
RAPPORT D'ACTIVITÉ

(en vue de solliciter une promotion à la première classe)

OLIVIER CATONI

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 deuxième classe recruté le premier septembre 2000, affecté au Département de Mathématiques et Applications de l'École Normale Supérieure (U.M.R. 8553).

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 1^{er} 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 :

“Étude asymptotique des algorithmes de recuit simulé”.

année 1991 :

Prix IBM jeunes chercheurs en mathématiques.

le 1^{er} 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 1^{er} 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 1^{er} septembre 2000 :

Promotion au grade de directeur de recherche.

le 1^{er} 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 1^{er} 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 quatre ans, avec reconduction de ma fonction de responsable.

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 "Stochastic 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. Azenkott - 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. (Zavidovique). 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 Transforms 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 statistique (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 inti-

tulé “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 Grafigne – Prisme, Université Paris V, Marie-Anne Gruet – INRA Biométrie, Pascal Massart – Université Paris XI, Alain Trounev – Université Paris XIII)

Organisation en collaboration avec Thierry Bodineau, Francis Comets, Dominique 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-Bayésiens 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 Grünwald, François Laviolette et John Shawe-Taylor, dans le cadre du réseau européen Pascal 2. Exposé intitulé « Robust PAC-Bayes bounds ». Video disponible sur internet : http://videolectures.net/pacbayesian_catoni_rpbb/

2. CONTRIBUTIONS SCIENTIFIQUES

2.1. 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.

2.2. THÉORIE STATISTIQUE DE L'APPRENTISSAGE.

2.2.1. *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 sont disponibles sous la forme de deux publications et deux prépublications : 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 [AC10] en collaboration avec Jean-Yves Audibert, dont la seconde version (78 pages) a été placée sur HAL et ArXiv en juillet dernier, portant sur le problème de la régression aux moindres carrés en révision favorable aux *Annals of Statistics*, ainsi

que la seconde version, complètement remaniée, d'un article proposant de nouveaux M-estimateurs de la moyenne et de la variance d'une variable aléatoire réelle présentant de meilleures performances que la moyenne et la variance empiriques dans le cas où l'échantillon n'est pas gaussien [Cat10] (53 pages). Mon cours à Saint-Flour est disponible sur ma page web, la monographie éditée par l'IMS est disponible sur ArXiv dans sa version publiée.

Cette dernière [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. Il y a donc toute une période pendant laquelle j'ai abandonné de fait la publication d'articles, préférant exposer les résultats de mes recherches dans des monographies, dont le format correspondait sans doute mieux à ce que j'avais en tête.

J'ai écrit d'autre part une présentation de la théorie statistique de l'apprentissage en quelques pages à destination de la brochure du CNRS « Images des mathématiques », qui pourrait, bien qu'elle soit un peu ancienne maintenant, servir d'introduction à ce rapport.

2.2.2. 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 télescope 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 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, malgré leur peu de sophistication mathématique (leur auteur relève d'une tradition plus proche de l'électrical engineering que des mathématiques), 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).

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'utilisation 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 j'ai essayé de montrer dans mon dernier ouvrage [Cat07] la souplesse et la polyvalence. (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 $[Z_1(\omega), \dots, Z_N(\omega)]$ de variables indépendantes (ou échangeables, ou partiellement échangeables, plusieurs types d'hypothèses sont possibles), et une famille $\ell : \Theta \times \mathcal{Z} \rightarrow \mathbb{R}_+$ 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 dernière 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 $\ell_{\theta}(Z_i) = \mathbb{1}[Y_i \neq f_{\theta}(X_i)]$ décrit l'erreur de classification commise par une règle de classification $f_{\theta} : \mathcal{X} \rightarrow \mathcal{Y}$. J'ai confié à mon thésard Pierre Alquier le soin d'étudier le cas d'une fonction de perte ℓ bornée générale. À la faveur d'une collaboration récente avec Jean-Yves Audibert [AC10], je me suis penché sur le cas où ℓ est non bornée et vérifie 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, ou, pour dire les choses avec emphase, « de la perception à une représentation symbolique »).

On souhaite minimiser en θ l'espérance du risque empirique $\mathbb{E}[r(\theta)]$. 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

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

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

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

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 (Minimum 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'analogie 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), ainsi que 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).

En ce qui concerne les bornes relatives, j'ai introduit il y a quatre ans 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 $\mathbb{E}[r(\theta)]$ 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.

Plus récemment, 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 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 sont connues. L'adaptation à des hypothèses

de moments polynomiaux plus faibles est traitée dans ma collaboration récente [AC10] avec Jean-Yves Audibert.

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 Cervonenkis.

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 tout ce qui a été démontré dans le cas inductif, comme cela est fait dans la section 3.1 de [Cat07]. 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 [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 obtenu ces résultats dans l'ordre inverse de celui choisi pour les exposer dans notre article.

Nous avons commencé par nous affranchir de toute hypothèse de moment exponentiel sur la fonction de risque. 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 5.5 [AC10, page 33] 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.

Dans ce cas, il est en effet possible de faire des calculs complètement explicites en utilisant des lois a priori et a posteriori gaussiennes sur l'espace des paramètres.

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 naturel d'utiliser la transformation qui revient tout simplement à tronquer le développement de Taylor de la fonction exponentielle à l'ordre deux, une idée simple qui n'avait pourtant pas été exploitée auparavant (si ce n'est par mes élèves à mon instigation).

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

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

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

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

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 ob-

tient l'inégalité suivante : Pour tout paramètres $\lambda \in \mathbb{R}_+$, $\alpha \in \mathbb{R}$, avec probabilité au moins $1 - \varepsilon$, pour tout $\theta_1 \in \Theta$,

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

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

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

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

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

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 2.1 *Supposons que*

$$\begin{aligned} & \mathbb{E}(\|X\|^4) < +\infty, \\ \text{et} \quad & \mathbb{E}[\|X\|^2 (\langle \theta, X \rangle - Y)^2] < +\infty. \end{aligned}$$

Soit ν_1, \dots, ν_d les valeurs propres de la matrice de Gram $Q = \mathbb{E}(XX^T)$ 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 $\varepsilon > 0$, il existe une taille d'échantillon n_ε telle que pour tout $n \geq n_\varepsilon$, avec probabilité au moins $1 - \varepsilon$,

$$\begin{aligned} R_\lambda(\hat{\theta}) & \leq R_\lambda(\theta^*) + 30 \tilde{\mathbb{E}}[(\langle \theta^*, X \rangle - Y)^2] \frac{D}{n} \\ & \quad + 1000 \sup_{v \in \mathbb{R}^d} \tilde{\mathbb{E}}_v[(\langle \theta^*, X \rangle - Y)^2] \frac{\log(3/\varepsilon)}{n}, \end{aligned}$$

où

$$\begin{aligned}\tilde{\mathbb{E}}(\mathbf{Z}) &= \frac{\mathbb{E}(\|(\mathbf{Q} + \boldsymbol{\lambda})^{-1}\mathbf{X}\|^2\mathbf{Z})}{\mathbb{E}(\|(\mathbf{Q} + \boldsymbol{\lambda})^{-1}\mathbf{X}\|^2)}, \\ \tilde{\mathbb{E}}_v(\mathbf{Z}) &= \frac{\mathbb{E}(\langle v, \mathbf{X} \rangle^2 \mathbf{Z})}{\mathbb{E}(\langle v, \mathbf{X} \rangle^2)}.\end{aligned}$$

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 alors ajouter des hypothèses un peu plus contraignantes sur la forme de l'ellipsoïde d'inertie des entrées et supposer que les sorties ont un moment d'ordre quatre :

$$\mathbb{E}[(\langle \theta^*, \mathbf{X} \rangle - Y)^4] < +\infty$$

pour établir une borne non asymptotique sur le risque de généralisation valable pour des niveaux de confiance $\varepsilon \geq 2/n$ et ayant une forme similaire à celle que nous venons de donner, c'est-à-dire en $\frac{d - \log(\varepsilon)}{n}$ aux constantes près.

Cette dernière borne a la propriété remarquable de montrer que la queue de distribution de l'excès de risque est sous exponentielle jusqu'au quantile $2/n$, et ce même en l'absence de tout moment exponentiel des sorties !

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. Nous avons, avec Jean-Yves Audibert, progressé durant cette dernière année à l'occasion de la révision de [AC10]. Elle nous a donné l'occasion d'introduire 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 $\psi : \mathbb{R} \rightarrow \mathbb{R}$ croissante bornée possédant la propriété fondamentale suivante :

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

Le point technique nouveau découvert cette année 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 [AC10, page 16], qui permet d'obtenir une convergence en $\frac{d - \log(\epsilon)}{n}$ 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 L_2 de la base dans le cas de la régression fonctionnelle) qui les impliquent, ainsi qu'expliqué dans le paragraphe 3.2 de [AC10, page 17].

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. Nous avons donc progressé depuis l'année dernière en transformant des propositions de nature essentiellement théoriques concernant la régression quadratique en algorithmes permettant d'obtenir en pratique de meilleurs résultats que l'estimateur des moindres carrés ordinaire dans certaines situations où le bruit n'est pas gaussien.

Un estimateur moins réaliste, l'application directe de la technique de troncature randomisée générale au cas quadratique, permet d'obtenir des vitesses en d/n sous des hypothèses encore plus faibles.

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.

Cela entraîne en particulier que 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 que je comparais un peu maladroitement semble-t-il au cadre de la statistique robuste. Cette première version de [Cat10], intitulée « High confidence estimates of the mean of heavy-tailed real random variables » a reçu, il faut le dire, un accueil très négatif des référés des Annales de l'IHP auxquelles je l'avais soumis. Ce refus complet d'admettre qu'il puisse être possible d'améliorer l'estimateur de la moyenne empirique, au lieu de me décourager, m'a incité au contraire à affûter

mes calculs d'une part et à compléter mon étude théorique par des résultats expérimentaux d'autre part. L'article [Cat10] que je resoumets en même temps que ce rapport sera je pense plus difficile à écarter dédaigneusement que sa première version. Profitant en partie des critiques des référés (dont certaines étaient pertinentes, même si, me semble-t-il, leur mauvaise volonté à admettre l'intérêt du sujet manifestait un certain manque de perspicacité), j'ai pu simplifier les estimateurs proposés, les remplaçant par des M-estimateurs plus classiques, améliorer la précision de l'ensemble des bornes théoriques et fournir des exemples pratiques montrant par l'expérience les améliorations possibles. J'ai pu ainsi montrer que dans le cas le pire, dans le modèle constitué de tous les échantillons de variance fixée, les quantiles des déviations de la moyenne empirique par rapport à l'espérance étaient sous optimaux pour un échantillon de taille cent et des niveaux de probabilité supérieurs à 90%. Le même type de sous optimalité, pour des niveaux de confiance plus élevés, peut être mis en évidence dans le modèle constitué par les échantillons de kurtosis fixée. Dans ces modèles très larges, il se trouve qu'il est possible de construire des M-estimateurs (utilisant dans un cas la valeur de la variance et dans l'autre celle de la kurtosis), dont les déviations sont du même ordre que celles 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).

Les résultats expérimentaux que j'ai obtenus très récemment sont encore plus frappants : il suffit de sortir du modèle gaussien en considérant des lois faites du mélange de deux gaussiennes pour obtenir une amélioration uniforme de la fonction quantile des déviations à la moyenne, avec, dans certain cas, une amélioration supérieure à 25% au niveau 90% pour des échantillons de taille 100. Il semble donc bien, même si la partie théorique de mon étude n'établit que des résultats plus faibles, que d'un point de vue pratique, la moyenne empirique ne soit pas même un estimateur admissible en dehors du modèle gaussien (puisque'il semble possible de construire un M-estimateur dont tous les quantiles soient essentiellement égaux ou meilleurs pour toute distribution de l'échantillon). L'autre résultat expérimental un peu inespéré semble montrer que la précision de l'estimateur empirique non biaisé classique de la variance est suffisante pour être réinjectée 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.

D'un côté, les lecteurs de ces lignes trouveront peut-être (comme les référés des Annales de l'IHP !) que cette question de l'estimation de la moyenne manque de panache : elle pourrait passer pour un problème élémentaire, résolu depuis des lustres, ne pouvant donner lieu qu'à des perfectionnements marginaux et peu enthousiasmants. D'un autre, il me semble que c'est tout de même une question fondamentale en statistique, et que toute amélioration dans ce domaine est bonne à prendre et potentiellement porteuse d'améliorations dans des situations plus com-

pliquées. Il me semble en particulier que ces M-estimateurs de la moyenne, qui ont de plus le mérite de pouvoir être calculés avec une précision suffisante par des schémas itératifs à convergence rapide (deux itérations, en pratique), devraient conduire à des algorithmes de filtrage améliorés dans le domaine du traitement du signal et des images. Ce qui me plaît aussi dans ce résultat, c'est qu'il pointe, d'après moi, l'intérêt qu'il y a pour les statistiques à sortir des modèles gaussiens d'une part et des études asymptotiques un peu trop hâtives, d'autre part, deux défauts que l'on peut reprocher aux statistiques traditionnelles, et à la modélisation stochastique en générale, qui a tendance, par habitude culturelle et facilité technique, à utiliser des modèles gaussiens à tort et à travers.

2.3. PROJETS EN COURS.

2.3.1. Bornes Pac-Bayésiennes. J'ai pu constater lors du Workshop organisé à Londres en mars dernier que la théorie PAC-Bayésienne gagnait en reconnaissance. D'un point de vue technique, nous avons pu, en collaboration avec Jean-Yves Audibert, franchir une nouvelle étape en trouvant le moyen, à travers le lemme 6.7 [AC10, page 51], ou à travers le lemme 4.3 [Cat10, page 19], qui améliore un peu les constantes, d'appliquer à des estimateurs classiques les techniques de perturbation du paramètre par une loi a posteriori qui sont à la base de l'approche PAC-Bayésienne. Après un retour aux sources de la statistique, avec l'estimation de la moyenne, j'envisage d'étendre l'étude expérimentale des nouveaux estimateurs tronqués mis en évidence, tant pour l'estimation de la moyenne que pour celle de la régression. Il serait intéressant par exemple d'explorer plus avant l'impact pratique du choix de la fonction d'influence, pour lequel la théorie laisse une certaine latitude. J'aimerais aussi explorer, tant du point de vue théorique que pratique, l'application de ce type d'estimateurs à des problèmes de filtrage. Dans un autre ordre d'idée, David McAllester m'encourage à tenter d'appliquer mon approche aux bornes PAC-Bayésiennes liées à la marge des Support Vector Machine, telles qu'elles sont exposées dans son article intitulé « Simplified PAC-Bayesian Margin Bounds » (COLT 2003). De plus, les applications à la régularisation L_1 (algorithme du Lasso) avec design aléatoire sont toujours à l'ordre du jour (en collaboration avec Pierre Alquier). Cela fait beaucoup de pistes, je ne progresserai certainement pas sur toute dans l'année qui vient, mais j'en explorerai à n'en pas douter certaines.

2.3.2. Linguistique computationnelle. A l'occasion du co-encadrement (en collaboration avec Edward Stabler, professeur à UCLA) du stage de M2 de Thomas Mainguy, dont j'encadrerai la thèse à partir de la fin septembre, je me suis familiarisé avec les enjeux de la linguistique computationnelle. L'objectif fixé pour la thèse est de proposer de nouveaux modèles de grammaires stochastiques inspi-

rés des grammaires minimalistes de Stabler dont on puisse estimer les catégories syntaxiques et les règles à partir d'un corpus de textes (en utilisant un critère de compression).

2.3.3. *Vision et apprentissage.* Comme expliqué dans mes rapports précédents, en liaison avec la création d'une équipe INRIA au DMA de l'ENS, je poursuis parallèlement aux recherches ci-dessus un programme de recherche plus appliqué concernant la classification d'images dites « naturelles ». Ce programme, est resté un peu en panne depuis quelques temps, mais ce n'est pas forcément un mal. En fait, après avoir testé des détecteurs de contours qui me paraissaient satisfaisants en pratique, je suis resté un peu perplexe quant à la façon de les utiliser. Le fait que la suite des opérations ait été repoussée du fait des sujets décrits ci-dessus n'est donc pas forcément une mauvaise chose. Mon agenda est un peu chargé, avec la perspective d'une collaboration avec David McAllester qui me presse de travailler sur les bornes de marge PAC-Bayésiennes pour les Support Vector Machine et la direction d'une thèse sur la linguistique computationnelle qui m'invite à me plonger dans les grammaires stochastiques. Néanmoins, je continue à accumuler des notes sur l'analyse d'images et la reconnaissance des formes, ce qui, bien évidemment ne me permet pas de remplir ce rapport d'activité d'une façon très convaincante, mais continue à faire pour moi de ce thème de recherche un sujet qui me tient à cœur. Disons simplement que, dans le cadre de l'exploitation des champs de directions que j'arrive à extraire des images, je suis à la recherche d'une notion de configurations symétriques qui ait la bonne fréquence d'apparition : suffisante pour pouvoir décrire des formes quelconques et suffisamment faible pour que la complexité de cette représentation n'explode pas.

Je reprends pour mémoire dans la suite de ces lignes la description du thème que j'avais faite l'an passé.

La classification d'images représente pour moi un défi complémentaire stimulant, parce qu'elle bouscule d'emblée le cadre de la classification supervisée auquel il est tentant de se restreindre dans les études théoriques sur l'apprentissage statistique. En effet, les images sont 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'en suis venu pour contourner cet obstacle à changer de méthode de travail et à me fixer comme objectif le développement d'une plate-forme logicielle de démonstration des méthodes que j'élabore, me permettant de faire précéder leur description théorique d'une validation expérimentale. Il faut dire aussi que j'ai fait ces dernières années de la fréquentation des images numériques l'un de mes loisirs préférés, à travers une activité de plus en plus soutenue de photographe amateur, qui m'a conduit à écrire un logiciel d'édition de photos numériques, mettant en œuvre des techniques de gestion des contrastes de mon cru, logiciel qui m'a conduit à déposer auprès de la Direction de la Politique Industrielle une déclaration d'invention à titre privé. (J'estimais en effet que la photographie ne faisait pas partie de mon activité professionnelle, point de vue que le CNRS, à travers sa DPI, a bien voulu partager, ce dont je lui suis reconnaissant). La fréquentation d'exemples variés d'images numériques est très stimulante, tester mes idées sur elles me permet de les soumettre très tôt à la sanction de l'efficacité pratique, et je me suis aperçu que cela m'apportait une aide précieuse pour guider ma réflexion.

La fabrication d'un logiciel prototype m'ouvre aussi la possibilité de partager mes résultats autrement : mes publications ont l'inconvénient d'être techniques, ce qui en restreint l'accès à un public étroit de mathématiciens motivés et compétents, alors qu'à travers un logiciel de démonstration, pour peu que l'on se donne la peine de rendre l'interface suffisamment conviviale, on peut espérer toucher un public d'utilisateurs plus large, qui n'a pas forcément besoin de prendre connaissance des complications techniques enfouies dans le code. Un logiciel prototype représente de ce fait un vecteur de transfert technologique vers l'industrie plus réaliste qu'un article de recherche.

Une heureuse coïncidence a voulu que ces considérations et cette évolution de ma pratique se conjuguent avec une proposition venue de la rue d'Ulm de

revenir au DMA (que j'avais quitté en 1998), pour y superviser la création d'une équipe INRIA consacrée à l'apprentissage statistique. C'est bien évidemment avec une grande joie que j'ai saisi cette opportunité : j'espère très vivement pouvoir contribuer à travers ce projet au rapprochement entre l'INRIA et le CNRS ainsi qu'à l'interface entre mathématiques et informatique.

Décrivons maintenant plus précisément l'état des lieux de mes recherches. Je n'ai pas encore publié sur la classification non supervisée. Néanmoins j'ai fait un effort de réflexion ainsi qu'un effort logiciel important et dispose d'une plateforme de traitement d'images opérationnelle, (mon expérience de programmation d'un logiciel d'édition photo m'a bien aidé, j'ai en particulier récupéré l'interface), sur laquelle j'ai pu valider 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. Disons tout de suite que j'associe dans mon propos classification non supervisée et codage par mélange de codes, ou encore modélisation stochastique par un mélange de lois. Essayons d'avancer quelques arguments en faveur de cette démarche qui peut paraître à première vue peu naturelle :

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 ambiguë, 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.

J'ai choisi pour réfléchir à ces problèmes de mise en correspondance et de classification non supervisée de m'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 capable. 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 de création d’équipe INRIA disponible 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).

Bien que la transformation multi-échelle G'_α que j’ai testée fonctionne en pratique, je n’en suis pas complètement satisfait. Je n’ai pas l’impression en effet (impression purement visuelle à ce stade), qu’elle détecte suffisamment d’intersections. Je pense donc plutôt arrêter les traitements génériques à la détection des contours et de leurs orientations, et enchaîner à ce stade là sur une phase d’apprentissage non supervisée dans laquelle une loi de codage des contours serait estimée

à partir de jeux de données spécifiques à l'application envisagée.

La partie validée expérimentalement (détection de contours et d'orientations) pourrait être isolée du reste et faire l'objet d'une publication, néanmoins, je préférerais l'inclure dans un ensemble plus vaste comportant au moins une phase d'apprentissage non supervisé.

J'ai en particulier en tête une façon particulière de construire des mélanges de loi en développant en « clusters » (ou en chaos si on préfère) un produit de mélanges de lois. Je l'ai appliquée pour construire mon modèle de contours et l'exposerai dans un cadre général si elle s'avère efficace dans des circonstances variées et que l'exemple des contours ne soit pas une réussite isolée.

2.3.4. Aspects institutionnels. La création au sein du DMA de l'ENS de l'équipe projet INRIA CLASSIC consacrée aux statistiques et à l'apprentissage machine a abouti comme prévu. J'en assume la responsabilité. Pour l'instant, nous avons demandé et obtenu un budget annuel de 7000 euros que nous n'avons pas commencé à dépenser. Cet affichage d'un lien avec l'INRIA me semble présenter deux avantages : manifester aux yeux des élèves de l'ENS le lien entre mathématiques et applications et fédérer les statistiques au sein du DMA.

3. ENSEIGNEMENT, FORMATION ET DIFFUSION DE LA CULTURE SCIENTIFIQUE

3.1. 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.

J'assurerai cette année (2010-2011) 10h de cours dans le cadre du cours de 50h consacré à l'apprentissage statistique et destiné à la filière math-info de la FIMFA de l'ENS (les autres modules de ce cours du second semestre de la première année étant assurés par S. Arlot, J.Y. Audibert, F. Bach et G. Stoltz).

Je participe aussi en compagnie de Gérard Biau et de Gilles Stoltz au premier trimestre à l'animation du groupe de travail sur l'estimation de la densité destiné aux élèves de deuxième année de la FIMFA.

3.2. FORMATION. J'ai coordonné en 1998 l'organisation d'un colloque intitulé « Théorie de l'Information, Statistique adaptative et Reconnaissance des formes » qui s'est tenu du 7 au 11 décembre au CIRM, à Marseille Luminy.

3.3. DIRECTION DE THÈSES. J'ai dirigé cinq thèses, celles de Cécile Cot, Gilles Blanchard, Jean-Philippe Vert, Jean-Yves Audibert et Pierre Alquier, dont une en co-direction avec Alain Trouvé (Gilles Blanchard).

J'ai co-dirigé l'année passée le stage de M2 de Thomas Mainguy (en collaboration avec E. Stabler d'UCLA), qui commence actuellement une thèse sous ma direction, consacrée à l'introduction de techniques statistiques en linguistique computationnelle. Il s'agit, comme décrit plus haut dans mes projets de recherche, de concevoir des modèles de grammaires stochastiques inspirés des grammaires minimalistes de Stabler et permettant un apprentissage des catégories syntaxiques à partir d'un corpus de textes en s'appuyant sur un critère de compression.

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 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 décorréélées 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.

Jean-Yves Audibert a soutenu le 29 juin 2004 une thèse portant sur l'agrégation d'estimateurs en norme L^2 , le contrôle empirique de la variance en classification PAC-Bayésienne 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 est aussi membre à temps partiel de l'équipe Willow de l'INRIA, située au Département d'Informatique de l'ENS.

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 ».

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.

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.

3.4. DIFFUSION DE L'INFORMATION SCIENTIFIQUE. Rédaction d'un article de six pages intitulé « Théorie statistique de l'apprentissage » pour la brochure du CNRS « Images des Mathématiques 2006 ».

4. TRANSFERT TECHNOLOGIQUE, RELATIONS INDUSTRIELLES ET VALORISATION

J'ai exercé au début de ma carrière au CNRS une activité de consultant (en collaboration avec mon directeur de thèse, Robert Azencott), qui a porté sur la conception d'un logiciel d'analyse de signaux transitoires (détection d'anomalies) pour la société Miriad. J'ai aussi participé à une étude menée pour Citroën à l'occasion de laquelle j'ai écrit un logiciel de planification d'emplois du temps, que j'ai ensuite repris dans le cadre d'une recherche plus théorique pour aboutir à un article [Cat98b] publié par le SIAM J. Control Optim. (1998), intitulé « Solving Scheduling Problems by Simulated Annealing ».

Comme expliqué dans mon projet de recherche, je suis en train de développer une plate-forme logiciel de test et de démonstration de mes travaux dans le

domaine de l'imagerie, qui devrait pouvoir servir au transfert technologique de certains aspects de mes recherches en cours ou à venir.

5. RESPONSABILITÉS COLLECTIVES ET MANAGEMENT DE LA RECHERCHE

J'ai mis en place à mon arrivée au LPMA, en 1998, un serveur de prépublications informatisé, dont la gestion a été ensuite progressivement confiée à son bibliothécaire, Monsieur Philippe Macé. J'ai aussi réalisé la migration de la base de données de ce laboratoire vers HAL. J'en ai été un temps correspondant valorisation, de plus j'ai fait partie de sa commission informatique, et me suis occupé des achats et des installations d'ordinateurs, du temps où nous n'avions pas d'ingénieur système pour le faire.

Je suis membre de la commission des thèses de l'Universités Paris 6 depuis mars 2007, ainsi que du comité éditorial de la collection Mathématiques et Applications publiée par Springer sous l'égide de la SMAI, depuis l'automne 2007.

Ma mutation au DMA en septembre 2008 avait pour objet la création d'une équipe INRIA. La création de l'équipe a eu lieu en juillet 2009 et sa transformation en équipe projet en juillet 2010. Je cite la décision de l'INRIA du 3 juillet 2010 : « Art 1 : Est créée auprès du centre de recherche INRIA Paris-Rocquencourt l'équipe projet INRIA CLASSIC - *Computational Learning, Aggregation, Supervised Statistical Inference, and Classification* au sein du thème *Optimisation, apprentissage et méthodes statistiques*, à compter du 1^{er} janvier 2010 et jusqu'au 31 décembre 2014. Cette équipe projet est commune avec l'ENS de Paris et le CNRS. Art 2 : Est nommé responsable de l'équipe-projet INRIA Monsieur Olivier Catoni. ». Comme expliqué plus haut, je vois surtout dans la formation de cette équipe un moyen d'afficher les liens entre mathématiques et applications et de promouvoir la filière math-info aux yeux des élèves de l'ENS.

6. MOBILITÉS

Géographique : de la rue d'Ulm à l'Université Paris 6, et retour dix ans après dans le but d'y créer une équipe INRIA consacrée à l'apprentissage statistique. Thématique : des probabilités appliquées à la théorie statistique de l'apprentissage, ainsi qu'à ses applications pratiques à l'analyse d'images. Je viens aussi de m'engager dans la direction d'une thèse de linguistique computationnelle.

LISTE DE PUBLICATIONS

- [AC10] Jean-Yves Audibert and Olivier Catoni. Risk bounds in linear regression through PAC-Bayesian truncation. *in revision for Annals of Statistics, second version (the first dating from 2009) preprint available on HAL and ArXiv <http://arxiv.org/abs/0902.1733>*, pages 1–78, 2010.
- [Cat88] Olivier Catoni. Grandes déviations et décroissance de la température dans les algorithmes de recuit. *C.R. Acad. Sci. Paris*, 1(307) :535–538, 1988.
- [Cat90a] Olivier Catoni. *Etude Asymptotique des algorithmes de recuit simulé (Asymptotics of simulated annealing algorithms)*. PhD thesis, Université Paris-Sud Orsay, 170 pages, 1990.
- [Cat90b] Olivier Catoni. Image restoration by stochastic dichotomic reconstruction of contour lines. In Frigessi A. Barone P. and Piccioni M., editors, *Stochastic Models, Statistical Methods, and Algorithms in Image Analysis, Lecture Notes in Statistics No 74*, pages 101–116. Springer-Verlag, 1990.
- [Cat91a] Olivier Catoni. Applications of sharp large deviations estimates to optimal cooling schedules. *Ann. Inst. Henri Poincaré*, 27(4) :463–518, 1991.
- [Cat91b] Olivier Catoni. Learning algorithms for pattern recognition on half-tone binary images. *unpublished*, pages 1–32, 1991.
- [Cat91c] Olivier Catoni. Sharp large deviations estimates for simulated annealing algorithms. *Ann. Inst. Henri Poincaré*, 27(3) :291–383, 1991.
- [Cat92a] Olivier Catoni. Exponential triangular cooling schedules for simulated annealing algorithms : A case study. In Ocone D. Karatzas I., editor, *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*, pages 74–89. Springer-Verlag, 1992.
- [Cat92b] Olivier Catoni. Rates of convergence for sequential annealing : a large deviation approach. In Robert Azencott, editor, *Simulated Annealing : Parallelization Techniques*, chapter 3, pages 25–35. John Wiley and Sons, 1992.
- [Cat92c] Olivier Catoni. Rough large deviation estimates for simulated annealing : Application to exponential schedules. *The Annals of Probability*, 20(3) :1109–1146, 1992.
- [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*, pages 1–6, 1994.
- [Cat96a] Olivier Catoni. The Legendre transform of two replicas of the Sherrington-Kirpatrick spin glass model. *Probab. Theory Relat. Fields*, 104 :369–392, 1996.
- [Cat96b] Olivier Catoni. Metropolis, Simulated Annealing and I.E.T. Algorithms : Theory and Experiments. *Journal of Complexity 12, special issue on the conference Foundation of Computational Mathematics, January 5-12 1997, Rio de Janeiro*, pages 595–623, December 1996.
- [Cat96c] Olivier Catoni. A New Inequality for the Free Energy of the Sherrington-Kirkpatrick Spin Glass Model. In P. Eichelsbacher and M. Löwe, editors, *Proceedings of the 1995 Workshop on Large Deviations and Statistical Mechanics, Oct. 20-21, SFB, Bielefeld, Germany*, pages 1–8. SFB-preprint-series, 1996.
- [Cat97] Olivier Catoni. A mixture approach to universal model selection. *LMENS-97-30* at <http://www.dmi.ens.fr/preprints>, pages 1–19, October 1997. First draft of [Cat99b].
- [Cat98a] Olivier Catoni. The Energy Transformation Method for the Metropolis Algorithm Compared with Simulated Annealing. *Probab. Theory Related Fields 110 (1998), no. 1.*, pages 69–89, 1998.
- [Cat98b] Olivier Catoni. Solving Scheduling Problems by Simulated Annealing. *SIAM J. Control Optim.* 36, no. 5, (electronic), pages 1539–1575, 1998.
- [Cat99a] Olivier Catoni. Simulated annealing algorithms and Markov chains with rare transitions. In *Séminaire de Probabilités XXXIII*, volume 1709 of *Lecture Notes in Math.*, pages 69–119. Springer, 1999. in French 1995, English revised translation at <http://www.dmi.ens.fr/preprints> 1997, published augmented revision 1999.
- [Cat99b] Olivier Catoni. Universal aggregation rules with sharp oracle inequalities. *essentially included in [Cat04b]*, pages 1–37, 1999. Revised and augmented from *A mixture approach to universal model selection*, available at <http://www.proba.jussieu.fr>.
- [Cat00] Olivier Catoni. Gibbs estimators. *essentially included in [Cat04b]*, pages 1–23, 2000. last revision available from the author’s webpage at <http://www.proba.jussieu.fr>.
- [Cat01] Olivier Catoni. Randomized estimators and empirical complexity for pattern recognition and least square regression. *included in [Cat04b]*, pages 1–33, 2001. preprint at <http://www.proba.jussieu.fr>.
- [Cat02] Olivier Catoni. Data compression and adaptive histograms. In F. Cucker and J.M. Rojas, editors, *Foundations of Computational Mathema-*

- tics, Proceedings of the Smalefest 2000*, pages 35–60. World Scientific, 2002.
- [Cat03a] Olivier Catoni. Laplace transform estimates and deviation inequalities. *Ann. Inst. H. Poincaré Probab. Statist.*, 39(1) :1–26, 2003. Revised from “Free energy estimates and deviation inequalities”, déc 1999, available at <http://www.proba.jussieu.fr>.
- [Cat03b] 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*, pages 1–72, 2003.
- [Cat04a] Olivier Catoni. Improved Vapnik Cervonenkis bounds. *preprint, included in revised form into [Cat07]*, pages 1–22, 2004.
- [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*, volume 1851 of *Lecture Notes in Mathematics*. Springer, 2004. pages 1–269.
- [Cat06] Olivier Catoni. Théorie statistique de l’apprentissage. *Images des Mathématiques 2006 — CNRS*, pages 31–39, 2006.
- [Cat07] Olivier Catoni. *Pac-Bayesian Supervised Classification : The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes Monograph Series*. Institute of Mathematical Statistics, 2007. pages i-xii, 1-163.
- [Cat09] Olivier Catoni. High confidence estimates of the mean of heavy-tailed real random variables. *preprint <http://arxiv.org/abs/0909.5366>*, pages 1–40, 2009.
- [Cat10] Olivier Catoni. Challenging the empirical mean and empirical variance : a deviation study. *preprint available on HAL and ArXiv*, pages 1–53, 2010. *second augmented, improved and completely rewritten version of [Cat09]*.
- [CC97] Olivier Catoni and Raphael Cerf. The exit path of a Markov chain with rare transitions. *ESAIM : Probability and Statistics*, pages 95–144, 1997.
- [CC98] Olivier Catoni and Cécile Cot. Piecewise constant triangular cooling schedules for generalized simulated annealing algorithms. *Ann. Appl. Probab.* 8, no. 2,, pages 375–396, 1998.
- [CCX00] Olivier Catoni, Dayue Chen, and Jun Xie. The loop erased exit path and the metastability of a biased vote process. *Stochastic Process. Appl.*, 86 :231–261, 2000.
- [CG89] Olivier Catoni and 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, pages 341–355. Hermes, 1989.

- [CT92] Olivier Catoni and Alain Trouvé. Parallel annealing by multiple trials : A mathematical study. In Robert Azencott, editor, *Simulated Annealing : Parallelization Techniques*, chapter 9, pages 129–143. John Wiley and Sons, 1992.