

# RAPPORT D'ACTIVITÉ 2003-2005

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 laboratoire de Probabilités et Modèles Aléatoires des Universités Paris 6 et 7 (U.M.R. 7599).

**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<sup>ème</sup> à l'E.N.S Ulm et 13<sup>ème</sup> 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<sup>ème</sup>.

### **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.

**1<sup>er</sup> 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 premier 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.

**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. 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. (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 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 Dufflo – Université de Marne-la-Vallée, Christine Graffigne – 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 ».

## 2. RECHERCHE SCIENTIFIQUE (SEPTEMBRE 2003 - SEPTEMBRE 2005)

Mon activité durant ces deux années a porté sur la théorie statistique de l'apprentissage, d'une part, et sur la reconnaissance des formes en imagerie, d'autre part.

**2.1. Vicissitudes éditoriales.** En ce qui concerne la théorie statistique de l'apprentissage, je me consacre depuis un moment déjà à l'approche dite PAC-Bayésienne initiée par David McAllester [1, 2]. Un exposé argumenté de cette approche et de ses enjeux est disponible sur ma page web, sous la forme d'un article de présentation écrit pour la brochure du CNRS « Images des Mathématiques 2006 ».

Les résultats obtenus durant ces deux années viennent à la suite de ceux exposés dans ma prépublication de l'été 2003 : A PAC-Bayesian approach to adaptive classification, soumise aux Annals of Statistics. J'ai reçu une réponse de l'éditeur en février 2005 exprimant « son enthousiasme » à l'idée de publier une présentation de l'approche PAC-Bayésienne dans Annals of Statistics ... et réclamant une refonte complète de l'exposition et une réduction à 45 pages du preprint qui en faisait 70. Cette réponse, ainsi que le refus de PTRF opposé à mon preprint d'octobre 2004

« Improved Vapnik Cervonenkis bounds » (jugé trop technique par un référé refusant de prendre le temps d'écrire un rapport), m'ont incité à remettre en cause une politique éditoriale par trop échevelée. De plus, pour compliquer un peu les choses, j'ai obtenu récemment une amélioration de l'inégalité de base de la classification PAC-Bayésienne, m'invitant à réécrire complètement ces deux articles devenus en quelque sorte obsolètes avant même d'avoir vu le jour (autrement que sous forme de prépublications). En résumé j'ai pris la décision suivante :

- réécrire complètement A PAC-Bayesian approach to adaptive classification, en utilisant des inégalités sans approximation de la transformée de Laplace, dont je vais parler ci-dessous. Cette prépublication devrait être disponible (sur ma page web et sur le serveur HAL) au moment où ce rapport sera lu. J'ai été contraint de réduire le contenu aux techniques PAC-Bayésiennes proprement dites, avec un volet inductif illustré par la classification sous les hypothèses de marge de Mammen et Tsybakov et un volet transductif centré sur les inégalités de Vapnik qui n'avaient pas trouvé preneur chez PTRF, mais qui me paraissent néanmoins intéressantes (ne serait-ce que parce qu'elles sont généralisées au cas indépendant non i.i.d. qui correspond au cas pratique de la collecte de données statistiques inhomogènes). Les applications aux Support Vector Machines devront donc être publiées ailleurs.
- envisager de publier une monographie plus extensive, reprenant le coeur de la théorie, mais aussi sa déclinaison dans le cas des Support Vector Machines, monographie qui pourrait provenir d'une  $n$ -ième refonte de mes notes de cours de M2 sur le sujet.

Tout ceci peut paraître un peu laborieux et assez brouillon, néanmoins ces vicissitudes ont eu le mérite de donner à la théorie le temps de mûrir et me permettront, je l'espère, d'en produire un exposé plus satisfaisant.

**2.2. Analyse d'images.** En ce qui concerne les applications à l'analyse d'images, programme auquel j'ai proposé à mon thésard Pierre Alquier et à mon ancien thésard Jean-Yves Audibert de participer, l'état des lieux est le suivant :

Dans le cadre de sa thèse sur la régression PAC-Bayésienne en norme  $L^2$ , Pierre Alquier a mis au point une méthode de sélection adaptative de variables explicatives en norme  $L^2$ . Une prépublication devrait bientôt être diffusée. Pourquoi la norme  $L^2$  ? parce que l'on peut construire dans cette norme des régions de confiance ayant une géométrie simple (convexe notamment), ce qui permet ensuite de sélectionner des variables par une méthode de projection sur ces régions de confiance. Nous comptons appliquer cet algorithme à la classification, suivant ainsi une approche très répandue, consistant à remplacer le taux d'erreur par un critère certes moins pertinent mais plus facile à mettre en œuvre (le boosting peut aussi s'interpréter ainsi, de même que les Support Vector Machines dans leurs déclinaisons pratiques admettant un taux d'erreur non nul).

Pourquoi concevoir un nouvel algorithme plutôt que d'employer une Support Vector Machine ? Pour deux raisons :

- Diminuer la charge de calcul, l'algorithme de Pierre Alquier effectue des projections en norme  $L^2$  (dans un espace virtuel dans le cas où on emploie un noyau positif comme dans les Support Vector Machines), ce qui est plus rapide que de résoudre un problème de minimisation quadratique sous contraintes ;
- Et surtout avoir une méthode de sélection adaptative d'un faible nombre de variables explicatives qui soit théoriquement fondée. En effet, les SVM ne

se prêtent pas naturellement à la sélection de variables, du fait qu'elles résolvent dans l'espace dual un problème de maximisation qui ne fait intervenir que les données sélectionnées pour être des « support vectors ». Une méthode travaillant dans l'espace direct, comme le boosting, est mieux adaptée au problème de la sélection de variables, mais ne dispose malheureusement pas d'une théorie très satisfaisante.

Nous avons un peu tâtonné sur le modèle de développement algorithmique (comment travailler à trois ?) Nous avons finalement décidé qu'il valait mieux laisser chacun faire ses expériences et suivre ses idées tout en faisant le point régulièrement ensemble. Ainsi :

- Pierre Alquier teste en ce moment ses méthodes de sélection de variables sur le problème de reconnaissance de caractères manuscrits utilisé par V. Vapnik pour tester les SVM (ce qui permet des comparaisons intéressantes) ;
- Jean-Yves Audibert étudie l'application de l'algorithme d'Alquier à la reconnaissance de visages dans un cadre expérimental identique à celui de Viola et Jones, où se pose le problème du déséquilibre entre les deux classes (il y a peu de chance pour qu'une sous-fenêtre donnée de l'image contienne un visage) ;
- je réfléchis à l'extraction de caractéristiques invariantes par transformations projectives du plan de l'image. Il me semble en effet indispensable de tenir compte de l'invariance projective si on veut pouvoir interpréter et classer des scènes tridimensionnelles (on peut par exemple penser à des applications à la navigation automobile assistée par ordinateur : les maisons, les panneaux, les autres voitures, la route, etc. sont vues en projection conique). L'ambition (peut-être irréaliste ?) serait de parvenir à un traitement purement statistique de la troisième dimension, sans autre modélisation que l'introduction de l'invariance projective, qui devrait favoriser l'identification de motifs plans, et donc une sorte de « découpage automatique » de la scène en facettes planes, les paramètres décrivant la transformation projective de chaque facette donnant des renseignements sur la géométrie 3D de la scène.

Je réfléchis aussi à un cadre expérimental moins supervisé, dans lequel on tenterait de classer directement des images au lieu de classer des imagerie cadrées (ou non) sur l'objet à reconnaître, comme c'est le cas dans les deux applications à la reconnaissance de caractères et à la détection de visages évoquées ci-dessus. Je ne peux pas non plus dire à ce stade si cette ambition est réaliste ou non.

Du point de vue logiciel, j'ai codé l'extraction d'un champ de directions limité à une grille (et lissé sur les bords de la grille dans la direction parallèle au bord, l'idée de coder les directions sur une grille est inspirée des travaux de Donoho sur les edgelets, qui montrent l'intérêt d'approcher un contour en utilisant une discrétisation plus fine dans une direction que dans l'autre).

Voici un ours et son champ de directions (calculé sur une grille) :



Les directions ont été regroupées en trois classes suivant leur orientation. La question consiste à savoir dans quelle mesure l'analyse des champs de directions permet de classer des images (avec éventuellement des informations de couleur supplémentaires). J'espère que je pourrai apporter des éléments de réponse à cette question dans mon prochain rapport ! Notons que la méthode de calcul des directions utilisée exploite directement des images couleurs (et peut donc par exemple représenter correctement les contours d'un triangle rouge de couleur  $(255, 0, 0)$  sur un fond vert de couleur  $(0, 255, 0)$ , alors que l'image en niveaux de gris correspondante est uniforme.)

L'algorithme de détection de directions mis au point se veut suffisamment robuste pour traiter des images de qualité médiocre et pouvoir par la même exploiter des « images de la vie de tous les jours » telles que celles que l'on peut se procurer sur le web, en enregistrant des émissions de télévision, à partir d'un DVD, d'un caméscope grand public etc. (l'image présentée provient de l'acquisition sous forme de flux mpeg2 d'un documentaire télévisé, traduit en une suite d'images compressées jpeg, la question était de savoir dans quelle mesure la détection était perturbée par les artefacts créés par les deux opérations de compression subies successivement par l'image).

**2.3. Apprentissage statistique PAC-Bayésien : des inégalités sans approximation de la transformée de Laplace.** L'approche de la classification supervisée que j'ai développée s'intéresse de façon privilégiée au contrôle du taux d'erreur, et plus précisément au lien entre le taux d'erreur observé sur l'échantillon servant à l'apprentissage d'une règle de classification et son espérance. Or ce taux d'erreur empirique suit une loi binomiale dont je n'avais pas exploité dans ma pré-publication de l'été 2003 toute la spécificité, me contentant d'utiliser une inégalité de déviation (ou plutôt un contrôle de la transformée de Laplace) valable pour n'importe quelle somme de variables aléatoires i.i.d. bornées.

Dans un premier temps, j'ai affiné mon étude en cessant de supposer que les observations étaient i.i.d., pour les supposer uniquement indépendantes. Cette extension permet de traiter le cas pratique dans lequel les données sont inhomogènes,

par exemple parce qu'elles proviennent d'expériences menées dans des endroits différents. Elle ne présente pas, dans l'approche que j'ai choisie de difficulté particulière par rapport au cas i.i.d.

Dans un second temps, je me suis rendu compte qu'il était possible d'utiliser le fait que la transformée de Laplace de l'opposé d'une variable de Bernoulli  $\sigma \in \{0, 1\}$ , à savoir  $\mathbb{E}[\exp(-\lambda\sigma)]$  était exactement sous Gaussienne, c'est-à-dire plus petite que  $\exp\left\{\frac{\lambda^2}{2N}\mathbb{E}(\sigma)[1 - \mathbb{E}(\sigma)]\right\}$ , du moins tant que  $\lambda \geq 0$  et  $\mathbb{E}(\sigma) \in [0, \frac{1}{2}]$ .

Ceci permet entre autre, dans l'approche transductive qui mène aux bornes de complexité de Vapnik, d'introduire un terme de variance au lieu de majorer comme le faisait Vapnik la variance d'une Bernoulli par sa moyenne.

Dans un troisième temps, au printemps 2005, j'ai réalisé une chose très simple qui m'avait échappé jusqu'à présent et que je n'ai d'ailleurs pas vu exploiter non plus par d'autres dans l'étude de la classification adaptative : la loi d'une variable de Bernoulli étant décrite par un seul paramètre, par exemple son espérance, sa transformée de Laplace est une fonction explicite de son espérance que l'on peut conserver tout au long des calculs sans faire d'approximation !

**2.3.1. Classification transductive et bornes de Vapnik.** Je me suis intéressé durant ces deux années aux bornes de Vapnik et à la façon dont on pouvait les prouver et en l'occurrence aussi les améliorer en utilisant les outils de l'inférence PAC-Bayésienne transductive. J'ai pu apporter plusieurs améliorations à l'inégalité prouvée par Vapnik. Décrivons cette inégalité pour pouvoir commenter ces améliorations. Supposons que nous observions un échantillon  $(X_i, Y_i)_{i=1}^N$  de  $N$  couples indépendants mais pas nécessairement équidistribués de données  $X_i$  appartenant à un espace mesurable  $\mathcal{X}$  munies de labels binaires  $Y_i \in \{0, 1\}$ . Supposons que l'on envisage d'appliquer à ces données une règle de classification  $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ , où le paramètre  $\theta$  varie dans un ensemble mesurable  $\Theta$  et où  $(\theta, x) \mapsto f_\theta(x) : \Theta \times \mathcal{X} \rightarrow \mathcal{Y}$  est mesurable. Intéressons nous au taux d'erreur

$$R(\theta) = \frac{1}{N} \sum_{i=1}^N \mathbb{P}[Y_i \neq f_\theta(X_i)]$$

(définition valable dans le cas non i.i.d., qui peut se simplifier dans le cas i.i.d., puisque tous les termes de la somme sont alors égaux). Intéressons nous aussi à sa contrepartie empirique :

$$r(\theta) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[Y_i \neq f_\theta(X_i)].$$

Vapnik a montré, dans le cas i.i.d., qu'il était possible de majorer uniformément  $R(\theta)$  par une fonction de  $r(\theta)$  faisant intervenir la taille de la trace des  $f_\theta$  sur un échantillon  $(X_1, \dots, X_{2N})$  de taille double. Ceci conduit à introduire la notation

$$\mathfrak{N}_k = \left| \left\{ [f_\theta(X_i)]_{i=1}^{(k+1)N} ; \theta \in \Theta \right\} \right|, \quad k \in \mathbb{N}.$$

Avec ces notations, l'inégalité prouvée par V. Vapnik dans [3, page 138] s'écrit

**Théorème (Vapnik).** *Dans le cas où les variables  $(X_i, Y_i)$  sont équidistribuées, avec probabilité au moins  $1 - \epsilon$ , pour tout  $\theta \in \Theta$ ,*

$$R(\theta) \leq r(\theta) + \frac{2d_V}{N} + \sqrt{\frac{4d_V r(\theta)}{N} + \frac{4d_V^2}{N^2}},$$

où

$$d_V = \log[\mathbf{E}(\mathfrak{N}_1)] + \log(4/\epsilon).$$

Nous en avons obtenu l'amélioration suivante

**Théorème.** *Dans le cas où les variables  $(X_i, Y_i)$  sont indépendantes, mais pas nécessairement i.i.d., avec probabilité au moins  $1 - \epsilon$ , pour tout  $\theta \in \Theta$ ,*

$$\begin{aligned} R(\theta) &\leq \inf_{k \in \mathbb{N}^*} \frac{k+1}{k} \inf_{\lambda \in \mathbb{R}_+} \frac{\exp\left[-\frac{\lambda}{N} \left[r(\theta) + \frac{1}{10N}\right] - \frac{d_k}{N}\right]}{[1 - \exp(-\frac{\lambda}{N})]} - \frac{r(\theta)}{k} \\ &\leq \inf_{k \in \mathbb{N}^*} \frac{k+1}{k \left(1 + \frac{2d_k}{N}\right)} \left[ r(\theta) + \frac{1}{10N} + \frac{d_k}{N} + \sqrt{\frac{2d_k r(\theta) [1 - r(\theta)]}{N} + \frac{d_k^2}{N^2}} \right] - \frac{r(\theta)}{k}, \\ d_k &= \mathbf{E}[\log(\mathfrak{N}_k) | (X_i, Y_i)_{i=1}^N] + \log \left[ \frac{2e \log(10N)}{\epsilon} \right] \\ &\quad - \frac{\log \log(10N)}{\log(10N)} + \log[k(k+1)]. \end{aligned}$$

Cette amélioration porte sur six points :

- suppression de l'hypothèse d'équidistribution ;
- un « vrai » terme d'entropie de Vapnik, avec intégration de  $\log(\mathfrak{N}_k)$  à l'extérieur du log et non à l'intérieur ;
- un terme de variance en  $r(\theta) [1 - r(\theta)]$  et non en  $r(\theta)$ , comme cela se voit dans l'approximation Gaussienne de notre borne ;
- utilisation d'un échantillon fantôme de taille quelconque et optimisation de cette taille ( $kN$  dans les formules) ;
- suppression du recours à un argument de symétrisation dans les preuves, qui fait gagner un facteur proche de deux dans le cas où  $r(\theta) = 0$  ;
- suppression de l'approximation Gaussienne et utilisation directe de l'inverse (explicite) de la transformée de Laplace d'une Bernoulli exprimée en fonction de son espérance.

Numériquement, dans le cas où le modèle de classification  $\{f_\theta ; \theta \in \Theta\}$  constitue une classe de Vapnik de dimension  $h$ , où la dimension de l'échantillon vaut  $N = 1000$ , où  $\epsilon = 0.01$  et où  $\inf_{\theta \in \Theta} r(\theta)$  vaut 0.2, atteint en  $\hat{\theta}$ , l'inégalité de Vapnik donne une borne supérieure à 0.61, qui ne permet pas d'assurer que le modèle utilisé soit capable de classer les données mieux qu'une classification aléatoire uniforme (qui donne un taux d'erreur de 0.5). Notre borne donne  $R(\hat{\theta}) \leq 0.4211$ , pour une taille optimale de l'échantillon fantôme de  $k = 15$  et une valeur du paramètre  $\lambda = 1010 \geq N$ . Son approximation Gaussienne donne une borne voisine de 0.4325, et il s'avère que toutes les améliorations mentionnées ci-dessus concourent à une majoration plus précise (mis à part le changement du type d'entropie, puisque dans cette application à une classe de Vapnik, nous majorons l'entropie dans les deux cas par la même borne, qui majore la taille de la trace du modèle sur l'échantillon étendu pour toutes les valeurs possibles de cet échantillon). En particulier la valeur optimale élevée du paramètre  $\lambda$  confirme que dans cette situation réaliste l'approximation Gaussienne de la transformée de Laplace n'est pas très précise,  $\lambda r(\theta)$  n'ayant pas un comportement Gaussien pour  $\lambda \simeq N$ .

L'exemple numérique évoqué ci-dessus nous paraît pertinent. Il concerne une situation « d'apprentissage faible » (weak learning) dans laquelle il n'est pas facile

de dire si les règles de classification envisagées sont significatives, au sens où elles produisent un taux d'erreur inférieur à 0.5 (et contiennent donc de l'information sur les classes). Être capable de distinguer des règles faiblement significatives nous semble être une étape importante dans la conception de méthodes d'agrégation efficaces (du type boosting). Dans cette perspective, nous pensons donc (en désaccord, semble-t-il avec les éditeurs de PTRF!) que l'amélioration de la valeur numérique des bornes de généralisation pour des échantillons de taille modérée par rapport à la dimension du modèle mérite des efforts techniques.

**2.3.2. Au delà des bornes de Vapnik : bornes localisées et bornes relatives.** Nous sommes récemment parvenus à un exposé unifié des cas inductif dans le cas d'un échantillon indépendant et transductif dans le cas d'un échantillon étendu partiellement échangeable, dans lequel les inégalités de départ sont les mêmes dans les deux cas, à condition de remplacer  $R(\theta)$  par  $\bar{r}(\theta) = \frac{1}{(k+1)N} \sum_{i=1}^{(k+1)N} \mathbb{1}[f_\theta(X_i) \neq Y_i]$ . La notion de variable partiellement échangeable est une restriction de la notion d'échangeabilité compatible avec le cas indépendant non identiquement distribué. Dans ce cas on considère comme échantillon fantôme  $k$  copies i.i.d. de l'échantillon observé (qui n'est pas lui-même i.i.d.), pour obtenir un échantillon étendu « faiblement échangeable », ce qui permet d'utiliser pour fabriquer des bornes une loi a priori sur les paramètres qui n'en est pas une, mais a au contraire le droit de dépendre des données de façon partiellement échangeable elle aussi.

Introduisons la fonction

$$\Phi_a(p) = -a^{-1} \log[1 - p(1 - \exp(-a))], \quad a \in \mathbb{R}, p \in [0, 1],$$

liée à la transformée de Laplace de l'opposée d'une variable de Bernoulli de paramètre  $p$ . C'est pour toute valeur de  $\lambda$  positive (resp. négative) une transformation convexe (resp. concave) croissante de l'intervalle unité dans lui-même.

Les inégalités de base s'énoncent de la façon suivante :

**Théorème** (cas inductif). *Pour tout échantillon indépendant, pour toute fonction mesurable  $\lambda : \Theta \rightarrow \mathbb{R}$ , pour toute loi de probabilité a priori sur les paramètres  $\pi \in \mathcal{M}_+^1(\Theta)$ ,*

$$\int_{\Omega} \exp \left[ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \int_{\theta \in \Theta} \lambda(\theta) \left[ \Phi_{\frac{\lambda}{N}} [R(\theta)] - r(\theta, \omega) \right] \rho(d\theta) - \mathcal{K}(\rho, \pi) \rho(d\theta) \right] \mathbb{P}(d\omega) \leq 1,$$

où  $\mathcal{K}(\rho, \pi)$  désigne la divergence de Kullback entre mesures de probabilités.

**Théorème** (cas transductif). *Considérons la représentation canonique dans laquelle la loi  $\mathbb{P}$  de  $(X_i, Y_i)_{i=1}^{(k+1)N}$  est définie sur l'espace produit  $\Omega = (\mathcal{X} \times \mathcal{Y})^{(k+1)N}$ . Pour toute loi  $\mathbb{P}$  partiellement échangeable, c'est-à-dire invariante par permutation circulaire des indices  $\{i + jN; j = 0, \dots, k\}$ , pour tout  $i = 1, \dots, N$ ; pour toute fonction mesurable  $\lambda : \Theta \times \Omega \rightarrow \mathbb{R}$  partiellement échangeable, c'est-à-dire invariante par permutation circulaire des ensembles d'indices de  $\omega \in \Omega$   $\{i + jN; j = 0, \dots, k\}$ , pour tout  $i = 1, \dots, N$ ; pour toute mesure de probabilités conditionnelle régulière  $\pi : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$  partiellement échangeable (en tant que fonction de  $\Omega$  à valeur mesure)*

$$\int_{\Omega} \exp \left[ \sup_{\rho \in \mathcal{M}_+^1(\Theta)} \int_{\theta \in \Theta} \lambda(\theta) \left[ \Phi_{\frac{\lambda}{N}} [\bar{r}(\theta, \omega)] - r(\theta, \omega) \right] \rho(d\theta) - \mathcal{K}(\rho, \pi) \rho(d\theta) \right] \mathbb{P}(d\omega) \leq 1,$$

Ceci posé nous décrivons donc maintenant les développements de notre travail dans le cas inductif, sachant qu'il en existe une traduction systématique au cadre transductif, due au parallélisme des deux théorèmes ci-dessus.

Un des développements postérieurs à 2003 de notre théorie PAC-Bayésienne a consisté à établir des résultats en espérance, qui n'offrent certes pas la même garantie que les résultats en déviation, mais ont l'intérêt de fournir des constantes plus faibles et donc nous l'espérons en pratique plus efficaces, quand les inégalités sont utilisées pour sélectionner des modèles ou fabriquer des estimateurs.

Les deux idées que nous avons avancées pour aller au delà des bornes de Vapnik (présentes dans notre présentation de 2003), ont trouvé ces deux dernières années les déclinaisons suivantes.

Des bornes localisées en espérance. On peut donner ainsi un encadrement assez précis de la performance d'un estimateur randomisé décrit par une loi a posteriori (qui peut toujours être une masse de Dirac dans le cas transductif où on travaille en fait toujours avec un nombre fini de règles de classification). Cette encadrement est fonction de la distance de l'estimateur à la loi de Gibbs a posteriori  $\pi_{\exp(-\beta r)}$  définie par sa densité

$$\frac{d\pi_{\exp(-\beta r)}(\theta)}{d\pi}(\theta) = \left[ \int_{\theta' \in \Theta} \exp[-\beta r(\theta')] \pi(d\theta') \right]^{-1} \exp[-\beta r(\theta)].$$

Il est donné par les théorèmes suivants, dans lesquels apparaît la fonction :

$$\Psi_{a,b}(p) = (1 - b)^{-1} [\Phi_a(p) - bp],$$

qui peut être inversée numériquement (et est comme  $\Phi_a$  une transformation convexe croissante de l'intervalle unité dans lui-même dès que  $0 \leq b \leq a^{-1} [1 - \exp(-a)]$ .)

**Théorème.** *Pour tous paramètres réels  $\beta$  et  $\lambda$  tels que  $0 \leq \beta \leq N [1 - \exp(-\frac{\lambda}{N})]$ , pour toute loi a posteriori  $\rho : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$  (i.e. toute probabilité conditionnelle régulière),*

$$\begin{aligned} \int_{\Omega} \left\{ \int_{\Theta} r d\rho - \frac{\mathcal{K}[\rho, \pi_{\exp(-\beta r)}]}{\beta} \right\} d\mathbb{P} &\leq \int_{\Omega} \left[ \int_{\Theta} R d\rho \right] d\mathbb{P} \\ &\leq \Psi_{\frac{\lambda}{N}, \frac{\beta}{N}}^{-1} \left\{ \int_{\Omega} \left[ \int_{\Theta} r d\rho + \frac{\mathcal{K}[\rho, \pi_{\exp(-\beta r)}]}{\lambda(1 - \frac{\beta}{\lambda})} \right] d\mathbb{P} \right\} \\ &\leq \frac{\lambda - \beta}{N [1 - \exp(-\frac{\lambda}{N})] - \beta} \int_{\Omega} \left[ \int_{\Theta} r d\rho + \frac{\mathcal{K}[\rho, \pi_{\exp(-\beta r)}]}{\lambda(1 - \frac{\beta}{\lambda})} \right] d\mathbb{P}. \end{aligned}$$

De plus

$$\begin{aligned} \int_{\Omega} \left[ \int_{\Theta} R d\rho \right] d\mathbb{P} &\leq \Psi_{\frac{\lambda}{N}, \frac{\beta}{N}}^{-1} \left\{ \frac{1}{\lambda - \beta} \int_{\beta}^{\lambda} \int_{\Theta} R d\pi_{\exp(-\gamma R)} d\gamma + \int_{\Omega} \frac{\mathcal{K}[\rho, \pi_{\exp(-\beta r)}]}{\lambda - \beta} d\mathbb{P} \right\} \\ &\leq \frac{1}{N [1 - \exp(-\frac{\lambda}{N})] - \beta} \left\{ \int_{\beta}^{\lambda} \int_{\Theta} R d\pi_{\exp(-\beta R)} d\gamma + \int_{\Omega} \mathcal{K}[\rho, \pi_{\exp(-\lambda r)}] d\mathbb{P} \right\}. \end{aligned}$$

Nous avons aussi obtenu des bornes en déviation, avec des constantes plus grandes et plus compliquées, qui montrent que les quantités intégrées par rapport à  $\mathbb{P}$  dans les inégalités précédentes se comportent essentiellement comme leurs espérances. Nous renvoyons pour plus de détails à [Cat05a]. On montre ainsi que  $\int_{\Theta} r d\pi_{\exp(-\beta r)}$  et  $\int_{\Theta} R d\pi_{\exp(-\beta)}$  sont très proches pour une large gamme de valeurs de  $\beta$ , ce qui montre que ces estimateurs de Gibbs ne souffrent pas de surapprentissage. L'intérêt de ces inégalités vient du fait qu'elles possèdent une expression universelle qui ne fait pas intervenir la complexité du modèle de classification utilisé. Cette complexité va uniquement se traduire dans la valeur de l'écart entre le taux d'erreur empirique de l'estimateur de Gibbs  $\int_{\Theta} r d\pi_{\exp(-\beta r)}$  et le minimum du risque empirique  $\inf_{\Theta} r$ .

Nous avons aussi affiné notre approche des bornes relatives, en intégrant certains des résultats de la thèse de Jean-Yves Audibert.

Les bornes relatives, portent sur la différence  $r(\theta) - r(\tilde{\theta})$  des taux d'erreur empiriques pour deux valeurs du paramètre, l'une étant destinée à être estimée et l'autre étant inconnue (on peut penser en particulier au cas où  $\tilde{\theta} \in \arg \min_{\Theta} R$ ). Elles sont nécessaires pour prendre en compte la structure des covariances du processus empirique  $\theta \mapsto r(\theta)$ , structure qui détermine comme on le sait le comportement du minimum  $\inf_{\Theta} r$  et sa relation avec  $\inf_{\Theta} R$ .

Là encore, nous nous sommes aperçu qu'il n'était pas indispensable d'approcher la transformée de Laplace de la différence de deux variables de Bernoulli, qui possède une expression exacte en fonction de l'espérance et de la variance de cette différence (dont la loi est portée par les trois points  $\{-1, 0, +1\}$  de la droite réelle).

Nous nous sommes enfin rendu compte qu'il était pratique d'énoncer des résultats en introduisant deux *fonctions de structure*, l'une « théorique », égale à

$$\varphi(x) = \sup_{\theta \in \Theta} M(\theta, \tilde{\theta}) - x[R(\theta) - R(\tilde{\theta})], \quad x \in \mathbb{R}_+$$

$$\text{où } M(\theta, \tilde{\theta}) = \frac{1}{N} \sum_{i=1}^N \mathbb{E} \left\{ \left| \mathbb{1}[f_{\theta}(X_i) \neq Y_i] - \mathbb{1}[f_{\tilde{\theta}}(X_i) \neq Y_i] \right| \right\}.$$

et l'autre empirique, faisant intervenir  $\bar{\theta} \in \arg \min_{\Theta} r$ ,

$$\psi(x) = \sup_{\theta \in \Theta} m(\theta, \bar{\theta}) - x[r(\theta) - \inf_{\Theta} r], \quad x \in \mathbb{R}_+,$$

$$\text{où } m(\theta, \bar{\theta}) = \frac{1}{N} \sum_{i=1}^N \left| \mathbb{1}[f_{\theta}(X_i) \neq Y_i] - \mathbb{1}[f_{\bar{\theta}}(X_i) \neq Y_i] \right|.$$

La fonction de structure théorique permet de borner le taux d'erreur *relatif* d'une loi a posteriori de la façon suivante (nous donnons ici la version linéarisée, et renvoyons à [Cat05a] pour une borne exprimée en fonction de la transformée de Laplace de la différence de deux variables de Bernoulli) :

**Théorème.** *Pour tous paramètres  $\beta$ ,  $\lambda$  et  $x$  tels que  $x \geq 0$  et  $0 \leq \beta < \lambda - x \frac{\lambda^2}{2N}$ , pour toute loi a posteriori  $\rho : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$ ,*

$$\mathbb{E} \left[ \int_{\Theta} R d\rho \right] \leq \inf_{\Theta} R + \left( \lambda - x \frac{\lambda^2}{2N} - \beta \right)^{-1} \\ \times \left\{ \int_{\beta}^{\lambda} [R - \inf_{\Theta} R] d\pi_{\exp(-\gamma R)} d\gamma + \mathbb{E} \left\{ \mathcal{K}[\rho, \pi_{\exp(-\lambda r)}] \right\} + \varphi(x) \frac{\lambda^2}{2N} \right\}.$$

Sous les hypothèses de marge de Tsybakov et Mammen, c'est-à-dire quand  $R(\theta) \geq \inf_{\Theta} R + cM(\theta, \tilde{\theta})$ , pour un exposant  $\kappa \geq 1$  et une constante  $c > 0$ , on obtient  $\varphi(x) \leq (1 - \kappa^{-1})(\kappa c x)^{-1/(\kappa-1)}$ . Si on est de plus dans une situation « paramétrique » où  $\int_{\Theta} R d\pi_{\exp(-\beta R)} - \inf_{\Theta} R \leq \frac{d}{\beta}$  pour une certaine *constante de dimension*  $d$ , alors

$$\mathbb{E} \left[ \int_{\Theta} R d\pi_{\exp(-\bar{\lambda} r)} \right] \leq (2 - \kappa^{-1})(\kappa c)^{-1/(2\kappa-1)} \left( \frac{8 \log(2)d}{N} \right)^{\kappa/(2\kappa-1)}, \\ \text{pour } \bar{\lambda} = 2^{-1} [8 \log(2)d]^{\frac{\kappa-1}{2\kappa-1}} (\kappa c)^{1/(2\kappa-1)} N^{\frac{\kappa}{2\kappa-1}}.$$

La question se pose alors de savoir s'il est possible d'obtenir une borne empirique de même forme, permettant d'estimer la valeur optimale de  $\lambda$ . Une réponse est donnée par le résultat suivant (qui possède une version un peu plus compliquée en déviation)

**Théorème.** *Pour tous paramètres réels  $x$ ,  $\alpha$  et  $\lambda$  tels que  $\alpha < N \sinh(\frac{\lambda}{N}) [1 - x \tanh(\frac{\lambda}{2N})]$ , pour toute loi a posteriori  $\rho : \Omega \rightarrow \mathcal{M}_+^1(\Theta)$ ,*

$$\mathbb{E} \left[ \int_{\Theta} R d\rho \right] \leq \inf_{\Theta} R + \mathbb{E} \left\{ \left[ 1 - \frac{N \sinh(\frac{\lambda}{N}) [1 - x \tanh(\frac{\lambda}{2N})] - \lambda}{N \sinh(\frac{\lambda}{N}) [1 - x \tanh(\frac{\lambda}{2N})] - \alpha} \right] \left[ \int_{\Theta} r d\rho - \inf_{\Theta} r \right] \right. \\ \left. + \frac{\mathcal{K}[\rho, \pi_{\exp(-\alpha r)}]}{N \sinh(\frac{\lambda}{N}) [1 - x \tanh(\frac{\lambda}{2N})] - \alpha} \right. \\ \left. + \frac{N \sinh(\frac{\lambda}{N}) \tanh(\frac{\lambda}{2N})}{N \sinh(\frac{\lambda}{N}) [1 - x \tanh(\frac{\lambda}{2N})] - \alpha} \left[ \psi(x) + \psi \left( \frac{\lambda - \alpha}{N \sinh(\frac{\lambda}{N}) \tanh(\frac{\lambda}{2N})} \right) \right] \right\}.$$

(On pourrait simplifier et affaiblir la borne en utilisant le fait que  $\tanh(y) \leq y \leq \sinh(y)$ ,  $y \in \mathbb{R}_+$ .) Sachant que l'on obtient aussi une borne du même type en déviation, on voit que l'on dispose d'une méthode pour estimer les meilleures valeurs des paramètres.

## 2.4. Production scientifique 2003-2005.

- [Cat03] Olivier Catoni. A PAC-Bayesian approach to adaptive classification. *submitted to the Annals of Statistics*, pages 1–72, 2003.
- [Cat04a] Olivier Catoni. Improved Vapnik Cervonenkis bounds. *preprint*, 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*. Number 1851 in Lecture Notes in Mathematics. Springer, 2004. pages 1–269.
- [Cat05a] Olivier Catoni. PAC-Bayesian inductive and transductive learning theorems. *revised from [Cat03] for the Annals of Statistics*, pages 1–37, 2005.

[Cat05b] Olivier Catoni. Théorie statistique de l'apprentissage. *Images des Mathématiques 2006* — CNRS, à paraître, pages 1–6, 2005.

### 3. ENSEIGNEMENT, FORMATION ET DIFFUSION DE LA CULTURE SCIENTIFIQUE

**3.1. Thèses.** Encadrement de la thèse de Pierre Alquier (débutée en septembre 2003). Pierre Alquier bénéficie d'une bourse du CREST et travaille comme je l'ai expliqué plus haut sur la régression PAC-Bayésienne en norme  $L^2$ , sur la conception d'algorithmes de sélection adaptative de variables explicatives et sur les applications de ces algorithmes à la classification d'images.

Encadrement de la thèse de Jean-Yves Audibert. 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 (dans l'esprit des bornes relatives évoquées ci-dessus), les bornes de généralisation sous des 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). Je continue depuis à travailler avec lui sur la classification d'images, comme évoqué ci-dessus.

**3.2. Enseignement.** février-juin 2004 : Cours « Classification et sélection de modèles » DEA Probabilités et Applications, filière finance (demi-module, 15 heures).

février-juin 2005 : Cours « Classification et sélection de modèles » (module de 25 heures), à destination des filières de statistique et de probabilités appliquées du M2 Probabilités et Applications, Paris 6.

**3.3. 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

Néant pour cette période.

### 5. ENCADREMENT, ANIMATION ET MANAGEMENT DE LA RECHERCHE

Maintenance, en 2003-2004, en collaboration avec Philippe Macé, Bibliothécaire du laboratoire de Probabilités et Modèles Aléatoires, du serveur électronique de prépublications du laboratoire, relié à la Cellule de Coordination Documentaire Nationale pour les Mathématiques Unité Mixte de Service 5638 - Université Joseph Fourier (Grenoble), CNRS. J'ai transféré le serveur de preprints sur HAL à l'automne 2004, comme demandé par notre tutelle CNRS, le fonctionnement du serveur HAL ne nécessite plus désormais mon intervention (les auteurs postent directement leurs prépublications sur HAL, nous ne fabriquons plus de page de garde pour la version électronique de leurs prépublications). Le jour où HAL diffusera une interface de chargement en batch, je m'occuperai de transférer les publications du laboratoire antérieures à l'automne 2004.

Participation à la commission informatique du Laboratoire de Probabilités et Modèles Aléatoires.

## RÉFÉRENCES

- [1] D. A. McAllester, Some PAC-Bayesian Theorems, *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (Madison, WI, 1998)*, 230–234 (electronic), ACM, New York, 1998 ;
- [2] D. A. McAllester, PAC-Bayesian Model Averaging, *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (Santa Cruz, CA, 1999)*, 164–170 (electronic), ACM, New York, 1999 ;
- [3] V. N. Vapnik, *Statistical learning theory*, Wiley, New York, 1998.