Rapport d'activité

portant sur la période allant
de décembre 2018 à juin 2021

Olivier Catoni [1]

Table des matières

1. CREST, Laboratoire de Statistiques, UMR 9194 du CNRS, bureau 3035, 5, avenue Henry Le Chatelier, TSA 96642, 91764 Palaiseau cedex, FRANCE.

1. Curriculum Vitæ

1.1. Cursus.

— Date de naissance : 21 avril 1965.
— Nationalité : Française.
— No INSEE : 1 65 04 75 073 159
— No d'agent C.N.R.S. : 8001114

Directeur de recherche de première classe recruté le premier septembre 2000, affecté au CREST (Centre de Recherche en Economie et Statistique) — U.M.R. 9194, depuis le premier janvier 2015.

juillet 1982 Baccalauréat série C, mention T.B. avec les félicitations du jury. Académie de Paris.
septembre 1982 à juin 1984 Classes de mathématiques supérieures et de mathématiques spéciales au lycée Louis–le–Grand à Paris.
juillet 1984 Reçu 29 ème à l'E.N.S Ulm et 13 ème à l'Ecole Polytechnique.
année 1984 – 1985 : Première année de scolarité à l'E.N.S., licence et maîtrise de mathématiques appliquées à l'Université Paris VI.
année 1985 – 1986

Deuxième année de scolarité à l'E.N.S.

— D.E.A. de Probabilités et Applications. Université Paris VI. (obtention des modules d'A.E.A., stage et inscription administrative l'année suivante).
— Agrégation de mathématique, rang 13 ème.

septembre 1986 à décembre 1986

Stage de D.E.A. à l'université Paris XI – Orsay, sous la direction de Robert Azencott sur le thème « Restauration d'images par des méthodes de champs markoviens ».

Obtention du D.E.A de Probabilités et applications de Paris VI avec la mention T.B.

janvier 1987 à juillet 1987

Stage aux laboratoires de Marcoussis, centre de recherche de la C.G.E. en intelligence artificielle. Participation à un projet Esprit de reconnaissance de la parole (projet I.K.A.R.O.S.). Ce stage a donné lieu à l'écriture de trois rapports internes qui cherchaient à replacer le problème du contrôle du processus de reconnaissance dans le cadre des chaînes de Markov contrôlées (Dynkin).

septembre 1987 Début d'une thèse sous la direction de Robert Azencott, Professeur à l'université Paris XI, portant sur « l'Étude asymptotique des algorithmes de recuit simulé ».

année 1988 – 1989

Nomination à un poste d'A.N.D. à l'Université Paris XI – Orsay, pour service dans le Magistère de Mathématiques Fondamentales et Appliquées et d'Informatique.

le 1er septembre 1989 :

Entrée au C.N.R.S. en qualité de chargé de recherche de deuxième classe affecté au laboratoire de mathématiques de l'École Normale Supérieure.

le 27 mars 1990 :

Soutenance d'une thèse nouveau régime à l'Université Paris-Sud, spécialité mathématiques, intitulée :

"Etude asymptotique des algorithmes de recuit simulé".

année 1991 :

Prix IBM jeunes chercheurs en mathématiques.

le 1er octobre 1993 :

Promotion au grade de chargé de recherche de première classe.

le 15 décembre 1997 :

Diplôme d'habilitation à diriger des recherches de l'Université Paris-Sud, Orsay, spécialité mathématiques, exposé de synthèse intitulé "Grandes déviations des chaînes de Markov à transitions exponentielles, métastabilité et applications algorithmiques".

le 1er octobre 1998 :

Affectation au laboratoire de Probabilités et Modèles Aléatoires, U.M.R. 7599 du C.N.R.S. (Universités Paris 6 et 7).

le 1er septembre 2000 :

Promotion au grade de directeur de recherche.

le 1er septembre 2008 :

Affectation au Département de Mathématiques et Applications de l'École Normale Supérieure, pour y superviser la création d'une équipe INRIA consacrée à la théorie statistique de l'apprentissage.

le 1er juillet 2009

Nomination par le centre de recherche INRIA Paris-Rocquencourt en qualité de responsable permanent de l'équipe CLASSIC « Convex Learning through Aggregation, Supervised Statistical Inference, and Classification » nouvellement créée au sein du DMA de l'ENS.

le 3 juillet 2010

Conversion de l'équipe CLASSIC en équipe-projet par l'INRIA, pour une durée de cinq ans, sous ma responsabilité.

le 1er octobre 2013

Promotion au grade de directeur de recherche de première classe.

le 31 décembre 2014

Arrivée à échéance de l'équipe-projet CLASSIC, qui ne sera pas renouvelée, en accord avec l'INRIA et le DMA. En effet, l'équipe CLASSIC avait pour rôle de fédérer les activités de recherche en statistique au sein du DMA. La faire déménager ailleurs à l'occasion de mon départ n'aurait pas eu de sens, et la faire survivre à un renouvellement de sa direction et de ses membres aurait été trop contraire aux traditions de l'INRIA. L'expérience fut néanmoins très positive, permettant un affichage fort des activités du DMA dans le domaine du machine learning auprès des élèves de l'ENS. J'espère qu'elle pourra être renouvelée et que d'autres collaborations entre le DMA et l'INRIA verront le jour.

le 1er janvier 2015

Départ du DMA, pour respecter la règle de renouvellement de ses membres, et affectation au Centre de Recherche en Économie et Statistique, UMR 9194 du CNRS.

1.2. Collaborations françaises et étrangères. Présentation d'un article intitulé "Détection de contours par seuillage adaptatif et restauration stochastique d'images binaires" au congrès "Pixim 1989" (collaboration avec Isabelle Gaudron-Trouvé), en septembre 1989 [CG89].

Séjour d'une semaine (fin mai 1990) à l'Istituto per le applicazioni del calcolo "Mauro Picone" dans le cadre de l'année intensive "Stochatic Models, Statistical Methods and Algorithms in Image Analysis" (Local Committee P. Barone, A. Frigessi), exposés sur les algorithmes de recuit simulé et sur la détection de contours. Participation aux proceedings [Cat90b].

Participation au séminaire "Stochastic Image Models and Algorithms" (R. Azencott - D. Geman, Oberwolfach, 15-21 juillet 1990) (exposés sur le recuit simulé et sur la restauration d'images bruitées.)

Service national (août 1990- août 1991) en tant que scientifique du contingent à l'E.T.C.A. (à ARCUEIL) dans le laboratoire ETCA/CREA/Systèmes de Perception. Participation au projet "Rétines programmables" développé conjointement par l'I.E.F. (U.R.A. 22 du C.N.R.S.) (Devos, Garda) et par l'E.T.C.A. (Zavido-

vique). Rédaction d'un article sur la reconnaissance des formes et la détection du mouvement par une rétine programmable, intitulé "Learning Algorithms for Pattern Recognition on Half-Tone Binary Images". Cet article propose un algorithme d'apprentissage où on maximise la distance de Kullback entre certaines marginales de deux images à différencier l'une de l'autre [Cat91b].

Exposé aux *Journées de Probabilités* (J. Azema et M. Yor, CIRM, Marseille Luminy, 22-26 octobre 1990) sur les algorithmes de recuit simulé.

Exposé au séminaire de l' *Institut für Statistik und Informatik, Universität Wien*, Autriche, sur invitation de G. Pflug (22-23 novembre 1990), sur le comportement asymptotique des algorithmes de recuit simulé.

Exposé et participation aux proceedings du *U.S.-French Workshop on Applied Stochastic Analysis (Rutgers University, 29 April - 2 May 1991)* organisé par Y. Karatzas et D. Ocone [Cat92a].

Séjour à l'Université de Bielefeld (Allemagne) sur invitation de F. Götze (septembre - octobre 1991). Conception et implantation sur transputers d'un algorithme de recuit parallèle avec suivi de la suite des températures conduisant à la solution finale calculée par l'algorithme. Etude théorique de la convergence de cet algorithme parallèle (travaux non publiés).

Participation au séminaire sur la méthode des répliques pour le calcul de l'énergie libre moyenne d'un verre de spin organisé par R. Azencott, M. Mézard et J.P. Nadal (année 1990 - 1991).

Participation au séminaire "From statistical physics to statistical inference and back", organisé par Peter Grassberger et Jean-Pierre Nadal à l'I.E.S. de Gargèse, (31 août, 12 septembre 1992).

Séjour à l'Université de Bielefeld (R.F.A.) du 10 au 22 mai 1993. Collaboration avec F. Götze.

Participation à l'organisation d'un groupe de travail "Mathématiques et réseaux de neurones formels" pendant deux années (R. Azencott, O. Catoni, A. Trouvé et L. Younes pour 1991-1992, R. Azencott, O.Catoni, I. Gaudron et A. Trouvé pour 1992-1993). Exposés sur la théorie de Vapnik Chervonenkis pour la reconnaissance des formes et l'estimation d'une régression.

Participation à l'European Science Foundation Network on Highly Structured Stochastic Systems, First Workshop, Cortona, 9-16 avril 1994, Italie, sur invitation d'A. Frigessi (Laboratoire de Statistique, Université de Venise), exposé intitulé "Energy Transforms for Metropolis Algorithms".

Participation à la l'Ecole d'Eté de Probabilités de Saint-Flour, 7-23 juillet 1994. Dans le cadre des exposés des participants, exposé sur la méthode des transformations itérées de l'énergie.

Participation à la "Twelfth Prague Conference on Information Theory, Statistical Decision Functions and Random Processes – August 29, September 2 1994". Exposé et publication d'une note dans les proceedings intitulée "Energy Trans-

forms for Metropolis and Simulated Annealing Algorithms" [Cat94] qui annonce les résultats de [Cat98a].

Ecole d'été de probabilités de Saint Flour (juillet 1995), participation en tant qu'auditeur. Exposé sur le modèle de verre de spin de Sherrington Kirkpatrick.

Workshop "Large Deviations and Statistical Mechanics" 20-21 octobre 1995 Bielefeld, Germany, organisé par Peter Eichelsbacher et Matthias Löwe. Participation en tant que conférencier invité. Communication dans les proceedings : "A New Inequality for the Free Energy of the Sherrington Kirkpatrick Spin Glass Model" [Cat96c] qui présente [Cat96a].

Troisième journée sur les "Algorithmes Stochastiques pour de grands systèmes", à l'Institut Henri Poincaré, Paris 5ième, le jeudi 16 novembre 1995, organisée par les groupes "Algorithmes et Automatique" des universités de Marne-la-Vallée et de Paris 11 (Orsay), "Probabilités Numériques" des universités de Créteil et de Marne-la-Vallée, "Réseaux de Neurones" du SAMOS de l'univ. Paris 1. Conférence invitée : "Comment utiliser l'algorithme de Metropolis et ses avatars (recuit simulé, transformations de l'énergie) pour résoudre des problèmes de planification."

Organisation avec L. Birgé (Paris VI) et P. Massart (Paris XI) à partir de 1994 à l'ENS Ulm d'un séminaire de Statistique et d'un groupe de travail sur l'estimation adaptative. 1994-1995 : Exposés sur les travaux d'Ornstein et Weiss sur les processus de Bernoulli et la théorie du codage. 1995-1996 : Deux exposés dans le groupe de travail sur les "Support Vector Machines" d'après Vapnik.

Collaboration avec Raphaël Cerf (laboratoire de Modélisation Stochastique et Statistique d'Orsay), pour l'étude du chemin de sortie des chaînes de Markov à transitions rares (printemps 1995) [CC97].

Collaboration avec C. Cot pour l'étude des suites de températures log-optimales constantes par paliers pour l'algorithme de recuit simulé (automne 1995) [CC98].

Participation à "Inhomogeneous Random Systems, Large Deviations and Hydrodynamic Limits" (Systèmes aléatoires inhomogènes, grandes déviations et limites hydrodynamiques), 24 janvier 1996, Ecole Polytechnique et CNRS, organisé par François Dunlop, Thierry Gobron et Ellen Saada, conférence invitée : "The Legendre Transform and the Replica Method : a New Inequality for the Sherrington Kirkpatrick Model".

Séminaire "Probabilités et Imagerie", Laboratoire Prisme, Université René Descartes, organisé par Christine Graffigne, exposé en deux parties (29-2 et 7-3 1996) "Chaînes de Markov à transitions rares et algorithmes d'optimisation".

Mini-workshop "Probabilistic Algorithms and Algorithmic Probability – Interacting Particle Systems", University of Nijmegen, The Netherlands, March 15, 1996, conférence invitée : "Solving Scheduling Problems by Simulated Annealing".

Conférencier invité des Journées SMAI-MAS Modélisation aléatoire et statis-

tique (23-25 septembre 1996, organisées par D. Michel – Toulouse et P. Cattiaux – Paris). Exposé sur les estimées de grandes déviations pour le recuit simulé généralisé.

Conférence dans la session image (organisée par J.-M. Morel – Paris et D. Mumford – Stanford) du congrès "Foundation of Computational Mathematics", IMPA, Rio de Janeiro, Brésil, 5-12 janvier 1997, intitulée "Metropolis, Simulated Annealing and Iterated Energy Transformation Algorithms : Theory and Experiments" (publiée dans le numéro spécial du Journal of Complexity consacré au congrès [Cat96b]).

Conférence au séminaire "Mathematische Stochastik" Oberwolfach 9-15 mars 1997 (organisé par J. Gärtner – Berlin, R.D. Gill – Utrecht et E. Mammen – Heidelberg), intitulée : "Stochastic optimization algorithms : speed-up methods".

Conférence invitée aux "Journées de Probabilités", Toulouse, 8-12 septembre 1997, organisées par D. Bacry, M. Ledoux, G. Letac, D. Michel, L. Saloff-Coste, comité scientifique, J. Azéma, M. Emery et M. Yor. Titre : "Mélanges adaptatifs de Modèles".

Deux exposés en région parisienne durant l'automne 1997 sur la sélection adaptative de modèles : le 22 octobre à l'Université Paris-Nord, le 27 octobre au Séminaire de statistique de l'ENS, deux autres durant l'hiver, au séminaire du laboratoire de Probabilités de Paris 6 (le 3 février 1998) sur la métastabilité d'un processus de vote majoritaire biaisé et au séminaire du laboratoire "Statistique et modèles aléatoires" (le 14 janvier 1998) de Paris 6/7 sur l'estimation adaptative d'un histogramme à pas variable.

Participation au colloque "Mathématiques pour la reconnaissance d'objets : Forme, Invariance et Déformation, Luminy 10-13 novembre 1997. Exposé intitulé "A mixture approach to statistical model selection".

Séjour de 15 jours à l'Université de Zürich, début mai 1998, sur invitation d'Erwin Bolthausen. Exposé intitulé "Statistical Mechanics and statistical inference".

Ecole d'été de probabilités de Saint Flour (août 1998), participation en tant qu'auditeur. Exposé sur l'estimateur de Gibbs.

Deux exposés en région parisienne durant l'automne, à l'IHP le 7 octobre et à Marne-la-Vallée le 13 novembre, sur l'estimation adaptative.

Coordination de l'organisation d'un colloque "Théorie de l'Information, Statistique adaptative et Reconnaissance des formes," qui s'est tenu du 7 au 11 déc. 1998 au CIRM, Marseille Luminy. (Comité d'organisation : Robert Azencott – ENS Cachan, Lucien Birgé – Université Paris VI, Olivier Catoni – Université Paris VI et ENS Paris, Marie Duflo – Université de Marne-la-Vallée, Christine Graffigne – Prisme, Université Paris V, Marie-Anne Gruet – INRA Biométrie, Pascal Massart – Université Paris XI, Alain Trouvé – Université Paris XIII)

Organisation en collaboration avec Thierry Bodineau, Francis Comets, Domi-

nique Picard, et Alexandre Tsybakov du séminaire "Statistique et Modélisation" du laboratoire de Probabilités et Modèles Aléatoires. Le programme de ce séminaire, depuis sa création, peut être consulté sur le site internet du laboratoire : `http://www.proba.jussieu.fr`

Participation en tant que conférencier invité au colloque *Computer vision and speech recognition : statistical foundations and applications*, Anogia, Crète, 3-9 juillet 1999, organisé par David Mumford et Basilis Gidas.

Exposé au séminaire du laboratoire de Statistique et Probabilités de l'Université Paul Sabatier de Toulouse, le 3 décembre 1999, sur invitation de Michel Ledoux, sur l'obtention d'inégalités de déviation "presque gaussiennes" pour les processus indépendants et les chaînes de Markov.

Exposé au séminaire de Probabilités du laboratoire de Probabilités et Modèles Aléatoires, le 25 janvier 2000, sur les *déviations presque gaussiennes*.

Exposé au séminaire du CMLA de l'ENS Cachan le 24 février, *Méthodes d'énergie libre pour la concentration de la mesure et la sélection d'estimateurs*.

Invitation de Felipe Cucker au Smale's Festschrift, Hong Kong, 13-17 July 2000, "Foundations of Computational Mathematics" (avec proceedings, voir publications).

Invitation au workshop on the Mathematical Foundations of Natural Language Modeling, October 30 – November 3, 2000, Institute for Mathematics and its Applications, University of Minnesota, Minneapolis, organisé par R Rosenfeld (CMU), S Khudanpur (JHU), M Johnson (Brown), F Jelinek (JHU), exposé intitulé "Non-asymptotic oracle inequalities, adaptive histograms and generalized n-grams".

Exposé aux journées de probabilités, CIRM, 11-15 septembre 2000, organisées par J. Azéma et M. Yor, intitulé "Inférence statistique, compression de données et inégalités de déviations".

Exposé à l'Université de Rennes, intitulé "Estimation de la transformée de Laplace, oracles et déviations", le 20 novembre 2000.

Exposé aux "Rencontres de statistiques mathématiques", CIRM, 11-15 décembre 2000, organisées par Oleg Lepski et Dominique Picard, intitulé "Aggregation of estimators and oracle inequalities".

Exposé au séminaire méthodes mathématiques du traitement d'images, organisé par Albert Cohen et Patrick Combettes, laboratoire d'analyse numérique, Paris 6, intitulé "Méthodes d'agrégation et complexité empirique en reconnaissance des formes", le 9 janvier 2001.

Conférencier à l'Ecole d'Eté de Probabilités XXXI, Saint-Flour 2001 (9-25 juillet) : « Statistical learning theory and stochastic optimization ».

Conférencier invité au colloque « Statistical Learning in Classification and Model Selection » EURANDOM, Eindhoven, The Netherlands, January 15-18, 2003, organisé par R. D. Gill (Universiteit Utrecht/EURANDOM), P. Grünwald

(CWI), A.W. van der Vaart (Vrije Universiteit Amsterdam / EURANDOM) et J. Lember (EURANDOM). Exposé intitulé « Localized PAC-Bayesian theorems and randomized estimators ».

Exposé au groupe de travail « Théorie de l'Information et Statistiques » organisé par E. Gassiat et S. Boucheron à l'Université Paris Sud, le 27 février 2003 : « Théorèmes PAC-Bayésiens locaux et estimateurs randomisés ».

Exposé au groupe de travail « Support Vector Machines », organisé par P. Reynaud, S. Boucheron et P. Massart à l'Université Paris Sud, le 28 mars 2003 : « Théorèmes PAC-Bayésiens et Support Vector Machines ».

Conférencier invité au colloque : « Journées de Probabilités », Toulouse, 8-12 septembre 2003, comité scientifique : J. Azéma, M. Emery, M. Yor, organisé par le LSP UMR C5583. Exposé intitulé « Théorèmes PAC-Bayesiens pour les Support Vector Machines ».

Conférencier invité au EU PASCAL Workshop on « Learning Theoretic and Bayesian Inductive Principles », organisé au Gatsby Computational Neuroscience Unit, University College, London (UK) du 19 au 21 Juillet 2004. Comité de programme : Z. Ghahramani, P. Grünwald, J. Langford, G. Lugosi, S. Mendelson, J. Shawe-Taylor. Exposé intitulé « Transductive PAC-Bayesian classification ».

Exposé au séminaire « Des Mathématiques » du Département de Mathématiques et Applications de l'École Normale Supérieure de Paris, le 1er juin 2005, intitulé « Classification PAC-Bayésienne et inégalités de Vapnik ».

Exposé au séminaire de statistique de Rennes, le 6 janvier 2006, intitulé « Apprentissage statistique : quelques théorèmes PAC-Bayésiens » (à l'invitation du groupe de recherche en statistique commun aux Universités de Rennes 1 et 2 et à l'Agrocampus de Rennes).

Conférencier invité à l'International Meeting on Empirical Processes and Asymptotic Statistics, Université de Rennes 1, 18-20 juin 2007, organisé par Philippe Berthet, exposé intitulé « Learning, information theory and thermodynamics ».

Conférencier invité du Workshop *Foundations and New Trends of PAC Bayesian Learning*, 22-23 March 2010, University College London, organisé par Jean-Yves Audibert, Matthew Higgs, Steffen Grünewald, François Laviolette et John Shawe-Taylor, dans le cadre du réseau européen Pascal 2. Exposé intitulé « Robust PAC-Bayes bounds ». Vidéo disponible sur internet :
http://videolectures.net/pacbayesian_catoni_rpbb/

Exposé au colloquium du laboratoire de mathématiques Paul Painlevé, Université de Lille 1, le vendredi 21 janvier 2011, (sur invitation de Thomas Simon) : *La moyenne empirique est-elle perfectible ?* (transparents disponibles sur ma page web).

Exposé au séminaire de Probabilités et Statistiques du Laboratoire J.A. Dieudonné, Université de Nice - Sophia Antipolis, le jeudi 12 mai 2011, (sur invitation

de Patricia Reynaud-Bouret) : *Petites perturbations des estimateurs et bornes PAC-Bayésiennes*, (transparent disponibles sur ma page web).

Invitation du Research Committee de la Royal Statistical Society, à l'*ordinary meeting* du mercredi 19 octobre 2011, en tant que second commentateur (second discussant) de l'article *Catching up faster by switching sooner : a predictive approach to adaptive estimation with an application to the AIC-BIC dilemma*, de T. van Erven, P. Grünwald et S. de Rooij, paru dans la série B (Statistical Methodology) du Journal of the Royal Statistical Society en juin 2012.

Exposé lors de la journée thématique *Image et Apprentissage* du séminaire *Méthodes Mathématiques du Traitement d'Images* du Laboratoire Jacques-Louis Lions (UMR 7598), le jeudi 5 juillet 2012, intitulé Bornes PAC-Bayésiennes, classification et clustering en grande dimension.

Invitation d'une semaine à l'Institute for Information Transmission Problems, à Moscou, par Yuri Golubev et Vladimir Spokoiny, du 25 au 30 novembre 2012, dans le cadre du programme « Structural methods of data analysis and optimization ». Exposé long (deux heures) intitulé « Unsupervised statistical learning through label aggregation ».

Exposés à Toulouse, le 26 mars 2013, aux séminaires de Probabilités et de Statistique de l'Institut de Mathématiques de Toulouse (Université Paul Sabatier), intitulés « Toric grammars, a new stochastic model » et « The statistics of Principal Component Analysis ».

Conférencier invité de l'*International Workshop on Statistical Learning*, Moscow, June 26-28, 2013, organisé à l'Institute for Information Transmission Problems par Philippe Rigollet et Yuri Golubev. Exposé intitulé « Dimension dependent and dimension free PAC-Bayes bounds for the Gram matrix ».

Invitation (du 25 au 27 juillet 2013) de Christophe Lampert, professeur à l'Institute for Science and Technology, Austria, (Computer Vision and Machine Learning Group), pour participer au suivi de la thèse de son étudiante, Anastasia Pentina, portant sur les fondements statistiques du « transfer learning ». Exposé au séminaire de l'Institut intitulé « PAC-Bayes bounds using Gaussian posterior distributions ».

Exposé au Séminaire Parisien de Statistique, intitulé « Markov substitute models and statistical inference in linguistics » (séance organisée par Estelle Kuhn et Mathilde Mougeot), le lundi 7 avril 2014, à l'Institut Henri Poincaré.

Participation par visio-conférence à l' IFCAM (Indo-French Centre for Applied Mathematics) Summer School on Applied Mathematics, Indian Institute of Science, Bangalore (July 2014), en tant que conférencier invité. J'ai fait trois heures d'exposé, réparties en deux séances, les 24 et 25 juillet, sur les bornes PAC-Bayésiennes appliquées à l'apprentissage statistique. Le matériel de cette visio-conférence est disponible sur ma page web : notes de cours, transparents et fichiers video combinant le défilement des transparents avec mes commentaires

audio.

Exposé le 4 mai 2015 au séminaire SMILE, Statistical Machine Learning in Paris, organisé par l'ENS Paris, l'INRIA, les Universités Paris 6 et 7, Télécom Paris-Tech, les Mines Paris-Tech, l'Université Paris Sud, l'Ecole des Ponts Paris-Tech, et l'Institut Curie. Titre : « PAC-Bayes bounds for the Gram matrix and least squares regression ».

Exposé le 12 octobre 2015, à l'Institut Henri Poincaré, au Séminaire Parisien de Statistiques, intitulé « Spectral clustering, reproducing kernels and Markov chains with exponential transitions ». (Séance organisée par Marc Hoffmann et Vincent Rivoirard.)

Exposé le 25 janvier 2016, intitulé « Least squares regression with a random design and Gram matrix estimates », au Séminaire de Statistique de l'ENSAE-CREST, organisé par Alexandre Tsybakov et Arnak Dalalyan.

Exposé à Lille, le 9 juin 2017, intitulé « Markov substitute models, an alternative to Hidden Markov Models », au séminaire de l'équipe INRIA Sequel.

Exposé en tant que conférencier invité du workshop « (Almost) 50 Shades of Bayesian Learning : PAC-Bayesian trends and insights », tenu à Long Beach, CA, USA, le 9 décembre 2017, dans le cadre de la conférence NIPS 2017, sous le titre « Dimension-free PAC-Bayesian bounds ».

Exposé dans le même workshop de présentation de l'article « Dimension-free PAC-Bayesian bounds for the estimation of the mean of a random vector », [CG17a] sélectionné par le comité de lecture du workshop.

Conférencier invité de la conférence « Élément de mathématique pour l'IA », organisée en l'honneur de Robert Azencott, les 14 et 15 mai 2019 à l'ENS Paris-Saclay alors encore installée sur son campus de Cachan. Exposé intitulé « Statistical syntax analysis for signal processing ».

Conférencier invité de la conférence en ligne « Conference on robustness and privacy », organisée les 22 et 23 mars 2021 par Cristina Butucea, Victor-Emmanuel Brunel, Nicolas Chopin, Arnak Dalalyan, Guillaume Lecué, Matthieu Lerasle, Vianney Perchet et Alexandre Tsybakov sous l'égide du CNRS. Exposé intitulé « Means and k-means : dimension free PAC-Bayesian bounds for robust estimators. »

2. Résumé des travaux effectués entre décembre 2018 et juin 2021

Mon activité principale durant la période concernée a porté sur la classification non supervisée et l'analyse syntaxique de signaux, en lien avec la direction de la thèse de Gautier Appert [App20], soutenue le 29 octobre 2020.

Je me suis intéressé à trois questions.

— L'étude de l'algorithme des k-means [AC21], [App20]. Cet algorithme joue un rôle majeur dans les domaines de la classification non supervisée, de la quantification vectorielle et de la compression de données avec pertes. Nous apportons deux types de contributions.
  — Une contribution à l'étude statistique de l'algorithme sous la forme de nouvelles bornes de généralisation valables en grande dimension et en dimension infinie. Le point de vue statistique consiste à comparer le comportement de l'algorithme appliqué à un échantillon statistique à son comportement en moyenne. Il permet de savoir si la classification obtenue rend compte de la loi des données ou si elle est instable et donnerait un résultat complètement différent sur un autre jeu de données tirées suivant la même loi. Les bornes indépendantes de la dimension du signal et les bornes en dimension infinie sont importantes pour montrer qu'une méthode statistique, et elles sont rares, échappe au fléau de la dimension et peut être appliquée directement à des données de grande dimension sans prétraitement.
  — Nous proposons aussi une nouvelle interprétation des k-means en terme d'entropie conduisant à des critères modifiés, critère robuste pour la classification de signaux non bornés et critère entropique spécifique pour la classification de bags of words, ou plus généralement de probabilités conditionnelles régulières.
— La classification non supervisée de parties d'un signal, ou fragmentation. C'est une problématique nouvelle et une étape essentielle dans l'élaboration d'une analyse syntaxique du signal, une théorie que nous avons l'ambition de faire émerger dans les années qui viennent. Alors que la méthode des k-means attribue un label à chaque observation, la fragmentation lui attribue un ensemble aléatoire de labels, correspondant à un découpage du signal en fragments. Cette modification somme toute naturelle permet de rapprocher la classification non supervisée des algorithmes de compression avec perte : on indexe des blocs fréquents et non plus des signaux entiers. Elle permet aussi de passer d'une représentation continue du signal à une représentation symbolique beaucoup plus riche, propice à une analyse syntaxique supplémentaire. En un mot, alors que l'analyse des signaux porte le plus souvent sur l'étude de structures de superposition, nous tentons d'étudier des structures de juxtaposition, que nous espérons pertinentes pour faire progresser l'état de l'art de l'intelligence artificielle.
  Plus concrètement et plus techniquement, il s'agit d'introduire un dictionnaire de fragments et de représenter les signaux par des assemblages de fragments disjoints, en tolérant comme pour les k-means un certain degré de distorsion. Dans la thèse de Gautier Appert et dans un article en préparation, nous établissons une borne de généralisation pour la fragmentation

dans laquelle le terme de complexité dépend du nombre de fragments du dictionnaire (ou plus précisément de leur surface cumulée) et non du nombre possible d'assemblages de ces fragments, comme cela aurait été le cas en appliquant la borne obtenue pour les k-means à la fragmentation vue comme un cas particulier des k-means dans lequel les centres ont une structure déterminée. Ici encore, les bornes sont indépendantes de la dimension et montrent que la méthode s'applique à des signaux d'une résolution arbitraire sans qu'il soit besoin de faire exploser la taille de l'ensemble d'apprentissage. De plus, les bornes ne dépendent que de la taille et du nombre de fragments, ouvrant la possibilité de travailler avec des fragments de forme arbitraire. Cela permet d'adapter les fragments aux données d'une manière beaucoup plus fine que si nous étions contraints pour éviter le sur-apprentissage à restreindre la géométrie des fragments.

— L'analyse syntaxique d'ensembles aléatoires de labels. Il s'agit de décrire les assemblages possibles de fragments obtenus à l'étape précédente à l'aide d'un langage généré par une grammaire. Plus ce langage sera petit, meilleure sera la compression de l'échantillon d'apprentissage. Plus la description de la grammaire sera petite, meilleure sera la fiabilité du langage pour décrire d'autres données suivant la même loi. Une première ébauche est proposée dans la thèse de Gautier Appert [App20]. Tant le critère de compression que la méthode de construction des grammaires sont en cours d'amélioration en vue d'une publication à venir. Nous avons aussi le projet de tester un modèle similaire pour l'analyse syntaxique des langues naturelles.

La thèse de Gautier Appert [App20] fournit une première version des trois étapes du programme démontrant me semble-t-il leur faisabilité ainsi que leur intérêt pratique à travers quelques simulations. La première question, concernant les k-means a été retravaillée dans une première prépublication soumise et disponible sur ArXiv [AC21]. Les deux points suivants sont en cours d'amélioration en vue d'être publiés chacun séparément.

En marge de ce programme de recherche principal, j'ai collaboré durant l'année 2020 avec Miquel Oliu-Barton et Bruno Ziliotto à l'étude des propriétés de paiement constant de certains jeux stochastiques escomptés à somme nulle. Un article intitulé Constant payoff in zero-sum stochastic games [COZ20] est accepté par les Annales de l'Institut Henri Poincaré : Probabilités et Statistiques.

Durant cette collaboration, j'ai pu apporter mes compétences concernant les chaînes de Markov à transitions rares, acquises durant la première partie de ma carrière, à la démonstration d'une conjecture de Sorin, Venel et Vigneral.

3. Rapport d'activité de décembre 2018 à juin 2021

3.1. Etude des k-means et de certaines de leurs généralisations. Il s'agit d'un travail en collaboration avec Gautier Appert présenté dans une prépublication soumise [AC21] ainsi que, sous une forme préliminaire, dans la première partie de la thèse de Gautier Appert [App20].

Il propose une extension des k-means adaptée à la quantification vectorielle de probabilités conditionnelles à valeurs dans une famille exponentielle. Ce cadre couvre en particulier la quantification de bags of words. Dans le cadre statistique où on effectue la quantification à partir d'un échantillon i.i.d., nous prouvons une borne en $\sqrt{k/n}$ à des termes logarithmiques près. Il s'agit d'une borne non asymptotique indépendante de la dimension, les données pouvant en particulier appartenir à un domaine borné d'un espace de Hilbert séparable de dimension infinie. Elle améliore la borne en $k/\sqrt{n}$ obtenue par Biau, Devroye et Lugosi [BDL08] et la borne en $\sqrt{k/n}$ à des termes logarithmiques près de Fefferman, Mitter et Narayanan [FMN16] en ce qui concerne les termes logarithmiques et le caractère explicite des constantes, avec une preuve plus directe et entièrement différente de la leur. Dans le cas des k-means classique, notre résultat s'énonce de la façon suivante.

Théorème 3.1 *Soit X un vecteur aléatoire à valeurs dans la boule*

$$\mathcal{B} = \left\{x \in H \,:\, \|x\| \leq B\right\}$$

d'un espace de Hilbert séparable H. Soit $(X_1, \ldots, X_n)$ un échantillon composé de n copies indépendantes de X. Supposons que $n \geq 2k$ et $k \geq 2$. Pour tout $\delta \in]0,1[$, avec probabilité au moins $1-\delta$, pour tout ϵ-minimiseur $\widehat{c} \in \mathcal{B}^k$ du risque empirique

$$\frac{1}{n}\sum_{i=1}^{n} \min_{j\in[\![1,k]\!]} \left\|X_i - \widehat{c}_j\right\|^2$$

$$\mathbb{E}_X\Big(\min_{j\in[\![1,k]\!]} \left\|X - \widehat{c}_j\right\|^2\Big) \leq \inf_{c\in H^k} \mathbb{E}_X\Big(\min_{j\in[\![1,k]\!]} \left\|X - c_j\right\|^2\Big) + 16B^2 \log\Big(\frac{n}{k}\Big)\sqrt{\frac{k\log(k)}{n}} + 4B^2\sqrt{\frac{2\log(\delta^{-1})}{n}} + \epsilon.$$

Nous avons obtenu cette vitesse optimale en $\sqrt{k/n}$ à des termes logarithmiques près au printemps 2020, en mettant en œuvre une nouvelle méthode de chaînage PAC-Bayésien en dimension infinie, alors que les autres preuves [BDL08], [FMN16] utilisent la complexité de Rademacher. Notre résultat a l'intérêt de montrer que

la quantification vectorielle à partir d'un échantillon reste pertinente en grande dimension et tolère l'utilisation d'un nombre de centres quasiment linéaire en la taille de l'échantillon. La nouvelle méthode de chaînage PAC-Bayésien de perturbations Gaussiennes du paramètre introduite ici a très certainement son utilité dans un cadre plus vaste et pourra être déclinée dans d'autres contextes.

3.2. Algorithme de fragmentation et classification locale. Dans la perspective de notre programme portant sur l'apprentissage non supervisé, la quantification vectorielle représente une étape importante de passage d'une représentation continue à une représentation discrète prémisse d'une classification non supervisée des observations. Néanmoins, dans notre programme, nous proposons d'utiliser une généralisation de la quantification selon le critère des k-means qui consiste à découper les signaux en fragments et à approcher chacun de ces fragments par un centre utilisé plusieurs fois pour approcher des signaux différents de l'échantillon d'apprentissage. Notre algorithme de fragmentation se présente ainsi comme un algorithme de compression avec perte dans lequel on maintient la distorsion de la représentation au dessous d'un certain seuil tout en minimisant la taille de la représentation dont le terme principal est constitué par la somme des tailles des fragments. Cette somme est d'autant plus faible que chaque fragment intervient dans la représentation d'un grand nombre de signaux de l'échantillon. L'algorithme indexe ainsi des fragments fréquents (à distorsion près), faisant penser à une version avec perte de l'algorithme de Lempel Ziv. Nous établissons des bornes de généralisation comparant la distorsion d'une fragmentation calculée sur un échantillon i.i.d. avec la distorsion moyenne par rapport à la loi de l'échantillon. Ces bornes ressemblent à celles que nous obtenons pour le critère des k-means. Elles sont indépendantes de la dimension de l'espace ambiant et décroissent comme la racine carrée du rapport c sur n, où n est la taille de l'échantillon et où le terme de complexité c qui remplace k est proportionnel à la somme des tailles des fragments. Ces bornes sont une indication du fait qu'il est possible de travailler directement avec des signaux haute résolution (du fait de l'indépendance de la dimension) et qu'il est possible aussi de ne pas restreindre la forme des fragments, le terme de complexité ne faisant intervenir que la somme de leurs tailles.

Ces résultats sur la fragmentation sont présentés dans la thèse de Gautier Appert [App20] et seront réorganisés et améliorés pour former le contenu d'un article signé en commun en cours de rédaction.

3.3. Compression à base de grammaires et classification par arbres syntaxiques. Notre algorithme de fragmentation permet de passer du continu au discret, de l'analogique au symbolique, chaque signal étant alors représenté par un ensemble aléatoire de symboles indexant chacun l'un de ses fragments. La

seconde étape consiste à effectuer une forme d'analyse syntaxique de cette représentation symbolique. Cette analyse syntaxique a pour but de rendre compte des interactions entre les symboles et pour effet de fournir pour chaque signal un arbre syntaxique effectuant une classification multi-échelles des fragments le constituant. L'idée qui sous-tend cette proposition est que la structure des langues naturelles, la syntaxe, reflète plus largement le type de traitements effectués par le cerveau sur tous types de signaux et peut de ce fait être utilisée non seulement pour le traitement automatique des langues naturelles mais aussi pour la modélisation de la perception. Cependant, le choix d'un modèle syntaxique est une question largement ouverte. Nous avions fait une proposition en ce sens en collaboration avec Thomas Mainguy, à l'occasion de la direction de sa thèse, en introduisant les modèles de substitution Markoviens. Cette approche semblait prometteuse à ceci près qu'elle introduisait des familles de modèles si larges que la question de la sélection du modèle semblait difficile à mettre en œuvre en pratique, sans utiliser des tailles d'échantillon d'apprentissage gigantesques ou des heuristiques difficiles à justifier. Avec Gautier Appert, après avoir consacré en vain beaucoup d'énergie à améliorer ce modèle statistique pour en faire un modèle polyvalent capable d'extraire de l'information de n'importe quel type de données, nous avons décidé de nous simplifier la vie en cherchant un modèle de compression de données au lieu de chercher un modèle statistique. Les deux sont liés, de tout modèle statistique on peut déduire une méthode de compression de données et de tout algorithme de compression on peut déduire un estimateur statistique, mais ce qui se modélise simplement dans l'un de ces deux cadres ne se modélise pas toujours simplement dans l'autre. Nous nous sommes aperçu que la compression de données offrait des possibilités de modélisation particulièrement souples. Nous nous sommes plus particulièrement intéressés à la compression fondée sur des grammaires (grammar based compression en anglais). L'algorithme de Lempel Ziv rentre dans cette catégorie et la littérature sur le sujet est tournée vers l'obtention de la compression la plus efficace possible. Nous avons adopté le point de vue inverse, considérant le taux de compression non pas comme une fin en soi, mais comme un moyen de sélectionner une grammaire significative. La technique générale que nous avons mise en œuvre consiste à définir une famille de grammaires et à construire d'une manière itérative une grammaire appartenant à cette famille en ajoutant à chaque étape une règle qui optimise la variation du taux de compression parmi un ensemble de règles possibles. Une première proposition est faite dans la thèse de Gautier Appert, dans laquelle on construit une première grammaire possédant des règles du type $p \rightarrow ab$, puis dans laquelle on construit une deuxième grammaire qui comprime une représentation factorisée des paires apparaissant dans la première grammaire. Expérimentalement, ce modèle donne des résultats prometteurs pour la reconnaissance des formes en imagerie, comme exposé dans la dernière partie de la thèse. Nous avons réfléchi à l'automne 2020 à des modèles de grammaire plus

élaborés. Nous comptons les tester prochainement dans le domaine du traitement des langues naturelles aussi bien que dans celui de la classification non supervisée d'images.

3.4. Propriété de paiement constant dans les jeux stochastiques escomptés à somme nulle. Il s'agit de montrer que dans un jeu à espace d'états fini et à paiements escomptés, lorsque les stratégies des deux joueurs sont optimales au sens où elles constituent un point selle d'un problème min-max, dans la limite où le taux d'escompte tend vers zéro, le paiement au temps t défini dans l'intervalle zéro-un à partir du taux d'escompte est linéaire en t. Les différentes quantités entrant en jeu peuvent s'exprimer en terme de matrices de Markov à transitions rares dépendant du taux d'escompte. Alors que la décomposition en cycles permet de comprendre qualitativement pourquoi le résultat est juste, la difficulté technique consiste à s'en passer pour ne pas avoir à se lancer dans des considérations inextricables concernant la possible périodicité des chaînes de Markov entrant en ligne de compte.

4. Activités antérieures, de décembre 2013 à décembre 2018

Mon activité scientifique depuis décembre 2013 s'est organisée autour de trois thèmes principaux,

- L'obtention de nouvelles bornes PAC-Bayésiennes dans divers contextes : des bornes de marge pour les Support Vector Machines, l'estimation de la matrice de Gram d'un échantillon aléatoire (et plus généralement l'estimation de la moyenne d'un échantillon de matrices aléatoires), la régression aux moindres carrés avec design aléatoire dans le cas de la dimension finie ainsi que dans le cas d'un design à valeurs dans un espace de Hilbert séparable (collaboration avec Ilaria Giulini). A l'occasion de la direction de la thèse d'Ilaria Giulini, je me suis aussi intéressé à l'obtention de bornes PAC-Bayésiennes pour la convergence de certains algorithmes de clustering spectral. Dans un travail en cours en lien avec la direction de la thèse de Gautier Appert, j'étudie des bornes PAC-Bayésiennes pour la convergence de l'algorithme des k-means et de certaines de ses variantes.
- L'étude de nouveaux modèles statistiques pour l'analyse syntaxique des langues naturelles à partir d'un corpus. Ce thème comprend la direction de la thèse de Thomas Mainguy.
- La conception et l'étude de modèles statistiques « syntaxiques » pouvant représenter tous types de données, langage, signal, données multi-capteurs, de même qu'il existe des algorithmes de compression sans perte (l'algorithme de Lempel-Zif, par exemple) applicables à tous types de données.

Cette nouvelle analyse statistique syntaxique modélise la combinaison répétée de deux types d'opérations. Premièrement le découpage des données en blocs fréquents et la classification non supervisée de ces blocs. A l'issue de cette première phase, les données sont représentées par un modèle discret de type bag of words. Deuxièmement, une analyse de type contextuelle des probabilités conditionnelles d'apparition d'un bloc conditionnellement à la présence de certains autres, permettant de regrouper les labels des blocs apparaissant dans les mêmes contextes pour créer des catégories syntaxiques. Le tout dans une approche hiérarchique où on forme et on étiquette des blocs de blocs, de blocs, etc. de plus en plus larges de façon à obtenir une classification arborescente des données. Ce programme de recherche comprend la direction en cours de la thèse de Gautier Appert.

4.1. Bornes PAC-Bayésiennes. Durant la période précédente, j'avais étudié des bornes PAC-Bayésiennes utilisant des lois de Gibbs a priori et a posteriori sur les paramètres [Cat07]. Durant la période concernée par ce rapport, j'ai plutôt analysé des bornes PAC-Bayésiennes fondées sur des perturbations Gaussiennes du paramètre. Cette voie a été ouverte par [LS02] et [McA03] en ce qui concerne les bornes de marge pour les Support Vector Machines. J'ai repris et complété leurs calculs dans des notes de cours qui devaient devenir [Cat15b].

J'y expose des résultats originaux qui prennent la suite de l'étude de la classification PAC-Bayésienne exposée dans [Cat07].

Tout d'abord, j'établis une correspondance entre les bornes PAC-Bayésiennes que j'avais obtenues dans [Cat07] et les bornes de Seeger. Je montre que, pour toute fonction de coût binaire, (telle que l'erreur de classification), la borne que j'ai proposée dans [Cat07] permet de déduire une borne du même type que celle de Seeger, avec une légère amélioration, le terme $\log(n/\epsilon)$ de la borne de Seeger étant remplacé par un terme en $\log(\log(n)^2/\epsilon)$ (comme établi dans [Cat15b, Proposition 20.4, page 293]). Je reprends ensuite les travaux de Langford, Shawe-Taylor et McAllester sur les inégalités de marge pour les Support Vector Machines en les précisant à la lumière de cette borne générique. On aboutit à une borne empirique de l'erreur de généralisation, obtenue par une technique de perturbation gaussienne. On peut utiliser cette borne empirique pour définir un estimateur, et majorer son risque de généralisation. Diverses variations sur ce thème sont possibles. On peut en particulier affaiblir la borne empirique minimisée pour retrouver le critère de *box constraint* utilisé classiquement pour choisir les pondérations des Support Vector Machines et en déduire un nouveau critère pour fixer leur paramètre de régularisation. Il convient pour cela d'étendre l'étude à la séparation par un hyperplan affine [Cat15b, Corollaire 20.1, page 301].

J'ai aussi appliqué l'utilisation de lois Gaussiennes sur les paramètres permet-

tant des calculs explicites à la régression aux moindres carrés avec design aléatoire. Dans un premier temps, dans des travaux en collaboration avec Jean-Yves Audibert [AC11a ; AC11b ; AC10], nous avons proposé et analysé un estimateur de régression robuste solution d'un problème min max. Ces travaux m'ont conduit à publier par la suite une étude spécifique sur l'estimation robuste de la moyenne d'une variable réelle [Cat12] qui est régulièrement citée dans les articles parus depuis sur l'estimation robuste, l'estimateur que je propose partageant avec les médianes de moyennes [LI11 ; Min15] la propriété d'avoir des déviations essentiellement sous-Gaussiennes dès que la variable à estimer possède une variance [Dev+16].

Ces premières études, antérieures à 2014, m'ont progressivement convaincu du rôle central joué par l'estimation de la matrice de Gram dans la conception et l'étude d'estimateurs de régression aux moindres carrés avec design aléatoire. J'ai développé ce point de vue pour obtenir une suite de résultats exposés dans [Cat16], soumis pour publication. Dans cette prépublication, je me concentre sur le cas de la dimension finie, ou plus précisément sur l'obtention de bornes sur le risque d'estimation dépendant de la dimension. Parallèlement, je proposai à Ilaria Giulini, dans une thèse préparée de septembre 2012 à septembre 2015, d'étudier la possibilité d'obtenir des bornes indépendantes de la dimension. La première partie de sa thèse établit ainsi des bornes indépendantes de la dimension pour l'estimation de la matrice de Gram ou plus généralement de l'opérateur de Gram dans un espace de Hilbert séparable. En s'appuyant sur ces bornes, Ilaria Giulini a pu justifier des méthodes robustes d'analyse en composantes principales de la représentation d'un échantillon aléatoire dans un espace de Hilbert à noyau reproduisant. Le passage des bornes dépendantes de la dimension aux bornes indépendantes de la dimension se fait en changeant la matrice de covariance des lois Gaussiennes a priori et a posteriori sur les paramètres.

Dans [Cat16], qui concerne donc les bornes dépendant de la dimension, j'expose quatre types de résultats.

Tout d'abord, je propose un estimateur robuste de la matrice de Gram (ainsi qu'une variante permettant d'estimer la matrice de covariance). Plus précisément, étant donné un échantillon i.i.d. $X_1, \dots, X_n \in \mathbb{R}^d$, et sa matrice de Gram

$$G = \mathbb{E}\big(XX^\top\big),$$

je propose un estimateur robuste de

$$\theta^\top G\theta = \mathbb{E}\big(\langle\theta, X\rangle^2\big),$$

dont l'erreur d'estimation peut être contrôlée uniformément pour tout $\theta \in \mathbb{S}_d$, la sphère unité de $\mathbb{R}^d$. De cet estimateur directionnel, on peut déduire un estimateur

$\widehat{G}$ de la matrice de Gram G tel que pour tout $\theta \in \mathbb{R}^d$,

$$\left|\theta^\top\big(\widehat{G} - G\big)\theta\right| \leq \theta^\top G\,\theta \;\mathbf{O}\left(\sqrt{\frac{\kappa[d + \log(\epsilon^{-1})]}{n}}\right), \tag{4.1}$$

où la notation $\mathbf{O}$ correspond à une borne non asymptotique avec des constantes numériques explicites. La constante κ doit être connue (elle est utilisée dans la construction de l'estimateur) et vérifier

$$\sup\Big\{\mathbb{E}\big(\langle\theta, X\rangle^4\big) \;:\; \theta \in \mathbb{R}^d, \mathbb{E}\big(\langle\theta, X\rangle^2 \leq 1\Big\} \leq \kappa < \infty. \tag{4.2}$$

Les autres constantes intervenant dans la borne étant numériques, le fait que κ soit fini est donc la seule condition sur la distribution des données.

Remarquons que la borne ci-dessus contrôle un type d'erreur d'estimation plus fin que la norme d'opérateur de la différence $\widehat{G} - G$ pour la métrique canonique de $\mathbb{R}^d$. On peut interpréter si on veut le résultat comme un contrôle de la norme d'opérateur pour la métrique définie par la matrice de Gram à estimer G elle-même. On peut en particulier remarquer que ce type de borne permet un contrôle de toutes les valeurs propres de G.

Le deuxième point abordé dans [Cat16] est celui de l'analyse de l'estimateur empirique usuel de la matrice de Gram,

$$\overline{G} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top.$$

Nous fondons notre étude sur l'analyse du rapport

$$\frac{\theta^\top \overline{G}\,\theta}{\theta^\top \widehat{G}\,\theta}$$

entre l'estimateur empirique et notre estimateur robuste. En dehors de cas pathologiques, dont il convient de contrôler la probabilité en ajoutant des hypothèses, les deux estimateurs se comportent de la même façon, tout au moins au premier ordre. Je propose différents résultats sous des hypothèses plus ou moins fortes. Je montre par exemple que, pour un échantillon de taille n telle que

$$n \geq \left[20\sqrt{\kappa d} + \left(\frac{5}{2} + \frac{1}{2(\kappa-1)}\right)\sqrt{2(\kappa-1)\big[\log(\epsilon^{-1}) + 0.73\,d\big]}\right]^2 = \mathbf{O}\Big(\kappa\big[d + \log(\epsilon^{-1})\big]\Big),$$

avec probabilité au moins $1 - 4\epsilon$, pour tout $\theta \in \mathbb{R}^d$,

$$-\theta^\top G\,\theta\left(\delta + \gamma_-\right) \le \theta^\top \overline{G}\,\theta - \theta^\top G\,\theta \le \theta^\top G\,\theta\,\frac{\delta + \gamma_+}{(1-\delta)(1-\gamma_+)_+}$$

où, pour deux exposants p et $q \in [1,2]$,

$$\begin{aligned}
\mu &= \sqrt{\frac{2(\kappa-1)}{n}\Big[\log(\epsilon^{-1}) + 0.73\,d\Big]} + 6.81\sqrt{\frac{2\kappa d}{n}},\\
\delta &= \frac{\mu}{1-2\mu} = \mathbf{O}(\mu) = \mathbf{O}\left(\sqrt{\frac{\kappa\Big[\log(\epsilon^{-1}) + d\Big]}{n}}\,\right),\\
\gamma_- &= \frac{2\Big[\log(\epsilon^{-1}) + 0.73\,d\Big]}{3(\kappa-1)n},\\
\gamma_+ &= \frac{1}{p+1}\left(\frac{2\Big[\log(\epsilon^{-1}) + 0.73\,d\Big]}{(\kappa-1)n}\right)^{p/2}\big(1+\delta\big)^{p+1}\\
&\qquad\times\left[\mathbb{E}\big(\|G^{-1/2}X\|^{2(p-1)}\big) + \frac{C_q\mathbb{E}\Big(\|G^{-1/2}X\|^{2q(p+1)}\Big)^{1/q}}{\epsilon^{1/q}n^{1-1/q}}\right],\\
\text{avec } C_q &= \frac{q^{q-1}}{2(q-1)^{q-1}(1-q/2)^{(2-q)/q}} \le 1.4.
\end{aligned}$$

J'indique ce résultat dans le détail pour montrer que j'obtiens des bornes non asymptotiques explicites avec des constantes numériques qui, sans prétendre à l'optimalité, ont des valeurs raisonnablement faibles.

On voit que les bornes obtenues pour les deux estimateurs $\widehat{G}$ et $\overline{G}$ sont asymptotiquement équivalentes quand la taille de l'échantillon n tend vers l'infini, étant toutes deux équivalentes au terme dominant δ. Tout l'intérêt d'avoir établi une borne non asymptotique réside dans la présence du terme du second ordre γ_+ dans la borne concernant $\overline{G}$. On voit bien en examinant γ_+ qu'il peut très bien dominer la borne pour un large évantail de valeurs de n et ne s'effacer devant δ que pour des tailles d'échantillons déraisonnablement grandes. Cette possibilité d'une plus faible performance de $\overline{G}$ s'observe en pratique, dans des situations où la distribution de $\langle\theta, X\rangle$ a une queue fournie dans certaines directions θ.

Le troisième point abordé dans [Cat16] consiste à établir un lien entre la régression linéaire aux moindres carrés avec design aléatoire de $Y \in \mathbb{R}$ sur $X \in \mathbb{R}^d$, où (X, Y) est un couple de variables aléatoires à valeurs dans $\mathbb{R}^{d+1}$ et l'estimation de la matrice de Gram de dimension $d+1$

$$G = \mathbb{E}\left[\begin{pmatrix} X \\ Y \end{pmatrix}\left(X^\top, Y\right)\right],$$

à partir d'un échantillon $(X_1, Y_1), \dots, (X_n, Y_n)$ composé de n copies indépendantes de $(X, Y) \in \mathbb{R}^{d+1}$.

On peut en effet remarquer que

$$\mathbb{E}\big[\big(Y - \langle \theta, X\rangle\big)^2\big] = \begin{pmatrix}\theta\\ -1\end{pmatrix}^\top G \begin{pmatrix}\theta\\ -1\end{pmatrix}, \qquad \theta \in \mathbb{R}^d.$$

Une façon de lier les deux problèmes consiste alors à considérer un estimateur $\check{G}$ de G tel qu'avec probabilité $1-\epsilon$, pour tout $\xi \in \mathbb{R}^{d+1}$,

$$\big|\xi^\top (G - \check{G})\xi\big| \le \delta\, \xi^\top G \xi$$

pour deux paramètre $\epsilon \in]0,1[$ et $\delta > 0$. Dans ce cas tout estimateur de régression $\widehat{\theta} \in \mathbb{R}^d$ vérifiant

$$\widehat{\theta} \in \arg\min_{\theta\in\mathbb{R}^d} \begin{pmatrix}\theta\\ -1\end{pmatrix}^\top \check{G} \begin{pmatrix}\theta\\ -1\end{pmatrix}$$

est tel qu'avec probabilité au moins $1-\epsilon$,

$$\mathbb{E}\big[\big(Y - \langle \widehat{\theta}, X\rangle\big)^2\big] - \inf_{\theta\in\mathbb{R}^d} \mathbb{E}\big[\big(Y - \langle \theta, X\rangle\big)^2\big] \le \frac{\delta^2}{(1-\delta)^2(1+\delta)} \begin{pmatrix}\widehat{\theta}\\ -1\end{pmatrix}^\top \check{G} \begin{pmatrix}\widehat{\theta}\\ -1\end{pmatrix}.$$

Autrement dit, la précision de l'estimation aux moindres carrés mesurée par l'excès de risque est le carré de la précision de l'estimation de la matrice de Gram de $(X, Y) \in \mathbb{R}^{d+1}$.

Cette propriété générale permet de transposer les résultats obtenus pour l'estimation de la matrice de Gram au problème de la régression aux moindres carrés avec design aléatoire. On peut ainsi d'une part proposer de nouveaux estimateurs robustes de la régression linéaire aux moindres carrés et étudier le minimiseur du risque empirique sous des hypothèses faibles de moments polynomiaux.

Si on se contente du principe de transfert ci-dessus, on obtient des majorations de l'excès de risque sous des hypothèses portant sur le couple de variables aléatoires (X, Y). En travaillant un peu plus, on peut se ramener à des hypothèses séparées sur la loi de X et sur la loi de $Y - \langle \theta_*, X\rangle$, où

$$\theta_* \in \arg\min_{\theta\in\mathbb{R}^d} \mathbb{E}\big[\big(Y - \langle \theta, X\rangle\big)^2\big].$$

En particulier pour l'étude de la régression robuste, la meilleure constante κ

$$\kappa = \sup\Big\{\mathbb{E}\big[\big(\zeta Y - \langle \theta, X\rangle\big)^4\big] \;:\; (\theta,\zeta) \in \mathbb{R}^{d+1}, \mathbb{E}\big[\big(\zeta Y - \langle \theta, X\rangle\big)^2\big] \le 1\Big\}$$

est majorée par

$$\sqrt{\kappa} \le \sqrt{\kappa_1} + \sqrt{\kappa_2},$$

où

$$\kappa_1 = \sup\Big\{\mathbb{E}\big(\langle\theta, X\rangle^4\big) \,:\, \theta \in \mathbb{R}^d, \mathbb{E}\big(\langle\theta, X\rangle^2\big) \leq 1\Big\},$$

$$\text{et } \kappa_2 = \begin{cases} \dfrac{\mathbb{E}\big[\big(Y - \langle\theta_*, X\rangle\big)^4\big]}{\mathbb{E}\big[\big(Y - \langle\theta_*, X\rangle\big)^2\big]^2}, & \mathbb{E}\big[\big(Y - \langle\theta_*, X\rangle\big)^2\big] > 0, \\ 0, & \text{sinon.} \end{cases}$$

On peut scinder de la même façon les hypothèses de moment pour l'étude du minimiseur du risque empirique.

Dans les deux cas, régression robuste ou minimisation du risque empirique, on obtient des bornes sur l'excès de risque d'ordre asymptotique

$$\mathbb{E}\big[\big(Y - \langle\theta_*, X\rangle\big)^2\big]\, \mathbf{O}_{n\to\infty}\left(\frac{(\kappa_1 + \kappa_2)\big[d + \log(\epsilon^{-1})\big]}{n}\right),$$

avec dans le cas du minimiseur du risque empirique des termes asymptotiquement du second ordre mais qui peuvent néanmoins dominer la borne pendant longtemps.

On peut alors montrer que cet ordre de grandeur est optimal dans le cas général, du moins pour le minimiseur du risque empirique. Cela peut paraître un peu surprenant, car cela contredit l'idée selon laquelle la convergence des moindres carrés empiriques est « en $d\,\sigma^2/n$ ».

En fait cette apparente contradiction est intéressante à évoquer, elle vient du fait qu'habituellement on fait une hypothèse de structure du type

$$Y = \langle\theta_*, X\rangle + W$$

où W est un bruit centré indépendant de X. Dans ce cas là, la convergence est bien « en $d\sigma^2/n$ ». Par contre, dans le cas général, la constante n'est pas forcément $d\,\sigma^2$, et peut en fait prendre une valeur arbitraire.

Nous avons pu préciser ce point en établissant la constante exacte dans la vitesse de convergence du minimiseur du risque empirique. Techniquement, nous montrons dans la proposition 4.12 (toujours dans [Cat16]) que, sous des hypothèses de moments polynomiaux faisant intervenir un exposant $q \in]1, 2[$,

$$\left|\frac{n}{C}\mathbb{E}\Big[\min\Big\{R(\widehat{\theta}) - R(\theta_*), C\exp\big(n^{2-q}\big)\Big\}\Big] - 1\right| \leq \mathbf{O}_{n\to\infty}\big(n^{-(q-1)/2}\big),$$

alors que

$$\mathbb{P}\Big[R(\widehat{\theta}) - R(\theta_*) \geq C\exp(n^{2-q})\Big] \leq \mathbf{O}_{n\to\infty}\left(\frac{1}{n\exp(n^{2-q})}\right),$$

où la constante exacte C vaut

$$C = \mathbb{E}\Big[\big(Y - \langle\theta_*, X\rangle\big)^2 \big\|\mathbb{E}\big(XX^\top\big)^{-1/2} X\big\|^2\Big].$$

Il est alors facile de voir que cette constante vaut bien $d\sigma^2$ dans le cas où le bruit $Y - \langle\theta_*, X\rangle$ est indépendant du design X. Il est aussi facile de construire des exemples dans lesquels $Y - \langle\theta_*, X\rangle$ et X sont dépendants et dans lesquels la constante C prend n'importe quelle valeur arbitraire. Cette remarque nous paraît particulièrement intéressante, dans la mesure où de nombreuses méthodes de sélection de modèles de régression reposent sur une pénalisation en $d\sigma^2/n$ des modèles en compétition, justifiée par le fait qu'il faut pénaliser les modèles proportionnellement à leur excès de risque. Notre étude montre que cette approche perd tout fondement dans le cas où on ne peut garantir que le bruit est indépendant du design.

Les algorithmes de régression robustes passant par une estimation de la matrice de Gram en dimension $d+1$ représentent une avancée par rapport à ceux que j'avais proposés précédemment en collaboration avec Jean-Yves Audibert [AC11a ; AC11b]. En particulier les algorithmes proposés alors nécessitaient la résolution d'un problème min max, comportant le risque algorithmique de rester piégé dans un minimum local. De plus les bornes établies alors pour la régression robuste supposaient le paramètre θ restreint à un domaine borné, alors que celles concernant la minimisation du risque empirique étaient asymptotiques, avec une constante sous-optimale.

J'ai aussi en parallèle à ces recherches dirigé la thèse d'Ilaria Giulini qui s'est attachée à obtenir des résultats indépendants de la dimension et de ce fait généralisables à un espace de Hilbert séparable de dimension infinie. Dans [Giu15a ; Giu15b], Ilaria Giulini montre dans le cadre évoqué ci-dessus de l'estimation de la matrice de Gram $G = \mathbb{E}\big(XX^\top\big)$ à partir d'un échantillon i.i.d. X_i, $1 \le i \le n$, qu'il existe un estimateur $\widehat{N}(\theta)$ de $\theta^\top G\theta$ vérifiant la propriété suivante. Pour tout $n \le 10^{20}$, avec probabilité au moins $1 - 2\epsilon$, pour tout $\theta \in \mathbb{R}^d$

$$\mathbb{1}\big(4\mu < 1\big)\Big|\widehat{N}(\theta) - \theta^\top G\theta\Big| \le \frac{\mu}{1-4\mu}\,\theta^\top G\theta,$$

où

$$\mu = \sqrt{\frac{2.07(\kappa-1)}{n}\left[\log(\epsilon^{-1}) + 4.3 + \frac{1.6\times\|\theta\|^2\,\mathbf{Tr}(G)}{N(\theta)}\right]} + \sqrt{\frac{2\kappa}{n}\times\frac{92\times\|\theta\|^2\,\mathbf{Tr}(G)}{N(\theta)}},$$

et où la constante κ est celle définie par l'équation (4.2, page 20). (La condition $n \le 10^{20}$ peut être supprimée à condition de remplacer la constante numérique

4.3 par un terme en $\log(\log(n))$.) On obtient donc une borne analogue à la borne dépendant de la dimension (4.1, page 20), où la dimension d de l'espace ambiant est remplacée par le terme $\dfrac{\|\theta\|^2 \operatorname{Tr}(G)}{N(\theta)}$ qui mesure le rapport entre l'énergie totale de la distribution et l'énergie dans la direction θ. En effet, pour n'importe quelle base orthonormée θ_i,

$$\frac{\|\theta_1\|^2 \operatorname{Tr}(G)}{N(\theta_1)} = \frac{1}{\mathbb{E}\big(\langle \theta_1, X\rangle^2\big)} \sum_i \mathbb{E}\big(\langle \theta_i, X\rangle^2\big).$$

L'exploitation de cette borne permet de proposer des estimateurs de la matrice de Gram plus stables que la matrice de Gram empirique et valables en grande dimension, voire en dimension infinie.

Dans un travail récent [CG17a ; CG17b] en collaboration avec Ilaria Giulini, nous nous affranchissons de l'hypothèse concernant la kurtosis et proposons un estimateur robuste de la matrice de Gram qui s'adapte à la variance directionnelle. Cet estimateur nécessite la connaissance d'une borne $T \geq \mathbb{E}\big(\|X\|^4\big)$ et vérifie avec probabilité au moins $1 - 2\delta$

$$\begin{aligned}\widehat{N}(\theta) \leq \mathbb{E}\Big(\langle \theta, X\rangle^2\Big) \leq \widehat{N}(\theta) \\ + 2\sqrt{\frac{\mathbb{E}\big(\langle \theta, X\rangle^4\big)}{n}} \left\{ 3.3 \left(\frac{T}{\mathbb{E}\big(\langle \theta, X\rangle^4\big)}\right)^{1/4} \right. \\ \left. + \sqrt{4\log\left(\frac{1}{2}\log\left(\frac{T}{\mathbb{E}\big(\langle \theta, X\rangle^4\big)}\right) + \frac{5}{2}\right) + 2\log(\delta^{-1})} \right\}.\end{aligned}$$

Nous examinons aussi plus généralement dans cet article le problème de l'estimation d'un vecteur ou d'une matrice aléatoire et proposons divers estimateurs robustes pour lesquels nous prouvons des bornes indépendantes de la dimension. Nous proposons en particulier des estimateurs robustes s'adaptant à la variance pour les vecteurs et à la variance directionnelle pour les matrices, comme dans le résultat ci-dessus. Cette adaptation à la variance est obtenue en introduisant une fonction d'influence asymétrique.

Dans [CG17a] nous proposons pour l'estimation de la moyenne d'un vecteur aléatoire $X \in \mathbb{R}^d$ un estimateur robuste particulièrement simple à calculer et possédant néanmoins des déviations sous-gaussiennes au premier ordre. Plus précisément, partant d'un échantillon $(X_1, \ldots, X_n)$ de n copies indépendantes de X, nous introduisons les variables tronquées

$$Y_i = \frac{\min\big\{\lambda\|X_i\|, 1\big\}}{\lambda\|X_i\|} X_i$$

et l'estimateur $\widehat{m} = \dfrac{1}{n}\sum_{i=1}^{n} Y_i$ de la moyenne $\mathbb{E}(X)$.

Supposant connue une borne

$$v \geq \sup_{\theta \in \mathbb{S}_d} \mathbb{E}\big(\langle \theta, X - \mathbb{E}(X)\rangle^2\big),$$

où $\mathbb{S}_d$ est la sphère unité de $\mathbb{R}^d$, nous choisissons $\lambda = 4\sqrt{\dfrac{2\log(\delta^{-1})}{1.2vn}}$ et montrons qu'avec probabilité au moins $1-\delta$,

$$\|\widehat{m} - \mathbb{E}(X)\| \leq \sqrt{\frac{2.4v\log(\delta^{-1})}{n}} + \sqrt{\frac{4\max\big\{\mathbb{E}\big(\|X-\mathbb{E}(X)\|^2\big), v\big\}}{n}} + \frac{C_p}{n^{p/2}},$$

où C_p est une quantité explicite, indépendante de d et de n, qui est finie dès que $\mathbb{E}\big(\|X\|^{p+1}\big)$ l'est. Il suffit donc que X ait un moment fini d'ordre supérieur à deux (on ne suppose pas p entier dans ce résultat) pour que notre estimateur robuste ait des déviations sous-Gaussiennes au premier ordre.

L'article [CG17b] se termine par une partie consacrée à la régression linéaire aux moindres carrés avec design aléatoire. Nous y examinons la possibilité d'obtenir des bornes indépendantes de la dimension d'ordre $\mathbf{O}\left(\sqrt{\dfrac{\log(\delta^{-1})}{n}}\right)$ (vitesse lente) où $\mathbf{O}\left(\dfrac{\log(\delta^{-1})}{n}\right)$ (vitesse rapide) sous des hypothèses faibles de moments polynomiaux concernant le design et le bruit. Nous pouvons obtenir des vitesses rapides soit pour une version modifiée du risque quadratique, soit en supposant que le paramètre de régression optimal a un support de dimension contrôlée (c'est-à-dire dans un scénario d'estimation dite « sparse »).

4.2. Nouveaux modèles statistiques inspirés de la linguistique. Les travaux de thèse de Thomas Mainguy [Mai14] que j'ai dirigés de septembre 2010 à décembre 2014, portaient sur l'estimation statistique de structures syntaxiques à partir d'un échantillon aléatoire de phrases tirées d'un corpus de textes écrits. Ils ont permis de faire émerger un nouveau modèle statistique, dont les applications potentielles dépassent le cadre de l'analyse automatique des langues naturelles. Ce modèle, que nous avons baptisé processus de substitution Markoviens, est présenté dans le deuxième chapitre de la thèse, le troisième chapitre étant consacré à explorer les liens qu'il entretient avec les grammaires sans contexte et à montrer comment des algorithmes de parsing peuvent être utilisés pour calculer explicitement la probabilité d'une phrase donnée. La description du modèle en lui-même,

cependant, n'est pas directement liée à la théorie des grammaires formelles, et repose uniquement sur une collection d'hypothèses d'indépendance conditionnelle. Le modèle, dépouillé de ses références à la linguistique et aux grammaires sans contexte, est décrit dans [CM16]. Il apparaît comme une généralisation (très substantielle néanmoins) de la notion de champ de Markov unidimensionnel, obtenue en élargissant le type d'hypothèses d'indépendance conditionnelle utilisé dans la définition d'un champ de Markov.

Etant donné un dictionnaire fini D et une mesure de probabilités P définie sur un domaine $\mathcal{D} \subset D^+ = \bigcup_{n=1}^{\infty} D^n$, l'ensemble des suites de mots du dictionnaire de longueur finie arbitraire, on définit un sous ensemble d'expressions $B \subset D^+$ comme formant un ensemble de substitution Markovien si et seulement si il existe une fonction exposant $\beta : B \times B \to \mathbb{R}$ telle que pour tous $x, z \in D^*$, et tous $(y, y') \in B$, tels que $\{xyz, xy'z\} \subset \mathcal{D}$,

$$P(xy'z) = P(xyz) \exp\big(\beta(y, y')\big).$$

Autrement dit, la probabilité de voir apparaître y' dans le contexte $x \cdot z$ est liée à la probabilité de voir apparaître y par une relation indépendante du contexte.

Etant donné une famille $\mathcal{B}$ de sous-ensembles de D^+ et un domaine $\mathcal{D} \subset D^+$, on introduit alors l'ensemble $\mathfrak{M}\big(\mathcal{D}, \mathcal{B}\big)$ des processus de substitution $\mathcal{B}$-Markoviens sur $\mathcal{D}$ comme l'ensemble des lois de probabilité sur $\mathcal{D}$ pour lesquelles tous les ensembles $B \in \mathcal{B}$ sont des ensembles de substitution Markoviens.

On peut alors décrire tous les supports possibles $\mathcal{C}$ des lois de $\mathfrak{M}(\mathcal{D}, \mathcal{B})$ et montrer que pour chacun de ces choix, le modèle

$$\mathfrak{M}_{\mathcal{C}}(\mathcal{D}, \mathcal{B}) = \big\{P \in \mathfrak{M}(\mathcal{D}, \mathcal{B}) \ : \ \mathrm{supp}(P) = \mathcal{C}\big\}$$

dans lequel on a fixé le support est une famille exponentielle (soit une famille de lois de Gibbs, si on préfère la terminologie de la mécanique statistique). L'énergie associée n'est pas explicite, mais cette propriété nous assure néanmoins que le modèle possède des propriétés statistiques intéressantes (en particulier l'estimateur du maximum de vraisemblance est un estimateur asymptotiquement efficace des paramètres).

On peut se demander pourquoi un modèle aussi simple à définir n'a pas émergé plus tôt dans la littérature. La raison est probablement à chercher dans le fait que la simulation d'un processus de substitution Markovien n'est pas directe. Elle nécessite l'utilisation d'une méthode de Monte-Carlo, c'est-à-dire d'une chaîne de Markov irréductible dont elle est l'unique loi invariante. Le calcul de l'estimateur du maximum de vraisemblance n'est pas non plus évident, mais peut aussi être entrepris par une technique de Monte-Carlo, comme montré dans la thèse. Bien qu'il soit en général impossible de spécifier explicitement l'énergie et les paramètres du

modèle, on peut néanmoins simuler d'après la loi du maximum de vraisemblance en utilisant la propriété suivant laquelle

$$P(xyz)P(x'y'z') = P(xy'z)P(x'yz'), \qquad y, y' \in B \in \mathcal{B}.$$

Cette propriété d'invariance de la vraisemblance par crossing-over permet d'appliquer une dynamique de crossing-over à un grand nombre de répliques de l'échantillon observé pour obtenir une approximation de la loi du maximum de vraisemblance.

De façon raccourcie et imagée, on peut dire que les modèles de substitution Markoviens sont les lois invariantes de dynamiques de crossing-over sur un échantillon de phrases.

Bien que nous n'ayons pas approfondi cet aspect des choses, cette propriété invite à explorer la pertinence des modèles de substitution Markoviens dans le domaine de la modélisation de séquences génétiques.

4.3. Clustering non supervisé. J'ai aussi durant la période visée par ce rapport réfléchi au clustering non supervisé, notamment à des variantes des algorithmes de clustering spectral, à l'occasion de la direction de la thèse d'Ilaria Giulini [Giu15a] et de la supervision du stage de recherche de l'Ecole Polytechnique de Xiayang Zhou (avril-août 2015). En particulier, dans le dernier chapitre de sa thèse, Ilaria Giulini prouve la convergence d'un algorithme de clustering spectral appliqué à un échantillon i.i.d. dans un espace de Hilbert vers un étiquetage des composantes connexes de la loi de l'échantillon, le nombre de classes étant estimé automatiquement par l'algorithme à partir d'une borne supérieure connue (mais qui peut en pratique être assez large).

A partir de 2015, j'ai entamé une réflexion concernant la transposition des principes sous-jacents aux modèles de substitution Markoviens dans le domaine de l'analyse de signaux. A travers cette transposition, j'aimerais dégager de nouveaux principes d'interprétation automatique de signaux fondés sur une analyse contextuelle apparentée au type d'analyse syntaxique réalisée par les modèles de substitution Markoviens sur la langue écrite. J'aimerais en effet explorer l'hypothèse selon laquelle la structure des langues naturelles pourrait refléter plus généralement le type de traitement effectué par le cerveau dans l'interprétation des données sensorielles qu'il reçoit. Plus précisément, il s'agit de rechercher par une analyse statistique non supervisée des structures remarquables dans des signaux.

Ainsi la thèse en cours de Gautier Appert est-elle consacrée à la mise en place et à la validation d'un nouveau modèle de classification non supervisée d'un échantillon aléatoire d'images numériques.

Notre approche consiste dans un premier temps à effectuer un découpage hiérarchique de chaque image en zones d'apparition fréquente et de forme arbitraire.

Ce découpage est dichotomique, chaque étape consistant à scinder une zone en deux. Ainsi, chaque zone est décrite non seulement par son label, mais aussi par le couple de labels des deux sous-zones qui la constituent. Ces couples de labels forment des réalisations d'un couple de variables aléatoires (X, Y). On peut en approcher la loi par un mélange de lois produits. Pour une telle loi de mélange, on impose plus précisément que les deux variables X et Y soient indépendantes conditionnellement à un label $f(X, Y)$ qui soit une fonction (déterministe) du couple (X, Y). Il s'agit donc d'une propriété d'indépendance conditionnelle. Elle conduit à remplacer le couple de labels (X, Y) par un label $f(X, Y)$ plus simple mais qui rend néanmoins compte de l'interaction entre X et Y. On peut en effet reconstituer la loi du couple (X, Y) à partir de la loi de $f(X, Y)$ et des lois conditionnelles de X et de Y sachant $f(X, Y)$. Le modèle utilisé ici pour le couple (X, Y) s'apparente aux modèles de substitution Markoviens que nous avions proposés pour l'analyse syntaxique des langues naturelles avec Thomas Mainguy lors de sa thèse. On peut en effet ici substituer à X un label X' tiré suivant la loi conditionnelle de X sachant $f(X, Y)$ sans changer la loi du couple (X, Y). Contrairement à ce qui avait été envisagé pour les modèles de substitution Markoviens, le label $f(X, Y)$ peut dépendre ici du contexte Y du label X et non pas uniquement de X.

Dans la thèse de Thomas Mainguy, il avait été proposé d'effectuer des tests multiples pour valider les hypothèses d'indépendance conditionnelle décrites ci-dessus. Cette approche ne se justifie que dans le cas où la loi des données appartient au modèle. Or nous souhaitons pouvoir classer des échantillons d'images qui n'ont pas forcément de structure syntaxique, c'est-à-dire dont la loi n'appartient pas forcément au modèle. Une façon de se tirer d'affaire consiste à estimer le modèle, présumé faux, en minimisant une fonction de perte. Cette fonction de perte mesure l'écart entre la loi des données et la loi du modèle estimé, qu'il soit petit ou grand. On se livre donc à une sorte de projection de la loi des données sur le modèle. Comment choisir la fonction de perte ? Nous avons été conduits après réflexion à proposer l'entropie relative de la loi du modèle par rapport à la loi des données. Le choix le plus étudié consiste à utiliser l'entropie relative dans l'autre sens. En effet l'estimateur du maximum de vraisemblance, très étudié et très utilisé, est relié à la minimisation de l'entropie relative de la loi des données par rapport à la loi du modèle et non l'inverse. L'entropie relative est une fonction de perte très asymétrique. D'après le théorème de grandes déviations de Sanov, l'entropie relative $\mathcal{K}(Q, P)$ d'une loi Q par rapport à une loi P mesure la probabilité d'observer un échantillon statistique i.i.d de loi P de mesure empirique proche de Q. Pour le dire plus simplement, si $\mathcal{K}(Q, P)$ est faible, Q est vraisemblable sous P, et en particulier son support est inclus dans celui de P. Si P est la loi des données et Q une loi du modèle, le modèle est donc vraisemblable. Cependant, d'autres configurations des données, non prédites par le modèle, peuvent aussi se présenter. Le modèle ne prédit donc qu'un partie des données. La situation classique en

statistique est plutôt la situation inverse, on cherche un modèle sous lequel les observations sont toujours vraisemblables, quitte à inclure dans le modèle des configurations supplémentaires. Autrement dit on cherche un modèle qui prédit toutes les configurations possibles des données, quitte à prédire par dessus le marché des configurations jamais observées.

Dans notre programme d'analyse syntaxique, nous ne souhaitons introduire dans le modèle que des structures syntaxiques observables dans les données, quitte à en omettre certaines, d'où notre choix de fonction de perte. Ce choix repose aussi sur des propriétés techniques de l'entropie. Pour traiter des images numériques et non plus des suites de mots comme dans l'analyse des langues, nous introduisons un modèle de bruit Gaussien sur le niveau de gris des pixels. Ce bruit est conservé par projection en utilisant l'entropie dans le sens où nous l'utilisons, alors qu'il se transformerait en un mélange de Gaussiennes si nous utilisions l'entropie dans l'autre sens. Plus généralement, si le bruit appartient à une famille exponentielle, le bruit projeté appartient aussi à la même famille.

Nous avons dégagé un formalisme dans lequel le même critère d'entropie peut être utilisé pour effectuer les deux étapes de la classification : le découpage en zones fréquemment observées et l'analyse syntaxique des zones. Pour obtenir cette approche unifiée, nous représentons les images comme des distributions conditionnelles de pixels, de manière à tout traduire en termes de mesures de probabilités. Ainsi nous pouvons fonder l'ensemble des étapes de la classification non supervisée sur un même principe.

Nous étudions en ce moment les algorithmes de minimisation dichotomique du critère évoqués au début de ce résumé. Les propriétés de décomposition de l'entropie relative permettent d'optimiser explicitement le critère par rapport à une partie des paramètres du modèle. Elles permettent aussi un calcul factorisé de l'évolution du critère à chaque étape. Comme dans le cas de l'algorithme de Lloyd pour les k-means, on ne peut espérer trouver qu'un minimum local.

Enfin une mise en œuvre est en cours. Elle est constituée d'une interface codée en R et de routines écrites en C++, intégrées au code R à l'aide du package Rcpp. Le code C++ inclut des directives OpenMP (Open Multi-Processing API) permettant une exécution en parallèle sur une architecture multi-cœurs.

5. Résumé des travaux antérieurs à 2014

5.1. Le début de ma carrière. Je ne décrirai pas en détail mon activité scientifique durant la période où j'étais chargé de recherche, pour laquelle je renvoie aux pages 8 à 23 de mon dossier de candidature à un poste de directeur de recherche d'avril 2000 (qui doit se trouver dans les archives du CNRS, mais peut aussi être téléchargé depuis ma page web).

Disons simplement pour résumer que j'ai commencé par étudier des algorithmes d'optimisation stochastique du type recuit simulé généralisé, et que j'ai fait aussi pendant cette période quelques incursions du côté de l'analyse d'images (débruitage, détection de contours, algorithmes de poursuite pendant mon service militaire) et du côté de l'étude des verres de spin.

D'un point de vue technique, je me suis essentiellement intéressé aux grandes déviations des trajectoires des systèmes métastables, homogènes (algorithmes de Metropolis généralisés) ou inhomogènes (recuit simulé généralisé) en temps, ainsi qu'à la transition de phase du modèle de verre de spin de Sherrington Kirkpatrick.

5.2. Théorie statistique de l'apprentissage.

5.2.1. Références. Cette présentation mettra l'accent sur mes contributions à la théorie statistique de l'apprentissage (qui ont débuté avant ma promotion DR, mais qui ont essentiellement été publiées depuis). Elles se sont incarnées dans les travaux personnels suivants,

- un cours à l'école d'été de probabilités de Saint-Flour, [Cat04b], publié chez Springer en 2004 (269 pages) ;
- des Lecture Notes, [Cat07], parues dans les « Lecture Notes - Monograph Series » de l'Institute for Mathematical Statistics (175 pages),
- un article [AC11a] en collaboration avec Jean-Yves Audibert (48 pages), proposant un estimateur de régression aux moindres carrés solution d'un problème min-max, publié aux Annals of Statistics, accompagné d'un supplément en ligne comme c'est maintenant l'usage pour ce journal ;
- un article [AC10], en collaboration avec Jean-Yves Audibert, portant sur l'estimation d'une régression pour une fonction de perte fortement convexe par un estimateur de Gibbs effectuant une troncature des valeurs extrêmes. Il s'agit d'obtenir des résultats sous des hypothèses plus faibles, (dites de marge généralisée) en utilisant un estimateur plus complexe à mettre en œuvre, mais intéressant sur le plan théorique pour la généralité de ses performances. Par opposition, l'article précédent propose une méthode avec un coût algorithmique raisonnable (50 fois le temps de calcul des moindres carrés ordinaires dans les simulations que nous avons effectuées), valide sous des hypothèses dépendant d'un plus grand nombre de paramètres caractérisant la loi jointe du couple (X,Y) sur lequel porte la régression.
  Cet article était en révision favorable chez Bernoulli, dans une version courte de 24 pages, mais fut rejeté au dernier moment à la suite d'une polémique concernant sa prétendue proximité avec [AC11a]. Polémique à mon sens inutile, les deux articles traitant de deux estimateurs complètement différents. L'estimateur décrit dans l'article refusé par Bernoulli n'en reste pas moins l'estimateur de régression aux moindres carrés qui converge à

vitesse optimale sous les hypothèses les plus faibles connues dans la littérature. Je suis certain que l'engouement actuel pour les méthodes à poids exponentiels permettra à cette étude de trouver un jour son public et n'ai pas renoncé à la publier un jour ailleurs que sur arXiv, intégrée à d'autres travaux.

— *Challenging the empirical mean and empirical variance : a deviation study*. Troisième version remaniée une dernière fois en 2011, d'un article proposant de nouveaux M-estimateurs de la moyenne et de la variance, présentant de meilleures déviations que la moyenne et la variance empiriques [Cat12] (48 pages).

— les notes de cours de ma contribution au cours d'apprentissage du cursus math-info de l'ENS (destiné aux élèves de première année). Ce cours a eu lieu trois années de suite, donnant lieu à trois séries de notes de cours (2011, 2012 et 2013) [Cat13]. Le chapitre de ces notes portant sur les bornes PAC-Bayésiennes pour la classification contient des résultats originaux depuis publiés dans un Festschrift en l'honneur d'A. Chervonenkis [Cat15b]. Ces résultats rapprochent mon point de vue de celui de M. Seeger et fournissent de nouvelles inégalités de marge pour les Support Vector Machines.

— *Toric grammars : a new statistical approach to natural language modeling* un article en collaboration avec mon étudiant, Thomas Mainguy, [CM13], présentant le modèle d'analyse statistique des langues naturelles étudié dans sa thèse. Il s'agit d'une première étude de linguistique computationnelle, présentant un modèle probabiliste de communication du langage d'un locuteur à un autre.

  La suite de la thèse de Thomas Mainguy [Mai14] introduit un modèle statistique sous-jacent, les processus de substitution Markoviens, et étudie les liens entre ce modèle, les grammaires toriques du début de la thèse et les grammaires sans contexte.

La monographie sur la classification PAC-Bayésienne [Cat07] est au départ issue d'un article soumis aux Annals of Statistics (sous le titre *A PAC-Bayesian approach to adaptive classification*) : j'étais réticent pour effectuer la réduction à 40 pages qui était requise, j'ai alors demandé à l'éditeur d'Annals of Statistics s'il ne valait pas mieux envisager une publication dans les Lecture Notes de l'IMS (qui édite aussi Annals of Statistics), suggestion qui a reçu son soutien. J'ai du coup inclus dans cette monographie des résultats supplémentaires. J'ai donc à une certaine période enchaîné la rédaction de deux monographies, au lieu de présenter mes résultats sous forme d'articles.

J'ai écrit d'autre part une présentation [Cat06] de la théorie statistique de l'apprentissage en quelques pages à destination de la brochure du CNRS « Images des mathématiques ».

Pendant toute cette période consacrée à l'élaboration d'une approche de l'ap-

prentissage statistique en lien avec la théorie de l'information et la mécanique statistique, j'ai dirigé trois thèses, dont les auteurs, Jean-Philippe Vert, Jean-Yves Audibert et Pierre Alquier, ont prolongé dans une partie de leurs travaux (notamment ceux issus de leur thèse) l'exploration des idées et des méthodes abordées dans mes propres publications et dans les discussions que nous avions ensemble.

5.2.2. Résumé des résultats publiés. J'ai abordé dans un premier temps la théorie statistique de l'apprentissage par le biais de la théorie de la compression sans perte. Elle possède une traduction statistique en terme de minimisation du risque cumulé. Mes premiers travaux ont consisté, en utilisant une méthode de télescopage dont la primeur revient à Andrew Barron, à adapter la théorie du codage à la minimisation du risque d'estimation non cumulé. Cette approche est décrite dans les premiers chapitres de mon cours à Saint-Flour. J'ai en particulier pu, en utilisant des idées proches de celles avec lesquelles je m'étais familiarisé en étudiant les verres de spin, réaliser une étude en moyenne des estimateurs de Gibbs, qui met l'accent sur l'estimation de la densité, mais dont on peut déduire aussi des résultats concernant la classification ou l'estimation d'une régression. (Un logiciel d'estimation d'une densité par des histogrammes adaptatifs est disponible sur ma page web).

Dans un deuxième temps, je me suis assez naturellement posé la question de l'étude des déviations du risque des estimateurs de Gibbs. Pendant la même période, j'ai pris connaissance des premiers travaux de David McAllester, qui m'ont paru ouvrir une voie particulièrement féconde. J'ai gardé la dénomination de « théorie PAC-Bayésienne » utilisée par McAllester, bien que d'autres auteurs aient par la suite développé des idées proches sous d'autres noms (je pense en particulier aux travaux très pertinents de Tong Zhang). D'autre part, à l'étranger, la dénomination « bornes PAC-Bayésiennes » a aussi prospéré en partant dans une direction un peu différente de la mienne, issue des bornes établies par Matthias Seeger, et restreinte au problème de la classification. Dans mes notes de cours [Cat13], issues en partie de discussions avec D. McAllester, puis dans [Cat15b], j'ai pu faire le lien entre les deux approches, en montrant que les bornes de Seeger, sous une forme légèrement améliorée, pouvait se déduire des miennes par l'optimisation d'un paramètre laissé libre.

Avec un peu de recul, il me semble que l'on peut qualifier la théorie PAC-Bayésienne de l'apprentissage sur le plan technique de la façon suivante :

L'objet de la théorie statistique de l'apprentissage est de réaliser des tâches de prédiction ou d'inférence sur des données « complexes ». Par données complexes, il faut entendre des données qui ne peuvent être décrites avec une exactitude raisonnable par un modèle paramétrique possédant un nombre de paramètres faible devant le nombre d'observations dont on dispose. Il est alors nécessaire d'utiliser des modèles approchés qui tiennent compte à la fois de l'information disponible et de la tâche à effectuer. La sélection de ce genre de modèles nécessite l'utili-

sation d'inégalités de déviation non asymptotiques portant sur des minimums de processus empiriques. L'approche PAC-Bayésienne apparaît dans ce cadre comme une sorte de pendant non asymptotique de l'approche des grandes déviations qui conduit au théorème de Gartner-Ellis : elle consiste à combiner trois ingrédients, à savoir le contrôle des déviations du processus empirique par sa *transformée de Laplace*, des techniques de *dualité* et d'*analyse convexe*. Cette économie de moyens permet une approche unifiée, fait espérer de meilleures constantes et conduit au développement d'une « technique de calcul entropique » dont la souplesse et la polyvalence, que j'ai essayé de montrer dans [Cat07], s'est depuis confirmée au fil des résultats obtenus par une communauté de contributeurs qui s'élargit au fil des ans — avec notamment en France des contributions d'A. Tsybakov et A. Dalalyan à la régression sparse fondées sur des inégalités d'entropie. (L'entropie intervient en tant que transformée de Legendre du logarithme de la transformée de Laplace.)

Prenons des notations pour préciser un peu les choses. Étant donné un échantillon $\big[Z_1(\omega), \dots, Z_N(\omega)\big]$ de variables indépendantes (ou échangeables, ou partiellement échangeables, plusieurs types d'hypothèses sont possibles), et une famille $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}_+$ de fonctions de perte, on considère le processus empirique

$$r(\theta, \omega) = \frac{1}{N} \sum_{i=1}^{N} \ell_\theta(Z_i).$$

Dans ma monographie [Cat07], je me suis concentré sur le cas de la classification, c'est-à-dire celui où $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ est un couple décrivant une « forme » et sa classification (en un nombre fini de classes) et où la fonction de perte $\ell_\theta(Z_i) = \mathbb{1}\big[Y_i \neq f_\theta(X_i)\big]$ décrit l'erreur de classification commise par une règle de classification $f_\theta : \mathcal{X} \to \mathcal{Y}$. J'ai alors confié à mon thésard Pierre Alquier le soin d'étudier le cas d'une fonction de perte ℓ bornée générale. Je me suis ensuite penché en collaboration avec Jean-Yves Audibert [AC10 ; AC11a], sur le cas d'une perte ℓ non bornée, vérifiant uniquement des hypothèses de moments polynomiaux. J'ai commencé par approfondir le cas de la classification parce qu'il présente des particularités et qu'il s'agit d'une question fondamentale de la théorie de l'apprentissage (le passage d'une représentation continue des objets à un étiquetage discret, passage « de la perception à une représentation symbolique », en quelque sorte).

On souhaite minimiser en θ l'espérance du risque empirique $\mathbb{E}\big[r(\theta)\big]$. Ceci nécessite de contrôler avec un certain degré d'uniformité les fluctuations de $r(\theta, \omega)$ avec l'aléa ω, causé par le caractère aléatoire de l'échantillon $(Z_i)_{i=1}^N$. L'approche PAC-Bayésienne fonde cette étude sur celle de la quantité

$$-\log \mathbb{E}\bigg[\int_\Theta \exp\big[-\lambda r(\theta, \omega)\big] \pi(d\theta)\bigg],$$

où $\pi \in \mathcal{M}_+^1(\Theta)$ est une mesure a priori sur l'espace des paramètres. En plus de l'analogie avec les grandes déviations à la Gartner-Ellis, l'approche PAC-Bayésienne s'inspire aussi de la mécanique statistique, les physiciens ayant l'habitude de dériver les propriétés macroscopiques d'un système microscopique en étudiant son énergie libre. Dans cette analogie, le processus empirique r jouerait le rôle de l'énergie microscopique d'un système de particules dont l'aléa ω décrirait les fluctuations microscopiques alors que θ décrirait les fluctuations du milieu, un peu comme c'est le cas dans l'étude des systèmes désordonnés. (L'élaboration de ce point de vue doit bien évidemment beaucoup à l'influence de mes travaux antérieurs sur les verres de spin et les grandes déviations des trajectoires des systèmes métastables.)

L'idée suivante consiste à comparer dans l'énergie libre l'utilisation de la probabilité de référence π avec des probabilités « a posteriori », en mettant le processus empirique $\theta \mapsto r(\theta, \omega)$ en dualité avec les mesures de probabilité aléatoires (dites « a posteriori ») sur l'espace des paramètres $\rho(\omega) \in \mathcal{M}_+^1(\Theta)$, via la transformée de Legendre

$$-\log\left[\int_\Theta \exp\bigl[-\lambda r(\theta)\bigr]\pi(d\theta)\right] = \inf_{\rho\in\mathcal{M}_+^1(\Theta)} \lambda \int_\Theta r(\theta)\rho(d\theta) + \mathcal{K}(\rho,\pi),$$

où $\mathcal{K}(\rho, \pi)$ désigne la divergence de Kullback (encore appelée entropie relative). Cette mise en dualité donne de la souplesse, permet de mesurer et de contrôler la complexité des modèles de façon générique, et conduit naturellement à des mesures empiriques de la complexité. Elle permet aussi des interprétations en terme de théorie du codage qui font le lien avec le principe de la description de longueur minimal (Mimimum Description Length principle) mis en avant dans les travaux de Rissanen. Techniquement elle permet de « convexifier » Θ en le remplaçant par $\mathcal{M}_+^1(\Theta)$ à la faveur de la dualité. Ces quelques équations suffisent à dresser le décor : log-Laplace, dualité et convexification, voici les trois piliers sur lesquels repose l'approche PAC-Bayésienne (de même que le versant Gartner-Ellis de la théorie des grandes déviations dans l'univers des théorèmes limites).

Partant de là, la théorie se développe dans plusieurs directions : la localisation des bornes, la localisation en deux étapes qui est l'analogue des « double mixture codes » en théorie du codage adaptatif, l'usage de bornes relatives (portant sur la différence entre les risques de deux estimateurs) et l'extension des bornes au cadre de l'inférence transductive. Les développements les plus récents (intervenus ces dernières années) concernent la localisation en deux étapes, les bornes relatives (sections 1.3.5, 1.3.6 et 2 de [Cat07]), une approche du cas transductif qui permet de transférer dans ce cadre les résultats concernant l'inférence inductive à partir d'inégalités de départ analogues, l'étude de la régression aux moindres carrés non bornée sous des hypothèses faibles de moments polynomiaux (et plus

généralement l'étude de fonctions de perte non bornées vérifiant une inégalité de marge généralisée), l'étude des déviations de l'estimation de la moyenne et de la variance d'une variable aléatoire dans le cas où l'on ne contrôle respectivement que la variance ou la kurtosis (ce qui autorise des distributions à queue lourde) ainsi que l'estimation de la matrice de Gram d'un design aléatoire. Dans une première période, j'ai utilisé la perturbation optimale du paramètre dans les inégalités PAC-Bayésiennes, perturbation correspondant à l'utilisation d'estimateurs définis par des lois de Gibbs (pondérations à poids exponentiels). Dans une seconde période, j'ai privilégié l'utilisation de perturbations explicites (le plus souvent gaussiennes), qui permettent moyennant quelques ingrédients techniques supplémentaires, d'étudier des estimateurs plus classiques (résultant de la minimisation d'un risque empirique).

En ce qui concerne les bornes relatives, j'ai introduit l'idée de comparer le risque d'un estimateur à celui d'une loi de Gibbs a priori (c'est-à-dire fondée sur l'espérance du risque $\mathbb{E}\big[r(\theta)\big]$ et non sur le risque empirique lui-même). Il se trouve que cela est techniquement possible, et conduit à définir la « température effective » d'un estimateur, comme celle de la loi de Gibbs a priori de même risque. Cette température peut être estimée entièrement à partir des données empiriques et conduit à un critère de choix d'estimateur qui atteint de manière adaptative la bonne vitesse sous des hypothèses de marge à la Tsybakov et des hypothèses de complexité paramétriques. On peut aussi dans ce cadre réaliser une localisation en deux étapes et comparer les estimateurs à des lois de Gibbs a priori « à deux étages », procédé plus satisfaisant pour effectuer une sélection de modèle parmi une famille de modèles paramétriques. Il est à noter que les hypothèses de marge n'interviennent pas dans la construction de l'estimateur, qui tient compte directement de la structure des covariances au voisinage du risque minimum via un terme de covariance empirique.

Par la suite, partant des résultats de la thèse de Jean-Yves Audibert, j'ai proposé une manière alternative d'exploiter des bornes relatives, qui repose sur un nouvel algorithme de portée générale permettant d'exploiter de manière presque optimale n'importe quelle famille d'intervalles de confiance portant sur les différences des risques. Dans cette nouvelle approche, la comparaison avec une loi de Gibbs a priori n'intervient que dans le contrôle des divergences de Kullback. La procédure d'estimation est plus compliquée que la comparaison directe avec le risque d'une loi de Gibbs a priori évoquée précédemment, il n'est donc pas certain que son efficacité pratique soit meilleure ; par contre elle permet d'atteindre la vitesse d'ordre optimale sous des hypothèses de marge à la Tsybakov et des hypothèses de complexité paramétriques un peu plus faibles, et donc plus satisfaisantes, et il est plus facile de donner un résultat avec des constantes complètement explicites. De plus on peut la mettre en œuvre dans le cadre d'une procédure de localisation partielle qui conduit à de nouveaux résultats théoriques concernant la sélection

de modèles sous hypothèses de marge et de complexité (section 2.2 de [Cat07] aboutissant au théorème 2.2.11 page 110). Enfin cet algorithme reposant sur des comparaisons entre risques empiriques permet de raffiner les techniques de double localisation proposées précédemment pour aboutir à une méthode de sélection doublement localisée qui soit plus pertinente, en particulier dans le cas de modèles emboîtés, la localisation du choix du modèle faisant intervenir non seulement le risque mais aussi un terme de variance dépendant des propriétés de marge des différents modèles. En effet, le passage à un modèle plus grand, et ceci reste vrai dans le cas de modèles emboîtés, peut détériorer la performance non seulement par une augmentation inappropriée de la taille du modèle, mais aussi par la détérioration de ses propriétés de marge, liées aux variances des différences de risque au voisinage du risque optimal : la localisation du choix du modèle par rapport à ce dernier critère est une amélioration apportée par le théorème 2.3.9 [Cat07, page 131] par rapport aux méthodes de double localisation que j'avais proposées précédemment.

Je me suis aussi attaché à montrer l'intérêt général de la théorie PAC-Bayésienne, en soulignant qu'elle s'appliquait, via une opération de randomisation mineure, à l'étude de n'importe quel estimateur. Dans sa version localisée, l'erreur d'estimation, liée à la complexité du modèle utilisé est contrôlée par une approximation observable de *l'information mutuelle* entre l'échantillon et le paramètre estimé, comme expliqué au paragraphe 1.3.1. de [Cat07]. Ceci met en lumière le fait que l'information mutuelle entre l'échantillon observé et le paramètre estimé joue dans la théorie PAC-Bayésienne un rôle analogue à celui joué par l'information de Fisher dans la théorie de Cramer-Rao de l'estimation sans biais en risque quadratique.

D'un point de vue technique, j'ai pu éliminer le recours à une approximation de la log-Laplace (du type Bernstein, que j'utilisais avant) et obtenir des contrôles empiriques de la variance directement en effectuant un changement de variable dans l'équation de déviation. Moyennant ce changement de variable, on peut faire des allers et retours entre variance théorique et variance empirique, qui permettent de prouver l'adaptativité du choix d'un estimateur par estimation de sa température effective (c'est le sujet de la section 2.1.4 de [Cat07]). Cette idée de changement de variable dans la log-Laplace se décline aussi de façon naturelle dans le cadre plus général de la régression avec fonction de perte bornée, comme cela apparaît dans la thèse de Pierre Alquier. Elle permet d'éviter de se retrouver avec un terme de variance à réestimer par une quantité empirique et simplifie de ce fait notablement les écritures tout en améliorant les résultats obtenus. Le cas d'une fonction de perte non bornée est évoqué dans la thèse de Pierre Alquier dans le cas où des hypothèses de moment exponentiel sont connues. L'adaptation à des hypothèses de moments polynomiaux plus faibles est traitée dans ma collaboration avec Jean-Yves Audibert [AC11a ; AC10].

Dans les prépublications que j'ai d'abord fait circuler, le cas transductif était

traité à part, et il fallait refaire toutes les démonstrations dans ce cas, malgré certaines analogies avec le cas inductif. Les améliorations apportées dans le traitement du cas inductif (l'abandon en particulier de l'approximation de la log-Laplace), ont eu comme avantage collatéral de permettre un traitement unifié avec le cas transductif. Par « cas transductif » il faut comprendre ici le cas où on utilise un échantillon fantôme (ou échantillon test), qu'il soit réellement observé ou qu'il intervienne simplement dans les calculs. Ce « cas transductif », dont l'importance a été pointée en premier par V. Vapnik, correspond à certaines situations réalistes du point de vue expérimental, et conduit d'autre part sur le plan théorique à la théorie de la complexité de V. Vapnik, entropie de Vapnik et dimension de Vapnik Chervonenkis.

Développons un peu la présentation de cette question : nous sommes ici dans le contexte de la classification supervisée, où on observe des couples (X_i, Y_i) de formes et d'étiquettes. Dans le cas transductif, on considère la réunion de deux échantillons, un échantillon d'apprentissage $(X_i, Y_i)_{i=1}^{N}$, supposé complètement observé, et un échantillon test $(X_i, Y_i)_{i=N+1}^{N+M}$, dont on observe au plus les formes $(X_i)_{i=N+1}^{M}$, et qui peut parfois simplement servir d'intermédiaire de calcul, les résultats étant réintégrés par rapport à l'échantillon test.

Alors que la théorie de Vapnik utilise le plus souvent un échantillon fantôme de même taille que l'échantillon d'apprentissage, nous avons mis en avant l'intérêt qu'il y avait à considérer des tailles M plus grandes que N. Ceci permet en effet de moduler la variance de l'échantillon test, et d'obtenir de cette façon de meilleures constantes.

Bien que le cas transductif, et en particulier les bornes de généralisation de Vapnik, mérite des développements spécifiques, il est possible de transposer au cas transductif tous les résultats du cas inductif obtenus dans les premiers chapitres de [Cat07], comme cela est fait dans la section 3.1. Ceci permet de mettre les bornes de Vapnik en perspective, en montrant qu'elles forment le cas non local et non relatif d'une famille de bornes de généralisation plus précises qu'il est possible de calculer quand on observe des formes test non étiquetées en plus de l'échantillon d'apprentissage et que les moyens de calcul le permettent.

Il n'existe pas de véritable hiérarchie entre toutes ces bornes, les améliorations apportées par les bornes les plus sophistiquées ne se faisant sentir que lorsque la taille de l'échantillon est suffisamment grande. L'inférence à partir d'échantillons de petite taille étant un enjeu central, il n'est pas inutile de consacrer des efforts aux bornes les plus simples, qui sont dans ce cas les meilleures (par simples, il faut entendre ici non locales et non relatives). C'est ce qui est fait dans les sections 3.2, 3.3 et 3.4 de [Cat07], (de conception antérieure à ce qui a été évoqué dans les paragraphes précédents,) qui montrent comment améliorer les bornes de Vapnik en combinant un certain nombre d'idées : ne pas faire d'approximation gaussienne

de la transformée de Laplace, utiliser un échantillon fantôme de taille optimisée, ne pas avoir recours à une technique de symétrisation, mais plutôt à une technique de réintégration par rapport à l'échantillon fantôme. Un petit exemple numérique montre que le gain est significatif pour des tailles de problèmes réalistes.

Dans [AC11a ; AC10], j'aborde avec Jean-Yves Audibert le cas des fonctions de perte non bornées et plus spécifiquement celui de la régression aux moindres carrés.

Nous avons commencé par nous affranchir de toute hypothèse de moment exponentiel sur la fonction de risque dans [AC10]. Nous avons pour cela utilisé un estimateur de Gibbs particulier, qui réalise une troncature de la différence des risques au voisinage du paramètre optimal. Le théorème 3.5 [AC10, page 21] montre que sous une simple hypothèse de marge, généralisation naturelle de celle introduite par A. Tsybakov dans le cas de la classification, et ne faisant intervenir que les moments d'ordre deux des différences des risques, cet estimateur atteint une vitesse en d/n où la dimension d est mesurée par le comportement de la transformée de Laplace du risque par rapport à une mesure a priori sur l'espace des paramètres.

Nous nous sommes ensuite rendu compte que l'approche PAC-Bayésienne pouvait être considérablement simplifiée dans le cas quadratique, ce qui nous a conduit à un nouvel estimateur, plus facile à calculer [AC11a].

On peut en effet, dans le cas du risque quadratique, mener explicitement à bien une approche perturbative dans laquelle on compare un estimateur se présentant sous la forme classique d'une fonction de l'échantillon observé avec un estimateur perturbé, dont la loi a posteriori est une loi gaussienne centrée sur l'estimateur classique. L'utilisation de perturbations gaussiennes avait aussi été introduite avec succès par Langford et Shawe-Taylor dans le calcul d'inégalités de marge pour les Support Vector Machines, je renvoie à [Cat13] pour plus de détails à ce sujet.

Décrivons plus en détail les idées de [AC10 ; AC11a]. Pour pouvoir travailler sous des hypothèses de moments polynomiaux, nous avons utilisé l'idée du changement de variable dans la transformée de Laplace. Dans ce cadre, il est avantageux d'utiliser la transformation obtenue en tronquant le développement de Taylor de la fonction exponentielle à l'ordre deux, une idée simple qui n'avait pourtant pas été exploitée auparavant.

Considérons un échantillon i.i.d. (X_i, Y_i) où $X_i \in \mathbb{R}^d$ et $Y_i \in \mathbb{R}$, et notons la différence des risques quadratiques

$$W_i(\theta, \theta') = \big(\langle \theta, X_i \rangle - Y_i\big)^2 - \big(\langle \theta', X_i \rangle - Y_i\big)^2.$$

Soit (X, Y) un couple de variables de même loi et indépendant le l'échantillon, Θ un convexe fermé de $\mathbb{R}^d$, λ un paramètre réel positif ou nul et

$$\theta^* \in \arg\min_{\theta \in \Theta} \mathbb{E}\big[\big(\langle \theta, X \rangle - Y\big)^2\big] + \lambda \|\theta\|^2$$

le minimiseur du risque de la régression ridge. Soit ρ_θ la loi gaussienne de variance βI centrée en θ. En mettant ensemble les deux idées précédentes, on obtient l'inégalité suivante : *Pour tous paramètres $\lambda \in \mathbb{R}_+$, $\alpha \in \mathbb{R}$, avec probabilité au moins $1-\epsilon$, pour tout $\theta_1 \in \Theta$,*

$$-n\log\left\{\int \rho_{\theta_1}(d\theta)\left(1-\alpha\mathbb{E}\big[W_i(\theta,\theta^*)\big]+\frac{\alpha^2}{2}\mathbb{E}\big[W_i(\theta,\theta^*)^2\big]\right)\right\} \leq \sum_{i=1}^n \log\left\{\int \rho_{\theta_1}(d\theta)\left(1+\alpha W_i(\theta,\theta^*)+\frac{\alpha^2}{2}W_i(\theta,\theta^*)^2\big]\right)\right\} + \mathcal{K}(\rho_{\theta_1},\rho_{\theta^*})-\log(\epsilon),$$

où $\mathcal{K}$ désigne la divergence de Kullback.

On peut alors utiliser le fait que $\log(1+x) \leq x$, se placer dans une base où la matrice de Gram est diagonale et tout calculer dans l'inégalité précédente pour obtenir un théorème non asymptotique complètement explicite concernant la convergence des estimateurs ridge de la forme

$$\widehat{\theta} \in \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \big(\langle\theta, X_i\rangle - Y_i\big)^2 + \lambda\|\theta\|^2.$$

Il en découle entre autres le théorème suivant, où on a noté

$$R_\lambda(\theta) = \mathbb{E}\big[\big(\langle\theta,X\rangle - Y\big)^2\big] + \lambda\|\theta\|^2$$

le risque associé à la régression ridge et où le paramètre λ peut être positif ou nul (ce qui couvre le cas des moindres carrés usuels quand de plus $\Theta = \mathbb{R}^d$).

Théorème 5.1 *Supposons que*

$$\mathbb{E}\big(\|X\|^4\big) < +\infty,$$
$$\textit{et} \quad \mathbb{E}\Big[\|X\|^2\big(\langle\theta,X\rangle - Y\big)^2\Big] < +\infty.$$

Soit $\nu_1,\dots,\nu_d$ les valeurs propres de la matrice de Gram $Q = \mathbb{E}\big(XX^\top\big)$ et D la dimension tronquée

$$D = \sum_{k=1}^d \frac{\nu_k}{\nu_k+\lambda}\mathbb{1}(\nu_k > 0) \leq d.$$

Pour tout $\epsilon > 0$, il existe une taille d'échantillon n_ϵ telle que pour tout $n \geq n_\epsilon$, avec probabilité au moins $1-\epsilon$,

$$R_\lambda(\widehat{\theta}) \le R_\lambda(\theta^*) + 30\widetilde{\mathbb{E}}\big[\big(\langle\theta^*, X\rangle - Y\big)^2\big]\frac{D}{n} + 1000 \sup_{v\in\mathbb{R}^d} \widetilde{\widetilde{\mathbb{E}}}_v\big[\big(\langle\theta^*, X\rangle - Y\big)^2\big]\frac{\log(3/\epsilon)}{n},$$

où, pour toute variable aléatoire Z,

$$\widetilde{\mathbb{E}}(Z) = \frac{\mathbb{E}(\|(Q+\lambda I)^{-1}X\|^2 Z)}{\mathbb{E}(\|(Q+\lambda I)^{-1}X\|^2)},$$
$$\widetilde{\widetilde{\mathbb{E}}}_v(Z) = \frac{\mathbb{E}(\langle v, X\rangle^2 Z)}{\mathbb{E}(\langle v, X\rangle^2)}.$$

Ce résultat montre en particulier que l'estimateur des moindres carrés, asymptotiquement, atteint toujours une vitesse en d/n (sans $\log(n)$ additionnel), sous des hypothèses de moments très faibles. Il n'y a en particulier besoin d'aucune autre hypothèse sur la forme de la loi du couple (X, Y), le théorème couvrant le cas d'un design aléatoire non borné et d'une sortie Y dont tous les moments exponentiels sont infinis. Il n'y a pas non plus besoin d'hypothèse sur le conditionnement de la matrice de Gram.

Par contre, sous des hypothèses aussi faibles, la taille d'échantillon à partir de laquelle ce régime asymptotique est atteint peut être arbitrairement grande (ce qui est inévitable, même dans le cas monodimensionnel de l'estimation de la moyenne de Y).

On peut aller plus loin dans la recherche d'un estimateur dont l'excès de risque soit sous-exponentiel, même dans le cas où le bruit n'a pas de moments exponentiels, en modifiant l'estimateur. Dans [AC11a], nous introduisons un algorithme de troncature des erreurs solution d'un problème min-max dont le temps de calcul est raisonnable et qui ne suppose pas d'hypothèse sur le conditionnement de la matrice de Gram. Ce nouvel estimateur est construit à partir d'une fonction d'influence $\psi : \mathbb{R} \to \mathbb{R}$ croissante bornée possèdant la propriété fondamentale suivante :

$$-\log\left(1 - x + \frac{x^2}{2}\right) \le \psi(x) \le \log\left(1 + x + \frac{x^2}{2}\right), \qquad x \in \mathbb{R}.$$

Le point technique nouveau permettant cette approche perturbative en régression consiste à approcher les valeurs de ψ en des points déterministes par l'espérance de ψ sous une perturbation aléatoire de son argument. Cette technique permet d'appliquer des bornes PAC-Bayésiennes à des estimateurs « classiques », c'est-à-dire non randomisés, les lois a posteriori sur les paramètres n'étant introduites que dans les calculs. Cela permet non seulement d'étudier des estimateurs à la fois plus élégants et plus traditionnels, mais aussi d'utiliser des lois a posteriori dépendant de quantités non observables. C'est cette dernière liberté qui nous a

permis de nous débarrasser des hypothèses sur le conditionnement de la matrice de Gram présentes dans la version de 2009 en introduisant dans les preuves des lois a posteriori gaussiennes ayant pour covariance la matrice de Gram (et non pas son approximation empirique). Ces développements nous ont conduit au Théorème 3.1 [AC11a, page 10], qui permet d'obtenir une convergence en $\left[d - \log(\epsilon)\right]/n$ de la fonction quantile du risque quadratique sous des hypothèses portant sur la kurtosis de certains moments polynomiaux du design et de l'erreur quadratique. Ces hypothèses sont plus faibles que les hypothèses utilisées classiquement dans la littérature (par exemple concernant la stabilité $\mathbb{L}^2$ de la base dans le cas de la régression fonctionnelle) qui les impliquent, ainsi qu'expliqué dans le paragraphe 3.2 de [AC11a, page 11].

Nous avons pu aussi proposer un schéma de calcul de cet estimateur tronqué et montrer qu'il permettait en pratique de diminuer le risque quadratique dans le cas d'un bruit non gaussien à queue lourde.

Le fait qu'il soit possible d'obtenir des estimateurs exponentiellement consistants en ne faisant que des hypothèses de moments polynomiaux sur le bruit était inattendu. Le fait que cela conduise à des performances expérimentales améliorées en présence d'un bruit non gaussien à queue lourde ne l'était pas moins.

Le même phénomène se produit dans le cas plus simple de l'estimation de la moyenne d'une variable aléatoire réelle. J'ai écrit un premier article sur le sujet en 2009, utilisant une méthode d'estimation itérative. Cette première version de [Cat12], intitulée « High confidence estimates of the mean of heavy-tailed real random variables » [Cat09] fut accueillie fraîchement par les rapporteurs. Le sujet consistant à trouver un estimateur de la moyenne possédant des déviations sous exponentielles sous une simple hypothèse de variance finie paraissait à l'époque incongru. Cette première incompréhension fut en fait fructueuse, puisqu'elle me permit d'améliorer l'estimateur proposé ainsi que les bornes qui l'accompagnent, au cours de deux révisions successives, pour aboutir à la publication de [Cat12]. Elle fut suivie un peu plus tard de réactions élogieuses, et d'une invitation à présenter ces résultats à Lille. La méthode est applicable dans de nombreux contextes, d'autres variantes sont apparues dans la littérature, Gabor Lugosi m'a félicité en m'assurant que le résulat lui était très utile : d'incongru le sujet est devenu à la mode !

Les principaux enseignements de [Cat12] sont les suivants. Il existe un M-estimateur dont les déviations de niveau de probabilité supérieur à 90% par rapport à la moyenne sont inférieures aux déviations de la moyenne empirique. Ce phénomène se produit pour certaines distributions de variance finie, pour des échantillons de taille 100 et plus, de manière prouvée par des bornes non asymptotiques. Le même type de sous optimalité, pour des niveaux de confiance plus élevés, peut être mis en évidence pour des distributions d'échantillons de kurtosis bornée. Dans ces modèles très larges, (variance bornée, ou kurtosis bornée), des M-estimateurs

appropriés (utilisant dans un cas la valeur de la variance et dans l'autre celle de la kurtosis), ont, pour toutes les distributions d'échantillons appartenant au modèle, des déviations du même ordre que les déviations de la moyenne empirique d'un échantillon gaussien (ce qui implique qu'il y aurait peu à gagner à considérer des modèles intermédiaires entre ces modèles très larges et le modèle gaussien, la moyenne empirique étant l'estimateur optimal dans le cas gaussien).

Dans le cas où l'on connaît une borne supérieure de la kurtosis, le M-estimateur de la moyenne que je propose utilise un estimateur de la variance, qui sert à fixer son paramètre d'échelle. Une dernière amélioration technique, obtenue durant l'année 2010-2011 à la faveur de la dernière révision de mon manuscrit, a consisté à mettre au point un schéma de troncature par blocs qui fournit d'emblée un estimateur de la variance asymptotiquement optimal, alors que dans les versions précédentes, l'estimation de la variance se faisait en trois étapes : une étape préliminaire, une étape d'estimation de la moyenne, suivie d'une troisième étape de réestimation de la variance au vu de l'estimation de la moyenne. De ces trois étapes, je suis donc parvenu à ne conserver que la première, sous une forme améliorée. L'obtention d'un estimateur de la variance possédant des déviations sous-exponentielles était à ma connaissance une question ouverte, comme expliqué page 28 de [Cat12].

Les résultats expérimentaux mettent en évidence un gain par rapport à la moyenne empirique encore plus net que ce que la théorie laisse présager : il suffit de sortir du modèle gaussien en considérant des lois composées du mélange de deux gaussiennes pour obtenir une amélioration uniforme de la fonction quantile des déviations par rapport à la moyenne, avec, dans certains cas, une amélioration supérieure à 25% au niveau 90% pour des échantillons de taille 100. Il semble donc bien que, d'un point de vue pratique, la moyenne empirique ne se comporte pas comme un estimateur admissible en dehors du modèle gaussien (puisqu'il semble possible de construire un M-estimateur dont tous les quantiles soient essentiellement égaux ou meilleurs pour toute distribution de l'échantillon). De plus les expériences semblent montrer que la précision de l'estimateur empirique non biaisé classique de la variance est suffisante pour le réinjecter purement et simplement dans le M-estimateur de la moyenne sans perte notable de performance par rapport au cas où la variance est supposée connue.

Enfin ces nouveaux M-estimateurs de la moyenne et de la variance peuvent être calculés avec une précision satisfaisante grâce à des schémas itératifs à convergence rapide (deux itérations, en pratique). Ce sont donc des candidats potentiels pour construire de nouveaux algorithmes de filtrage dans le domaine du traitement du signal et des images.

5.3. Linguistique computationnelle.

Après avoir co-encadré au printemps 2010, en collaboration avec Edward Stabler,

professeur à UCLA, le stage de M2 de Thomas Mainguy, et m'être à cette occasion familiarisé avec les enjeux de la linguistique computationnelle, j'ai entrepris d'encadrer la thèse de cet étudiant. L'objectif fixé était de proposer des modèles statistiques des langues naturelles permettant d'extraire des structures syntaxiques à partir de l'analyse d'un corpus. Il s'agissait de rapprocher les approches statistiques et linguistiques dans le domaine de l'analyse des langues naturelles.

Durant son stage, Thomas Mainguy a étudié les grammaires minimalistes de Stabler. Il a proposé pour elles un algorithme de parsing top down ainsi qu'une loi de probabilité sur les règles permettant de définir un modèle de grammaires minimalistes stochastiques. Durant sa première année de thèse, nous sommes partis à la recherche d'un modèle statistique du langage mieux adapté à l'estimation à partir d'un corpus que les grammaires stochastiques étudiées durant le stage, qui sont dotées d'une structure *cachée* d'une complexité dissuasive.

Durant l'année 2010-2011, nous avons pu faire l'inventaire des écueils sur lesquels menace de s'échouer toute approche statistique de la langue. Ces écueils, bien décrits par Noam Chomsky, ne nous ont pas épargnés dans nos efforts pour forger des modèles d'inspiration Markovienne. Nous avons tout d'abord passé du temps à envisager les généralisations possibles des modèles d'arbres de contextes, qui peuvent eux-mêmes être vus comme une version adaptative des modèles de n-grammes, dans laquelle on adapte la longueur n du n-gramme à son contenu. Nos réflexions ont porté sur la possibilité d'introduire dans les n-grammes une interruption (unique, pour ne pas trop augmenter la complexité du modèle). Traduit en terme de contexte, cela signifie que l'on accepte à l'instant t de faire dépendre la loi du mot w_t soit d'une fonction $f(w_1, \ldots, w_{t-1})$ du passé immédiat, soit d'une fonction $f(w_1, \ldots, w_s)$ d'un passé $s < t - 1$ plus lointain. Cela nous a conduit à envisager des modèles où le contexte n'est plus une fonction déterministe du passé mais une variable aléatoire dépendant du passé. Il nous a semblé que cela devrait permettre de paramétrer de façon économe des modèles intermédiaires entre les modèles à arbre de contextes et les modèles de Markov cachés, que l'on pourrait nommer modèles à contextes aléatoires. Le choix aléatoire d'un contexte peut en effet s'interpréter comme le choix d'un type particulier d'état caché. A l'époque, nous pensions intituler la thèse « Modèles à contextes aléatoires » (random context models), « et applications en linguistique computationnelle ». Ces modèles à contextes aléatoires, nous les voulions différents des mélanges d'arbres de contextes, ou plus généralement d'autres modèles de contextes. En effet, dans les modèles de mélange de contextes, le contexte est bien tiré au hasard, mais suivant une loi qui cherche à approcher le meilleur modèle déterministe, et sera typiquement de ce fait concentrée autour d'un seul modèle, dans une perspective PAC-Bayésienne de petite perturbation par une loi de Gibbs d'un modèle déterministe. Dans les modèles à contextes aléatoires, nous envisagions pour la distribution du contexte aléatoire des lois plus générales, pouvant se répartir de

façon équilibrée sur plusieurs choix de contextes.

Cette approche n'a pas résisté à nos tentatives de simulations : nous avons dû nous rendre aux arguments de Chomsky, la structure récursive des langues naturelles ne peut être décrite efficacement à l'aide d'un modèle markovien.

A l'automne 2011, cet échec m'a fait apercevoir une autre voie, qui consiste à appréhender la structure globale des phrases. Cela nous a conduit à envisager un modèle d'emboîtement aléatoire de sous-expressions à l'intérieur d'une structure de phrase globale, en d'autres termes un modèle de construction de phrases par copier-coller à partir d'une collection de phrases type. Après avoir introduit un système d'étiquettes indiquant comment recoller les morceaux, nous nous sommes vite aperçu que les expressions que nous obtenions, séquences composées de mots et d'étiquettes appariées, coïncidaient avec les règles de production d'une grammaire sans contexte. Ainsi en passant de modèles à contextes locaux, les modèles Markoviens, à des modèles à contextes globaux, nous retombions sur les grammaires sans contexte, connues des linguistes, ce qui était plutôt rassurant après tout.

Néanmoins, contrairement à l'habitude, ces grammaires sans contexte, nous les construisions en fragmentant un texte, alors que l'emploi classique d'une grammaire consiste à générer des phrases, et donc potentiellement un texte si on répète l'opération, à partir de règles de production. Cette idée de fragmenter un texte (nous désignons ici comme texte un ensemble non ordonné de phrases, échantillon destiné à l'apprentissage d'une grammaire), nous a conduit à modéliser un texte par une mesure de comptage sur ses phrases. De là nous en sommes venus à modéliser le produit de la fragmentation d'un texte par une mesure de comptage sur des expressions constituées d'une suite de mots et d'étiquettes appariées (les symboles non terminaux de notre grammaire). Nos étiquettes appariées sont constituées d'un crochet ouvrant ou fermant suivi d'un nombre entier, elles sont plus précisément de la forme $[\, i, i \in \mathbb{N}$ ou $]i, i \in \mathbb{N} \setminus \{0\}$, $[0$ jouant le rôle particulier de symbole initial. Pour des raisons algorithmiques et d'uniformité, nous considérons des mesures de comptage invariantes par permutations circulaires, pour aboutir à une nouvelle notion de grammaire, les *grammaires toriques* (nous avons choisi ce nom à cause du rôle des permutations circulaires).

Une grammaire torique, telle que nous la définissons, peut être vue comme une mesure de comptage sur les règles de production d'une grammaire sans contexte. Une grammaire torique obtenue par fragmentation d'un texte vérifie des relations supplémentaires entre les poids des étiquettes appariées. A la fragmentation, on peut faire correspondre la transformation inverse, appelons-la coalescence, qui constitue à recomposer un texte en appariant les étiquettes. En enchaînant une coalescence aléatoire et une fragmentation aléatoire, on obtient un noyau markovien sur les grammaires toriques. L'ensemble des grammaires toriques n'est pas fini, mais on peut montrer néanmoins que cette chaîne de Markov sur les grammaires

toriques possède deux propriétés :

— la chaîne est faiblement réversible, au sens où la transition inverse d'une transition de probabilité positive est aussi de probabilité positive ;
— l'ensemble des états accessibles depuis un état donné est fini, si bien que la chaîne découpe l'ensemble des grammaires toriques en une partition composée de classes récurrentes de cardinal fini.

Ces réflexions nous ont conduits à proposer dans [CM13] un modèle de transmission inter-locuteurs prenant la forme d'une chaîne de Markov sur les mesures empiriques sur les phrases, obtenue en enchaînant une fragmentation et une coalescence. Dans la [CM13, proposition 7.2, page18] nous établissons que cette chaîne de communications inter-locuteurs est faiblement réversible et que l'ensemble des états accessibles depuis n'importe quel état de départ est borné (du fait de la présence d'un certain nombre d'invariants, dont la fréquence d'apparition des mots fait partie).

Ce modèle permet de définir un estimateur, de la forme

$$\widehat{Q}(P) = \lim_{k\to+\infty} \frac{1}{k} \sum_{t=1}^{k} q^t(P, \cdot),$$

où P est la mesure empirique de l'échantillon observé, et où $\widehat{Q}(P)$ est une mesure de probabilité sur les mesures empiriques de tous les échantillons composés de n phrases arbitraires. Il s'agit donc d'une sorte de noyau de convolution appliqué à la mesure empirique tout entière : dans un estimateur à noyau classique, on utilise un noyau de convolution pour remplacer la masse de Dirac en chaque observation par une probabilité chargeant un voisinage de cette observation. Ici, on perturbe aussi la mesure empirique, comme dans un estimateur à noyau, mais la perturbation implique des interactions entre les observations. Il s'agit en quelque sorte d'un « sample level kernel estimate ».

Cette approche nous semble novatrice, dans la mesure où elle permet de donner une définition du langage engendré par une grammaire torique complètement différente du langage engendré par la grammaire sans contexte qui lui sert de support. D'une part, l'emploi d'une mesure de comptage empêche d'utiliser une même règle un nombre arbitraire de fois, d'autre part, les allers et retours entre coalescence et fragmentation permettent de produire des structures syntaxiques qui ne sont pas directement accessibles par réécritures successives du symbole initial (qui est $[_0$ avec nos notations), c'est-à-dire par une unique opération de coalescence, pour employer la terminologie que nous avons introduite.

Les simulations présentées dans [CM13] ne portent que sur de petits corpus. Elles sont néanmoins très encourageantes. Elles montrent en particulier que le modèle produit des phrases qui, à quelques fautes d'accord près, gardent une cohérence globale et ressemblent de ce fait à celles qu'aurait pu prononcer un locuteur

humain. Des simulations sur des corpus de grande taille seraient souhaitables mais demanderaient de développer un code plus optimisé.

La suite de la thèse de Thomas Mainguy permet d'établir un lien entre les grammaires toriques et un nouveau modèle statistique, celui des processus de substitution Markoviens. Un processus de substitution Markovien est décrit par la liste de ses ensembles de substitution et des mesures de substitution associées. Un ensemble de substitution B est un ensemble d'expressions que l'on peut substituer les unes aux autres suivant une mesure de substitution $q_B \in \mathcal{M}_+^1(B)$ indépendamment du contexte. En d'autres termes, si $\mathbb{P}$ est la loi d'un processus de substitution Markovien, et si B est un ensemble de substitution,

$$\mathbb{P}(xyz) = \left(\sum_{y' \in B} \mathbb{P}(xy'z) \right) q_B(y),$$

pour toutes suites de mots x, y, z, telles que $y \in B$. On peut montrer que les processus de substitution Markoviens de support donné forment une famille exponentielle. Les ensembles de substitution définissent un graphe sur l'ensemble des phrases et ce modèle exponentiel est paramétré par la probabilité de chacune de ses composantes connexes ainsi que par un sous-ensemble « libre » de mesures de substitution à partir desquelles les autres peuvent se calculer. A partir des ensembles de substitution d'un processus de substitution Markovien, on peut construire des règles de réécriture sans contexte, et décrire certains ensembles de substitution comme les langages engendrés par des grammaires sans contexte. On peut aussi construire un processus de fracturation-coalescence (split and merge process) qui laisse invariant la loi des processus de substitution Markoviens associés à une famille donnée d'ensembles de substitution. Plus généralement, les processus de fragmentation-coalescence associés aux grammaires toriques apparaissent comme une forme accélérée de processus de crossing over, où les phrases d'un corpus sont hybridées entre elles.

A l'aide de ces outils, la thèse de Thomas Mainguy parvient à aborder les problèmes de sélection de modèle et d'estimation des paramètres pour les processus de substitution Markoviens d'une façon cohérente et très prometteuse pour l'avenir de cette théorie. Malgré le manque de simulations à grande échelle, qui restent à mettre en œuvre, on peut espérer que les processus de substitution Markoviens pourront apporter une contribution pertinente à la modélisation des langues naturelles dans la mesure où ils fédèrent, dans un cadre statistique représentant une extension naturelle des modèles de champ de Markov, la détection et la description de structures déjà validées par les linguistes, et décrites par eux à l'aide de modèles de grammaires formelles dérivés du modèle simplifié des grammaires sans contexte.

5.4. Vision et apprentissage. Je reprends dans ce paragraphe des réflexions faites dans mes rapports précédents sur l'analyse d'images.

La classification d'images est un défi aux multiples facettes. L'un des mérites du sujet est d'obliger à sortir du cadre de la classification supervisée auquel il est tentant de se restreindre dans les études théoriques sur l'apprentissage statistique. Les images sont en effet des données de bien trop grande dimension, sous leur forme brute de matrices de pixels, pour qu'on puisse en faire quoi que ce soit sans prétraitement. J'aborde le prétraitement en le considérant comme un problème d'apprentissage non supervisé, préalable à une phase ultérieure d'apprentissage supervisé.

Je me suis fixé un objectif en terme de données à traiter : celui de la classification d'images numériques de notre environnement quotidien, que les spécialistes de la vision appellent souvent « images naturelles ». Il s'agit des images que tout le monde peut recueillir avec un appareil photo numérique au cours de ses vacances, celles qui peuplent internet, ou celles encore fournies par le cinéma ou la télévision. Ces images interviennent dans des applications telles que la classification automatique de contenus vidéo ou de pages internet illustrées de photographies numériques. Elles interviennent aussi dans le domaine de la navigation assistée par ordinateur (navigation de véhicules ou de robots industriels). Elles ont l'avantage d'être très faciles d'accès tout en ayant le privilège d'être difficiles à traiter. On peut légitimement penser que des méthodes qui se révéleraient efficaces sur ce type de corpus pourraient trouver des applications dans d'autres domaines de l'imagerie (tels, par exemple, que la cartographie ou l'imagerie médicale).

Alors que la classification supervisée se présente du point de vue mathématique comme un problème de régression d'une variable discrète sur une variable explicative de grande dimension, problème auquel il est assez naturel d'associer la minimisation d'une fonction de perte telle que l'erreur de classification, la classification non supervisée poursuit des objectifs plus difficiles à cerner. Une approche purement théorique, comme celle que j'ai développée ces dernières années concernant le risque de généralisation et son lien avec des mesures d'entropie et l'information mutuelle entre paramètre et échantillon, me semble difficile à mener *in abstracto*. J'ai pensé contourner cet obstacle en développant une plate-forme logicielle d'expérimentation, en lien avec une pratique amateur de la photographie, qui m'a conduit à inclure dans ce logiciel des fonctions d'édition de photos numériques, mettant en œuvre des techniques de gestion des contrastes de mon cru. J'ai déposé à ce propos auprès de la Direction de la Politique Industrielle une déclaration d'invention.

J'ai pu valider à l'aide de ce logiciel certains traitements avec succès.

Je me suis fixé comme objectif la prise en compte des invariants projectifs pour la classification d'images. Concrètement, il s'agit de construire une représentation des données qui facilite la classification d'une scène plane (des photographies

de tableaux, par exemple) quelle que soit la position de l'appareil de prise de vue par rapport à la scène. Cette exigence fait partie du minimum requis pour la classification de scènes ordinaires, composées d'objets comportant des faces plus ou moins planes, (telles que des façades d'immeubles, les murs, les meubles d'une pièce, etc.) prises sous des angles pour lesquels la perspective cavalière (liée à l'invariance affine) ne suffit pas.

Le lecteur de ces lignes se demande peut-être ce que la classification non supervisée a à voir avec ces préoccupations. Essayons d'esquisser quelques liens.

La classification non supervisée, au bout du compte, peut être rapprochée des techniques de quantification, c'est-à-dire des techniques qui permettent de passer de mesures continues, en l'occurrence l'intensité lumineuse mesurée en chaque pixel, à une représentation discrète. Ce passage du continu au discret est rendu à mon sens nécessaire par le besoin de résoudre des problèmes de mise en correspondance. Autrement dit et très concrètement, d'une image à l'autre, les pixels représentant le même objet, ou le même détail d'un objet, ne se correspondent pas de façon évidente, si bien que le fait de savoir quels pixels des deux images doivent être rapprochés pour être comparés nécessite de faire des choix discrets, issus d'une classification locale des différentes zones des deux images (des expériences psychotechniques bien connues mettent en évidence d'autre part le fait que le cerveau, confronté à certaines scènes d'interprétation ambigüe, est capable de basculer brutalement d'une interprétation à une autre en fonction du contexte, ce qui incite aussi à penser que l'interprétation d'une scène nécessite de faire des choix discrets, un peu comme un voyageur se trouvant à un carrefour et devant choisir entre plusieurs routes). La classification non supervisée apparaît ainsi à mes yeux (c'est un pari personnel, et non une vérité scientifique que je serais susceptible de démontrer rigoureusement !) comme une étape incontournable dans la *réduction de la dimension* de la représentation des données.

Une approche de ces problèmes de mise en correspondance et de classification non supervisée consiste à s'appuyer sur la *théorie du codage sans perte* qui fait le lien entre *longueur de code* et log vraisemblance d'un modèle probabiliste appelé dans ce cadre *code idéal*. Cette théorie, fondée par Shannon, est décrite au début de mon cours à Saint-Flour. Définir des lois de probabilités sur des données peut ainsi servir à définir une façon de les coder (à travers des techniques de codage du type Shannon-Fano-Elias), en l'absence même de tout projet de modéliser un phénomène fréquentiel : bien que le lien avec une modélisation fréquentielle existe, à travers le fait que le code le plus court est celui formé à partir de la distribution de probabilités des données, d'autres codes, fondés sur des lois plus simples, peuvent néanmoins se révéler intéressants et efficaces. La classification apparaît alors très naturellement dans ce cadre en y introduisant des *modèles de mélange*.

Le mélange de lois est une technique de modélisation très naturelle dès que l'on a affaire à des données hétérogènes. Il permet d'exprimer de façon probabiliste des

disjonctions, permettant de modéliser (i.e. coder) des données pouvant prendre deux formes différentes (ou plus) — de même les modèles produits permettent de représenter de façon probabiliste des conjonctions — si bien qu'en envisageant des mélanges de lois produits, on obtient un outil de modélisation très souple. Une fois un mélange de lois construit, on peut associer à chaque composante un label et effectuer une classification en calculant dans le modèle de mélange la loi a posteriori des différentes classes. Cette classification peut constituer un changement de représentation, ou être ajoutée aux autres paramètres décrivant les données pour enrichir leur représentation.

J'ai obtenu les résultats suivants, premiers pas vers l'invariance projective. Je me suis attaché à concevoir des méthodes dont le temps de calcul soit linéaire par rapport à la taille des images, de façon à pouvoir traiter des images qui pèsent quelques millions de pixels, sans que la méthode explose.

Représentation multiéchelle des contours. Les droites sont des invariants projectifs, ainsi que leurs intersections, il est donc assez naturel de rechercher dans les images des contours, et plus particulièrement des contours rectilignes, ou encore des tangentes remarquables aux contours et leurs intersections. J'ai appliqué les principes précédemment décrits aux contours, définissant la loi d'une classe de contours multiéchelle comme mélange de lois à échelles données. La méthode s'est révélée très efficace en pratique, permettant de passer d'une photographie en niveaux de gris à une sorte de « dessin au trait » très satisfaisant du point de vue visuel. Ce traitement pourrait avoir des applications en lui même (par exemple dans la production de dessins animés ou de bandes dessinées, comme aide au dessin de décors réalistes, ou pour aider à l'animation des personnages — le fait de s'aider de l'analyse de séquences filmées ayant été utilisé dès les débuts du dessin animé ; la méthode pourrait aussi servir dans le domaine de la cartographie).

En extrayant des contours, on réduit une dimension, celle des niveaux de gris, à deux valeurs, tout en gardant beaucoup d'information sur la géométrie de la scène photographiée. De plus les contours occupent un lieu de dimension réduite dans l'image. Pour ces deux raisons, ce changement de représentation apparaît comme un pas significatif vers une représentation « parcimonieuse ».

Une fois la détection multiéchelle des contours validée, je me suis posé la question de la détection de leur orientation. Cela peut se faire efficacement à l'aide d'un opérateur de convolution. Là encore il pourrait y avoir des applications directes à la stéréoscopie ou à l'analyse du mouvement.

Je me suis alors posé la question de l'analyse des intersections de contours. Ceci m'a conduit à explorer une notion de champ dual, défini à partir du champ G des directions normales aux contours par la formule :

$$G'_\alpha(t) = \int_s \rho_\alpha(t, ds)\langle t - s, G(s)\rangle G(s),$$

où s parcourt les points de l'image et où $\rho_\alpha(t, ds)$ est un noyau de convolution effectuant une moyenne locale sur les sites de contours à une échelle donnée α. Ces champs duaux étant obtenus par convolution avec un noyau régularisant, possèdent une stabilité qui augmente avec l'échelle. De plus, ils possèdent des zéros isolés dont certains coïncident avec des intersections de contours. Des illustrations concernant le calcul des contours et du champ dual sont disponibles dans le projet rédigé à l'époque de la création d'équipe INRIA CLASSIC (qui se trouve sur ma page web, — les mêmes paramètres ont été utilisés pour toutes les images présentées).

Ces étapes sont validées par un logiciel prototype avec une interface conviviale (prise en charge des principaux formats d'images, traitement en batch possible, on peut aussi faire subir aux images des transformations projectives, ainsi que diverses opérations standard, comme la conversion en niveaux de gris d'une image couleur, la visualisation d'histogrammes, le réglage de la luminosité et du contraste, ce qui fait gagner du temps en dispensant l'expérimentateur de jongler avec plusieurs logiciels : ceci peut paraître anodin, mais permet en pratique de tester plus d'images dans le même temps).

Tous ces traitements, détection de contours, calculs d'orientations, champ dual G'_α, pourront servir d'ingrédients dans la sélection d'une représentation des images suffisamment stable et invariante pour servir de base à des méthodes de clustering non supervisé pertinentes pour l'analyse automatique de scènes visuelles.

6. Enseignement, formation et diffusion de la culture scientifique

6.1. Enseignement. J'ai donné depuis l'année universitaire 1999-2000 et jusqu'à l'année 2006-2007 dans le DEA (puis dans le M2) du Laboratoire de Probabilités et Modèles Aléatoires, un cours de troisième cycle avancé dont j'ai chaque année fait évoluer le contenu en rapport avec l'actualité de mes travaux de recherche. J'avais aussi donné un tel cours en 1995 à l'Université Paris Sud, dont les notes de cours [Cat99a] ont été publiées par le Séminaire de Probabilités.

Durant l'année universitaire 2010-2011, j'ai participé au premier semestre, en compagnie de Gérard Biau et de Gilles Stoltz, à l'animation d'un groupe de travail portant sur l'estimation de la densité, destiné aux élèves de deuxième année de la FIMFA de l'ENS. Au second semestre, j'ai assuré 10h de cours dans le cadre du cours de 50h consacré à l'apprentissage statistique destiné aux étudiants de première année de la filière math-info de la FIMFA (les autres modules de ce cours étant assurés par S. Arlot, J.Y. Audibert, F. Bach et G. Stoltz). Ces deux enseignements ont été reconduits en 2011-2012. Le groupe de travail de statistique de l'automne 2011, contrairement à celui de l'année précédente a porté sur des thèmes variés, illustrés par la présentation d'articles de recherche par les étudiants.

Le contenu du cours d'apprentissage du second semestre a par contre été revu à la baisse, chaque thème conservé étant accompagné de plus d'exemples concrets d'applications. Ce cours a été reconduit en 2013 sous un format différent, 30 heures de cours et 20 heures de TD, de façon à insister encore plus sur les applications et le maniement des techniques enseignées. J'ai assuré 8h de cours (accompagnées de 4h de TD animées par Rémi Lajugie), portant sur les inégalités de concentration et les inégalités de marge PAC-Bayésiennes pour les support vector machines dans cet enseignement collégial aux côtés de Francis Bach et Guillaume Obozinski, du département d'informatique de l'ENS. Un groupe de travail de statistique, animé en collaboration avec Gérard Biau, et portant sur la lecture d'articles variés, s'est tenu à l'automne 2013.

Comme indiqué à la rubrique collaborations françaises et étrangères, je suis intervenu durant l'été 2014 par visio-conférence pour donner trois heures de cours sur la théorie statistique de l'apprentissage dans une école d'été organisée par l'Indo-French Center for Applied Mathematics à Bangalore. La partie la plus innovante des notes de cours correspondant doit paraître dans un Festschrift en l'honneur d'A. Chervonenkis.

D'avril à la mi-août 2015, j'ai encadré le stage de recherche de l'Ecole Polytechnique (niveau M1) de Xiayang Zhou, concernant les algorithmes de clustering spectral.

De mai à septembre 2016, j'ai dirigé le stage de M2 de Gautier Appert, comptant aussi pour sa scolarité à l'ENSAE, stage qui se prolonge par une thèse sous ma direction. Le stage a porté sur la classification non supervisée de fragments (patch) d'images.

D'avril à août 2019, j'ai encadré le stage de M2 d'Aurélien Brouillaud portant sur l'analyse des modèles d'arbres de contextes (Master Vision et Apprentissage).

6.2. Direction de thèses. J'ai dirigé huit thèses, celles de Cécile Cot, Gilles Blanchard, Jean-Philippe Vert, Jean-Yves Audibert, Pierre Alquier, Thomas Mainguy, Ilaria Giulini et Gautier Appert, dont une en co-direction avec Alain Trouvé (celle de Gilles Blanchard).

La thèse de Gautier Appert, entamée en septembre 2016 a été soutenue le 29 octobre 2020. Elle a pour titre *Information k-means, fragmentation and syntax analysis. A new approach to unsupervised machine learning.*

En voici le résumé.

« Le critère de l'information k-means étend le critère des k-means en utilisant la divergence de Kullback comme fonction de perte. La fragmentation est une généralisation supplémentaire permettant l'approximation de chaque signal par une combinaison de fragments.

Nous proposons un nouvel algorithme de fragmentation pour les signaux numériques se présentant comme un algorithme de compression avec perte.

A l'issue de ce traitement, chaque signal est représenté par un ensemble aléatoires de labels, servant d'entrée à une procédure d'analyse syntaxique, conçue comme un algorithme de compression sans perte.

Cet algorithme, fondé sur deux principes appliqués itérativement, la factorisation et le réétiquetage de configurations fréquentes, produit pour chaque signal un arbre syntaxique fournissant une classification hiérarchique des composantes du signal.

Nous avons testé la méthode sur des images en niveaux de gris, sur lesquelles il a été possible de détecter des configurations translatées ou transformées par une rotation. Ceci donne l'espoir d'apporter une réponse à la reconnaissance invariante par transformations fondée sur un critère de compression très général.

D'un point de vue mathématique, nous avons prouvé deux types de bornes. Tout d'abord, nous avons relié notre algorithme de compression à un estimateur implicite d'un modèle statistique lui aussi implicite, à travers un lemme, prouvant que le taux de compression et le niveau de distorsion de l'un sont reliés à l'excès de risque de l'autre. Ce résultat contribue à expliquer la pertinence de nos arbres syntaxiques.

Ensuite, nous établissons des bornes de généralisation non asymptotiques et indépendantes de la dimension pour les différents critères des k-means et critères de fragmentation que nous avons introduits. Nous utilisons pour cela des inégalités PAC-Bayésiennes appliquées dans des espaces de Hilbert à noyau reproduisant.

Par exemple dans le cas des k-means classiques, nous obtenons une borne en $\mathcal{O}(k\log(k)/n)^{1/4}$ qui fournit la meilleure condition suffisante de consistance, à savoir que l'excès de risque tend vers zéro quand $k\log(k)/n$ tend vers zéro. Grâce à une nouvelle méthode de chaînage PAC-Bayésien, nous prouvons aussi une borne en $\mathcal{O}\big(\log(k/n)\sqrt{k\log(k)/n}\big)$. »

La thèse d'Ilaria Giulini, entamée en septembre 2012 et soutenue le 24 septembre 2015, a été financée par le Réseau de Recherche Doctoral de Mathématiques de l'Île-de-France (RDM-IdF). Intitulée Generalization bounds for random samples in Hilbert spaces, elle porte sur l'analyse de données en grande dimension.

Le premier chapitre établit des bornes PAC-Bayésiennes indépendantes de la dimension pour l'estimation robuste de l'opérateur de Gram d'un échantillon i.i.d. à valeurs dans un espace de Hilbert séparable de dimension finie ou infinie. Des extensions à l'estimation de l'espérance d'une matrice (ou d'un opérateur) symétrique aléatoire à partir d'un échantillon sont proposées, ainsi qu'à l'estimation de la matrice de covariance d'un échantillon.

Le second chapitre porte sur l'estimation de la matrice de Gram par la matrice de Gram empirique. Il prouve des bornes non asymptotiques dont le terme domi-

nant est du même ordre que dans le cas de l'estimation robuste, mais qui possèdent un second terme dépendant d'hypothèses de moment plus fortes qui peut ne passer au second plan que pour des tailles d'échantillon très grandes lorsque les données ont des queues de distribution lentement décroissantes.

Le troisième chapitre traite de l'analyse en composantes principales d'un échantillon à valeurs dans un espace de Hilbert séparable. L'objectif est d'approcher la projection sur les premiers vecteurs propres de l'opérateur de Gram inconnu. Ilaria Giulini propose des versions robustes de l'estimation de ces projections accompagnées de bornes de convergence non asymptotiques. Elle propose aussi de remplacer la projection sur les premiers vecteurs propres par une troncature plus régulière des valeurs propres (en l'occurrence Lipschitzienne), permettant de prouver des vitesses de convergence qui ne dépendent pas comme dans le cas de la projection du trou spectral entre la dernière valeur propre retenue et la suivante.

Le quatrième chapitre étudie le problème de la classification non supervisée d'un échantillon i.i.d. à valeurs dans un espace de Hilbert séparable par clustering spectral. Le problème du clustering spectral est mis en relation avec l'estimation d'un opérateur de Gram dans un espace de Hilbert à noyau reproduisant. De nouveaux algorithmes robustes sont proposés, ainsi que des bornes supérieures portant sur la vitesse de convergence. La projection sur les premiers vecteurs propres du Laplacien est remplacée par un changement de représentation utilisant un Laplacien itéré correspondant à un opérateur Markovien. Cette modification de l'algorithme permet en particulier d'estimer automatiquement le nombre de classes à partir d'une borne supérieure.

J'ai co-dirigé au printemps 2010 le stage de M2 de Thomas Mainguy (en collaboration avec E. Stabler d'UCLA), puis dirigé seul sa thèse [Mai14], commencée en septembre 2010 et soutenue le jeudi 11 décembre 2014. Elle s'intitule Processus de substitution Markoviens, un modèle statistique pour la linguistique.

Il s'agit, comme décrit plus haut dans mes projets de recherche, de concevoir des modèles statistiques permettant d'apprendre des structures syntaxiques à partir d'un corpus de textes.

Le premier chapitre, dont est tiré [CM13], propose, à travers la notion de grammaire torique, une dynamique sur les échantillons de phrases dont les distributions invariantes permettent de définir un modèle de langage. On peut aussi utiliser la dynamique comme modèle de transmission de la langue d'un locuteur à un autre.

Le second chapitre introduit une famille de lois de probabilités plus explicites sur les phrases, les processus de substitution Markoviens, définis par des propriétés d'indépendance conditionnelle qui généralisent la notion de champ de Markov unidimensionnel. Ce modèle est aussi présenté dans [CM16]. Il possède la propriété intéressante de former une famille exponentielle. On peut le simuler à l'aide de dynamiques analogues à celles décrites dans le premier chapitre. Thomas Mainguy

traite aussi dans ce chapitre de la question de la sélection d'un tel modèle à partir de tests multiples concernant les hypothèses d'indépendance conditionnelle sous-jacentes.

Le troisième et dernier chapitre analyse les liens entre processus de substitution Markoviens et grammaires sans contexte. Il montre comment utiliser des algorithmes de parsing pour calculer la probabilité d'une phrase, ainsi que des dynamiques invariantes accélérées. La question du calcul du maximum de vraisemblance dans un modèle de substitution Markovien est traitée. Elle se pose dans la mesure où la représentation du modèle sous la forme d'une famille exponentielle n'est pas explicite. Néanmoins on peut approcher la loi du maximum de vraisemblance en utilisant une dynamique de crossing-over sur un grand nombre de répliques de l'échantillon observé, en exploitant le fait que les modèles de substitution Markoviens sont les lois invariantes de certaines dynamiques de crossing-over.

La thèse de Pierre Alquier, soutenue le 8 décembre 2006 porte sur la conception d'un algorithme de sélection de variables explicatives pour la régression en norme $\mathbb{L}^2$ par projections successives sur des régions de confiance, l'adaptation de cette démarche au cas de l'estimation de densité et la généralisation au cas d'une fonction de perte quelconque de l'approche PAC-Bayésienne de la classification adaptative. Elle aborde en particulier l'obtention directe d'un terme de variance empirique par une méthode de changement de variable dans la transformée de Laplace, un changement de variable particulièrement bien adapté à l'introduction d'hypothèses de moment étant entre autres proposé. Pierre Alquier a testé tout au long de sa thèse ses résultats sur des exemples de reconstruction de courbes classiques dans la littérature. Les algorithmes qu'il propose et qu'il étudie conduisent en particulier à de nouveaux types de Support Vector Machines, et permettent de généraliser les méthodes d'estimation adaptative par seuillage de coefficients au cas où les variables explicatives ne sont pas décorrélées entre elles. Cette thèse a conduit à trois publications parues (Pierre ayant bien entendu continué à publier depuis). Il a été recruté comme Maître de Conférences à l'Université Paris VII en septembre 2007, puis comme Lecturer at University College Dublin en 2012.

Jean-Yves Audibert a soutenu le 29 juin 2004 une thèse portant sur l'agrégation d'estimateurs en norme $\mathbb{L}^2$, le contrôle empirique de la variance en classification PAC-Bayesienne par l'introduction de bornes relatives, la prise en compte d'hypothèses de marge et d'entropie polynomiale, ainsi que sur une approche PAC-Bayésienne de la méthode du chaining (nécessaire pour obtenir la meilleure vitesse de convergence possible sous des hypothèses de complexité non paramétriques). Il a été depuis recruté au CERTIS (Centre d'Enseignement et de Recherche en Technologies de l'Information et Systèmes) de l'École Nationale des Ponts et Chaussées (site de Marne-la-Vallée). Jean-Yves a aussi été membre à temps patiel de l'équipe Willow de l'INRIA, située au Département d'Informatique de l'ENS. Il a rejoint à l'automne 2011 un fonds d'investissement privé pour y développer

des outils statistiques de finance quantitative.

La thèse de Jean-Philippe Vert, soutenue le 30 mars 2001, étudie des modèles de Markov à mémoire variable issus de la théorie de la compression sans perte et traite de leur application à l'analyse statistique de bases de données textuelles. L'objectif de cette application est de définir des distances entre textes rédigés en langue naturelle à partir de critères statistiques, dans le but de structurer le contenu d'une base de données et d'en faciliter l'interrogation. À la suite de sa thèse, Jean-Philippe Vert est parti en séjour post-doctoral dans un laboratoire de bio-informatique japonais - Kanehisa Laboratory, Bioinformatics Center, Institute for Chemical Research, Kyoto University, pour être ensuite recruté par le Centre de Géostatistique de l'École des Mines de Paris pour y fonder en octobre 2002 un groupe de recherche en « Computational Biology » qu'il dirige depuis tout en collaborant par ailleurs en tant que directeur adjoint avec une équipe de l'Institut Curie.

La thèse de Gilles Blanchard, soutenue le 5 janvier 2001, traite du mélange et de l'agrégation d'estimateurs en reconnaissance des formes et de leurs applications aux arbres de décision. Elle traite en particulier de l'estimation adaptative d'un arbre de décision par des méthodes pseudo-Bayésiennes de mélange et étudie certains algorithmes de boosting. Gilles Blanchard est actuellement professeur à l'Université de Potsdam.

La thèse de Cécile Cot, soutenue le 17 décembre 1998, porte sur des techniques d'accélération des algorithmes de Metropolis et de recuit simulé sur un réseau et leur application au traitement d'images. Après une étude des algorithmes de recuit simulé constants par paliers, elle aborde deux techniques d'accélération : l'optimisation répétée par blocs de variables et une technique d'optimisation hiérarchique. Elle se termine par un travail plus spécifique au traitement d'images portant sur la restauration dichotomique des lignes de niveaux qui sert entre autres à tester les méthodes d'accélération introduite dans la thèse.

7. Responsabilités collectives et management de la recherche

Je suis membre de la commission des thèses de l'Universités Paris 6 depuis mars 2007.

8. Mobilités

Thématiques : des probabilités appliquées à la théorie statistique de l'apprentissage, ainsi qu'à ses applications pratiques à l'analyse d'images et à la linguistique computationnelle.

Références

[App20] Gautier Appert. « Information k-means, fragmentation and syntax analysis. A new approach to unsupervised machine learning ». Thèse de doct. EDMH, CREST, oct. 2020. URL : https://tel.archives-ouvertes.fr/tel-03015285.

[LM18] Gabor Lugosi et Shahar Mendelson. *Near-optimal mean estimators with respect to general norms*. preprint. 2018. arXiv : 1806.06233.

[Dev+16] Luc Devroye, Matthieu Lerasle, Gabor Lugosi et Roberto Imbuzeiro Oliveira. « Sub-Gaussian mean estimators ». In : *Ann. Stat.* (2016). to appear.

[FMN16] Charles Fefferman, Sanjoy Mitter et Hariharan Narayanan. « Testing the manifold hypothesis ». In : *Journal of the American Mathematical Society* 29.4 (2016), p. 983-1049.

[Giu16] Ilaria Giulini. *Robust Principal Component Analysis in Hilbert spaces*. 2016. arXiv : 1606.00187v1.

[Giu15a] Ilaria Giulini. « Generalization bounds for random samples in Hilbert spaces ». Thèse de doct. Ecole Normale Supérieure, INRIA, France, sept. 2015. URL : http://pages.saclay.inria.fr/ilaria.giulini/PhD_thesis_Giulini.pdf.

[Giu15b] Ilaria Giulini. *Robust dimension-free Gram operator estimates*. preprint. 2015. arXiv : 1511.06259.

[Min15] Stanislav Minsker. « Geometric Median and Robust Estimation in Banach Spaces ». In : *Bernoulli* 21.4 (2015), p. 2308-2335.

[Mai14] Thomas Mainguy. « Markov Substitute Processes, (a statistical model for linguistics) ». Thèse de doct. Université Pierre et Marie Curie, ENS, INRIA, France, déc. 2014. URL : https://tel.archives-ouvertes.fr/tel-01127344.

[LI11] Matthieu Lerasle et Roberto Imbuzeiro Oliveira. *Robust empirical mean estimators*. preprint. 2011. arXiv : 1112.3914.

[BDL08] Gérard Biau, Luc Devroye et Gábor Lugosi. « On the performance of clustering in Hilbert spaces ». In : *IEEE Transactions on Information Theory* 54.2 (2008), p. 781-790. URL : https://hal.archives-ouvertes.fr/hal-00290855.

[McA03] David A. McAllester. « Simplified PAC-Bayesian margin bounds ». In : *Learning Theory and Kernel Machines, COLT 2003*. Sous la dir. de B. Schölkopf et M.K Warmuth. Lecture Notes in Artificial Intelligence. 2003, p. 203-215.

[LS02] J. Langford et J. Shawe-Taylor. « PAC-Bayes & Margins ». In : *Advances in Neural Information Processing Systems*. Cambridge : MIT Press, 2002, p. 423-430.

Liste de publications

[AC21] Gautier Appert et Olivier Catoni. *New bounds for k-means and information k-means*. Soumis. 2021, p. 1-35. arXiv : 2101.05728v2.

[COZ20] Olivier Catoni, Miquel Oliu-Barton et Bruno Ziliotto. « Constant payoff in zero-sum stochastic games ». In : *Ann. Inst. H. Poincaré Probab. Statist.* (2020). to appear, p. 1-13. url : https://imstat.org/journals-and-publications/annales-de-linstitut-henri-poincare/annales-de-linstitut-henri-poincare-accepted-papers/.

[CG17a] O. Catoni et I. Giulini. « Dimension free PAC-Bayesian bounds for the estimation of the mean of a random vector ». In : *the Nips 2017 Workshop : (Almost) 50 shades of Bayesian learning : PAC-Bayesian trends and insights*. 2017, p. 1-4. eprint : https://bguedj.github.io/nips2017/pdf/PAC-Bayes_2017_paper_1.pdf. url : https://bguedj.github.io/nips2017/50shadesbayesian.html.

[CG17b] Olivier Catoni et Ilaria Giulini. *Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression*. submitted preprint. 2017, p. 1-40. arXiv : 1712.02747v2.

[Cat16] Olivier Catoni. *PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design*. submitted preprint. 2016, p. 1-81. arXiv : 1603.05229.

[CM16] Olivier Catoni et Thomas Mainguy. *Markov substitute processes : a new model for linguistics and beyond*. 2016, p. 1-22. arXiv : 1603.07850v1.

[Cat15a] Olivier Catoni. « Comment: Transductive PAC-Bayes Bounds Seen as a Generalization of Vapnik–Chervonenkis Bounds ». In : *Measures of Complexity: Festschrift for Alexey Chervonenkis*. Sous la dir. de Vladimir Vovk, Harris Papadopoulos et Alexander Gammerman. *(discussion of Making Vapnik-Chervonenkis Bounds Accurate by Léon Bottou)*. Cham : Springer International Publishing, 2015, p. 157-160.

[Cat15b] Olivier Catoni. « PAC-Bayes Bounds for Supervised Classification ». In : *Measures of Complexity: Festschrift for Alexey Chervonenkis*. Sous la dir. de Vladimir Vovk, Harris Papadopoulos et Alexander Gammerman. Cham : Springer International Publishing, 2015, p. 287-302.

[Cat14] Olivier Catoni. *PAC-Bayes learning bounds*. Lecture Notes. IFCAM Summer school, Bangalore, 2014, p. 1-29. url : http://www.math.ens.fr/~catoni/homepage/homepage-en.html.

[Cat13] Olivier Catoni. *Statistical learning*. Lecture notes, see author's homepage. (the three versions have some parts in common, and will be merged at some point). 2011, 2012, 2013, p. 1-34, 1-34, 1-28.

[CM13] Olivier Catoni et Thomas Mainguy. *Toric grammars : a new statistical approach to natural language modeling*. 2013, p. 1-41. arXiv : 1302.2569.

[Cat12] Olivier Catoni. « Challenging the empirical mean and empirical variance: a deviation study ». In : *Ann. Inst. Henri Poincaré* 48.4 (2012). Second augmented, improved and completely rewritten version of [Cat09], p. 1148-1185.

[AC11a] J.-Y. Audibert et O. Catoni. « Robust linear least squares regression ». In : *Ann. Stat.* 39.5 (2011), p. 2766-2794.

[AC11b] J.-Y. Audibert et O. Catoni. « Supplement to "Robust linear least squares regression." » In : *Ann. Stat.* (2011), p. 1-19. doi : 10.1214/11-AOS918SUPP.

[AC10] Jean-Yves Audibert et Olivier Catoni. *Linear regression through PAC-Bayesian truncation*. preprint. (revised in 2011). 2010, p. 1-40. arXiv : 1010.0072v2.

[Cat09] Olivier Catoni. *High confidence estimates of the mean of heavy-tailed real random variables*. preprint. 2009, p. 1-40. arXiv : 0909.5366.

[Cat07] Olivier Catoni. *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. T. 56. IMS Lecture Notes Monograph Series. pages i-xii, 1-163. Institute of Mathematical Statistics, 2007.

[Cat06] Olivier CATONI. « Théorie statistique de l'apprentissage ». In : *Images des Mathématiques 2006 — CNRS* (2006), p. 31-39.

[Cat04a] Olivier CATONI. *Improved Vapnik Cervonenkis bounds*. preprint. Included in revised form into [Cat07]. 2004, p. 1-22.

[Cat04b] Olivier CATONI. *Statistical Learning Theory and Stochastic Optimization, Lectures on Probability Theory and Statistics, École d'Été de Probabilités de Saint-Flour XXXI – 2001*. T. 1851. Lecture Notes in Mathematics. pages 1–269. Springer, 2004.

[Cat03a] Olivier CATONI. *A PAC-Bayesian approach to adaptive classification*. preprint. submitted first to the Annals of Statistics, eventually the starting point of [Cat07], published in the IMS Lecture Notes series. 2003, p. 1-72.

[Cat03b] Olivier CATONI. « Laplace transform estimates and deviation inequalities ». In : *Ann. Inst. H. Poincaré Probab. Statist.* 39.1 (2003), p. 1-26.

[Cat02] Olivier CATONI. « Data compression and adaptive histograms ». In : *Foundations of Computational Mathematics, Proceedings of the Smalefest 2000*. Sous la dir. de F. CUCKER et J.M. ROJAS. World Scientific, 2002, p. 35-60.

[Cat01] Olivier CATONI. *Randomized estimators and empirical complexity for pattern recognition and least square regression*. preprint. Included in [Cat04b]. 2001, p. 1-33. URL : http://www.proba.jussieu.fr.

[Cat00] Olivier CATONI. *Gibbs estimators*. preprint. Essentially included in [Cat04b]. 2000, p. 1-23. URL : http://www.proba.jussieu.fr.

[CCX00] Olivier CATONI, Dayue CHEN et Jun XIE. « The loop erased exit path and the metastability of a biased vote process ». In : *Stochastic Process. Appl.* 86 (2000), p. 231-261.

[Cat99a] Olivier CATONI. « Simulated Annealing Algorithms and Markov Chains with Rare Transitions ». In : *Séminaire de Probabilités XXXIII*. T. 1709. Lecture Notes in Math. *(in French 1995, English revised translation at http://www.dmi.ens.fr/preprints 1997, published augmented revision 1999)*. Springer, 1999, p. 69-119.

[Cat99b] Olivier CATONI. *Universal aggregation rules with sharp oracle inequalities*. preprint. essentially included in [Cat04b]. Revised et augmented from [Cat97], 1999, p. 1-37.

[Cat98a] Olivier CATONI. « Solving Scheduling Problems by Simulated Annealing ». In : *SIAM J. Control Optim.* 36.5 (1998). (electronic), p. 1539-1575.

[Cat98b] Olivier Catoni. « The Energy Transformation Method for the Metropolis Algorithm Compared with Simulated Annealing ». In : *Probab. Theory Related Fields* 110.1 (1998), p. 69-89.

[CC98] Olivier Catoni et Cécile Cot. « Piecewise constant triangular cooling schedules for generalized simulated annealing algorithms ». In : *Ann. Appl. Probab.* 8.2 (1998), p. 375-396.

[Cat97] Olivier Catoni. *A mixture approach to universal model selection*. preprint LMENS-97-30. First draft of [Cat99b]. Oct. 1997, p. 1-19. URL : http://www.dmi.ens.fr/preprints.

[CC97] Olivier Catoni et Raphael Cerf. « The exit path of a Markov chain with rare transitions ». In : *ESAIM: Probability and Statistics* (1997), p. 95-144.

[Cat96a] Olivier Catoni. « A New Inequality for the Free Energy of the Sherrington-Kirkpatrick Spin Glass Model ». In : *Proceedings of the 1995 Workshop on Large Deviations and Statistical Mechanics, Oct. 20-21, SFB, Bielefeld, Germany*. Sous la dir. de P. Eichelsbacher et M. Löwe. SFB-preprint-series, 1996, p. 1-8.

[Cat96b] Olivier Catoni. « Metropolis, Simulated Annealing and I.E.T. Algorithms: Theory and Experiments ». In : *Journal of Complexity 12, special issue on the conference Foundation of Computational Mathematics, January 5-12 1997, Rio de Janeiro* (déc. 1996), p. 595-623.

[Cat96c] Olivier Catoni. « The Legendre transform of two replicas of the Sherrington-Kirpatrick spin glass model ». In : *Probab. Theory Relat. Fields* 104 (1996), p. 369-392.

[Cat94] Olivier Catoni. « Energy Transforms for Metropolis and Simulated Annealing Algorithms ». In : *Proceedings of the Twelfth Prague Conference on Information Theory, Statistical Decision Functions and Random Processes - Aug. 29, Sept. 2*. 1994, p. 1-6.

[Cat92a] Olivier Catoni. « Exponential Triangular Cooling Schedules for Simulated Annealing Algorithms: A Case Study ». In : *Applied Stochastic Analysis- Proceedings of a US-French Workshop, Rutgers University, New Brunswick, N.J., April 29- May 2, 1991, Lecture Notes in Control and Information Sciences 177*. Sous la dir. d'Ocone D. Karatzas I. Springer-Verlag, 1992, p. 74-89.

[Cat92b] Olivier Catoni. « Rates of Convergence for Sequential Annealing: a Large Deviation Approach ». In : *Simulated Annealing: Parallelization Techniques*. Sous la dir. de Robert Azencott. John Wiley et Sons, 1992. Chap. 3, p. 25-35.

[Cat92c] Olivier CATONI. « Rough Large Deviation Estimates for Simulated Annealing: Application to Exponential Schedules ». In : *The Annals of Probability* 20.3 (1992), p. 1109-1146.

[CT92] Olivier CATONI et Alain TROUVÉ. « Parallel Annealing by Multiple Trials: A Mathematical Study ». In : *Simulated Annealing: Parallelization Techniques*. Sous la dir. de Robert AZENCOTT. John Wiley et Sons, 1992. Chap. 9, p. 129-143.

[Cat91a] Olivier CATONI. « Applications of Sharp Large Deviations Estimates to Optimal Cooling Schedules ». In : *Ann. Inst. Henri Poincaré* 27.4 (1991), p. 463-518.

[Cat91b] Olivier CATONI. *Learning Algorithms for Pattern Recognition on Half-Tone Binary Images*. unpublished. 1991, p. 1-32.

[Cat91c] Olivier CATONI. « Sharp Large Deviations Estimates for Simulated Annealing Algorithms ». In : *Ann. Inst. Henri Poincaré* 27.3 (1991), p. 291-383.

[Cat90a] Olivier CATONI. « Etude Asymptotique des algorithmes de recuit simulé (Asymptotics of simulated annealing algorithms) ». Thèse de doct. Université Paris-Sud Orsay, 170 pages, 1990.

[Cat90b] Olivier CATONI. « Image Restoration by Stochastic Dichotomic Reconstruction of Contour Lines ». In : *Stochastic Models, Statistical Methods, and Algorithms in Image Analysis, Lecture Notes in Statistics No 74*. Sous la dir. de Frigessi A. BARONE P. et Piccioni M. Springer-Verlag, 1990, p. 101-116.

[CG89] Olivier CATONI et Isabelle GAUDRON. « Détection de contours par seuillage adaptatif et restauration stochastique d'images binaires ». In : *Proceedings of the Second Annual Conference on Computer Graphics in Paris, Pixim 89, André Gagalowicz ed., ACM SIGGRAPH FRANCE*. Hermes, 1989, p. 341-355.

[Cat88] Olivier CATONI. « Grandes déviations et décroissance de la température dans les algorithmes de recuit. » In : *C.R. Acad. Sci. Paris* 1.307 (1988), p. 535-538.