ACI DYNAMICAL : BILAN DÉTAILLÉ DU PROJET 2004-2007

Table des matières

1. Rappel des objectifs initiaux du projet

Le projet DynamicAl vise à élaborer des concepts, des méthodes et des techniques pour un ensemble de problématiques à l'interface des mathématiques et de l'informatique : analyse dynamique des algorithmes et des structures de données avec des applications particulières en compression, et en segmentation des données.
Les buts poursuivis sont principalement informatiques et les méthodes utilisées de nature mathématique. Le projet repose sur une approche originale qui fait intervenir des méthodes et des outils mathématiques qui ne sont pas utilisés dans les interfaces "classiques" mathématiques-informatique, notamment les systèmes dynamiques. Le projet est structuré autour de deux équipes. L'une, localisée à Dijon, regroupe des mathématiciens de Dijon, l'École polytechnique, Jussieu, Santiago du Chili. L'autre est localisée à Caen et regroupe des informaticiens de Caen et Nantes.

Les axes de recherches relèvent à la fois de l'algorithmique (analyse en moyenne et en distribution des algorithmes et des structures de données) et des mathématiques (systèmes dynamiques, méthodes probabilistes et statistiques).

Le projet repose sur l'analyse dynamique des algorithmes et des structures de données. Le but est d'obtenir l'analyse en moyenne puis en distribution de paramètres essentiels pour quantifier leur efficacité : le coût en bits d'un algorithme, la taille des structures de données, la hauteur pour les arbres

digitaux ... Les systèmes dynamiques discrets apparaissent comme un outil central dans ces études, une compréhension plus fine de leurs propriétés statistiques et spectrales s'avère nécessaire.
Les objectifs du projet étaient structurés autour de trois axes :

- **étude des algorithmes arithmétiques.** Développement de méthodes systématiques pour aborder la complexité tant arithmétique que booléenne ("en bits") des principaux algorithmes euclidiens, notamment par le calcul des constantes de structure, des variances, et de lois limites. Détermination de l'impact des méthodes dynamiques et probabilistes sur l'algorithmique de la réduction des réseaux (LLL), en partant des petites dimensions.
- **algorithmique du texte et problèmes de mots.** Analyse des occurrences de motifs dans le cadre général des sources dynamiques, études des structures d'arbres dans le cadre des algorithmes de recherche, de tri, de compression.
- **propriétés spectrales et statistiques des systèmes dynamiques.** Étude du spectre des opérateurs de transfert associés à des systèmes dynamiques en dimension supérieure, propriétés statistiques fines et estimation dans les systèmes dynamiques.

2. Rapport final

Les activités de l'ACI depuis septembre 2004 ont donné lieu à 31 publications dans des revues ou conférence internationales à comité de lecture, 5 membres de l'ACI ont soutenu une thèse, 2 membres de l'ACI ont soutenu une habilitation à diriger des recherches et une thèse a débuté en 2005.

2.1. **Analyse dynamique d'algorithmes arithmétiques.**

Il s'agit d'étudier la famille des algorithmes d'Euclide, de divers points de vue : les algorithmes eux-mêmes admettent énormément de variantes, avec des propriétés algorithmiques souvent intéressantes. Les paramètres qu'on veut étudier sont aussi très divers, et les analyses –en moyenne, variance, distribution– peuvent être aussi variées. Le cadre de l'analyse dynamique est un cadre général, mais les algorithmes, qui paraissent avoir à première vue à peu près tous la même structure, donnent pourtant lieu à des systèmes dynamiques qui possèdent des propriétés extrêmement diverses.

2.1.1. *Arithmétique en petite dimension.* Les travaux de l'ACI DynamicAl ont permis des avancées significatives dans la compréhension et l'analyse des algorithmes arithmétiques par la méthode "d'analyse dynamique". Les résultats obtenus sont les suivants.

- Mise en place de l'analyse en distribution ([4, 5, 12, 25, 23, 6, 36]) : les principaux paramètres (y compris le coût en bit) des algorithmes d'Euclide sont asymptotiquement gaussiens, les coûts additifs vérifient, de plus, des théorèmes de la limite locale.

- Analyse en moyenne de diverses variantes de l'algorithme d'Euclide ([19, 18]) : algorithme de Lehmer, algorithme LSB basé sur la division sur les bits de poids faibles.
- Calcul de constantes ([24, 41]) : obtention d'approximations prouvées, et obtenues en temps polynômial, des constantes (entropie, constante de Hensley) qui apparaissent dans les calculs d'espérance et de variance des paramètres de l'algorithme).
- Algorithmes contraints ([32, 12]) : les réels dont le développement en fraction continue obéit à certaines contraintes interviennent dans beaucoup de problèmes d'algorithmique arithmétique, par exemple dans l'analyse d'Euclide soustractif. Pour un ensemble de contraintes large, la dimension de Haussdorff de l'ensemble de ces réel est caractérisée par la solution d'un système différentiel qui fait intervenir la valeur propre dominante d'opérateurs de transfert.
- Algorithmes d'Euclide rapides ([11]) : Knuth et Schönhage ont proposé un algorithme de calcul du pgcd qui utilise les mêmes idées que celles de Lehmer, et est fondé sur une idée «diviser pour régner». Des phénomènes de retour arrière rendent cet algorithme très complexe, à la fois à programmer et à analyser. Le comportement moyen de cet algorithme est en $O(n \log^2 n \log\log n)$, avec des informations précises sur la constante du O.
- Version métrique du théorème des trois distances ([13]) : V.I. Arnold a défini une mesure de «l'aléa» d'un générateur pseudo-aléatoire, et propose d'étudier cette mesure $s(u, v)$ dans le cas d'une progression arithmétique modulaire $x_{i+1} = i \cdot u \pmod{v}$. La valeur moyenne de ce paramètre s est asymptotique à $(2/3) + 1/(4 \log 2) \sim 1.027$, avec un terme de reste explicite.

2.1.2. *En dimension supérieure.* Un réseau euclidien est l'ensemble des combinaisons linéaires entières d'un système libre de vecteurs de $\mathbf{R}^n$; ce système de vecteurs est une base du réseau. Un réseau possède une infinité de bases. Le problème de la réduction consiste à trouver, parmi toutes les bases d'un réseau, une base ayant de bonnes propriétés euclidiennes, avec des vecteurs "assez courts" et "assez orthogonaux". Il s'agit là d'un problème difficile (au sens de la théorie de la complexité), qui a trouvé, en 1982, avec l'algorithme LLL inventé par Lenstra, Lenstra et Lovàsz, une première solution satisfaisante. Cet algorithme n'est toujours pas bien compris : on ne connaît pas sa complexité dans le pire des cas, ni sa complexité "en moyenne". Étant données les applications essentielles de la réduction des réseaux dans beaucoup de domaines importants, et particulièrement en cryptographie, une compréhension fine du fonctionnement des algorithmes de la famille de LLL, si différent en théorie et en pratique, constitue une activité "en amont" importante. L'étude de la complexité de cet algorithme par des méthodes dynamiques fait intervenir des systèmes dynamiques en dimension supérieure, pour lesquels les propriétés spectrales ne sont pas suffisamment connues. Les

résultats obtenus concernent le comportement fin de l'algorithme de Gauss (LLL en dimension 2), [2] et les comportements limites des réseaux aléatoires pour le modèle uniforme ([1]). Des premiers résultats pour d'autres modèles aléatoires (plus réalistes en particulier dans le cadre de la cryptographie) ont étés obtenus dans le cadre d'une thèse débutée en septembre 2005 ([42][30, 31]). Cette même thèse a mis également en avant des ressemblances intéressantes entre l'algorithme de Gauss et l'algorithme de Gosper de comparaison par fractions continues.

2.2. **Algorithmique du texte, problèmes de mots, théorie de l'information.** Il est très important d'étudier et analyser précisément des algorithmes et des structures de données qui traitent des masses de données de grande taille (données issues du génome, pages web, fichiers de *log*, dictionnaires). Souvent l'étude d'algorithmes traitant de telles données n'est effectuée que dans le pire des cas ou sur un modèle peu réaliste de génération des données. Il ne reflète donc pas forcément le comportement observé en pratique. L'analyse en moyenne cherche à remédier à ces faiblesses, en considérant un modèle probabiliste suffisamment réaliste. Les sources dynamiques sont une possibilité de modèle qui généralise les sources sans mémoire et les chaînes de Markov.

- **Motifs cachés et les motifs généralisés.** Les "motifs cachés" apparaissent dans divers contextes (bio-informatique, sécurité). Dans [21], pour des textes produits par des sources sans mémoire, on obtient des estimations précises de la moyenne et de la variance du nombre d'occurrences de ces motifs. Ce nombre suit aussi, asymptotiquement, une loi gaussienne. Ces résultats ont ensuite été généralisés ([9]) à des sources dynamiques, modèle plus réaliste que les sources sans mémoire.
- **Structures d'arbres.** Les structures d'arbres (arbre de recherche, arbre de suffixes, arbre binaire ...) apparaissent dans de nombreux algorithmes (tri, compression, recherche). Les résultats obtenus concernent l'analyse en moyenne des principaux paramètres (hauteur, taille ...) des arbres (à clés répétée, de suffixes ...) sous des hypothèses probabilistes sans mémoire et markoviennes ([8, 3, 20]). Les arbres de contextes apparaissent d'une part dans le cadre d'algorithmes de compression proposés par Riessanen, d'autre part, comme structure support d'information pour les chaînes de Markov à portée variable (VLMC). Ces structures sont étudiées d'un point de vue probabiliste dans [22, 39], des inégalités exponentielles permettent d'estimer statistiquement la structure d'arbre sous-jacente aux VLMC.

2.3. **Propriétés statistiques et spectrales des systèmes dynamiques.** Cette partie - plus théorique - concerne la mise en place d'outils visant à une meilleure compréhension statistique et spectrale des systèmes dynamiques. Les résultats obtenus concernent :

– la généralisation des propriétés spectrales des opérateurs de transfert associés à des systèmes dynamiques en dimension supérieure et vérifiant des hypothèses d'hyperbolicité ([7, 10]). De tels outils trouveront probablement des applications dans l'étude des algorithmes qui font apparaître des systèmes dynamiques en dimension supérieure (par exemple LLL).
– l'exploitation d'inégalités de type exponentielle ou d'inégalités de variances pour l'estimation dans les systèmes dynamiques ([28, 29, 17, 16]).

2.4. **Perspectives.** Les travaux de recherche menés dans le cadre de l'ACI DynamicAl se poursuivent d'ores et déjà dans les directions suivantes.

– Compréhension des simulations des systèmes dynamiques : études des systèmes dynamiques discrétisés, exploitation des méthodes d'analyse d'algorithmes pour décrire les orbites des points rationnels des systèmes dynamiques.
– Algorithme LLL : le groupe vise à obtenir une analyse en moyenne précise de l'algorithme LLL, dans un modèle réaliste, en mêlant approches probabilistes et dynamiques. Il voudrait aussi mieux comprendre le système dynamique sous-jacent à l'algorithme LLL. Ce projet ambitieux, qui réunit des chercheurs de Versailles, Montpellier, Amiens, Paris, Lyon, et Caen, avec des spécialistes en probabilités, systèmes dynamiques et cryptographie a été sélectionné dans le cadre de l'appel à projet "non thématique" de l'ANR en 2007 : c'est le projet LAREDA.

3. Bilan financier

La subvention accordée dans le cadre de l'ACI a été dépensée essentiellement en frais de mission et a ainsi permis la mise en place de groupes de travail sur des thématiques spécifiques, l'organisation de rencontres et la participation des membres de l'ACI à des congrès nationaux et internationaux. Citons par exemple :

– la mise en place d'un groupe de travail sur les algorithmes de calcul de pgcd basés sur la division sur les bits de poids faibles qui s'est réuni à Caen et à Dijon,
– la mise en place d'un groupe de travail sur l'analyse des algorithmes d'Euclide rapides qui s'est réuni à Paris et à Caen,
– la mise en place d'un groupe de travail sur l'algorithme LLL : aspects géométriques, dynamiques et probabilistes qui s'est réuni à Paris, et a permis la mise sur pied du futur projet LAREDA
– la mise en place d'un groupe de travail sur les problèmes d'auto-corrélation en algorithmique du texte qui s'est réuni à Caen,
– l'organisation de rencontres de l'ACI à Dijon,
– la participation des membres de l'ACI à des congrès nationaux et internationaux notamment ALEA05, ALEA06, ALEA07, AofA05, AofA06, ANALCO, Math-Info, Latin06, MAS06, ...

Le détail des dépenses pour les groupes de Dijon et Caen sont fournis en annexe.

Références

Articles parus ou acceptés.

[1] A. Akhavi, J.-F. Marckert, A. Rouault, *On the reduction of a random basis"*, à paraître dans les Proceedings de SIAM-ALENEX/ANALCO'07.

[2] A. Akhavi, C.Moreira Dos Santos, *Another view of the Gaussian algorithm*, in Proccedings of LAtin American Informatics LATIN'04 (Buenos Aires, 2004) Lecture Notes in Computer Science volume 2976, pp 474–487, Springer.

[3] M. Archibald, J. Clément, *Average depth in a binary search tree with repeated keys*, Fourth Colloquium on Mathematics and Computer Science Algorithms, Trees, Combinatorics and Probabilities, (2006).

[4] V. Baladi, B. Vallée, *Euclidean algorithms are Gaussian.* J. Number Theory 110 (2005), no. 2, 331–386.

[5] V. Baladi, B. Vallée, *Exponential decay of correlations for surface semi-flows without finite Markov partitions.* Proc. Amer. Math. Soc. 133 (2005), no. 3, 865–874.

[6] V. Baladi et A. Hachemi, *A local limit theorem with speed of convergence for euclidean algorithms and diophantine costs* accepté pour publication aux Ann. IHP.

[7] V. Baladi et M. Tsujii, *Spectra of differentiable hyperbolic maps*, à paraitre proceedings "Traces in number theory, geometry and quantum fields", Bonn 2005, Vieweg,

[8] F. Bassino, J. Clément, G. Seroussi, A. Viola, *Optimal prefix codes for pairs of geometrically-distributed random variables*, 2006 IEEE International Symposium on Information Theory.

[9] J. Bourdon, B. Vallée, *Pattern Matching Statistics on Correlated sources*, Proceedings of LATIN'06, LNCS 3887, (2006), 224–237.

[10] J. Buzzi, V. Maume-Deschamps, *Decay of correlations on towers for potentials with summable variation*, Discrete and Continuous Dynamical Systems, 12, (2005), no 4, 639-656.

[11] E. Cesaratto, J. Clément, B. Daireaux, L. Lhote, V. Maume-Deschamps, B. Vallée *Analysis of fast versions of the Euclid Algorithm*, à paraître dans les Proceedings de SIAM-ALENEX/ANALCO'07

[12] E. Cesaratto, B. Vallée *Hausdorff dimension of real numbers with bounded digit averages*, Acta Arith., 125 (2006), 115-162.

[13] E. Cesaratto, A. Plagne, B. Vallée, *On the non-randomness of modular arithmetic progressions : a solution to a problem by V. I. Arnold* Comptes-Rendus du Colloquium on Mathematics and Computer Science : Algorithms, Trees, Combinatorics and Probability, P. Chassaing et al., ed., Discrete Mathematics and Theoretical Computer Science (2006).

[14] F. Chazal, V. Maume-Deschamps, *General Markov Dynamical sources : applications to information theory*, Discrete Mathematics and Theoretical Computer Science, 6, (2004), no 2, 283-314.

[15] F. Chazal, V. Maume-Deschamps, B. Vallée, *Erratum to "Dynamical sources in Information Theory : Fundamentals Intervals and Word Prefixes" by B. Vallée.* Algorithmica 38 (2004), no. 4, 591-596.

[16] J-R. Chazottes, P. Collet, B. Schmitt *Statistical consequences of Devroye inequality for processes. Applications to a class of non-uniformly hyperbolic dynamical systems* à paraître à Nonlinearity

[17] P. Collet, S. Martinez, B. Schmitt *Asymptotic Distribution of tests for expanding maps of the interval.* Ergodic Theory and Dynamical Systems 24, (2004), 1–16.

[18] B. Daireaux, V. Maume-Deschamps, B. Vallée *The Lyapunov Tortoise and the dyadic Hare.* 2005 international Conference on Analysis of Algorithms, Discrete Mathematics and Theoretical Computer Science, proc. AD, (2005), 71-94.

[19] B. Daireaux, B. Vallée, *Dynamical analysis of the parametrized Lehmer-Euclid algorithm.* Combin. Probab. Comput. 13 (2004), no. 4-5, 499–536.

[20] J. Fayolle, MD. Ward, *Analysis of the Average Depth in a Suffix Tree under a Markov Model*, Proceedings of the 2005 International Conference on the Analysis of Algorithms, DMTCS, (2005), 95–104.

[21] P. Flajolet, W. Szpankowski, B. Vallée, *Hidden word statistics* Journal de l'ACM, Vol 53, No1, (2006), 281–352.

[22] A. Galves, V. Maume-Deschamps, B. Schmitt, *Exponential inequalities for empirical probabilistic trees* accepté pour publication ÈSAIM prob. stat.

[23] A. Hachemi, *Un Théorème de la Limite locale pour des Algorithmes Euclidiens*, Acta Arith. 117 (2005), 265-276.

[24] L. Lhote, *Computation of a class of Continued Fraction Constants*, Proceedings of Alenex-ANALCO04 (2004), 199–210.

[25] L. Lhote, B. Vallée, *Gaussian laws for the main parameters of the Euclid Algorithms*, à paraître à Algorithmica (2007) [60 p]

[26] S. Martinez, *Some Bounds on the Coupon Collector Problem.* Random Structures and Algorithms 25, (2004), 208–226.

[27] S. Martinez, J. San Martin, *Classification of killed one-dimensional diffusions.* Annals of Probability 32, (2004), 530–552.

[28] V. Maume-Deschamps, *Exponential inequalities and estimation of conditional probabilities.* dans *Dependence in probability and statistics*, Lect. notes in Stat., Springer, Vol. 187 Bertail, Patrice ; Doukhan, Paul ; Soulier, Philippe (Eds.), (2006).

[29] V. Maume-Deschamps, *Exponential inequalities and functional estimations for weak dependent data ; applications to dynamical systems,* Stochastics and Dynamics, 6, (2006), no. 4 , 535-560.

[30] B. Vallée, A. Vera, *Lattice reduction in two dimensions : analyses under realistic probabilistic models*, Comptes-Rendus du Colloque AofA'07 (Analysis of Algorithms), à paraître dans DMTCS, 2007 (36 pages)

[31] B. Vallée, A. Vera, *Probabilistic behaviour of lattice reduction algorithms*, à paraître dans les Comptes-rendus du Colloque LLL+25, Springer (40 pages)

Thèses, HDR.

[32] E. Cesaratto, *Dimensión de Hausdorff y Esquemas de representación de números. (Dimensions de Hausdorff associées à des systèmes de numération)*, Thèse de l'université de Buenos Aires (Argentine), soutenue en mars 2005.

[33] F. Chazal, *Quelques contributions en approximation géométrique et topologique, en analyse dynamique d'algorithmes et à l'étude des feuilletages non spiralants*, Mémoire d'Habilitation à Diriger des Recherches, soutenue en juin 2005 à l'université de Bourgogne.

[34] B. Daireaux, *Analyse des algorithmes d'Euclide : une approche dynamique*, Thèse de l'université de Caen, soutenue en juin 2005.

[35] J. Fayolle *Compression de données sans perte et combinatoire analytique*, Thèse de l'université de Paris VI, soutenue en mars 2006.

[36] A. Hachemi, *Analyse dynamique d'algorithmes euclidiens et théorèmes limites*, Thèse de l'université de Paris VII, soutenue en juillet 2007.

[37] L. Lhote, *Algorithmes du pgcd et fouille de données : le point de vue de l'analyse dynamique*, Thèse de l'université de Caen, soutenue en septembre 2006.

[38] V. Maume-Deschamps, *Document de synthèse d'Habilitation à diriger des recherches, Propriétés statistiques des systèmes dynamiques ; applications en analyse d'algorithmes*, soutenue le 14 septembre 2005 à l'université de Bourgogne.

Articles soumis ou en préparation.

[39] B. Hafidi, V. Maume-Deschamps, *Model selection for probabilistic trees* en préparation.

[40] E. Cesaratto, Remarks on the paper "Euclidean Algorithms are gaussian" by V. Baladi and B. Vallée, soumis

[41] E. Cesaratto, L. Lhote, en préparation.

[42] A. Vera, *Thèse* en préparation.