Annexe A : rapports de pré-soutenance

Sébastien Gerchinovitz

Cette annexe contient les rapports de pré-soutenance de mes rapporteurs de thèse :

1. **Arnak Dalalyan**, Professeur au CREST/ENSAE
2. **Claudio Gentile**, Professeur à l'Università degli Studi dell'Insubria (Varèse)

Arnak Dalalyan
ENSAE ParisTech
3, avenue Pierre Larousse
92245 MALAKOFF Cedex
Tél : +33 (0)1 41 17 65 33
Fax : +33 (0)1 41 17 38 52

Rapport sur la thèse « Prédiction de suites individuelles et cadre statistique classique : étude de quelques liens autour de la régression parcimonieuse et des techniques d'agrégation » présentée par Mr Sébastien GERCHINOVITZ

Paris, le 17 novembre, 2011

La thèse expose les travaux du candidat dans le domaine de l'apprentissage statistique. Ces travaux portent sur l'étude théorique des méthodes de prédiction séquentielle et leurs liens avec les approches parcimonieuses dans le cadre de la régression.

Le document est structuré en 6 chapitres et une annexe. Le premier chapitre contient un résumé en français des résultats présentés dans cette thèse. Le deuxième chapitre est une introduction mathématique aux problèmes étudiés. Le candidat en profite pour dresser un état de l'art sur les thèmes abordés dans cette thèse. Dans les trois chapitres suivants, le candidat expose avec beaucoup de clarté ses contributions avec les démonstrations complètes. La thèse contient également une annexe – dédiée aux résultats techniques qui constituent les ingrédients principaux des démonstrations – et une bibliographie. D'une manière générale, le document est très bien structuré et parfaitement bien rédigé.

Le premier chapitre a pour titre : *Vue d'ensemble des résultats*. Il s'agit d'un résumé en français, sans démonstration technique, des résultats exposés dans ce manuscrit.

Le deuxième chapitre, intitulé *Mathematical introduction* est consacré à un rappel des différentes variantes de la problématique de la prévision avec avis d'experts, à l'introduction des principaux algorithmes de prévision (poids exponentiels, gradient exponentiel, régression ridge, lasso) utilisés dans les chapitres suivants. Le candidat présente également des versions adaptatives de ces algorithmes et donne, avec les démonstrations, des résultats théoriques sur les bornes supérieures du regret (ce dernier correspond au risque ou plutôt à l'excès du risque dans la terminologie statistique). Bien que ces résultats soient bien connus, certaines démonstrations présentées diffèrent un peu de celles que l'on peut trouver dans la littérature et permettent d'améliorer légèrement les constantes dans les inégalités sur les bornes de regret.

Le troisième chapitre, *Sparsity regret bounds for individual sequences in online linear regression,* contient à mon avis les résultats les plus intéressants de ce document. Le problème étudié est celui de la régression linéaire séquentielle en grande dimension et sous l'hypothèse de parcimonie. Il s'agit d'un problème statistique où on cherche le meilleur

prédicteur dans l'ensemble des prédicteurs qui s'écrit comme des combinaisons linéaires d'un grand nombre de prédicteurs de base (appelés également des avis d'experts). L'hypothèse de parcimonie signifie alors qu'il existe un prédicteur dans cet ensemble défini comme une combinaison linéaire ne faisant intervenir qu'un petit nombre d'avis d'experts. Dans le cadre stochastique, Dalalyan et Tsybakov ont prouvé que l'algorithme des poids exponentiels satisfait une inégalité oracle exacte dans le contexte de parcimonie, sans aucune condition sur les corrélations entre les avis des experts. Les résultats du chapitre 3 abordent ce problème en utilisant des techniques des suites individuelles et permettent d'améliorer les résultats de Dalalyan et Tsybakov en prouvant des inégalités oracles exactes pour une variante de l'algorithme par poids exponentiels qui, contrairement à celui employé par Dalalyan et Tsybakov, ne dépend ni de l'intensité du bruit ni de la norme L_1 du meilleur prédicteur sparse. Cela constitue une avancée méthodologique extrêmement importante, car l'intensité du bruit et la norme L_1 du meilleur prédicteur ne sont pas disponibles en pratique. L'idée, très astucieuse, qui a permis d'obtenir ces améliorations est de modifier l'algorithme par poids exponentiels en tronquant les combinaisons linéaires des avis d'experts à partir d'un certain seuil. D'un point de vue théorique, le prix à payer pour cette adaptation est la présence de quelques termes supplémentaires dans le résidu de l'inégalité oracle, mais tous ces termes sont d'un ordre de grandeur plus petit que le terme principal.

Le quatrième chapitre, *Adaptive and optimal online linear regression on l_1-balls,* est dédié au problème de régression linéaire séquentielle lorsque le paramètre appartient à une boule L_1. Dans le cadre stochastique, ce problème a été étudié par Tsybakov en 2003 et est connu sous le nom d'agrégation convexe. Pour commencer, un résultat sur la borne supérieure du regret cumulé est établi et il est montré que cette borne supérieure est de bon ordre de grandeur à un facteur logarithmique près. Ces résultats mettent en évidence deux régimes de convergence très différents : sous le premier régime la vitesse est proportionnelle à la racine carrée de *T (vitesse lente)*, alors que sous le second elle est logarithmique en *T (vitesse rapide)*, où *T* désigne le nombre d'observations. (Ce comportement été déjà observé par Tsybakov en 2003.) Cependant, l'algorithme utilisé dans la démonstration souffre d'une complexité de calcul exponentielle et n'est pas adaptatif par rapport à un certain nombre de paramètres (les bornes de l'intervalle contenant les variables explicatives et la variable à expliquer, le rayon de la boule L_1). La suite du chapitre est consacrée à l'étude des algorithmes adaptatifs d'une complexité de calcul raisonnable. Une nouvelle technique appelée Lipschitzification de la perte est introduite et il est montré qu'en combinaison avec l'algorithme de l'exponentiel du gradient elle conduit vers un algorithme adaptatif dont le regret est d'ordre de grandeur optimal sous le premier régime.

Dans le cinquième chapitre, intitulé *Minimax rates of internal and swap regrets,* d'autres types de regret, appelés regret interne et regret swap, sont considérés dans le cadre stochastique ou déterministe. Les résultats les plus notables de ce chapitre concernent les vitesses minimax dans le cadre aléatoire i.i.d. et améliorent les résultats existants dans la littérature en supprimant un facteur logarithmique prouvé inutile. Comme dans les autres chapitres, un état de l'art complet est dressé qui permet de bien situer les contributions du candidat par rapport à ce qui était déjà connu. Il y a également un résultat sur la borne inférieure du risque minimax dans cadre déterministe qui mérite d'être mentionné car il révèle une différence essentielle entre le regret interne et le regret swap : pour le premier, le cas des pertes aléatoires i.i.d. est presqu'aussi difficile que le cas des pertes déterministes quelconques, tandis que pour le second le cas déterministe est considérablement plus compliqué.

Enfin, sous le titre *Aggregation of nonlinear models*, le sixième et dernier chapitre explore le comportement de l'algorithmes des poids exponentiels dans le cadre d'agrégation des modèles quelconques pour le modèle de régression à design aléatoire. L'algorithme considéré est très similaire à celui proposé par Leung et Barron dans le cadre de la régression à design fixe pour agréger des estimateurs des moindres carrés dans des modèles linéaires. En utilisant des techniques de concentration de Birgé et Massart, une inégalité oracle est obtenue aussi bien en espérance qu'avec grande probabilité. Modulo une

constante multiplicative supérieure à un, le résultat obtenu est inaméliorable car le terme résiduel est d'ordre de grandeur optimal.

En conclusion, Monsieur Sébastien Gerchinovitz a travaillé sur des problèmes aussi intéressants que variés et a obtenu des résultats remarquables. L'ensemble du travail témoigne d'une grande richesse de matière et d'idées ce qui n'est pas aisé dans un domaine à l'interface de plusieurs branches de mathématiques. Le manuscrit est extrêmement bien rédigé ce qui reflète la capacité du candidat d'exposer de façon claire et concise des résultats complexes et très pointus.

Cette thèse mérite sans conteste d'être soutenue en vue de l'obtention du titre de Docteur en Sciences.

Pr Arnak DALALYAN

Università degli Studi dell'Insubria
Dipartimento di Informatica e Comunicazione
Prof. Claudio Gentile

Assessment on Doctoral Thesis by Sebastien Gerchinovitz

Title: **Prediction de suites individuelles et cadre statistique classique: etude de quelques liens autour de la regression parcimonieuse et des techniques d'agregation**

Assessor: **Claudio Gentile**

Date: **November 14th, 2011**

1. **Content**
2. **Overall Evaluation**
3. **Some questions**

Italy - 21100 Varese - via Mazzini, 5 - Tel. +39 0332 218934 - Fax +39 0332 218909

1. Content

This thesis deals with the problem of learning from either individual sequences (expert setting, linear regression), or more standard stochastic settings under fixed and random design. In a nutshell, the author revisits known results in the former setting (sometimes yielding significant improvements), then adapts them to the latter, yielding further improved results.
In fact, the content of this dissertation is more complex: the topics touched upon range from adaptive algorithms for sparse linear regression, including minimax rates, to internal and swap regret for either individual sequences or stochastic scenarios. The breakdown into chapters helps illustrating the actual content.

Chapter 1 is an excellent primer that helps introducing the main goals and the results this thesis builds on. The main objectives and results on learning on individual sequences are recalled. The chapter presents the problem of online learning with expert advice (the Weighted Majority algorithm, Vovk's aggregating algorithm, refinements thereof), with a special emphasis on on-the-fly parameter tuning. Then the more general setup of online PAC-Bayes analysis is sketched, which leads to improved regret bounds. The chapter then surveys the online linear regression setting (with side information or *covariates*) under square loss. The sequential ridge regression predictor is presented, along with the main related tuning issues. An adaptive EG-like algorithm that works for general differential convex losses is also described, which is similar in spirit to the "self-confident" algorithms available in the literature. Convexity and averaging leads to bounds on the generalization error in stochastic batch settings with random design. Specific presentation efforts are devoted to regression problems with sparse targets via sparsity oracle inequalities.

Chapter 2 is chiefly taken from a Colt 2011 publication by the author. It essentially deals with sparse online linear regression for individual sequences. A preliminary algorithm, SeqSEW, is presented which relies on prior knowledge of relevant parameters of the problem. This algorithm is strongly influenced by the (heavy-tailed) sparsity enforcing priors of the PAC-Bayes analysis carried out by Dalalyan and Tsybakov. The assumptions on the above prior knowledge are progressively removed to obtain a parameterless algorithm. Adaptation of the above to the fixed and random design settings yields algorithms which are also adaptive to the unknown variance of the noise.

Chapter 3 is taken from an Alt 2011 publication by the author. Minimax rates for linear regression over the L1 ball are established which depend in a detailed way on the interplay between the horizon T, the (L1-)radius of the ball norm, the dimension of the input space and the (L-infinity) length of the covariate vectors. Then an efficient variant of the EG algorithm is presented that is optimal in the "short horizon" regime. It is also claimed that a similar adaptation works for online ridge regression, though no details are provided (see "Some questions" below). As a little digression, it is also presented a

technique (called Lipschizification by the author) that is able to achieve improvements for the square loss as well as handle losses other than square. Again, emphasis is put on the ability of the algorithms to self-tune themselves.

Chapter 4 is taken from an author's contribution to the 42emes journee de Statistique. The main scope of this chapter is comparing different kinds of regrets (external, internal, swap, and generalizations), again under adversarial (individual sequences) and stochastic (i.i.d.) assumptions on the sequence of loss vectors. Specifically, the author derives exact minimax rates on internal and swap regrets in the i.id. setting, showing that the optimal internal regret rate is independent of the input dimension (the number of "experts", in this case). In the adversarial setting a new sharper (in some sense) lower bound on the swap regret is shown. This serves to highlight a major difference between external and swap regret rates. Whereas external regret behaves similarly on i.i.d. and individual sequences, this is not the case for swap regret, where a gap between the two regimes is exponential in the input dimension. Finally, a general technique based on convex duality (similar techniques have independently been developed by other authors) is presented that can be used to prove (non-constructively) regret bounds in the individual sequence setting for more general notions of regrets ((psi,phi) – regrets, which include makespan, and all of the above).

Finally, Chapter 5 faces tasks related to the aggregation of nonlinear models within a standard selecting vs. mixing dilemma. A large portion of this chapter is work in progress, partially presented at the StatMathAppli 2011 workshop. In particular, Birge' and Massart's generalized linear Gaussian framework is adopted, where one has at his disposal a family of nonlinear models in a separable Hilbert space, and the goal is to estimate the target almost as well as the best linear least squares estimator, each one being associated with a model in the family. The aggregation is based on exponential weights, where the temperature parameter sets the trade-off between selection (infinite temperature) and aggregation (finite temperature). The author derives high probability oracle inequalities for aggregation and devise a lower bound analysis which is suggestive that in the special case of linear models aggregation is more robust than selection, and might bring similar benefits for nonlinear models as well.

2. Overall evaluation

As for presentation, this dissertation is scholarly written and very well organized. When going over it, I had the impression of a surprisingly deep and insightful understanding its author has on these subjects. English writing is careful (sometimes a bit redundant), only a minor amount of typos, the list of references is thorough.
As for content, this thesis is supported by a number of publications that makes it meet any standard requirements for a PhD dissertation. Most contributions are very technical and valuable, others are slightly incremental in nature.

In short, based on this dissertation, I'm in favour of Gerchinovitz' oral defence to take place.

3. Some questions

On Chapter 3

- In this chapter, an efficient variant of the EG algorithm is presented which is optimal in the "short horizon" regime. A claim is made that a similar adaptation also works for online ridge regression over L1 in the long horizon regime. Is this adaptation retaining computational efficiency ?

- What kind of consequences are then produced in the stochastic *batch* setting ? The author might want to take a look at Liang and Srebro's ICML 2010 paper.

On Chapter 4

- Being a computer scientist, rather than a mathematician, I tend to be more sensitive to computational efficiency issues. The fact that one is able to exhibit an explicit algorithm for the (psi,phi) – regret is no more appealing than a nonconstructive proof of existence, if the algorithm itself is far from being efficient: any hope to get reasonably efficient versions of it (for specific choices of phi and psi)?

- [Vov98] and [FS97] both have a more refined lower bounding argument on external regret which is not based on randomization of the input sequence. I'm wondering if this more refined argument can be adapted to work on other kinds of regrets, too.

Claudio Gentile
Università dell'Insubria, DICOM

Contact details:

Claudio Gentile
DICOM, Università dell'Insubria, Varese, Italy
Email: claudio.gentile@uninsubria.it
Ph: +39 0332 218934
Fax: +39 0332 218909
Mobile: +39 339 2210124
Skype: claudio.gentile000
Web: http://www.dicom.uninsubria.it/~cgentile/