Dossier de candidature à un poste de Maître de conférences

Sébastien Gerchinovitz

Détails du poste

Référence GALAXIE : n°4073
Section CNU : 26, 25 et 27
Profil : Statistiques du risque et Fouilles de données
Localisation : Université Paris 7 (Denis Diderot)

Contact

Adresse professionnelle :	Département de Mathématiques et Applications École Normale Supérieure 45 rue d'Ulm 75230 Paris Cedex 05
Téléphone :	01 69 15 57 84 ou 06 84 76 80 24
Courriel :	sebastien.gerchinovitz@ens.fr
Page web :	www.math.ens.fr/∼gerchino

Table des matières

Thèmes de recherche - Mots clés

- Apprentissage statistique, prévision de suites déterministes arbitraires
- Régression parcimonieuse en grande dimension
- Agrégation PAC-bayésienne, sélection de modèles
- Calibration d'algorithmes et adaptativité

1 Curriculum vitæ

Nom :	**Sébastien Gerchinovitz**
Date de naissance :	03/03/1985
Nationalité :	Française
État civil :	Célibataire
Titre :	Docteur en Sciences (Spécialité *Mathématiques*) ; Qualifié en section 26 du CNU.

1.1 Situations actuelle et passées

Fév. 2012 Pré-sélectionné pour le concours CNRS en mathématiques (n° 01/03).

2011-2012 **ATER à temps partiel** à l'Université Paris-Sud 11, Orsay.

2008-2011 **Allocataire-Moniteur** à l'Université Paris-Sud 11, Orsay.
Thèse effectuée au Département de Mathématiques et Applications de l'École normale supérieure, Paris.

1.2 Cursus universitaire

2008-2011 **Doctorat en mathématiques** au Département de Mathématiques et Applications, École normale supérieure, Paris (rattachement à l'Université Paris-Sud 11) sous la direction de Gilles Stoltz.

Titre : Prédiction de suites individuelles et cadre statistique classique : étude de quelques liens autour de la régression parcimonieuse et des techniques d'agrégation.

Rapporteurs : Arnak Dalalyan et Claudio Gentile.

Soutenance à l'École normale supérieure le 12/12/2011. *Mention très honorable.*

Jury : Pierre Alquier, Olivier Catoni, Arnak Dalalyan, Pascal Massart, Gilles Stoltz et Alexandre Tsybakov.

2007-2008 **Master 2 de Probabilités et Statistiques** à l'Université Paris-Sud 11, Orsay, *mention bien.*

- Stage de recherche de 5 mois à l'École normale supérieure et à l'INRIA Paris-Rocquencourt sous la direction de Gilles Stoltz et Vivien Mallet, sur le sujet *Prévision d'ensemble en qualité de l'air avec agrégation séquentielle lacunaire de modèles.*
- Enseignements suivis : Calcul stochastique (J.-F. Le Gall), Processus markoviens (J.-F. Le Gall), Concentration et sélection de modèles (P. Massart), Méthodes asymptotiques en statistique (C. Durot), Apprentissage statistique (P. Massart), Information et statistique (E. Gassiat).

2005-2008 **Diplôme d'ingénieur de l'École Centrale Paris**, Châtenay-Malabry.
Option de troisième année : Mathématiques Appliquées, *mention très bien.*

1.3 Enseignements suivis pendant ma thèse

Août 2011 **An introduction to high-dimensional statistics**, cours donné par M. Wainwright au *Workshop StatMathAppli 2011*, Fréjus (1 semaine).

Print. 2011	**Sparse recovery problems**: **beyond finite dictionaries**, cours doctoral donné par V. Koltchinskii à l'ENSAE.
Print. 2011	**Uniform probability inequalities with application in high-dimensional statistical models**, cours doctoral donné par S. van de Geer à l'ENSAE.
Août 2010	**Completely random measures, hierarchies, and nesting** ainsi que **Nonparametric estimation under shape restrictions**, cours donnés par M. Jordan et J. Wellner au *Workshop StatMathAppli 2010*, Fréjus (1 semaine).
Aut. 2009	**Classification et statistiques en grande dimension**, cours de Master 2 donné par V. Rivoirard et T. Mary-Huard à l'Université Paris-Sud 11.
Août 2009	**École d'été Stats in the Château** sur les thèmes *Inverse problems and high-dimensional estimation*, cours donnés par L. Cavalier et V. Chernozhukov, Jouy-en-Josas (1 semaine).

1.4 Compétences informatiques

Languages	C++, Python
Logiciels	Matlab, Scilab, Mathematica, R
Bureautique	LaTeX, XHTML

1.5 Langues

Anglais	écrit et parlé couramment
Allemand	niveau intermédiaire
Espagnol	notions

2 Enseignement

J'ai exercé mes activités d'enseignement en tant que moniteur à l'IUT de Sceaux (2008–2011), puis en tant qu'ATER à temps partiel à la Faculté des sciences d'Orsay (2011–2012).

2.1 ATER à temps partiel à la Faculté des sciences d'Orsay (2011–2012)

- **TD de statistiques (Master 1 Mathématiques Fondamentales et Appliquées)** associés au cours d'Elisabeth Gassiat, 65h au second semestre.
  - *Contenu* : tests et intervalles de confiance, estimation par maximum de vraisemblance, tests non-paramétriques, modèle linéaire, théorie de la décision (risque minimax, théorème de Le Cam), estimation bayésienne, lemme de Neyman-Pearson, exhaustivité, efficacité, introduction au bootstrap.
  - *Modalités* : 60h de séances d'exercices et 6h de TP sous Matlab (illustration du modèle linéaire). Responsable de la conception des feuilles de TD et TP, conception et correction des devoirs maison, et participation à l'élaboration des sujets de partiel et d'examen.

- **TD de mathématiques de la modélisation (Licence 1 cursus Biologie)** associés aux cours d'Édouard Maurel-Segala et de Hans Rugh, 30h au premier semestre.
  - *Contenu* : fonctions, suites, séries, intégration, probabilités discrètes, variables aléatoires à densité.
  - *Responsabilités* : correction des interrogations écrites et devoirs maison de mes étudiants, participation à la correction des partiel et examen de la promotion entière.

2.2 Moniteur à l'IUT de Sceaux (2008–2011)

- **TD de mathématiques générales (DUT GEA 1ère année)** associés au cours de Michelle Lauton, 36h les trois années, au premier semestre.
  - *Contenu* : fonctions, dérivation, suites, modélisation de problèmes économiques.
  - *Responsabilités* : participation à l'élaboration des feuilles de TD, correction des devoirs maison de mes étudiants, participation à l'élaboration et à la correction des sujets de partiel et d'examen.

- **TD de mathématiques financières (DUT GEA 1ère année)** associés au cours de Michelle Lauton, 18h la première année, au second semestre.
  - *Contenu* : introduction au calcul associé à des produits financiers de la vie quotidienne (prêts, assurance).
  - *Responsabilités* : participation à l'élaboration des feuilles de TD, correction des devoirs maison de mes étudiants, participation à l'élaboration et à la correction des sujets de partiel et d'examen.

- **Cours/TD de soutien en mathématiques (DUT GEA 1ère année)**, 28h les deux dernières années, au premier semestre.
  - *Contenu* : soutien en mathématiques réservé aux étudiants en difficulté.
  - *Responsabilités* : autonomie complète pour l'élaboration des cours et feuilles de TD, ainsi que la confection et correction des interrogations écrites et du sujet d'examen.

3 Publications

Articles publiés dans des actes de conférences internationales avec comité de lecture (très sélectives)

[1] S. Gerchinovitz. **Sparsity regret bounds for individual sequences in online linear regression**, *JMLR Workshop and Conference Proceedings* 19 : 377–396, 2011. (actes de la conférence COLT 2011)

Note : Une version étendue de 42 pages correspondant au chapitre 3 de ma thèse a été soumise à *Journal of Machine Learning Research*.

[2] S. Gerchinovitz and J.Y. Yu. **Adaptive and optimal online linear regression on ℓ^1-balls**. *In Kivinen, Jyrki et al. (ed.),* Algorithmic Learning Theory*, volume 6925 of* Lecture Notes in Computer Science, pages 99–113, 2011. (actes de la conférence ALT 2011)

Note : Une version étendue de 31 pages correspondant au chapitre 4 de ma thèse a été soumise (sur invitation) au journal *Theoretical Computer Science*.

Les articles [1] et [2] seront fournis au jury lors de l'audition. Ils sont également disponibles – ainsi que le manuscrit de ma thèse – sur la page web :

http://www.math.ens.fr/∼gerchino.

4 Communications orales

4.1 Conférences internationales

Oct. 2011 **Conférence ALT 2011**, Espoo, Finlande.
Adaptive and optimal online linear regression on ℓ^1-balls.

Juil. 2011 **Conférence COLT 2011**, Budapest, Hongrie.
Sparsity regret bounds for individual sequences in online linear regression.

4.2 Congrès nationaux

Mai 2011 **43èmes Journées de Statistique**, Tunis.
Bornes de sparsité en suites individuelles dans un cadre de régression linéaire séquentielle.

Mai 2010 **42èmes Journées de Statistique**, Marseille.
Vitesse minimax du regret interne en prédiction de suites individuelles.

4.3 Séminaires et groupes de travail

Sauf mention contraire, les exposés suivants portaient sur le thème : *Régression linéaire séquentielle pour des suites déterministes arbitraires. Liens avec le cadre statistique classique.*

Mars 2012 **Séminaire de Probabilités et Statistiques**, Université de Lorraine, Nancy.

Mars 2012 **Séminaire SAMM**, Université Paris 1.

Fév. 2012 **Séminaire du MODAL'X**, Université Paris-Ouest Nanterre.

Fév. 2012 **Séminaire de Mathématiques Appliquées**, Université de Nantes.

Fév. 2012 **Séminaire à l'École Centrale de Lyon**, Écully.

Jan. 2012 **Séminaire de Statistiques de l'IMT**, Université Paul Sabatier, Toulouse.

Jan. 2012 **Séminaire sur la non stationnarité et la gestion des risques**, Université de Cergy-Pontoise.

Jan. 2012 **Séminaire de Statistiques du MAP5**, Université Paris 5.

Nov. 2011 **Séminaire de Probabilités et Statistiques du LAGA**, Université Paris 13.

Oct. 2011 **Séminaire de Probabilités et Statistiques**, Université Montpellier 2.

Août 2011 **Workshop StatMathAppli 2011**, Fréjus.
Aggregation of nonlinear models.

Mars 2011 **Séminaire de Statistiques du CREST**, ENSAE, Malakoff.

Fév. 2011 **Séminaire d'apprentissage statistique SMILE**, ENS Paris.

Déc. 2010 **Séminaire hebdomadaire de l'équipe SEQUEL**, INRIA Lille.

Août 2010 **Workshop StatMathAppli 2010**, Fréjus.
Minimax rate of internal regret in prediction of individual sequences.

J'ai aussi donné des exposés aux séminaires des doctorants des universités Paris 6, Paris 7, Paris-Sud 11, à Télécom ParisTech, et au séminaire informel de l'équipe Probabilités-Statistiques du DMA (ENS Paris).

5 Activités scientifiques connexes

5.1 Rapporteur d'articles scientifiques

J'ai été rapporteur pour les journaux et conférences suivants :

Journaux IEEE Transactions on Information Theory (1 article)
Annals of Statistics (1 article)

Conférences COLT 2011 (1 article)
AISTATS 2012 (1 article court)
COMPSTAT 2010 (3 résumés)

5.2 Déplacements scientifiques à l'étranger (hors communications orales)

Mars 2011 **Séjour de recherche chez Nicolò Cesa-Bianchi**, Università degli Studi di Milano (1 semaine).

Juil. 2009 **Conférence COLT 2009**, Montréal (1 semaine).

5.3 Vulgarisation mathématique

2009-2010 **Encadrement d'un club MATh.en.JEANS**, Lycée René Cassin, Arpajon.
Sujet choisi par les élèves : *Statistiques et prévision météorologique*.

5.4 Informations diverses

- Membre de la Société Française de Statistique.
- Membre non permanent du projet ANR EXPLO-RA.

6 Résumé des travaux de recherche

J'ai effectué ma thèse intitulée

"Prédiction de suites individuelles et cadre statistique classique : étude de quelques liens autour de la régression parcimonieuse et des techniques d'agrégation"

sous la direction de Gilles Stoltz à l'École normale supérieure, Paris (rattachement à l'Université Paris-Sud 11, Orsay). Ma soutenance a eu lieu le 12 décembre 2011 devant le jury composé de :

M. Pierre	ALQUIER	CREST et Université Paris 7	Examinateur
M. Olivier	CATONI	CNRS et École normale supérieure	Examinateur
M. Arnak	DALALYAN	CREST et ENSAE	Rapporteur
M. Pascal	MASSART	Université Paris-Sud 11	Président
M. Gilles	STOLTZ	CNRS et École normale supérieure	Directeur
M. Alexandre	TSYBAKOV	CREST, ENSAE et Université Paris 6	Examinateur

au vu du rapport également écrit par Claudio Gentile (Università degli Studi dell'Insubria, Varèse).

Dans cette thèse, je me suis intéressé à deux types de problèmes d'apprentissage, tous deux du domaine de la prévision :

- Le cadre principal de cette thèse est celui de la prévision de suites déterministes arbitraires (ou *suites individuelles*). Au confluent entre les statistiques, la théorie de l'apprentissage (*machine learning*), la théorie de l'information et la théorie des jeux, ce cadre recouvre des problèmes d'apprentissage séquentiel où l'on ne fait aucune hypothèse de stochasticité sur la suite $(y_t)_{t\geqslant 1}$ des données à prévoir. Les algorithmes séquentiels qui en résultent bénéficient de garanties déterministes – valables dans le pire des cas – et sont donc en ce sens très robustes.
- Nous nous sommes également intéressés aux liens étroits entre la prévision de suites individuelles et des cadres satistiques plus classiques comme le modèle de régression avec *design* fixe ou aléatoire, où les données observées $(Y_t)_{t\geqslant 1}$ sont cette fois modélisées de façon stochastique.

Nous avons abordé au total quatre problèmes voisins, dont trois du domaine de la régression. Les résultats obtenus sont présentés aux chapitres 3 à 6, et deux ont fait l'objet de publications dans des conférences internationales très sélectives : [Ger11] pour le chapitre 3 et [GY11] pour le chapitre 4. Nous exposons ci-après à très grands traits les principales contributions de cette thèse, toutes au croisement du cadre des suites individuelles et du cadre statistique classique.

6.1 Régression linéaire séquentielle pour des suites déterministes arbitraires

Les chapitres 3 et 4 considèrent comme cadre principal celui de la régression linéaire séquentielle pour des suites déterministes arbitraires. Ce cadre peut être décrit comme suit.

Un statisticien doit prévoir de façon séquentielle, à chaque date $t = 1, 2, \ldots$, la valeur $y_t \in \mathbb{R}$ d'une suite inconnue d'observations. Pour ce faire, il dispose d'un vecteur $\boldsymbol{x}_t \triangleq (x_{j,t})_{1\leqslant j\leqslant d}$ de prévisions élémentaires $x_{j,t} \in \mathbb{R}$, à partir desquelles il formule sa propre prévision $\widehat{y}_t \in \mathbb{R}$ (il peut utiliser les données passées $(\boldsymbol{x}_s, y_s)_{1\leqslant s\leqslant t-1}$ pour combiner les $x_{j,t}$). La qualité des prévisions est évaluée avec la perte carrée. L'objectif du statisticien est de prévoir *à terme* presque aussi bien que le meilleur prédicteur linéaire $\boldsymbol{x} \in \mathbb{R}^d \mapsto \boldsymbol{u}\cdot\boldsymbol{x} \triangleq \sum_{j=1}^d u_j x_j$, où $\boldsymbol{u} \in \mathbb{R}^d$, i.e., de vérifier, uniformément sur toutes les suites déterministes $(\boldsymbol{x}_t, y_t)_{1\leqslant t\leqslant T}$, une inégalité de la forme suivante (qualifiée de *borne de regret*) :

$$\sum_{t=1}^{T}\bigl(y_t - \widehat{y}_t\bigr)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\bigl(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\bigr)^2 + \Delta_{T,d}(\boldsymbol{u})\right\}\,, \tag{6.1}$$

pour un terme de regret $\Delta_{T,d}(\boldsymbol{u})$ aussi petit que possible et, en particulier, sous-linéaire en T. (Par soucis de clarté, on omet les dépendances de $\Delta_{T,d}(\boldsymbol{u})$ en les amplitudes $\max_{1\leqslant t\leqslant T}\|\boldsymbol{x}_t\|_\infty$ et $\max_{1\leqslant t\leqslant T}|y_t|$.)

6.1.1 Bornes de parcimonie en régression linéaire séquentielle (chap.3 ou [Ger11])

Dans le cadre ci-dessus, un exemple d'algorithme séquentiel (i.e., une procédure pour produire les prévisions $\widehat{y}_t$ en fonction des données disponibles) est donné par l'algorithme *ridge séquentiel*, qui est une extension au cadre déterministe de l'algorithme *ridge* étudié trente ans plus tôt dans le cadre stochastique. [AW01] et [Vov01] ont montré que cette version séquentielle assure un regret d'ordre au plus $d \ln T$.

La borne $d \ln T$, qui est l'analogue de la vitesse paramétrique d/T de l'estimateur des moindres carrés en statistique, est sous-linéaire en T en faible dimension $d \ll T/\ln T$, mais est inutile en grande dimension $d \gtrsim T/\ln T$. En revanche, tout comme dans le cadre statistique classique, nous avons montré qu'il est possible d'assurer un petit regret sous une hypothèse supplémentaire qualifiée d'*hypothèse de parcimonie*, à savoir s'il existe une combinaison linéaire $\boldsymbol{u}^* \in \mathbb{R}^d$ parcimonieuse (*sparse* en anglais, i.e., avec $s \ll T/\ln(dT)$ coordonnées non nulles) dont la perte cumulée est petite.

La première contribution du chapitre 3 a en effet été d'étendre des idées du cadre statistique classique au cadre déterministe afin de construire un algorithme séquentiel dont le regret est au plus de l'ordre de $s \ln(dT)$, qui est donc sous-linéaire en T sous l'hypothèse de parcimonie $s \ll T/\ln(dT)$. Pour ce faire, nous avons considéré un algorithme procédant par pondération exponentielle : à l'instant t, la prévision $\widehat{y}_t$ produite est un mélange convexe de toutes les prévisions linéaires $\boldsymbol{u} \cdot \boldsymbol{x}_t$ (avec $\boldsymbol{u} \in \mathbb{R}^d$) préalablement seuillées à un seuil $B > 0$, et où la probabilité utilisée pour le mélange est de type Gibbs – cf. figure 6.1.

Notre algorithme SeqSEW (*Sequential Sparse Exponential Weighting*) est inspiré de l'algorithme *Sparse Exponential Weighting* introduit par [DT08, DT11] dans le cadre stochastique. Notre première contribution a été de montrer que, moyennant quelques adaptations clés (seuillage séquentiel des prévisions), cet algorithme bénéficiait en fait de garanties déterministes de la forme (6.1) avec $\Delta_{T,d}(\boldsymbol{u})$ au plus de l'ordre de $\Delta_{T,d}(\boldsymbol{u}) \lesssim \|\boldsymbol{u}\|_0 \ln\bigl(dT \|\boldsymbol{u}\|_1\bigr)$, où $\|\boldsymbol{u}\|_0$ désigne le nombre de coordonnées non nulles de $\boldsymbol{u}$ et où $\|\boldsymbol{u}\|_1 \triangleq \sum_{j=1}^d |u_j|$. Nous appellons de telles bornes de regret des *bornes de parcimonie*. Notre preuve repose sur une analyse PAC-bayésienne séquentielle inspirée des travaux de [Cat04, Aud09] et exploite la forme à queue lourde de la loi a priori π_τ introduite par [DT08].

Paramètres : seuil $B > 0$, température inverse $\eta > 0$ et résolution $\tau > 0$ à laquelle on associe la loi a priori π_τ sur $\mathbb{R}^d$ défini par

$$\pi_\tau(\mathrm{d}\boldsymbol{u}) \triangleq \prod_{j=1}^d \frac{(3/\tau)\,\mathrm{d}u_j}{2\bigl(1+|u_j|/\tau\bigr)^4} \ .$$

Initialisation : $p_1 \triangleq \pi_\tau$.

A chaque tour de prévision $t \geqslant 1$,

1. Recevoir la donnée $\boldsymbol{x}_t \in \mathbb{R}^d$ et prévoir $\widehat{y}_t \triangleq \displaystyle\int_{\mathbb{R}^d} \bigl[\boldsymbol{u} \cdot \boldsymbol{x}_t\bigr]_B \, p_t(\mathrm{d}\boldsymbol{u})$,
   où $[x]_B \triangleq \max\bigl\{-B, \min\{B, x\}\bigr\}$;
2. Recevoir l'observation $y_t \in \mathbb{R}$ et calculer la probabilité a posteriori p_{t+1} sur $\mathbb{R}^d$ via l'expression (W_{t+1} est une constante de renormalisation)

$$p_{t+1}(\mathrm{d}\boldsymbol{u}) \triangleq \frac{1}{W_{t+1}} \exp\left(-\eta \sum_{s=1}^t \Bigl(y_s - \bigl[\boldsymbol{u} \cdot \boldsymbol{x}_s\bigr]_B\Bigr)^2\right) \pi_\tau(\mathrm{d}\boldsymbol{u}) \ .$$

FIGURE 6.1 – Définition de l'algorithme $\mathrm{SeqSEW}_\tau^{\mathrm{B},\eta}$.

La deuxième contribution du chapitre 3 concerne la calibration des paramètres de l'algorithme SeqSEW. La version la plus simple de notre algorithme (cf. figure 6.1) est en effet calibrée à partir de quantités inconnues en pratique : pour prouver nos bornes de parcimonie, B, η et τ sont choisis en fonction des amplitudes finales des observations $\max_{1 \leqslant t \leqslant T} |y_t|$ et des prévisions élémentaires $\max_{1 \leqslant t \leqslant T} \|\boldsymbol{x}_t\|_\infty$. Nous avons alors proposé une calibration séquentielle des paramètres de notre algo-

rithme qui permet d'obtenir une borne de parcimonie similaire, mais de façon totalement automatique (i.e., à l'instant t, les paramètres B, η et τ sont remplaçés par des paramètres B_t, η_t et τ_t choisis uniquement en fonction des données disponibles à l'instant t). Comme expliqué ci-après, cette adaptativité dans le cadre séquentiel déterministe permet d'obtenir des résultats d'adaptativité dans le cadre stochastique.

La troisième contribution du chapitre 3 a trait à l'application de nos bornes de parcimonie *déterministes* au cadre stochastique. On se place maintenant dans le modèle de régression avec *design* aléatoire : le statisticien à accès à T copies indépendantes $(X_1, Y_1), \dots, (X_T, Y_T)$ de $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ de loi inconnue. On suppose $\mathbb{E}[Y^2] < \infty$ et on note $f : \boldsymbol{x} \in \mathbb{R}^d \mapsto \mathbb{E}[Y|X = \boldsymbol{x}]$ la fonction de régression. En posant de plus $\varepsilon_t \triangleq Y_t - f(X_t)$, on a $Y_t = f(X_t) + \varepsilon_t$ et $\mathbb{E}[\varepsilon_t|X_t] = 0$ presque surement pour tout $t = 1, \dots, T$. L'objectif est d'estimer la fonction de régression $f : \mathbb{R}^d \to \mathbb{R}$ inconnue à partir de la seule donnée de l'échantillon $(X_1, Y_1), \dots, (X_T, Y_T)$.

Bien que l'échantillon $(X_t, Y_t)_{1 \leqslant t \leqslant T}$ soit observé entièrement dès le début de la tâche de prévision, nous le traitons séquentiellement, de la date 1 à la date T, au moyen d'une version totalement automatique de notre algorithme séquentiel SeqSEW. Cet algorithme vérifie une borne de parcimonie déterministe, i.e., une borne de regret de la forme (6.1) avec $\Delta_{T,d}(\boldsymbol{u}) \lesssim \|\boldsymbol{u}\|_0 \ln\bigl(dT \|\boldsymbol{u}\|_1\bigr)$. Cette borne est valable presque sûrement. En supposant le bruit $\varepsilon_t \triangleq Y_t - f(X_t)$ indépendant de X_t et gaussien de variance inconnue σ^2 (des formes plus générales de bruit sont traitées via une analyse générique), cette borne déterministe implique, par intégration, une borne en espérance de la forme

$$\mathbb{E}\Bigl[\bigl(f(X) - \widehat{f}_T(X)\bigr)^2\Bigr] \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \mathbb{E}\Bigl[\bigl(f(X) - \boldsymbol{u} \cdot X\bigr)^2\Bigr] + \frac{c \|\boldsymbol{u}\|_0 \left(\|f\|_\infty^2 + \sigma^2 \ln T\right)}{T} \ln\bigl(dT \|\boldsymbol{u}\|_1\bigr) \right\} + \mathcal{R} \,,$$

où l'espérance est prise par rapport à tout l'aléa (i.e., $(X_1, Y_1, \dots, X_T, Y_T, X)$), où $c > 0$ est une constante absolue, où le terme de reste $\mathcal{R}$ est généralement négligeable devant le deuxième terme de l'accolade, et où l'estimateur $\widehat{f}_T : \mathbb{R}^d \to \mathbb{R}$ de la fonction de régression est défini pour tout $\boldsymbol{x} \in \mathbb{R}^d$ par $\widehat{f}_T(\boldsymbol{x}) \triangleq \frac{1}{T} \sum_{t=1}^T \int_{\mathbb{R}^d} \bigl[\boldsymbol{u} \cdot \boldsymbol{\varphi}(\boldsymbol{x})\bigr]_{B_t} p_t(\mathrm{d}\boldsymbol{u})$.

La borne de risque précédente sur l'estimateur $\widehat{f}_T$ est une *inégalité oracle de parcimonie* similaire à celle de [DT11, Proposition 1] à des facteurs logarithmiques près, mais est obtenue de façon adaptative car notre algorithme – totalement automatique – n'est pas calibré en fonction de la variance inconnue σ^2 du bruit. Nous avons ainsi montré que des techniques de suites individuelles s'avèrent utiles à des fins d'adaptativité dans le cadre statistique classique.

6.1.2 Régression linéaire séquentielle optimale et adaptative sur des boules ℓ^1 (chapitre 4 ou [GY11])

Au chapitre 4, nous avons considéré un objectif de régression linéaire séquentielle voisin : il s'agissait de prévoir presque aussi bien que le meilleur prédicteur linéaire $\boldsymbol{x} \mapsto \boldsymbol{u} \cdot \boldsymbol{x}$ de norme $\|\boldsymbol{u}\|_1 \triangleq \sum_{j=1}^d |u_j|$ bornée, i.e., pour un rayon $U > 0$ et un horizon de temps $T \geqslant 1$ (fixés ou pas), de minimiser le regret

$$\sum_{t=1}^T (y_t - \widehat{y}_t)^2 - \min_{\|\boldsymbol{u}\|_1 \leqslant U} \left\{ \sum_{t=1}^T (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 \right\} \,. \tag{6.2}$$

L'objectif de comparaison à des boules ℓ^1 a plusieurs intérêts, notamment :

- Cela étend la tâche d'*agrégation convexe* (étudiée par [Nem00, Tsy03] en statistique).
- Une majoration du regret (6.2) par $f_{T,d}(U)$ pour tout $U > 0$ (avec $f_{T,d} : \mathbb{R} \to \mathbb{R}$ décroissante) implique une borne *régularisée* de la forme

$$\sum_{t=1}^T (y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T \bigl(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t\bigr)^2 + f_{T,d}(\|\boldsymbol{u}\|_1) \right\} \,.$$

  De telles bornes ont été obtenues en statistique pour l'estimateur Lasso par [MM11].

Notre première contribution a été de déterminer l'ordre de grandeur du regret minimax : il s'agit de la quantité $\inf\sup\{\text{regret}\}$, où le regret est défini par (6.2), où le supremum est pris sur toutes les suites déterministes $(\boldsymbol{x}_t, y_t)_{1\leqslant t\leqslant T}$ bornées par $\max_{1\leqslant t\leqslant T}\|\boldsymbol{x}_t\|_\infty \leqslant X$ et $\max_{1\leqslant t\leqslant T}|y_t| \leqslant Y$, et où l'infimum est pris sur tous les algorithmes séquentiels $(\widehat{y}_t)_{t\geqslant 1}$. Cette quantité minimax, qui mesure la meilleure performance possible (en termes de regret dans le pire des cas) dépend des paramètres du problème X, Y, T, d et U. Comme dans le cadre statistique classique [Tsy03], nous avons montré un phénomène de transition d'une vitesse lente à une vitesse rapide : sur un premier régime, le regret minimax croît en $\sqrt{T}$ et sur un second régime, le regret minimax croît en $\ln T$.

La deuxième contribution du chapitre 4 concerne l'adaptation efficace en les paramètres du problème. Certains algorithmes utilisés pour l'analyse minimax sont en effet algorithmiquement coûteux en grande dimension d et sont calibrés en fonction de X, Y, T, d ou U. Nous avons alors proposé un algorithme à poids exponentiels de complexité algorithmique linéaire en d à chaque instant t, calibré séquentiellement en fonction des données seulement (notamment à l'aide des techniques de [CBMS07]), et dont le regret est quasi-optimal sur le premier régime (vitesse en $\sqrt{T}$), le deuxième régime pouvant être traité de façon similaire.

6.2 Autres liens avec le cadre statistique classique

6.2.1 Vitesses minimax des regrets interne et *swap* (chapitre 5)

Au chapitre 5, nous avons étudié d'autres formes de regret qui jouent un rôle important en théorie des jeux. Le cadre séquentiel diffère du précédent car la perte est linéaire et la décision du statisticien appartient au simplexe $\mathcal{X}_K \triangleq \left\{\boldsymbol{x}\in\mathbb{R}_+^K, \quad \sum_{i=1}^K x_i = 1\right\}$. A chaque date $t\in\mathbb{N}^* \triangleq \{1,2,\ldots\}$, les trois mêmes étapes se succèdent :

1. Le statisticien choisit un vecteur de poids $\boldsymbol{p}_t = (p_{i,t})_{1\leqslant i\leqslant K} \in \mathcal{X}_K$ entre K actions.
2. Chaque action $i = 1,\ldots,K$ encourt une perte $\ell_{i,t}\in[0,1]$ et l'environnement révèle le vecteur de pertes $\boldsymbol{\ell}_t \triangleq (\ell_{i,t})_{1\leqslant i\leqslant K}$.
3. Le statisticien encourt la *perte linéaire* $\boldsymbol{p}_t\cdot\boldsymbol{\ell}_t \triangleq \sum_{i=1}^K p_{i,t}\ell_{i,t}$.

Dans ce cadre, le *regret interne* et le *regret swap* sont de la forme

$$\sum_{t=1}^T \boldsymbol{p}_t\cdot\boldsymbol{\ell}_t - \min_{\varphi\in\Phi}\sum_{t=1}^T \varphi(\boldsymbol{p}_t)\cdot\boldsymbol{\ell}_t\ , \tag{6.3}$$

où Φ est l'ensemble des applications linéaires $\varphi : \mathcal{X}_K\to\mathcal{X}_K$ qui préservent le simplexe $\mathcal{X}_K$ (pour le regret *swap*) ou un ensemble plus petit (pour le regret interne).

Nous avons étudié les vitesses (ou ordres de grandeur) minimax des regrets interne et *swap* en environnement stochastique ou déterministe, i.e., quand la suite $(\boldsymbol{\ell}_t)_{1\leqslant t\leqslant T}$ est aléatoire i.i.d. ou déterministe arbitraire. Ces vitesses étaient partiellement connues depuis les travaux de [Sto05, BM07b]. Nous avons complété ces résultats :

1. En déterminant les vitesses minimax exactes des regrets interne et *swap* en environnement stochastique (respectivement de l'ordre de $\sqrt{T}$ et $\sqrt{T\ln K}$) ;
2. En exhibant une borne inférieure pour le regret *swap* en environnement déterministe de l'ordre de $\sqrt{TK}$, donc proche de la borne supérieure $\sqrt{TK\ln K}$ connue ; cette borne inférieure est obtenue via des arguments stochastiques et l'inégalité de Pinsker ;
3. En développant une technique stochastique qui permet de retrouver les meilleures bornes supérieures connues en environnement déterministe et de majorer d'autres formes de regret.

6.2.2 Agrégation de modèles non linéaires (chapitre 6)

Dans ce dernier chapitre, qui traite à nouveau de régression, on présente des résultats partiels sur des estimateurs par pondération exponentielle – comme aux chapitres 3 et 4 – mais dans un cadre de sélection de modèles.

La version finie-dimensionnelle du cadre étudié au chapitre 6 est le modèle de régression gaussienne avec *design* fixe : le statisticien observe le vecteur $(Y_1, \dots, Y_n) \in \mathbb{R}^n$ donné par

$$Y_i = s_i + \sigma \xi_i \in \mathbb{R} \ , \quad 1 \leqslant i \leqslant n \ ,$$

où les variables aléatoires $\xi_1, \dots, \xi_n$ sont i.i.d. de loi $\mathcal{N}(0,1)$, où $\sigma > 0$ est le niveau de bruit supposé connu, et où $s = (s_1, \dots, s_n) \in \mathbb{R}^n$ est un vecteur déterministe inconnu.

L'objectif du statisticien est d'estimer s en fonction de $Y = (Y_1, \dots, Y_n) \in \mathbb{R}^n$. La performance d'un estimateur $\widetilde{s} \in \mathbb{R}^n$ est évaluée via son risque quadratique (empirique) $\|\widetilde{s} - s\|_n^2$, où l'on pose $\|u\|_n^2 \triangleq n^{-1} \sum_{i=1}^n u_i^2$ pour tout $u \in \mathbb{R}^n$.

Afin d'estimer le vecteur s, le statisticien a accès à une famille au plus dénombrable $(S_m)_{m \in \mathcal{M}}$ de parties non vides de $\mathbb{R}^n$ (appelées *modèles non linéaires* ci-après) ; il dispose alors des estimateurs des moindres carrés $\widehat{s}_m \in \operatorname{argmin}_{t \in S_m} \|Y - t\|_n^2$, $m \in \mathcal{M}$. Dès lors, un objectif classique est le suivant : étant donnée la famille $\big(\widehat{s}_m\big)_{m \in \mathcal{M}}$ d'estimateurs des moindres carrés associés aux modèles non linéaires $S_m \subset \mathbb{R}^n$, comment estimer le vecteur s presque aussi bien que le meilleur des estimateurs $\widehat{s}_m$, $m \in \mathcal{M}$?

Pour ce faire, nous avons considéré un estimateur $\widetilde{s}$ qui mélange les $\widehat{s}_m$ par *pondération exponentielle*, de même que [LB06] dans le cas où les modèles S_m étaient linéaires (i.e., des sous-espaces vectoriels de $\mathbb{R}^n$). Plus précisément, notre estimateur est une variante bayésienne de la procédure de sélection de modèles de [Mas07] : il est de la forme $\widetilde{s}^{(\eta)} = \sum_{m \in \mathcal{M}} \widehat{\rho}_m^{(\eta)} \widehat{s}_m$, où

$$\widehat{\rho}_m^{(\eta)} = \frac{\exp\Big[-\eta\big(\|Y - \widehat{s}_m\|_n^2 + \operatorname{pen}^{(\eta)}(m)\big)\Big]}{\sum_{m' \in \mathcal{M}} \exp\Big[-\eta\big(\|Y - \widehat{s}_{m'}\|_n^2 + \operatorname{pen}^{(\eta)}(m')\big)\Big]} \ , \quad m \in \mathcal{M} \ ,$$

où $\eta > 0$ est un paramètre de l'algorithme et où la pénalité $\operatorname{pen}^{(\eta)} : \mathcal{M} \to \mathbb{R}+$ est choisie en fonction de la variance σ^2 du bruit et d'une *dimension généralisée* des modèles S_m, tout comme dans [Mas07].

En utilisant une formule de dualité sur la divergence de Kullback-Leibler prouvée par [Cat04] et des arguments de concentration développés par [BM01, Mas07] pour une procédure de sélection de modèles, nous avons montré que notre procédure d'agrégation de modèles vérifie une inégalité de type oracle en espérance et avec grande probabilité. Nous avons par exemple prouvé le résultat suivant : pour une pénalité $\operatorname{pen}^{(\eta)}(m)$ suffisamment grande (en fonction notamment d'une dimension généralisée du modèle S_m), l'estimateur $\widetilde{s}^{(\eta)} = \sum_{m \in \mathcal{M}} \widehat{\rho}_m^{(\eta)} \widehat{s}_m$ vérifie, pour tout $s \in \mathbb{R}^n$ et tout $z > 0$, avec probabilité au moins égale à $1 - \Sigma^2 e^{-z}$,

$$\Big\| \widetilde{s}^{(\eta)} - s \Big\|_n^2 \leqslant C \inf_{m \in \mathcal{M}} \left\{ d^2(s, S_m) + \operatorname{pen}^{(\eta)}(m) + \frac{\ln \Sigma}{\eta} + \frac{\sigma^2}{n}(z+1) \right\} - \mathcal{J}(\widehat{\rho}^{(\eta)}) \ , \tag{6.4}$$

où $C > 1$ est une constante, où Σ mesure la complexité de la famille de modèles $(S_m)_{m \in \mathcal{M}}$, et où $\mathcal{J}(\rho) \triangleq \sum_{m \in \mathcal{M}} \rho_m \|\widehat{s}_m - s\|_n^2 - \big\| \sum_{m \in \mathcal{M}} \rho_m \widehat{s}_m - s \big\|_n^2 \geqslant 0$.

Ces travaux étendent donc ceux de [LB06] au cas de modèles S_m non linéaires (à constante multiplicative près). Par ailleurs, ils pointent un lien naturel entre l'agrégation de modèles et la sélection de modèles, puisque notre inégalité de type oracle est valide pour un continuum d'estimateurs $\big\{\widetilde{s}^{(\eta)} : \eta > 0\big\}$ qui s'étend de l'agrégation de modèles classique (où η est au plus de l'ordre de n/σ^2) à la sélection de modèles (où $\eta = +\infty$).

7 Projet de recherche

Dans les années à venir, je souhaiterais poursuivre l'étude des liens entre les deux cadres que j'ai abordés pendant ma thèse :

- le cadre statistique classique, où la suite des observations $(Y_t)_{t\geqslant 1}$ est modélisée de façon stochastique, et où l'on cherche des garanties en espérance ou avec grande probabilité ;
- le cadre des *suites individuelles*, qui est un cadre séquentiel déterministe où l'on ne fait aucune hypothèse de stochasticité sur la suite $(y_t)_{t\geqslant 1}$ des données à prévoir.

Les questions décrites ci-après illustrent la multiplicité des apports mutuels que l'on peut espérer pointer entre le cadre statistique classique et celui des suites individuelles, et auxquels je souhaiterais notamment m'atteler ces prochaines années. Certains problèmes se situent dans le prolongement direct de mes travaux de thèse (sections 7.1, 7.2 et début de la section 7.3), tandis que d'autres en sont beaucoup plus éloignés (fin de la section 7.3 et section 7.4). Je serais par ailleurs ravi de m'investir dans d'autres projets, suivant les souhaits et les besoins de l'équipe dans laquelle j'effectuerai mes recherches.

7.1 Régression linéaire séquentielle parcimonieuse

Au chapitre 3 de ma thèse (cf. aussi [Ger11]), nous avons importé la notion statistique d'*inégalité oracle de parcimonie* dans un cadre de suites déterministes arbitraires et avons traité des problèmes d'adaptativité (dans un cadre déterministe dans un premier temps, puis, en corollaire, dans un cadre statistique classique). Ces résultats peuvent être prolongés de la façon suivante ; la deuxième piste de recherche est en cours d'étude. Nous reprenons les notations de la section 6.1.

Peut-on modifier l'algorithme séquentiel SeqSEW pour produire des combinaisons linéaires parcimonieuses ? Le prédicteur séquentiel SeqSEW vérifie des bornes parcimonie, mais ses prévisions $\widehat{y}_t$ ne sont en général pas – au premier abord du moins – des combinaisons linéaires parcimonieuses des prévisions de base $x_{j,t}$, i.e., des prévisions de la forme $\widehat{y}_t = \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t$ avec $\|\widehat{\boldsymbol{u}}_t\|_0 \ll T$. En fait, les prévisions $\widehat{y}_t$ de l'algorithme SeqSEW ont une forme un peu plus élaborée car elles font intervenir une opération de seuillage avant le mélange convexe ; en utilisant les notations de la figure 1 (cf. section 6.1.1), elles sont en effet de la forme :

$$\widehat{y}_t \triangleq \int_{\mathbb{R}^d} \big[\boldsymbol{u} \cdot \boldsymbol{x}_t\big]_B \, p_t(\mathrm{d}\boldsymbol{u}) \ , \tag{7.1}$$

où p_t est une probabilité sur $\mathbb{R}^d$ de type Gibbs construite à partir des données passées $(\boldsymbol{x}_s, y_s)_{1\leqslant s\leqslant t-1}$, et où l'opération de seuillage $[\,\cdot\,]_B$ est définie par $[z]_B \triangleq \max\big\{-B, \min\{B, z\}\big\}$ pour tous $z \in \mathbb{R}$ et $B > 0$.

En grande dimension, produire des prévisions qui sont des combinaisons linéaires parcimonieuses des prévisions de base pourrait pourtant être utile d'un point de vue statistique (à des fins de sélection de variables) et algorithmique (pour diminuer l'espace mémoire nécessaire). On pourrait envisager de modifier notre prédicteur SeqSEW en remarquant que la probabilité a priori $\pi_\tau(\mathrm{d}\boldsymbol{u})$ choisie sur $\mathbb{R}^d$ (et donc, dans une moindre mesure, les probabilités a posteriori $p_t(\mathrm{d}\boldsymbol{u})$ associées) charge(nt) davantage les combinaisons linéaires $\boldsymbol{u}$ approximativement parcimonieuses. Dans le modèle de régression avec *design* fixe, [DT09] remarquent ainsi sur des simulations que leur algorithme exponentiel sélectionne correctement les variables pertinentes (pourvu qu'une troncature raisonnable soit appliquée aux composantes de la combinaison linéaire produite) ; voir [DT09, section 5.2.1]. Dans

notre cadre séquentiel, on pourrait envisager d'étudier si de telles propriétés sont vraies (d'un point de vue pratique ou théorique) pour une modification appropriée de l'algorithme SeqSEW. Autrement dit, peut-on approcher les prévisions $\widehat{y}_t$ définies par (7.1) par des prévisions de la forme $\widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t$ avec $\|\widehat{\boldsymbol{u}}_t\|_0$ petit (par ex., $\|\widehat{\boldsymbol{u}}_t\|_0 \ll t$) et telles que la perte cumulée encourue soit proche ?

Peut-on prouver des bornes de parcimonie pour des algorithmes séquentiels parcimonieux ? Une autre piste de recherche, actuellement en cours d'étude, consiste à tenter de prouver des bornes de parcimonie pour des algorithmes séquentiels dont on sait qu'ils produisent des combinaisons parcimonieuses. Un exemple de tel algorithme est donné par une variante séquentielle de l'estimateur Lasso [1] introduit dans le cadre statistique classique par [Tib96, DJ94] ; cette variante produit la prévision $\widehat{y}_t = \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t$, où $\widehat{\boldsymbol{u}}_t$ est donné par

$$\widehat{\boldsymbol{u}}_t^{(\lambda)} \in \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^d} \left\{ \sum_{s=1}^{t-1} (y_s - \boldsymbol{u}\cdot\boldsymbol{x}_s)^2 + \lambda \|\boldsymbol{u}\|_1 \right\}$$

pour un paramètre de régularisation $\lambda \geqslant 0$ à calibrer judicieusement.

Afin de prouver une borne de parcimonie sur cet algorithme (ou l'une de ses variantes), j'ai pour l'instant tenté d'adapter des arguments employés par la plupart des travaux statistiques sur l'estimateur Lasso dans le cadre stochastique. Une technique de preuve récurrente consiste ainsi à reformuler le caractère optimal de $\widehat{\boldsymbol{u}}_t^{(\lambda)}$ puis à en déduire une borne déterministe sur le risque de l'estimateur Lasso ; cf. [CT07, vdG08, BRT09] par exemple. Sous des hypothèses classiques sur la matrice de Gram motivées par la théorie statistique du *compressed sensing* et stipulant essentiellement que les covariables sont presque orthogonales, la borne déterministe mentionnée précédemment implique une inégalité oracle de parcimonie avec grande probabilité pourvu que le paramètre de régularisation soit choisi suffisamment grand (généralement de l'ordre de $\sigma\sqrt{\ln(d)/T}$, si le bruit est Gaussien de variance σ^2). La technique de preuve précédente, qui est un argument déterministe appliqué sur un événement de grande probabilité, peut être adaptée directement à notre cadre séquentiel déterministe, mais ne permet pour l'instant que de prouver des bornes de parcimonie avec une vitesse lente en T (de l'ordre de $\sqrt{T}$ au lieu de la vitesse rapide $\ln T$ escomptée).

Obtenir une vitesse rapide est crucial, non seulement car il s'agit de la vitesse optimale (nous avions déjà cette vitesse avec l'algorithme SeqSEW), mais aussi car cela permettrait d'en déduire une borne non triviale (en $\sqrt{T}$ dans le pire des cas) pour une variante séquentielle de l'algorithme Elastic-Net (régularisation $\ell^1 + \ell^2$), et ce sans hypothèse sur la matrice de Gram, tout comme [HvdG11] dans le cadre stochastique. Au vu des premières bornes avec vitesse lente obtenues, il semble donc qu'un argument spécifique au cadre séquentiel déterministe soit nécessaire.

Pour ce faire, il pourrait certainement être utile de modifier légèrement la forme des prévisions $\widehat{\boldsymbol{u}}_t^{(\lambda)}$ en ajoutant au critère de minimisation le même terme technique $(\boldsymbol{u}\cdot\boldsymbol{x}_t)^2$ que [AW01] avaient employé pour étendre l'algorithme *ridge* du cadre stochastique au cadre déterministe séquentiel. Le point de départ de leur analyse devrait aussi pouvoir être utilisé dans notre cas.

Notons par ailleurs qu'en plus de produire des combinaisons parcimonieuses, l'algorithme Lasso et ses variantes ont l'avantage de pouvoir être implémentés avec un coût algorithmique faible. Cela contraste ainsi avec notre prédicteur théorique SeqSEW, qui pourrait certes être approché numériquement par des méthodes de Langevin Monte-Carlo étudiées par [DT09] dans le cadre stochastique, mais qui ne jouit pour l'instant pas de garanties théoriques quant à la précision de cette approximation.

De surcroît, l'implémentation du type LARS [EHJT04] disponible pour l'algorithme Lasso (mais qui conviendrait aussi pour la modification avec $(\boldsymbol{u}\cdot\boldsymbol{x}_t)^2$ suggérée précédemment) permet de calculer

1. Notons qu'on pourrait également tenter de prouver des bornes de parcimonie pour les algorithmes séquentiels de [LLZ09, SST09, Xia10, DSSST10]. Ces algorithmes produisent en pratique des combinaisons qui sont souvent parcimonieuses, mais ce fait n'est à notre connaissance pas encore démontré, contrairement à l'estimateur Lasso.

le chemin entier de régularisation de façon efficace, i.e., l'ensemble de la courbe $\lambda \in \mathbb{R}_+ \mapsto \widehat{\boldsymbol{u}}_t^{(\lambda)}$ pour un coût algorithmique total comparable à celui de l'estimateur des moindres carrés. Cette propriété pourrait s'avérer fort utile afin de calibrer automatiquement le paramètre λ pour déduire des résultats d'adaptativité, d'abord dans notre cadre séquentiel déterministe, puis, en corollaire, dans le cadre statistique classique (adaptation en la variance du bruit notamment).

7.2 Regret interne

Plusieurs questions relatives au regret interne et au regret *swap* – que j'ai étudiés au chapitre 5 de ma thèse – sont encore ouvertes. Elles feront probablement l'objet d'une collaboration avec Gilles Stoltz.

Nous reprenons ci-après le cadre de la section 6.2.1. Comme rappelé dans cette section, nous avons étudié les vitesses (ou ordres de grandeur) minimax des regrets interne et *swap* en environnement stochastique ou déterministe. Ces vitesses étaient partiellement connues depuis les travaux de [Sto05, BM07b] et nous avons complété ces résultats – cf. figure 7.1 ci-dessous.

	environnement	regret interne	regret *swap*
bornes sup	$(\boldsymbol{\ell}_t)_{1\leqslant t\leqslant T}$ déterministe	$\sqrt{T\ln K}$	$\sqrt{TK\ln K}$
	$(\boldsymbol{\ell}_t)_{1\leqslant t\leqslant T}$ i.i.d.	$\boldsymbol{\sqrt{T}}$	$\boldsymbol{\sqrt{T\ln K}}$
bornes inf	$(\boldsymbol{\ell}_t)_{1\leqslant t\leqslant T}$ i.i.d.	$\sqrt{T}$	$\sqrt{T\ln K}$
	$(\boldsymbol{\ell}_t)_{1\leqslant t\leqslant T}$ déterministe	$\sqrt{T}$	$\boldsymbol{\sqrt{TK}}$

FIGURE 7.1 – Bornes supérieures et inférieures sur les vitesses minimax des regrets interne et *swap* en environnement stochastique ou déterministe. Les vitesses nouvellement obtenues figurent en caractères gras.

Quelques questions restent cependant en suspens : on remarque ainsi un facteur logarithmique $\sqrt{\ln K}$ manquant entre les bornes inférieure et supérieure connues sur le regret interne en environnement déterministe (respectivement de l'ordre de $\sqrt{T}$ et $\sqrt{T\ln K}$). Ce facteur logarithmique est-il nécessaire ? Nous avons prouvé qu'il est inutile pour des suites i.i.d. (la vitesse minimax correspondante est de l'ordre de $\sqrt{T}$, donc indépendante de la dimension ambiante K), mais nous ne savons pas encore si le facteur $\sqrt{\ln K}$ est nécessaire pour des suites individuelles. Nous mentionnons ci-après quelques pistes pour s'attaquer à la borne inférieure ou à la borne supérieure [2] du *regret interne minimax en environnement déterministe*, qui est plus formellement défini par

$$\inf_{S} \sup_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T\in[0,1]^K} R_T^{\mathrm{int}}\big(S,(\boldsymbol{\ell}_t)_{1\leqslant t\leqslant T}\big)\ , \tag{7.2}$$

où l'infimum est pris sur toutes les stratégies $S=(\boldsymbol{p}_t)_{t\geqslant 1}$ du statisticien, i.e., toutes les suites de fonctions boréliennes $\boldsymbol{p}_t : [0,1]^{K(t-1)} \to \mathcal{X}_K$ (le poids choisi à l'instant t est fonction des pertes passées $\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_{t-1}\in[0,1]^K$), où le supremum est pris sur tous les vecteurs de pertes $\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T\in[0,1]^K$ et où $R_T^{\mathrm{int}}\big(S,(\boldsymbol{\ell}_t)_{1\leqslant t\leqslant T}\big)$ est une quantité de la forme (6.3).

2. Notons que chaque direction pourrait être utile si la vitesse minimax du regret interne pour des suites individuelles était strictement comprise entre $\sqrt{T}$ et $\sqrt{T\ln K}$.

Notons qu'une question similaire est en suspens pour le regret *swap* en environnement déterministe (dont la vitesse minimax est encadrée par $\sqrt{TK}$ et $\sqrt{TK \ln K}$). Les pistes suivantes sont néanmoins sans doute plus adaptées au cas du regret interne.

Amélioration de la borne inférieure du regret interne minimax en environnement déterministe ? Afin d'améliorer la borne inférieure d'ordre $\sqrt{T}$ sur le regret interne minimax prouvée dans [Sto10], Gilles Stoltz avait suggéré – tout comme dans le cas du regret externe – une réduction à des suites de pertes i.i.d. et un recourt judicieux au lemme de Fano, qui est un outil clé en théorie de l'estimation statistique. Notre étude de la vitesse minimax du regret interne dans le cadre stochastique (laquelle est exactement de l'ordre de $\sqrt{T}$) montre malheureusement que cette piste ne nous sera pas utile ici. Si l'on souhaite se réduire à des suites stochastiques, il sera ainsi nécessaire de considérer des suites plus élaborées que des suites i.i.d. (et même que des mélanges de suites i.i.d.). Par exemple, des suites i.i.d. par morceaux (qui ne sont pas stationnaires et donc sans doute plus difficiles à contrôler du point de vue du regret interne) pourraient peut-être permettre d'améliorer la borne inférieure.

Une autre piste m'a récemment été suggérée par Claudio Gentile (Università degli Studi dell'Insubria, Varèse) : peut-on adapter les techniques de borne inférieure sur le regret externe de [FS97, Vov98] au cas du regret interne ? Ces techniques sont fines ([FS97] obtient notamment une formule de récurrence exacte pour le regret externe minimax en classification binaire avec perte absolue), mais elles s'avéreront sans doute plus délicates à employer dans le cas du regret interne, car ce dernier ne vérifie pas une propriété d'additivité très utile pour l'étude du regret externe. Cette voie mérite néanmoins d'être explorée.

Amélioration de la borne supérieure du regret interne minimax en environnement déterministe ? Une autre direction de recherche consiste à améliorer la borne supérieure d'ordre $\sqrt{T \ln K}$ sur le regret interne minimax en environnement déterministe prouvée par [CBL03, SL05]. Pour ce faire, je souhaiterais dans un premier temps tenter d'exploiter la technique stochastique (non constructive) que j'ai développée en deuxième partie du chapitre 5 de ma thèse. Cette technique repose sur la dualité minimax/maximin : on a montré au chapitre 5 via une version du théorème minimax de von Neumann que le regret interne minimax en environnement déterministe (à gauche ci-dessous) est égal à sa contrepartie maximin (à droite), i.e.,

$$\inf_S \sup_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T \in [0,1]^K} R_T^{\mathrm{int}}\big(S, (\boldsymbol{\ell}_t)_{1\leqslant t\leqslant T}\big) = \sup_{\mathbb{Q}\in\mathcal{M}_1^+([0,1]^{KT})} \inf_S \ \mathbb{E}_{\mathbb{Q}}\Big[R_T^{\mathrm{int}}\big(S, (\boldsymbol{\ell}_t)_{1\leqslant t\leqslant T}\big)\Big] \ ,$$

où nous avons utilisé les mêmes notations qu'en (7.2), où le supremum $\sup_{\mathbb{Q}}$ s'étend sur toutes les probabilités $\mathbb{Q}$ sur $[0,1]^{KT}$ (muni de sa tribu borélienne), et où dans l'espérance, les vecteurs de pertes $\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_T \in [0,1]^K$ sont supposés aléatoires de loi jointe $\mathbb{Q}$.

L'égalité précédente permet pour l'instant de retrouver la meilleure borne connue sur le regret interne obtenue par [SL05]. Pour ce faire, nous avons choisie une stratégie S (qui peut dépendre de la loi $\mathbb{Q}$ dans la quantité maximin) minimisant à l'instant t la perte instantanée moyenne du statisticien conditionnellement au passé. Cette stratégie très simple permet de retrouver, via des arguments élémentaires de concentration de martingales (inégalité de Hoeffding-Azuma), la meilleure borne connue $\sqrt{T \ln K}$. En vue de supprimer le facteur logarithmique $\sqrt{\ln K}$ (si cela est possible), il sera vraisemblablement utile d'étudier une stratégie S plus régulière, procédant par exemple par pondération exponentielle, comme dans notre étude du regret interne avec des suites i.i.d.. La non-stationarité des suites de pertes dans notre cadre maximin plus général nécessitera néanmoins vraisemblablement de recourir à des arguments plus fins, et notamment à l'inégalité de Bernstein pour les martingales.

Transformation de la technique stochastique (non constructive) en un algorithme explicite et efficace ? Notons enfin que la technique stochastique évoquée précédemment est utile

d'un point de vue théorique (puisqu'elle permet de majorer le regret minimax), mais qu'elle n'est pas constructive. Nous souhaiterions donc nous pencher sur la construction d'algorithmes $S = (\boldsymbol{p}_t)_{t\geqslant 1}$ explicites (et efficaces) atteignant les bornes supérieures nouvellement prouvées, pour le regret interne si sa borne supérieure $\sqrt{T \ln K}$ est effectivement améliorable, mais aussi pour d'autres formes de regret comme le regret *makespan*[3]. Un algorithme de pondération exponentielle inspiré de celui utilisé dans le cas i.i.d. pourrait peut-être convenir. Bien sûr, une construction générique traitant d'emblée le regret généralisé défini au chapitre 5 (lequel inclut regrets externe, interne, *swap* et *makespan*) serait idéale.

7.3 Agrégation de modèles non linéaires

Le dernier chapitre de ma thèse présente des travaux en cours sur l'agrégation de modèles non linéaires, et plusieurs questions importantes sont encore en suspens. Certaines sont dans la droite lignée de mes travaux de thèse, d'autres en sont beaucoup plus éloignées. Nous reprenons les notations de la section 6.2.2.

Peut-on prouver des inégalités de type oracle *exactes* ? Les inégalités de type oracle que nous avons obtenues sont de type *non exactes*, i.e., avec une constante multiplicative C devant l'infimum strictement supérieure à 1 (cf. (6.4) en section 6.2.2). Est-ce une conséquence de l'approche par concentration — qui donne néanmoins des bornes avec grande probabilité — ou de la généralité des modèles ? En particulier, quand les modèles sont linéaires, il pourrait être intéressant de retrouver via une analyse unifiée les bornes plus fines de [LB06] (de type oracle *exactes*, i.e., avec constante 1) et de [BM07a] obtenues respectivement pour des méthodes d'agrégation de modèles et de sélection de modèles. Au vu de la complexité du cadre très général considéré, nous projetons aussi d'étudier la possibilité d'obtenir des inégalités de type oracle *exactes* dans des cas particuliers de modèles S_m non linéaires classiques, par exemple, les ellipsoïdes de Besov et les réseaux de neurones.

Ces exemples pourraient d'ailleurs permettre de mieux comparer les procédures d'agrégation de modèles et de sélection de modèles, des calculs explicites ayant déjà été effectués par [Mas07] dans le cas simple des ellipsoïdes de Besov pour la procédure de sélection de modèles.

Comment calibrer le paramètre de température inverse η ? Nous souhaiterions également aborder la question – cruciale en pratique – de la calibration du paramètre de température inverse η de notre estimateur. Rappelons que ce dernier est de la forme $\widetilde{s}^{(\eta)} = \sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\eta)} \widehat{s}_m$, où

$$\widehat{\rho}_m^{(\eta)} = \frac{\exp\Big[-\eta\big(\|Y-\widehat{s}_m\|_n^2 + \mathrm{pen}^{(\eta)}(m)\big)\Big]}{\sum_{m'\in\mathcal{M}} \exp\Big[-\eta\big(\|Y-\widehat{s}_{m'}\|_n^2 + \mathrm{pen}^{(\eta)}(m')\big)\Big]}\ , \quad m\in\mathcal{M}\ , \tag{7.3}$$

où $\eta > 0$ est un paramètre de l'algorithme et où la pénalité $\mathrm{pen}^{(\eta)} : \mathcal{M} \to \mathbb{R}+$ est choisie en fonction d'une *dimension généralisée* des modèles S_m et en fonction de la variance σ^2 du bruit, tout comme dans [Mas07].

Au chapitre 6, nous déduisons de notre inégalité de type oracle deux corollaires, l'un pour $\eta = +\infty$ (qui correspond à la procédure de sélection de modèles de [Mas07]), l'autre pour η de l'ordre de n/σ^2. Les deux estimateurs correspondants utilisent la connaissance a priori de la variance σ^2 du bruit, qui est généralement inconnue en pratique. Dans le cas $\eta = +\infty$, l'adaptation à σ^2 a d'abord été traitée par [BM07a, AM09] pour la régression par histogrammes via l'*heuristique de pente*. Quant aux valeurs

3. Le regret *makespan* est utile pour modéliser des problèmes de planification de tâches ou de répartition de charges. Au moyen de la technique stochastique évoquée précédemment, nous avons raffiné la borne supérieure connue d'ordre $\ln(K)\sqrt{T}$ sur sa vitesse minimax en une borne de l'ordre de $\sqrt{T \ln K}$.

plus petites de $\eta \approx n/\sigma^2$, l'adaptation à σ^2 a été abordée par [Gir08] pour des modèles linéaires. Cependant, il n'est pas clair que les choix $\eta = +\infty$ ou $\eta \approx n/\sigma^2$ mentionnés ci-dessus soient optimaux, et l'adaptativité en σ^2 dans le cadre général de modèles non linéaires n'a pas encore été traitée. Deux questions importantes sont donc encore ouvertes. Tout d'abord, peut-on identifier un choix optimal de η (en un sens raisonable) au moins sur des exemples classiques ? Ensuite, si un tel choix optimal (et théorique) est identifié, est-il possible de calibrer notre estimateur de façon totalement automatique et quasi-optimale ?

Une première direction de recherche pourrait consister à appliquer des techniques de calibration séquentielle robuste proches de ceux que nous avons utilisées dans [Ger11]. On pourrait notamment tenter de les mêler à des idées propres à l'heuristique de pente de [BM07a, AM09].

Existence d'une transition de phase ? Une autre direction de recherche, beaucoup plus éloignée de mes travaux de thèse, m'a été suggérée par Pascal Massart (Université Paris-Sud 11). Au vu de la forme de la loi de mélange $(\widehat{\rho}_m^{(\eta)})_{m\in\mathcal{M}}$ (de type Gibbs), on pourrait s'attendre à l'existence d'une *transition de phase* comme en mécanique statistique : existe-t-il une valeur critique η_c séparant deux intervalles de valeurs de η, l'un sur lequel la procédure d'agrégation possède de bonnes performances, et l'autre sur lequel ces performances sont moins bonnes ? C'est du moins ce que suggèrent dans le cadre séquentiel – quand les performances sont mesurées en termes du regret – certains essais de calibration empirique de l'algorithme des poids exponentiels que j'avais effectués pendant mon stage de Master 2 (en prévision séquentielle de la qualité de l'air). En effet, en pratique, le regret de cet algorithme variait brusquement autour d'une certaine valeur de η.

Notons qu'une telle étude pourrait peut-être ensuite permettre d'identifier un choix optimal de η via une détection de la valeur critique η_c en fonction des données seulement. En tout état de cause, ce travail nécessitera d'acquérir de nouvelles connaissances en mécanique statistique.

7.4 Calibration séquentielle et bornes de type oracle en suites individuelles

Le problème suivant – relativement éloigné des travaux présentés dans mon manuscrit de thèse – fait actuellement l'objet d'une collaboration avec Nicolò Cesa-Bianchi (Università degli Studi di Milano), Pierre Gaillard (École normale supérieure) et mon directeur de thèse, Gilles Stoltz.

Paramètres : espace de décision convexe $\mathcal{D}$, espace d'observation $\mathcal{Y}$, fonction de perte $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$, et entier $K \geqslant 1$ (nombre d'experts).

A chaque date $t \in \mathbb{N}^*$,

1. l'environnement choisit les avis d'experts $a_{i,t} \in \mathcal{D}$ pour tout $i \in \{1, \ldots, K\}$; ils sont révélés au statisticien ;
2. le statisticien prend une décision $\widehat{a}_t \in \mathcal{D}$ qu'il révèle à l'environnement ;
3. l'environnement choisit et révèle l'observation $y_t \in \mathcal{Y}$;
4. le statisticien encourt la perte $\ell\big(\widehat{a}_t, y_t\big)$ et chaque expert $i \in \{1, \ldots, K\}$ encourt la perte $\ell\big(a_{i,t}, y_t\big)$.

FIGURE 7.2 – Prévision avec avis d'experts (cas d'un nombre fini d'experts).

Le cadre considéré est celui de la prévision avec avis d'experts (cas d'un nombre fini d'experts) décrit en figure 7.2. Dans ce cadre, et pour des fonctions de perte $\ell : \mathcal{D}\times\mathcal{Y} \to \mathbb{R}$ bornées et convexes en leur premier argument, on connaît depuis plus d'une décennie des procédures d'agrégation optimales

au sens minimax. En particulier, le regret dans le pire des cas des stratégies optimales correspondantes ne peut pas être amélioré (même d'un facteur multiplicatif). En revanche, les travaux plus récents de [FS97, ACBG02, ANN04, CBMS07, HK08] ont montré qu'il existe des algorithmes qui, dans des cas favorables (donc loin du pire des cas considéré pour la quantité minimax), possèdent des performances bien meilleures. Les bornes associées ont été qualifiées de bornes du premier ou second ordre.

Dans ce cadre, un problème encore ouvert – initialement formulé par [CBMS07] – consiste en l'obtention de bornes du second ordre de type oracle, i.e., des bornes de regret du second ordre qui sont un analogue séquentiel des inégalités de type oracle en sélection de modèles. Plus précisément, pour des fonctions de perte $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$ bornées et convexes en leur premier argument, il s'agirait de prouver des bornes de regret de la forme

$$\sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) \leqslant \min_{1 \leqslant i \leqslant K} \left\{ \sum_{t=1}^{T} \ell\big(a_{i,t}, y_t\big) + \gamma_1 \sqrt{Q_{i,T} \ln K} \right\} + \gamma_2 \, E \ln K \ , \tag{7.4}$$

où $\gamma_1, \gamma_2 > 0$ sont des constantes, où $E \triangleq \max_{1 \leqslant t \leqslant T} \max_{1 \leqslant i,j \leqslant K} \big| \ell(a_{i,t}, y_t) - \ell(a_{j,t}, y_t) \big|$ désigne l'étendue des pertes jusqu'à la date T, et où $Q_{i,T}$ est une quantité du second ordre, par exemple, $Q_{i,T} = \sum_{t=1}^{T} \ell^2(a_{i,t}, y_t)$ ou, mieux, un terme de variance empirique

$$Q_{i,T} = \sum_{t=1}^{T} \big(\ell(a_{i,t}, y_t) - \mu_{i,T} \big)^2 \ , \quad \text{avec} \quad \mu_{i,T} \triangleq \frac{1}{T} \sum_{t=1}^{T} \ell(a_{i,t}, y_t) \ .$$

Une borne de la forme (7.4) permettrait de réaliser un compromis de type biais-variance entre les experts : la perte cumulée $\sum_{t=1}^{T} \ell(a_{i,t}, y_t)$ du i-ème expert joue le rôle d'une erreur d'approximation, alors que la quantité $\gamma_1 \sqrt{Q_{i,T} \ln K}$ est une mesure de la difficulté séquentielle d'estimation (et joue donc le rôle d'un terme de variance).

Les exemples de quantités du second ordre $Q_{i,T}$ mentionnés ci-dessus ont été introduits par [CBMS07, HK08], mais ces deux travaux ne prouvent une borne de la forme (7.4) qu'au prix d'une très forte connaissance a priori sur la suite des données à prévoir (en l'occurrence, pour obtenir la borne (7.4), il convient de calibrer leurs algorithmes en fonction de la quantité $Q_{i_T^*,T}$, où i_T^* réalise le minimum dans (7.4) ; dans le cas usuel où $Q_{i_T^*,T}$ est inconnu, leurs bornes sont plus faibles). Notre objectif est donc de prouver une borne du type (7.4) pour un algorithme n'utilisant pas une telle forte connaissance a priori. Pour ce faire, nous nous sommes tournés vers de nouvelles techniques de calibration séquentielle. Nous nous sommes pour l'instant intéressés à une variante de l'algorithme PROD de [CBMS07] calibrée avec des techniques inspirées de [BM07b]. Ce travail est actuellement en cours.

8 Personnes référentes

8.1 Mathématiciens connaissant bien mon travail de recherche

- **Gilles Stoltz** (directeur de thèse)
  E-mail : gilles.stoltz@ens.fr
- **Arnak Dalalyan** et **Claudio Gentile** (rapporteurs de ma thèse)
  E-mails : arnak.dalalyan@ensae.fr et claudio.gentile@uninsubria.it
- **Pascal Massart** (président du jury de ma thèse)
  E-mail : pascal.massart@math.u-psud.fr
- **Nicolò Cesa-Bianchi** (à qui j'ai rendu visite à Milan en mars 2011)
  E-mail : nicolo.cesa-bianchi@unimi.it

8.2 Collègues connaissant bien mon travail d'enseignant

- **Elisabeth Gassiat** (responsable du cours de statistiques en M1 Mathématiques à Orsay).
  E-mail : Elisabeth.Gassiat@math.u-psud.fr
- **Edouard Maurel-Segala** (responsable du cours de mathématiques de la modélisation en L1 Biologie à Orsay).
  E-mail : edouard.maurel-segala@math.u-psud.fr

Bibliographie

[ACBG02] P. Auer, N. Cesa-Bianchi, and C. Gentile. Adaptive and self-confident on-line learning algorithms. *J. Comp. Sys. Sci.*, 64 :48–75, 2002.

[AM09] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10(Feb) :245–279, 2009.

[ANN04] C. Allenberg-Neeman and B. Neeman. Full information game with gains and losses. In *Proceedings of the 15th International Conference on Algorithmic Learning Theory (ALT'04)*, pages 264–278, 2004.

[Aud09] J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. *Ann. Statist.*, 37(4) :1591–1646, 2009.

[AW01] K. S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Mach. Learn.*, 43(3) :211–246, 2001.

[BM01] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc.*, 3 :203–268, 2001.

[BM07a] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Relat. Fields*, 138 :33–73, 2007.

[BM07b] A. Blum and Y. Mansour. From external to internal regret. *J. Mach. Learn. Res.*, 8 :1307–1324, 2007.

[BRT09] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4) :1705–1732, 2009.

[Cat04] O. Catoni. *Statistical learning theory and stochastic optimization.* Springer, New York, 2004.

[CBL03] N. Cesa-Bianchi and G. Lugosi. Potential-based algorithms in on-line prediction and game theory. *Mach. Learn.*, 51(3) :239–261, 2003.

[CBMS07] N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Mach. Learn.*, 66(2/3) :321–352, 2007.

[CT07] E. Candes and T. Tao. The Dantzig selector : statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6) :2313–2351, 2007.

[DJ94] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3) :425–455, 1994.

[DSSST10] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT'10)*, pages 14–26, 2010.

[DT08] A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Mach. Learn.*, 72(1-2) :39–61, 2008.

[DT09] A. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT'09)*, pages 83–92, 2009.

[DT11] A. Dalalyan and A. B. Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, 2011. In press. Available at http://hal.archives-ouvertes.fr/hal-00461580/.

[EHJT04] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32(2) :407–499, 2004.

[FS97] S. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.*, 55(1) :119–139, 1997.

[Ger11] S. Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *JMLR Workshop and Conference Proceedings*, 19 (COLT 2011 Proceedings) :377–396, 2011.

[Gir08] C. Giraud. Mixing least-squares estimators when the variance is unknown. *Bernoulli*, 14(4) :1089–1107, 2008.

[GY11] S. Gerchinovitz and J.Y. Yu. Adaptive and optimal online linear regression on ℓ^1-balls. In J. Kivinen, C. Szepesvári, E. Ukkonen, and T. Zeugmann, editors, *Algorithmic Learning Theory*, volume 6925 of *Lecture Notes in Computer Science*, pages 99–113. Springer Berlin/Heidelberg, 2011.

[HK08] H. Hazan and S. Kale. Extracting certainty from uncertainty : regret bounded by variation in costs. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT'08)*, pages 57–67, 2008.

[HvdG11] M. Hebiri and S. van de Geer. The Smooth-Lasso and other $\ell^1 + \ell^2$-penalized methods. *Electron. J. Stat.*, 5 :1184–1226, 2011.

[LB06] G. Leung and A. R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8) :3396–3410, 2006.

[LLZ09] J. Langford, L. Li, and T. Zhang. Sparse online learning via truncated gradient. *J. Mach. Learn. Res.*, 10 :777–801, 2009.

[Mas07] P. Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007.

[MM11] P. Massart and C. Meynet. The Lasso as an ℓ^1-ball model selection procedure. *Electron. J. Stat.*, 5 :669–687, 2011.

[Nem00] A. Nemirovski. *Topics in Non-Parametric Statistics*. Springer, Berlin/Heidelberg/New York, 2000.

[SL05] G. Stoltz and G. Lugosi. Internal regret in on-line portfolio selection. *Mach. Learn.*, 59 :125–159, 2005.

[SST09] S. Shalev-Shwartz and A. Tewari. Stochastic methods for ℓ^1-regularized loss minimization. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09)*, pages 929–936, 2009.

[Sto05] G. Stoltz. *Incomplete information and internal regret in prediction of individual sequences*. PhD thesis, Paris-Sud XI University, 2005.

[Sto10] G. Stoltz. Agrégation séquentielle de prédicteurs : méthodologie générale et applications à la prévision de la qualité de l'air et à celle de la consommation électrique. *Journal de la Société Française de Statistique*, 151(2) :66–106, 2010.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1) :267–288, 1996.

[Tsy03] A. B. Tsybakov. Optimal rates of aggregation. In *Proceedings of the 16th Annual Conference on Learning Theory (COLT'03)*, pages 303–313, 2003.

[vdG08] S. A. van de Geer. High-dimensional generalized linear models and the Lasso. *Ann. Statist.*, 36(2) :614–645, 2008.

[Vov98] V. Vovk. A game of prediction with expert advice. *J. Comput. System Sci.*, 56(2) :153–173, 1998.

[Vov01] V. Vovk. Competitive on-line statistics. *Internat. Statist. Rev.*, 69 :213–248, 2001.

[Xia10] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, 11 :2543–2596, 2010.