troisième hypothèse (hypothèse 10). Pour simplifier les notations, nous notons $\mathbb{G}_{<\gamma>}$ l'opérateur de transfert multidimensionnel $\mathbb{G}_{<\gamma>,\log 2,[\overline{c}_\gamma]}$.
L'égalité 8.4 relie la somme $S_{\gamma,m}$ à l'opérateur multidimensionnel de la manière suivante

$$S_{\gamma,m} = \int_{\overline{I}} \mathbb{G}^{m/K}_{<\gamma>}[\overline{f}](t)dt.$$

La décomposition spectrale de l'opérateur multidimensionnel appliquée à l'égalité précédente montre que la somme $S_{\gamma,m}$ a la forme attendue

$$S_\gamma = \kappa_\gamma \cdot \lambda_\gamma^m (1 + O(\theta_\gamma^m)) \quad \text{et} \quad \lambda_\gamma > 1,$$

où les constantes κ_γ, λ_γ et θ_γ sont liées aux objets spectraux dominants de l'opérateur multidimensionnel :

$$\lambda_\gamma = \lambda(\gamma, \log 2, [\overline{c}_\gamma])^{1/K}, \qquad \kappa_\gamma = \int_{\overline{I}} \mathbb{P}_\gamma[\overline{f}](x)dx \qquad \theta_\gamma = r_\gamma,$$

avec r_γ le rayon spectral sous dominant.
Il reste à montrer que $\lambda_{\gamma+1} < \lambda_\gamma$. Pour cela, il suffit d'établir qu'il existe un réel $\theta_1 < 1$ tel que

$$S_{\gamma+1,m} \leq \theta_1^m S_{\gamma,m}.$$

Cette propriété découle du lemme suivant.

Lemme 20 *Il existe $\theta_2 < 1$ une constante ne dépendant que du système dynamique, de γ et de la densité initiale telle que, pour toute fonction f positive sur $I^{\gamma+1}$,*

$$\int_I \mathbb{G}_{<\gamma+1>}[f](x,t)dt \leq \theta_2 \mathbb{G}_{<\gamma>}[F](x) \quad \textit{où} \quad F(x) = \int_I f(x,t)dt.$$

En choisissant $f = \mathbb{G}^{k-1}_{<\gamma+1>}[g]$, on montre très rapidement par récurrence l'inégalité

$$\int_I \mathbb{G}_{<\gamma+1>}[f](x,t)dt \leq \theta_2^k \mathbb{G}^k_{<\gamma>}[g](x).$$

Comme S_γ, m est liée à l'intégrale de $\mathbb{G}^{m/K}_{<\gamma>}$, les sommes $S_{\gamma,m}$ vérifient clairement l'inégalité $S_{\gamma+1,m} \leq \theta_2^{m/K} S_{\gamma,m}$ (on pose $\theta_1 = \theta_2^{1/K}$).
Le lemme prouve que les sources dynamiques fortement markoviennes satisfont l'hypothèse 10 associée au seuil fixe et au théorème 20 (page 169) .

Preuve du lemme 20. Le lemme 20 se démontre en appliquant deux fois le lemme (trivial) qui suit.

Lemme 21 *Soit $(\alpha_i)_{i\in I}$ et $(g_i)_{i\in I}$ deux suites réelles et positive avec $\alpha_i \leq 1$. S'il existe en élément $i_0 \in I$, et un réel $A > 0$ tels que*

$$\alpha_{i_0} < 1 \quad \textit{et} \quad g_{i_0} \geq A \sum_{i\in I} g_i,$$

alors

$$\sum_{i\in I} \alpha_i g_i \leq (1 - (1 - \alpha_{i_0})A) \sum_{i\in I} g_i.$$

Nous considérons tout d'abord la relation

$$\int_I \mathbb{G}_{<\gamma+1>}[f](x,t)dt = \sum_{h\in\mathcal{H}^\gamma} 2^{c_\gamma(h)} J_h(x) A_h(x)$$

avec $c_\gamma(h)$ le coût du mot associé à la branche multidimensionnelle h, $J_h(x)$ le jacobien de h et

$$A_h(x) = \sum_{k\in\mathcal{H}} 2^{c_\gamma(h)-c_{\gamma+1}(h,k)} \int_{I_k} f(h(x),y)dy.$$

Pour les h tel que $c_\gamma(h)$ est strictement positif, il existe un intervalle (associé à un branche k) pour lequel $c_\gamma(h) - c_{\gamma+1}(h,k)$ est négatif strictement. Alors A_h satisfait les hypothèse du lemme 21 avec $\alpha_{i_0} = 1/2$ et $A = L\cdot B\cdot D$ où D est la constante de distorsion rencontrée précédemment, $B = \min f / \max f$ et $L = \min_{h\in\mathcal{H}} |h'|$. Si $\theta_3 = (1 - L\cdot B\cdot D/2)$, alors les A_h tels que $c_\gamma(h) \geq 1$ vérifient

$$A_h \leq \theta_3 \int_I f(h(x),y)dy.$$

Nous appliquons une deuxième fois le lemme 21, non plus sur les A_h mais sur la somme des A_h. Cette somme satisfait les hypothèses du lemme avec $\alpha_{i_0} = \theta_3$,

$$\forall h, \qquad A_h \leq \int_I f(h(x),y)dy$$

et

$$\exists h, \qquad A_h \leq \theta_3 \int_I f(h(x),y)dy.$$

Les g_i du lemme sont aussi donnés par

$$g_h = 2^{c_\gamma(h)} J_h(x) \int_I f(h(x),y)dy$$

et vérifient

$$g_h \geq \frac{1}{2^K}\frac{1}{|\mathcal{E}|^\gamma} BDP^\gamma \sum_h 2^{c_\gamma(h)} J_h(x) \int_I f(h(x),y)dy$$

où

$$P \leq \frac{\min_{h\in\mathcal{H},x\in J_h} |h'(x)|}{\max_{h\in\mathcal{H},x\in J_h} |h'(x)|}.$$

Le lemme donne alors l'existence d'un réel θ_2 tel que

$$\int_I \mathbb{G}_{<\gamma+1>}[f](x,t)dt \leq \theta_2 \sum_h 2^{c_\gamma(h)} J_h(x) \int_I f(h(x),y)dy = \theta_2 \mathbb{G}_{<\gamma>}[g](x)$$

avec $g = \int_I f(x,t)dt$ et θ_2 le réel

$$\theta_2 = 1 - \frac{L\cdot P^\gamma}{2^{K+1}\cdot|\mathcal{E}|^\gamma} B^2\cdot D^2.$$

Ceci termine la preuve du lemme 20. Nous venons donc de montrer que les sources dynamiques fortement markoviennes satisfont la condition 10 associée au seuil de fréquence fixe.■

8.3 Conclusion

Dans cette partie nous avons présenté le cadre général des sources dynamiques. Ces sources regroupent à la fois des sources simples comme les sources de Bernoulli et des sources complexes à mémoire non bornée. Malgré cela, nous disposons d'outils dont les opérateurs de transfert pour les étudier. Les sources dynamiques forment donc un bon intermédiaire entre les sources très générales et les sources analysables.

Nous avons montré que ces sources satisfont deux des trois hypothèses de la partie précédente. Ainsi, le nombre moyen de motifs fréquents pour les "bases dynamiques" est polynomial pour un seuil de fréquence proportionnel (théorème 18 page 166) et exponentiel pour un seuil fixe (théorème 20 page 169). Les bases dynamiques satisfont également l'hypothèse 9 du théorème 19 (page 167) pour le seuil logarithmique. De fait, pour un seuil au moins logarithmique, le nombre de motifs fermés est équivalent au nombre de motifs fréquents.

Nous terminons avec le modèle de bases de données qu'imposent les sources dynamiques. L'avantage est qu'il englobe un grand nombre de modèles simples théoriques déjà utilisés par plusieurs auteurs. Toutefois, avec ce modèle, les attributs dont les colonnes sont éloignées sont irrémédiablement peu corrélés. De plus l'aspect séquentiel du procédé, qui applique à chaque étape la même corrélation est très peu probable en pratique. Une première amélioration possible est de considérer plusieurs sources dynamiques qui interviendraient à des positions précises ou aléatoires. Je pense qu'il est tout à fait possible de considérer une chaîne de Markov pour chaque colonne modulo certaines conditions sur les matrices de transitions. Une autre idée de modèle serait de considérer des graphes de dépendances entre les attributs et de mettre un modèle sur ces graphes.

Conclusion de la Partie B

Résumé des résultats. Il s'agit d'une des premières analyses en moyenne dans le domaine de la fouille de données. Nous proposons d'abord un modèle général de base de données, dans lequel nous obtenons une estimation du nombre moyen de motifs fréquents et fermés. Ce modèle de base de données "informationnelle" considère que les mots-lignes sont produits indépendamment par la même source. Dans ce modèle, nous obtenons trois résultats, chacun étant associé à un seuil de fréquence et à une hypothèse sur la source. Sous une hypothèse 8, le nombre moyen de motifs γ-fréquents, pour un seuil γ linéaire en le nombre d'objets, est polynomial en le nombre d'attributs (théorème 18). C'est la première fois qu'un comportement polynomial du nombre de motifs fréquents est mis en évidence en fouille de données. Sous une hypothèse 9 et pour un seuil intermédiaire, le nombre de motifs fréquents est asymptotiquement équivalent au nombre moyen de motifs fermés. Ce résultat a déjà été observé pour des bases faiblement corrélées du type panier de la ménagère. Finalement, pour un seuil fixe, le nombre moyen de motifs γ-fréquents pour γ fixe, est exponentiel en le nombre d'attributs et polynomial en le nombre d'objets.

Tous ces résultats sont obtenus, pour une source a priori très générale, qui satisfait à des hypothèses "raisonnables". La source n'est pas vraiment explicitée, et reste une boite noire qui émet des symboles. Dans cette partie, nous avons exhibé un certain type de sources générales qui satisfont les hypothèses utiles à nos résultats. Les sources sans mémoire vérifient (avec quelques restrictions) toutes les hypothèses demandées, mais conduisent à des modèles de bases de données complètement non corrélées. Le modèle des chaînes de Markov ajoute des corrélations locales entre attributs qui se suivent dans la base de données, et ce modèle de source vérifie la première et la troisième hypothèse. Le modèle des sources dynamiques, qui contient les deux modèles précédents, permet de travailler avec des attributs dont les corrélations ne sont plus seulement locales, mais elles restent dépendantes de l'ordre des attributs dans la base. Ce modèle vérifie, lui aussi, la première et la troisième hypothèse.

Perspectives. Le modèle de "base dynamique" est trop séquentiel ce qui rend l'ordre des attributs trop important. On peut penser à une autre approche qui conduirait à un modèle qu'on pourrait baptiser "base de données arborescente". Supposons un instant que le problème de fouille de données soit (en partie) résolu. Il conduit alors à un graphe de dépendance orienté. On peut alors suivre la même démarche que dans [BV06] et construire une hybridation d'une source et d'un graphe. L'apparition de motifs dans un seul mot produit par une telle source hybride a déja été analysée. Si nous parvenons à traiter le même problème pour des mots émis en parallèle par la même source (hybride), nous pourrons résoudre les problèmes de cette partie de manière plus satisfaisante. En particulier, nous supprimons le côté artificiellement séquentiel des bases dynamiques.

La démarche d'analyse en moyenne d'algorithmes semble presque complètement inexistante dans le domaine de la fouille de données. La question de l'analyse réaliste des algorithmes d'extraction reste donc entièrement ouverte. L'analyse de l'algorithme APRIORI, fondé sur la notion

de motifs appelés motifs "candidats", nécessite que le comportement moyen de tels motifs soit d'abord élucidé. Ce comportement moyen pourrait probablement être mieux compris grâce à des méthodes assez semblables à celles que nous avons développées ici. Nous pensons que l'analyse des algorithmes en profondeur est, sous un modèle simple, susceptible d'une approche générique, puisque ces algorithmes ont une structure standard d'algorithmes récursifs. D'autres paramètres sont aussi très intéressants à étudier, car ils ont un rôle central dans les algorithmes du domaine : citons par exemple la taille de la bordure négative qui correspond aux traverses minimales d'un hypergraphe associé à la base, la taille de la plus grande traverse, le nombre de motifs fréquents de longueur donnée, la taille du plus grand motif, ...

Conclusion générale

Cette thèse regroupe des analyses dans deux domaines algorithmiques bien distincts : l'arithmétique et la fouille de données. Nous avons déjà expliqué, à fin de chaque partie, nos conclusions et décrit nos perspectives, dans chacun de ces deux domaines. Nous voudrions adopter ici, dans cette conclusion générale, un point de vue plus transverse, qui soit aussi plus relié aux aspects méthodologiques de notre travail.

Bien que cette thèse soit partagée en deux parties –une pour chaque domaine–, nous avons adopté, dans chacune d'elles, le point de vue de l'analyse dynamique, et nous avons contribué à faire évoluer ce domaine, encore naissant. Nous avons d'abord développé le cadre de l'analyse dynamique en distribution, dans le domaine de l'arithmétique. Nous montrons ainsi le côté générique de l'approche initiée par Baladi et Vallée. Nous avons aussi, dans la deuxième partie de la thèse, étendu le domaine potentiel de l'analyse dynamique, en définissant un modèle de base de données, fondé sur des idées dynamiques.

Notre travail repose finalement sur une double activité de modélisation et d'analyse. Dans un domaine fortement mathématisé comme celui de l'arithmétique, la modélisation n'est pas l'activité centrale ; elle est finalement assez facile, et est, en tous cas, bien "balisée". On peut alors se focaliser sur des approches probabilistes fines, et obtenir des résultats très précis. Dans un domaine "nouveau", comme celui de la fouille de données, peu est formalisé (peu est formalisable, peut-être ?), et l'activité de modélisation est centrale et difficile. Les modèles réalistes sont sans doute trop complexes pour qu'on puisse espérer y prouver des comportements probabilistes très fins.

Du point de vue méthodologique et transverse de l'analyse dynamique, et de manière adaptée à chacun des domaines, il reste beaucoup à faire. C'est pour cela que nous souhaitons continuer nos recherches dans chacun de ces domaines, car nous gardons ainsi la possibilité de persévérer dans ces deux activités scientifiques, modéliser et prouver, qui sont, pour nous, complémentaires.

Bibliographie

[AAP00] R. C. Agarwal, C. C. Aggarwal, and V. V. V. Prasad. Depth first generation of long patterns. In *Proceedings of the sixth ACM SIGKDD, ACM Press*, pages 108–118, 2000.

[AAP01] R. C. Agarwal, C. C. Aggarwal, and V. V. V. Prasad. A tree projection algorithm for generation of frequent item sets. In *Journal of Parallel and Distributed Computing 61-3 Special issue on high-performance data mining*, pages 350–371, 2001.

[AIS93] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, USA*, pages 207–216, 1993.

[ALL01] M. Ahues, A. Largillier, and V. Limaye. *Spectral computations for bounded operators*. Chapman & Hall/CRC, 2001.

[AMS+96] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Verkamo. Fast discovery of association rules. *Advances in Knowledge Discovery and Data Mining*, pages 307–328, 1996.

[AS94] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Intl. Conference on Very Large Data Bases (VLDB'94), Santiago de Chile, Chile*, pages 487–499, 1994.

[AV00] A. Akhavi and B. Vallée. Average bit–complexity of Euclidean algorithms. In *Proceedings of ICALP'00, Lecture Notes in Computer Science (1853), Springer*, pages 373–387, 2000.

[Bab78] K. I. Babenko. On a problem of Gauss. *Soviet Mathematical Doklady 19 (1)*, pages 136–140, 1978.

[BAG99] R. J. Bayardo, R. Agrawal, and D. Gunopulos. Constraint-based rule mining in large, dense databases. In *Proc. of the 15th Int'l Conf. on Data Engineering, Sydney, Austrialia*, pages 188–197, 1999.

[BBR00] J.-F. Boulicaut, A. Bykowski, and C. Rigotti. Approximation of frequency queries by means of free-sets. In *Principles of Data Mining and Knowledge Discovery (PKDD'00), Lyon, France*, pages 75–85, 2000.

[BDV02] J. Bourdon, B Daireaux, and B. Vallée. Dynamical analysis of α-Euclidean algorithms. *Journal of Algorithms, 44*, pages 246–285, 2002.

[BK85] R. P. Brent and H.T. Kung. A systolic vlsi array for integer gcd computation. In *ARITH-7, Proceedings of the Seventh Symposium on Computer Arithmetic (ed. par K. Hwang), IEEE CS Press*, pages 118–125, 1985.

[BK02] C. Borgelt and R. Kruse. Induction of association rules : Apriori implementation. In *15th Conference on Computational Statistics, Berlin, Germany*, pages 395–400, 2002.

[BMUT97] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. In *SIGMOD Conference*, pages 255–264, 1997.

[BNV01] J. Bourdon, M. Nebel, and B. Vallée. On the stack-size of general tries. *Theoretical informatics ans Applications*, 35 :163–185, 2001.

[Bou01] J. Bourdon. Size and path-length of patricia tries : Dynamical sources context. *Random Structures and Algorithms*, pages 289–315, 2001.

[Bre76] R.P. Brent. Analysis of the binary Euclidean algorithm. *Algorithms and Complexity, New directions and recent results (ed. par J.F. Traub), Academic Press*, pages 321–355, 1976.

[Bri03] K. Briggs. A precise computation of the gauss-kuz'min-wirsing constant. Technical report, Preliminary report, 2003.

[BS04] Anne Berry and Alain Sigayret. Representing a concept lattice by a graph. *Discrete Applied Mathematics*, 144(1-2) :27–42, 2004.

[BTP+00] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal. Mining minimal non-redundant association rules using frequent closed itemsets. In *International Conference on Deductive and Object Databases (DOOD'00)*, pages 972–986, 2000.

[BTP+02] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal. Pascal : un algorithme d'extraction des motifs fréquents. *Technique et science informatiques*, 21(1) :65–95, 2002.

[Bum82] R. T. Bumby. Hausdorff dimension of Cantor sets. *Journal Reine Angew. Math.*, 331 :192–206, 1982.

[Bum85] R. T. Bumby. Hausdorff dimension of set arising in in number theory. *Number Theory (New-York, 1983-1984), Lecture Notes in Mathematics, 1135, Springer,*, pages 1–8, 1985.

[BV02] J. Bourdon and B. Vallée. Generalized pattern-matching statistics. *Colloquium on Mathematics and Computer Science : Algorithms, Trees, Combinatorics and Probability*, pages 249–265, 2002.

[BV04] V. Baladi and B. Vallée. Distributional analyses of Euclidean algorithms. In *Proceedings of Alenex–ANALCO'04*, pages 170–184, 2004.

[BV05] V. Baladi and B. Vallée. Euclidean algorithms are gaussian. *Journal of Number Theory, 110, no 2*, pages 331–386, 2005.

[BV06] J. Bourdon and B. Vallée. Pattern matching statistics on correlated sources. *Proceedings of LATIN'06, LNCS 3887*, pages 224–237, 2006.

[Ces06] E. Cesaratto. Remarks and extensions on the paper "Euclidean algorithm are gaussian", by baladi v. and vallée b., communications personnelles. 2006.

[CFV01] J. Clément, P. Flajolet, and B. Vallée. Dynamical sources in information theory : A general analysis of trie structures. *Algorithmica*, 29(1) :307–369, 2001.

[CHR83a] S. CHRISTODOULAKIS. Estimating block transfers and join sizes. In *ACM SIGMOD*, pages 40–54, May 1983.

[CHR83b] S. CHRISTODOULAKIS. Estimating record selectivities. *Information Systems*, 8(2) :105–115, 1983.

[CHR84] S. CHRISTODOULAKIS. Implications of certains assumptions in database performance evaluation. *ACM Transactions on Database Systems*, 9(2) :165–186, June 1984.

[CMD04] F. Chazal and V. Maume-Deschamps. Statistical properties of general markov dynamical sources : applications to information theory. *Discrete Mathematics and Theoretical Computer Science*, 6(2) :283–314, 2004.

[CMV03] F. Chazal, V. Maume, and B. Vallée. Systèmes dynamiques et algorithmique. In F. Chyzak Ed., editor, *Algorithms Seminar*, pages 121–150. INRIA Res. Report 5003, 2003.

[Dai04] B. Daireaux. *Analyse des algorithmes d'Euclide : une approche dynamique*. PhD thesis, Thèse Université Caen, 2004.

[DC04] N. Dexters and T. Calders. Theoretical bounds on the size of condensed representations. In *ECML-PKDD 2004 Workshop on Knowledge Discovery in Inductive Databases (KDID'04), Pisa, Italy*, pages 25–36, 2004.

[Del54] H. Delange. Généralisation du théorème d'ikehara. *Annales Scientifiques de l' ENS, 71*, pages 213–242, 1954.

[DFV97] H. Daudé, P. Flajolet, and B. Vallée. An average-case analysis of the Gaussian algorithm for lattice reduction. *Combinatorics, Probability and Computing, 6*, pages 397–433, 1997.

[Dix70] J. D. Dixon. The number of steps in the Euclidean algorithm. *Journal of Number Theory 2*, pages 414–422, 1970.

[DJLM02] L. De Raedt, M. Jaeger, S.D. Lee, and H. Mannila. A theory of inductive query answering (extended abstract). In *IEEE International Conference on Data Mining (ICDM'02), Maebashi City, Japan*, pages 123–130, 2002.

[DLMDV06] B. Daireaux, L. Lhote, V. Maume-Deschamps, and B. Vallée. Analysis of fast versions of the Euclid algorithm. *soumis à ANTS*, 2006.

[DMDV05] B. Daireaux, V. Maume-Deschamps, and B. Vallée. The Lyapounov tortoise and the dyadic hare. In Discrete Mathematics and Theoretical Computer Science, editors, *Proceedings of AofA'05*, pages 71–94, 2005.

[Dol98] D. Dolgopyat. On decay of correlations in anosov flows. *Annals of Mathematics, 147*, pages 357–390, 1998.

[DT95] J. Demetrovics and V. Thi. Some remarks on generating armstrong and inferring functionnal dependencies relation. *Acta Cybernetica*, 12(2) :167–180, 1995.

[DV04] B. Daireaux and B. Vallée. Dynamical analysis of the parametrized Lehmer-Euclid algorithm. *Combinatorics, Probability, Computing*, pages 499–536, 2004.

[Dyr97] C. E. Dyreson. *Uncertainty Management in Information Systems*, chapter A Bibliography on Uncertainty Management in Information Systems. Kluwer Academic Publishers, 1997.

[EE85] W. Ellison and F. Ellison. *Prime Numbers*. Hermann, Paris, 1985.

[FGP01] Philippe Flajolet, Xavier Gourdon, and Daniel Panario. The complete analysis of a polynomial factorization algorithm over finite fields. *Journal of Algorithms*, 40(1) :37–81, 2001.

[FH 9] C. Friesen and D. Hensley. The statistics of continued fractions for polynomials over a finite field. In *Proceedings of the American Mathematical Society, Vol 124,*, pages 2661–2673, 1996 (9).

[Fin03] S. R. Finch. *Mathematical Constants*. Cambridge University Press, 2003.

[Fla06] P. Flajolet. Notes de cours. Technical report, INRIA Rocquencourt, Paris, 2006.

[FPSS96] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Knowledge discovery and data mining : Towards a unifying framework. In *International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, USA*, pages 82–88, 1996.

[FS96] P. Flajolet and R. Sedgewick. *An introduction to the analysis of algorithms*. Addisson Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1996.

[FS99] P. Flajolet and R. Sedgewick. *Analytic Combinatorics (1999)*. Book in preparation, see also INRIA Research Reports 1888, 2026, 2376, 2956, 1999.

[FV00] P. Flajolet and B. Vallée. Continued fractions, comparison algorithms, and fine structure constants. In *Constructive, Experimental et Non-Linear Analysis, Michel Thera, Proceedings of Canadian Mathematical Society, Vol 27*, pages 53–82, 2000.

[Gau07] C. F. Gauss. Recherches arithmétiques. *Blanchard, Paris, 1953*, 1807.

[GB02] J. Grzymala-Busse. *Handbook of Data Mining and Knowledge Discovery*, chapter Discretization of numerical attributes. Oxford University Press, 2002.

[GGR02] Minos N. Garofalakis, Johannes Gehrke, and Rajeev Rastogi. Querying and mining data streams : you only get one look a tutorial. In *SIGMOD Conference*, page 635, 2002.

[GGV01] F. Geerts, B. Goethals, and J. Van den Bussche. A tight upper bound on the number of candidate patterns. In *IEEE International Conference on Data Mining (ICDM'01), San Jose, USA*, pages 155–162, 2001.

[GMKT97] D. Gunopulos, H. Mannila, R. Khardon, and H. Toivonen. Data mining, hypergraph transversals, and machine learning. In *ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS'97), Tucson, USA*, 1997.

[GMS97] D. Gunopulos, H. Mannila, and S. Saluja. Discovering all most specific sentences by randomized algorithms. In *ICDT*, pages 215–229, 1997.

[Goe03] B. Goethals. Survey on frequent pattern mining, 2003.

[Goo41] I. J. Good. The fractional dimension of continued fractions. *Proc. Camb. Phil. Soc.*, 37 :199–228, 1941.

[Gro55] A. Grothendieck. Produits tensoriels topologiques et espaces nucléaires. *Memoirs of the American Mathematical Society, 16*, 1955.

[Hei69] H. Heilbronn. On the average length of a class of continued fractions. *Number Theory and Analysis, ed. by P. Turan, New-York, Plenum*, pages 87–96, 1969.

[Hen92] D. Hensley. Continued fraction Cantor sets, Hausdorff dimension and functional analysis. *Journal of Number Theory, Vol 40*, pages 336–358, 1992.

[Hen94] D. Hensley. The number of steps in the Euclidean algorithm. *Journal of Number Theory, Vol 49*, pages 142–182, 1994.

[Hen96] D. Hensley. A polynomial time algorithm for the Hausdorff dimension of a continued fraction Cantor set. *Journal of Number Theory*, 58(1) :9–45, 1996.

[Hen04] D. Hensley. *Continued Fractions*. World Scientific, to appear, 2004.

[HGN00] J. Hipp, U. Güntzer, and G. Nakhaeizadeh. Mining association rules : deriving a superior algorithm by analysing today's approaches. In *Principles of Data Mining and Knowledge Discovery (PKDD '00), Lyon, France*, 2000.

[HPY00] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation : A frequent-pattern tree approach. In *ACM SIGMOD International Conference on Management of Data (SIGMOD'00), Dallas, USA*, pages 1–12, 2000.

[Hwa94] H.-K. Hwang. *Théorèmes limite pour les structures combinatoires et les fonctions arithmétiques*. PhD thesis, Ecole Polytechnique, France, 1994.

[Hwa98] H.-K. Hwang. On convergence rates in the central limit theorems for combinatorial structures. *European Journal of Combinatorics, Vol 19*, pages 329–343, 1998.

[Jeb95] T. Jebelean. A double digit Lehmer Euclid algorithm for finding the gcd of long integers. *Journal of Symbolic Computation*, 19(1-3) :145–157, 1995.

[JGU04] O. Jenkinson, L.F. Gonzalez, and M. Urbański. On transfer operators for continued fractions with restricted digits. *Proceedings of the London Mathematical Society*, 2004.

[JP99] O. Jenkinson and M. Pollicott. Computing the dimension of dynamically defined sets i : e_2 and bounded continued fractions, preprint. Technical report, Institut Mathématiques de Luminy, 1999.

[JR05] B. Jeudy and F. Rioult. *Post-proceedings of the International Workshop on Knowledge Discovery in Inductive Databases (KDID'04) co-located with the ECML-PKDD'04*, chapter Database Transposition for Constrained (Closed) Pattern Mining, pages 89–107. Springer, 2005.

[KK88] J. Knopfmacher and A. Knopfmacher. The exact length of the Euclidean algorithm in $F_q[X]$. *Mathematika, Vol 35*, pages 297–304, 1988.

[Knu71] D.E. Knuth. The analysis of algorithms. In Gauthier-Villars, editor, *Actes du congrès des mathématiciens*, volume 3, pages 269–274, 1971.

[Knu97] D.E. Knuth. *The art of Computer programming, Volume 1,*. 3rd edition, Addison Wesley Longman Publishing Co., 1997.

[Knu98a] D.E. Knuth. *The art of Computer programming, Volume 2,*. 3rd edition, Addison Wesley, Reading, Massachussets, 1998.

[Knu98b] D.E. Knuth. *The art of Computer programming, Volume 3,*. 2rd edition, Addison Wesley, Reading, Massachussets, 1998.

[Kuz28] R. O. Kuz'min. Sur un problème de gauss. *Atti del Congresso Internationale dei Matematici, Vol 6, Bologna*, pages 83–89, 1928.

[Lé29] P. Lévy. Sur les lois de probabilité dont dépendent les quotients complets et incomplets d'une fraction continue. *Bull. Soc. Math. France*, 57 :178–194, 1929.

[Lam45] D. Lamé. Note sur la limite du nombre de divisions dans la recherche du plus grand commun diviseur entre deux nombres entiers. *C. R. Acad. Sc.*, 19 :867–870, 1845.

[Leh38] D. H. Lehmer. Euclid's algorithm for large numbers. *Am. Math. Mon.*, 45 :227–233, 1938.

[Lho04] L. Lhote. Computation of a class of continued fraction constants. *Proceedings of Alenex–ANALCO04*, pages pp 199–210, 2004.

[LLSW05] Jinyan Li, Haiquan Li, Donny Soh, and Limsoon Wong. A correspondence between maximal complete bipartite subgraphs and closed patterns. In *PKDD*, pages 146–156, 2005.

[LM94] A. Lasota and M Mackey. *Chaos, Fractals and Noise ; Stochastic Aspects of Dynamics*. Applied Mathematical Science 97, Springer, 1994.

[LM04] G. Le Mahec. Utilisation des arbres radicaux pour les algorithmes de data-mining sur grille de calcul (mémoire de dea). Master's thesis, Université de Picardie, 2004.

[LRS05a] L. Lhote, F. Rioult, and A. Soulet. Average number of frequent and closed patterns in random databases. In *Conférence d'Apprentissage (CAp'05), Nice, France*, pages 345–360, 2005.

[LRS05b] L. Lhote, F. Rioult, and A. Soulet. Average number of frequent (closed) patterns in bernouilli and markovian databases. In *IEEE International Conference on Data Mining (ICDM'05), Houston, USA*, pages 713–716, 2005.

[LV06a] L. Lhote and B. Vallée. Gaussian laws for the main parameters of the Euclid algorithms. *article invité pour Algorithmica*, 2006.

[LV06b] L. Lhote and B. Vallée. Sharp estimates for the main parameters of the Euclid algorithm. *Proceedings of LATIN'06, LNCS 3887*, pages 689–702, 2006.

[May79] D. H. Mayer. Spectral properties of certain composition operators arising in statistical mechanics. *Commun. Math. Phys.*, page pp 68, 1979.

[May91] D. H. Mayer. *Continued fractions and related transformations*. T. Bedford, M. Keane, and C. Series, 1991.

[MR86] H. Mannila and K.-J. Räihä. Inclusion dependencies : Application to logical database tuning. In *International Conference on Data Engineering, Los Angeles, USA*, 1986.

[MT97] H. Mannila and H. Toivonen. Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, 1(3) :241–258, 1997.

[Nak81] H. Nakada. Metrical theory for a class of continued fraction transformations and their natural extensions. *Tokyo J. Math.*, 2(4) :399–426, 1981.

[Nie87] H. Niederreiter. Continued fraction for formal power series, pseudorandom numbers, and linear complexity. *Contribution to general algebra,* **5** *(Salzburg, 1986)*, pages 221–233, Hölder-Pichler-Tempsky, Vienna, 1987.

[Nie88] H. Niederreiter. Sequences with almost perfect linear complexity profile. *Advances in Cryptology : Proc. EUROCRYPT'87*, pages 37–51, 1988.

[PBTL99a] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices. *Information Systems*, 24(1) :25–46, 1999.

[PBTL99b] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Closed set based discovery of small covers for association rules. In *BDA'99*, pages 361–381, 1999.

[PGG04] Paul W. Purdom, Dirk Van Gucht, and Dennis P. Groth. Average-case performance of the apriori algorithm. *SIAM Journal on Computing*, 33(5) :1223–1260, 2004.

[Phi70] W. Philipp. Some metrical results in number theory ii. *Duke Mathematics*, 38 :447–488, Errata p.788, 1970.

[PHM00] J. Pei, J. Han, and R. Mao. Closet : An efficient algorithm for mining frequent closed itemsets. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'00), Dallas, USA*, pages 21–30, 2000.

[Rie78] G. J. Rieger. über die mittlere schrittanzahl bei divisionalgorithmen. *Math. Nachr.*, pages 157–180, 1978.

[Rue78] D. Ruelle. Thermodynamic formalism. *Addison Wesley*, 1978.

[Rue85] Rainer A. Rueppel. Linear complexity and random sequences. In *EUROCRYPT*, pages 167–188, 1985.

[Rue86] R. A. Rueppel. *Analysis and Design of Stream Ciphers.* Springer-Verlag, Berlin, 1986.

[SA96] R. Srikant and R. Agrawal. Mining quantitative association rules in large relational tables. In *Proceedings of the 1996 ACM SIGMOD international conference on Management of data, Montreal, Canada*, pages 1–12, 1996.

[SC05] A. Soulet and B. Crémilleux. An efficient framework for mining flexible constraints. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining, (PAKDD'05), Hanoi, Vietnam*, 2005.

[Sch71] A. Schonhage. Schnelle berechnung von kettenbruchentwicklungen. *Acta Informatica*, pages 139–144, 1971.

[Sha92] J. Shallit. Real numbers with bounded partial quotients. a survey. *L'Enseignement Mathématique*, 38 :151–187, 1992.

[Sha93] J. Shapiro. *Composition operators and classical function theory.* Tracts in Mathematics, Springer-Verlag, 1993.

[Sha01] Claude E. Shannon. A mathematical theory of communication. *Mobile Computing and Communications Review*, 5(1) :3–55, 2001.

[Sig02] A. Sigayret. *Data Mining : une approche par les graphes.* PhD thesis, Université Blaise Pascal, Clermont II, 2002.

[SON95] A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. In *VLDB'95*, 1995.

[Ste67] J. Stein. Computational problems associated with racah algebra. *Journal of Computational Physics*, 1 :397–405, 1967.

[SZ04] D. Stehlé and P. Zimmermann. A binary recursive gcd algorithm. *Proceedings of ANTS'04, LNCS 3076*, pages 411–425, 2004.

[Toi96] Hannu Toivonen. Sampling large databases for association rules. In *International Conference on Very Large Data Bases (VLDB'96), Mumbai, India*, pages 134–145. Morgan Kaufman, 1996.

[Val97a] B. Vallée. Algorithms for computing signs of 2×2 determinants : dynamics and average-case algorithms. *Proceedings of the 8 th Annual European Symposium on Algorithms, ESA'97, LNCS 1284, Springer Verlag*, pages 486–499, 1997.

[Val97b] B. Vallée. Opérateurs de Ruelle-Mayer généralisés et analyse des algorithmes d'Euclide et de Gauss. *Acta Arithmetica*, 81.2 :101–144, 1997.

[Val98a] B. Vallée. Dynamics of the Binary Euclidean Algorithm : Functional analysis and operators. *Algorithmica*, 22(4) :660–685, 1998.

[Val98b] B. Vallée. Fractions continues à contraintes périodiques. *Journal of Number Theory*, 72 :183–235, 1998.

[Val00a] B. Vallée. Digits and continuants in Euclidean algorithms. ergodic versus tauberian theorems. *Journal de Théorie des Nombres de Bordeaux, JTNB*, 12 :531–570, 2000.

[Val00b] B. Vallée. A unifying framework for the analysis of a class of Euclidean algorithms. *Proceedings of LATIN'00, Lecture Notes in Computer Science*, 1776 :343–354, 2000.

[Val01] B. Vallée. Dynamical sources in information theory : Fundamental intervals and word prefixes. *Algorithmica*, 29 :262–306, 2001.

[Val03] B. Vallée. Dynamical analysis of a class of Euclidean algorithms. *Theoretical Computer Science*, 297 :447–486, 2003.

[Val06] B. Vallée. Euclidean dynamics. *Discrete and Continuous Dynamical Systems*, 15(1), 2006.

[WHP03] J. Wang, J. Han, and J. Pei. Closet+ : Searching for the best strategies for mining frequent closed itemsets. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, USA*, 2003.

[Wir74] E. Wirsing. On the theorem of Gauss–Kusmin–Lévy and a Frobenius–type theorem for function spaces. *Acta Arithmetica*, 24 :507–528, 1974.

[Yap96] C.K. Yap. *Fundamental Problems in Algorithmic Algebra*. Princeton University Press, 1996.

[YK75] A.C. Yao and D.E. Knuth. Analysis of the subtractive algorithm for greatest common divisors. *Proc. Nat. Acad. Sc. USA*, 72 :4720–4722, 1975.

[Zak00] Mohammed Javeed Zaki. Scalable algorithms for association mining. *IEEE Trans. Knowl. Data Eng.*, 12(2) :372–390, 2000.

[ZG01] M. Zaki and K. Gouda. Fast vertical mining using diffsets. Technical Report 01-1, 2001 11, RPI, 2001.

[ZH02] M. Zaki and C.-J. Hsiao. Charm : An efficient algorithm for closed itemset mining. In *Second SIAM International Conference on Data Mining (SDM'02), Arlington, USA*, 2002.

[ZPOL97] M. Zaki, S. Parthasarathy, M. Ogihara, and W. Li. New algorithms for fast discovery of association rules. In David Heckerman, Heikki Mannila, Daryl Pregibon, Ramasamy Uthurusamy, and Menlo Park, editors, *In 3rd Intl. Conf. on Knowledge Discovery and Data Mining*, pages 283–296. AAAI Press, 12-15 1997.

[ZRRF99] D.A. Zighed, S. Rabaseda, R. Rakotomalala, and F. Fescheta. Discretization methods in supervised learning. *Encyclopedia of Computer Science and Technology*, 40 :35–50, 1999.