FACULTÉ DES SCIENCES D'ORSAY

Prédiction de suites individuelles et cadre statistique classique

Étude de quelques liens autour de la régression parcimonieuse et des techniques d'agrégation

Sébastien Gerchinovitz

Thèse soutenue le 12 décembre 2011 à l'École normale supérieure devant le jury :

M. Pierre	ALQUIER	Examinateur
M. Olivier	CATONI	Examinateur
M. Arnak	DALALYAN	Rapporteur
M. Pascal	MASSART	Examinateur
M. Gilles	STOLTZ	Directeur
M. Alexandre	TSYBAKOV	Examinateur

Introduction

Dans cette thèse, on s'est intéressé à deux types de problèmes d'apprentissage, tous deux du domaine de la prévision.

1. Prévision de suites déterministes arbitraires

   Problèmes d'apprentissage séquentiel où l'on ne peut pas faire d'hypothèse stochastique sur la suite $(y_t)_{t \geqslant 1}$ des données à prévoir.

   Cela conduit à des algorithmes de prévision très robustes.

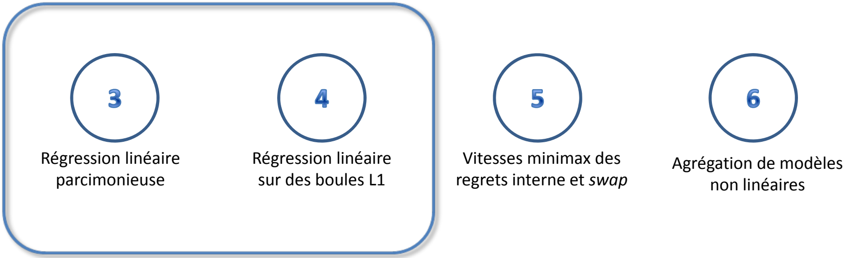
2. Cadre statistique classique

   Les données $(Y_t)_{t \geqslant 1}$ sont modélisées de façon stochastique.

Ces deux cadres entretiennent des liens étroits.

Chapitres traités

Nous avons abordé plusieurs problèmes voisins, dont trois dans le domaine de la régression.



Nous présenterons essentiellement les résultats des chapitres 3 et 4 :

- cadre principal : suites individuelles ;
- liens avec le cadre statistique classique.

1. Régression linéaire séquentielle
   - Brefs rappels statistiques
   - Cadre séquentiel déterministe
   - Exemple d'algorithme séquentiel
2. Régression linéaire séquentielle parcimonieuse
   - Grande dimension : même problème qu'en statistique
   - Algorithme séquentiel et bornes associées
3. Régression linéaire séquentielle sur des boules ℓ^1
   - Cadre et objectif de prévision
   - Vitesse minimax
   - Adaptativité en les paramètres du problème
4. Autres liens avec le cadre statistique
   - Agrégation de modèles non linéaires
   - Vitesse minimax des regrets interne et swap

Brefs rappels statistiques

Considérons le modèle de régression linéaire gaussienne avec *design* fixe : le statisticien observe $(X_1, Y_1), \ldots, (X_T, Y_T) \in \mathbb{R}^d \times \mathbb{R}$ tels que

$$Y_t = \sum_{j=1}^{d} u_j^* X_{t,j} + \varepsilon_t \ , \qquad 1 \leqslant t \leqslant T \ ,$$

où les vecteurs $X_1, \ldots, X_T \in \mathbb{R}^d$ sont déterministes et où $\varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$.

- Le vecteur $\boldsymbol{u}^* \in \mathbb{R}^d$ est inconnu.
- Le statisticien a seulement accès à $(X_1, Y_1), \ldots, (X_T, Y_T)$.

Exemples d'objectifs du statisticien :

- estimation : estimer $\boldsymbol{u}^* \in \mathbb{R}^d$;
- prévision/débruitage : estimer $\left(\sum_{j=1}^{d} u_j^* X_{t,j}\right)_{1 \leqslant t \leqslant T}$.

Erreur quadratique moyenne

Objectif : estimer $\left(\sum_{j=1}^{d} u_j^* X_{t,j}\right)_{1 \leqslant t \leqslant T} \in \mathbb{R}^T$, i.e., construire $\widehat{\boldsymbol{u}} \in \mathbb{R}^d$ de petite erreur quadratique moyenne (EQM)

$$R(\widehat{\boldsymbol{u}}) \triangleq \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{j=1}^{d} u_j^* X_{t,j} - \sum_{j=1}^{d} \widehat{u}_j X_{t,j} \right)^2 .$$

Un estimateur classique est l'estimateur des moindres carrés :

$$\widehat{\boldsymbol{u}} \in \operatorname*{argmin}_{\boldsymbol{u} \in \mathbb{R}^d} \frac{1}{T} \sum_{t=1}^{T} \left(Y_t - \sum_{j=1}^{d} u_j X_{t,j} \right)^2 .$$

L'EQM de cet estimateur vérifie $\mathbb{E}\left[R(\widehat{\boldsymbol{u}})\right] \leqslant d\sigma^2 / T$. Cette erreur est faible en petite dimension $d \ll T$.

Grande dimension et parcimonie

$\mathbb{E}\left[R(\widehat{\boldsymbol{u}})\right] \leqslant d\sigma^2/T$: erreur faible en petite dimension $d \ll T$.

En grande dimension $d > T$:

- si la matrice de *design* $(X_{t,j})_{t,j} \in \mathbb{R}^{T\times d}$ est de rang maximal, $\mathbb{E}\left[R(\widehat{\boldsymbol{u}})\right] = \sigma^2$ (sur-apprentissage) ;
- la tâche de prévision est néanmoins possible sous des hypothèses de parcimonie.

Hypothèse de parcimonie : on suppose $\boldsymbol{u}^*$ parcimonieux, i.e.,

$$\|\boldsymbol{u}^*\|_0 \triangleq \left|\{j : u_j^* \neq 0\}\right| = s \ll T \; .$$

Si on connaissait le support $J^* \triangleq \{j : u_j^* \neq 0\}$ de $\boldsymbol{u}^*$, on pourrait appliquer l'estimateur des moindres carrés relativement à

$$\left\{\boldsymbol{u} \in \mathbb{R}^d, \; \forall j \notin J^*, u_j = 0\right\} \; .$$

Pour cet estimateur idéal, l'EQM serait au plus de l'ordre de $s/T \ll 1$.

Moindres carrés régularisés

En pratique, le support de $\boldsymbol{u}^*$ est inconnu, mais on peut imiter l'estimateur précédent en régularisant les moindres carrés :

$$\widehat{\boldsymbol{u}} \in \operatorname*{argmin}_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \frac{1}{T} \sum_{t=1}^{T} \left(Y_t - \sum_{j=1}^{d} u_j X_{t,j} \right)^2 + \mathrm{reg}(\boldsymbol{u}) \right\} .$$

$\mathrm{reg}(\boldsymbol{u})$	$\mathbb{E}\left[R(\widehat{\boldsymbol{u}})\right]$	Hypothèses sur $(X_{\bullet,j})_j$	Coût algorithmique
$\lVert\boldsymbol{u}\rVert_0$	$\frac{s \ln(d/s)}{T}$	aucune	combinatoire
$\lVert\boldsymbol{u}\rVert_1$	$\frac{s \ln d}{T}$	$X_{\bullet,j}$ presque orthogonaux	minimisation convexe

Algos : régularisation ℓ^0 [BM01, BTW07], régularisation ℓ^1 [CT07, vdG08, BRT09], pondération exponentielle [DT08, AL11, RT11].

Remarque : extensions possibles du cadre

Plusieurs extensions sont souvent considérées dans la littérature.

1. On a postulé une relation linéaire entre X_t et Y_t. De façon équivalente, on pourrait considérer le modèle

$$Y_t = \sum_{j=1}^{d} u_j^* \varphi_j(X_t) + \varepsilon_t \;, \qquad 1 \leqslant t \leqslant T \;,$$

où le dictionnaire $(\varphi_1, \ldots, \varphi_d)$ est constitué de fonctions non linéaires $\varphi_j : \mathbb{R}^d \to \mathbb{R}$ (éléments d'une base de fonctions par ex.).

2. Si ce modèle linéaire n'est pas vérifié, on peut considérer le modèle

$$Y_t = f(X_t) + \varepsilon_t \;, \qquad 1 \leqslant t \leqslant T \;,$$

où la fonction de régression $f : \mathbb{R}^d \to \mathbb{R}$ est inconnue.

On peut chercher à estimer f par la meilleure comb. linéaire des φ_j.
Dans ce cas, $\mathbb{E}\big[R(\widehat{\boldsymbol{u}})\big] \lesssim \text{erreur d'approximation} + \sigma^2 \, s \ln(d)/T$.

Cadre séquentiel déterministe

1. Cadre déterministe

   • On supprime les hypothèses de modélisation stochastique : auparavant, la suite $(Y_t)_{t \geqslant 1}$ était stochastique.

   • Maintenant, la suite $(y_t)_{t \geqslant 1}$ est déterministe arbitraire et on cherche des garanties déterministes. Cela conduit à des algorithmes de prévision très robustes.

2. On ajoute une contrainte séquentielle : les données y_t sont observées séquentiellement.

Agrégation de prévisions : à chaque date t, le statisticien dispose d'un vecteur de prévisions élémentaires $\boldsymbol{x}_t = (x_{t,j})_{1 \leqslant j \leqslant d} \in \mathbb{R}^d$ qu'il peut combiner pour prévoir l'observation $y_t \in \mathbb{R}$.

Quelques références historiques : [Fos91, CBLW96, KW97, AW01, Vov01].

Protocole et objectif de prévision

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

1. L'environnement révèle le vecteur de prévisions élémentaires $\boldsymbol{x_t} \in \mathbb{R}^d$.
2. Le statisticien formule sa prévision $\widehat{y}_t \in \mathbb{R}$ à l'aide des prévisions élémentaires $x_{t,j}$ et des données passées $(\boldsymbol{x_s}, y_s)$, $1 \leqslant s \leqslant t-1$.
3. L'environnement révèle l'observation $y_t \in \mathbb{R}$ et le statisticien encourt la perte carrée $(y_t - \widehat{y}_t)^2$.

Objectif : sur le long terme, prévoir presque aussi bien que le meilleur prédicteur linéaire $\boldsymbol{x} \mapsto \boldsymbol{u} \cdot \boldsymbol{x} \triangleq \sum_{j=1}^d u_j x_j$, $\boldsymbol{u} \in \mathbb{R}^d$, i.e., vérifier

$$\sum_{t=1}^T (y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \underbrace{\sum_{t=1}^T (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2}_{\text{erreur d'approx.}} + \underbrace{\Delta_{T,d}(\boldsymbol{u})}_{\text{erreur d'estim. séq.}} \right\},$$

où le terme de regret $\Delta_{T,d}(\boldsymbol{u})$ est petit (sous-linéaire en T).

Protocole et objectif de prévision

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

1. L'environnement révèle le vecteur de prévisions élémentaires $\boldsymbol{x}_t \in \mathbb{R}^d$.
2. Le statisticien formule sa prévision $\widehat{y}_t \in \mathbb{R}$ à l'aide des prévisions élémentaires $x_{t,j}$ et des données passées $(\boldsymbol{x}_s, y_s)$, $1 \leqslant s \leqslant t-1$.
3. L'environnement révèle l'observation $y_t \in \mathbb{R}$ et le statisticien encourt la perte carrée $(y_t - \widehat{y}_t)^2$.

Objectif : sur le long terme, prévoir presque aussi bien que le meilleur prédicteur linéaire $\boldsymbol{x} \mapsto \boldsymbol{u} \cdot \boldsymbol{x} \triangleq \sum_{j=1}^{d} u_j x_j$, $\boldsymbol{u} \in \mathbb{R}^d$, i.e., vérifier

$$\frac{1}{T}\sum_{t=1}^{T}\big(y_t - \widehat{y}_t\big)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{ \underbrace{\frac{1}{T}\sum_{t=1}^{T}\big(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\big)^2}_{\text{erreur d'approx.}} + \underbrace{\frac{\Delta_{T,d}(\boldsymbol{u})}{T}}_{\text{erreur d'estim. séq.}} \right\},$$

où le terme de regret $\Delta_{T,d}(\boldsymbol{u})$ est petit (sous-linéaire en T).

Exemple : l'algorithme *ridge* séquentiel

L'algorithme *ridge*, initialement étudié par [HK70] en statistique, a été étendu au cadre déterministe séquentiel par [AW01] et [Vov01].

Pour un paramètre $\lambda > 0$, l'algorithme *ridge* séquentiel produit à l'instant t la prévision $\widehat{y}_t = \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t$, où

$$\widehat{\boldsymbol{u}}_t \in \operatorname*{argmin}_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{s=1}^{t-1} \left(y_s - \boldsymbol{u} \cdot \boldsymbol{x}_s\right)^2 + \lambda \left\|\boldsymbol{u}\right\|_2^2 + \left(\boldsymbol{u} \cdot \boldsymbol{x}_t\right)^2 \right\} .$$

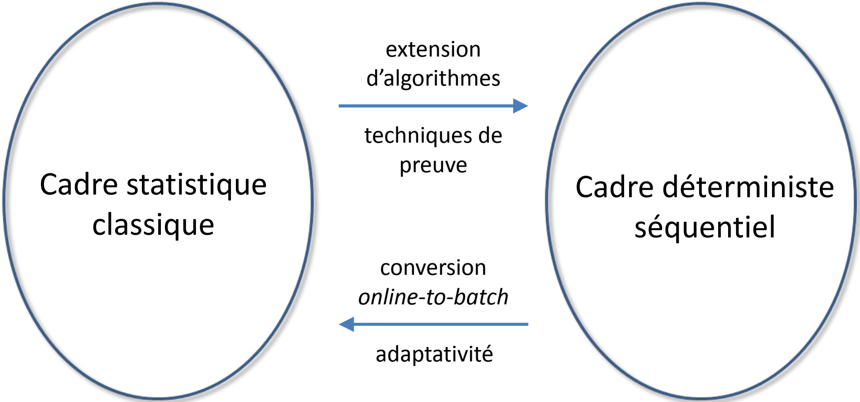
Cet algorithme vérifie, pour toute suite $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_T, y_T) \in \mathbb{R}^d \times \mathbb{R}$,

$$\sum_{t=1}^{T} \left(y_t - \widehat{y}_t\right)^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} \left(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t\right)^2 + \lambda \left\|\boldsymbol{u}\right\|_2^2 + d\, C_y \ln T \right\} + \ldots$$

La vitesse $d \ln T$ correspond à la vitesse paramétrique d/T dans le cadre statistique.

Liens entre les cadres statistique et déterministe

Protocoles de prévision et hypothèses associées radicalement différents, mais liens étroits entre les cadres statistique et déterministe.



On illustre ces liens pour la régression parcimonieuse et la régression sur des boules ℓ^1.

Grande dimension : même problème qu'en statistique

Rappel : l'algorithme *ridge* séquentiel encourt un regret au plus de l'ordre de $d \ln T$. Ce regret est sous-linéaire en T quand $d \ll T$.

En grande dimension $d > T$, on peut toujours espérer atteindre un regret sous-linéaire s'il existe $\boldsymbol{u}^* \in \mathbb{R}^d$ parcimonieux et de petite perte cumulée.

En effet, en utilisant l'algorithme *ridge* séquentiel non pas sur $\mathbb{R}^d$, mais sur l'e.v. engendré par le support J^* inconnu de $\boldsymbol{u}^*$, i.e.,

$$\left\{ \boldsymbol{u} \in \mathbb{R}^d, \ \forall j \notin J^*, u_j = 0 \right\} ,$$

on obtiendrait un regret au plus de l'ordre de $\|\boldsymbol{u}^*\|_0 \ln T$. Ce regret est sous-linéaire sous l'hypothèse de parcimonie $\|\boldsymbol{u}^*\|_0 \ll T/(\ln T)$.

Bornes de parcimonie

Dans la suite, on montre qu'il est possible d'atteindre des bornes proportionnelles à $\|\boldsymbol{u}^*\|_0$ (à des facteurs log près), i.e., on prouve des bornes de la forme

$$\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\left(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\right)^2 + \left(\|\boldsymbol{u}\|_0 + 1\right) g_{T,d}\left(\|\boldsymbol{u}\|_1\right)\right\} ,$$

où g croît au plus logarithmiquement en T, d et $\|\boldsymbol{u}\|_1 \triangleq \sum_{j=1}^{d}|u_j|$.
On appelle de telles bornes des bornes de parcimonie.

Par intégration, ces bornes déterministes impliquent des inégalités oracle de parcimonie dans le cadre statistique classique, approximativement de la forme

$$\mathbb{E}\left[R(\widehat{\boldsymbol{u}})\right] \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{R(\boldsymbol{u}) + C\,\frac{\|\boldsymbol{u}\|_0 \ln d}{T}\right\} .$$

Algorithme SeqSEW (*Sequential Sparse Exponential Weighting*)

Paramètres: seuil B, température inverse η et paramètre d'échelle τ.

A chaque date $t \geqslant 1$, l'algorithme $\mathrm{SeqSEW}_{\tau}^{B,\eta}$ produit la prévision

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

où $[z]_B = \max\big\{-B, \min\{B, z\}\big\}$ désigne une opération de seuillage, et où la probabilité p_t sur $\mathbb{R}^d$ est définie par

$$p_t(\mathrm{d}\boldsymbol{u}) \triangleq \frac{1}{W_t} \exp\left(-\eta \sum_{s=1}^{t-1} \Big(y_s - \big[\boldsymbol{u} \cdot \boldsymbol{x}_s\big]_B\Big)^2\right) \pi_\tau(\mathrm{d}\boldsymbol{u})$$

pour une constante de renormalisation W_t.

La probabilité a priori π_τ sur $\mathbb{R}^d$, introduite par [DT07] dans le cadre stochastique, favorise la parcimonie :

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

L'agrégation par pondération exponentielle a été développée parallèlement :

- en *machine learning* depuis [LW94, Vov90] ;
- en statistique depuis [Cat99, Cat04].

Le choix d'une probabilité a priori encourageant la parcimonie est plus récent (cf. [JS05, See08] par ex.). Notre probabilité a priori π_τ est celle de l'algorithme SEW de [DT08, DT11] dans le cadre stochastique.

Dans ces travaux, on montre que :

- l'algorithme de [DT11] fonctionne essentiellement pour des raisons déterministes ;
- en le calibrant (et en le seuillant) séquentiellement, on obtient des résultats adaptatifs dans le cadre stochastique.

Borne de parcimonie

On suppose pour simplifier que le statisticien a accès à l'avance à deux bornes B_y et B_x :

$$y_1, \ldots, y_T \in [-B_y, B_y] \qquad \text{et} \qquad \|\boldsymbol{x}_1\|_\infty, \ldots, \|\boldsymbol{x}_T\|_\infty \leqslant B_x \; .$$

Théorème (G.)

Sous les hypothèses précédentes, l'algorithme $\mathrm{SeqSEW}_\tau^{\mathrm{B},\eta}$ *calibré avec* $B = B_y$, $\eta = 1/(8B_y^2)$ *et* $\tau = 4B_y/(\sqrt{dT}B_x)$ *vérifie*

$$\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\left(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\right)^2 + 32\,\|\boldsymbol{u}\|_0\,B_y^2 \ln\left(1 + \frac{\sqrt{dT}B_x\,\|\boldsymbol{u}\|_1}{4B_y\,\|\boldsymbol{u}\|_0}\right)\right\} + 16B_y^2$$

Il s'agit d'une borne de parcimonie comme définie précédemment.

Preuve : recourt à une borne PAC-bayésienne séquentielle [Aud09] et exploitation de la forme à queue lourde de la loi a priori π_τ [DT07].

Borne de parcimonie

On suppose pour simplifier que le statisticien a accès à l'avance à deux bornes B_y et B_x :

$$y_1, \ldots, y_T \in [-B_y, B_y] \qquad \text{et} \qquad \|\boldsymbol{x}_1\|_\infty, \ldots, \|\boldsymbol{x}_T\|_\infty \leqslant B_x \ .$$

Théorème (G.)

Sous les hypothèses précédentes, l'algorithme $\mathrm{SeqSEW}_\tau^{B,\eta}$ *calibré avec* $B = B_y$, $\eta = 1/(8B_y^2)$ *et* $\tau = 4B_y/(\sqrt{dT}B_x)$ *vérifie*

$$\sum_{t=1}^T (y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d} \left\{ \sum_{t=1}^T \big(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\big)^2 + 32\,\|\boldsymbol{u}\|_0\, B_y^2 \ln\left(1 + \frac{\sqrt{dT}B_x\,\|\boldsymbol{u}\|_1}{4B_y\,\|\boldsymbol{u}\|_0}\right)\right\} + 16B_y^2$$

Calibration automatique : on peut prouver une borne similaire à l'aide de paramètres B_t, η_t et τ_t calibrés uniquement en fonction des données, i.e.,

$$\max_{1\leqslant s\leqslant t-1} |y_s| \qquad \text{et} \qquad \max_{1\leqslant s\leqslant t-1} \|\boldsymbol{x}_s\|_\infty \ .$$

Borne de parcimonie

On suppose pour simplifier que le statisticien a accès à l'avance à deux bornes B_y et B_x :

$$y_1, \ldots, y_T \in [-B_y, B_y] \qquad \text{et} \qquad \|\boldsymbol{x}_1\|_\infty, \ldots, \|\boldsymbol{x}_T\|_\infty \leqslant B_x \;.$$

Théorème (G.)

Sous les hypothèses précédentes, l'algorithme $\mathrm{SeqSEW}_\tau^{\mathrm{B},\eta}$ *calibré avec* $B = B_y$, $\eta = 1/(8B_y^2)$ *et* $\tau = 4B_y/(\sqrt{dT}B_x)$ *vérifie*

$$\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\left(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\right)^2 + 32\,\|\boldsymbol{u}\|_0\,B_y^2 \ln\left(1 + \frac{\sqrt{dT}B_x\,\|\boldsymbol{u}\|_1}{4B_y\,\|\boldsymbol{u}\|_0}\right)\right\} + 16B_y^2$$

Raffinement : on peut remplacer le terme $\sqrt{dT}B_x$ par $\sqrt{\sum_{j=1}^{d}\sum_{t=1}^{T} x_{t,j}^2}$ (quantité connue ou inconnue), qui fait apparaître la trace de la matrice de Gram empirique, plus standard en statistique.

Application au cadre statistique classique

Cadre : modèle de régression avec *design* aléatoire. On observe un échantillon i.i.d. $(X_1, Y_1), \ldots, (X_T, Y_T) \in \mathbb{R}^d \times \mathbb{R}$ donné par

$$Y_t = f(X_t) + \varepsilon_t \ , \quad 1 \leqslant t \leqslant T \ ,$$

où $(X_t, \varepsilon_t)_t$ est i.i.d. et $\mathbb{E}[\varepsilon_t | X_t] = 0$. L'objectif est d'estimer la fonction de régression $f : \mathbb{R}^d \to \mathbb{R}$ inconnue.

Méthode : l'échantillon $(X_1, Y_1), \ldots, (X_T, Y_T)$ est traité séquentiellement via l'algo. $\mathrm{SeqSEW}_{\tau}^{B_t, \eta_t}$ avec $\tau = 1/\sqrt{dT}$, qui vérifie la borne déterministe :

$$\sum_{t=1}^{T} \big(Y_t - \underbrace{\widetilde{f}_t(X_t)}_{=\widehat{y}_t}\big)^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} (Y_t - \boldsymbol{u} \cdot X_t)^2 + 64 \max_{1 \leqslant t \leqslant T} Y_t^2 \ \|\boldsymbol{u}\|_0 \ \ln(\ldots) \right\} + \ldots$$

où $\widetilde{f}_t : \mathbb{R}^d \to \mathbb{R}$ est construit à partir de $(X_s, Y_s)_{s \leqslant t-1}$ selon

$$\widetilde{f}_t(\boldsymbol{x}) \triangleq \int_{\mathbb{R}^d} \big[\boldsymbol{u} \cdot \boldsymbol{x}\big]_{B_t} \, p_t(\mathrm{d}\boldsymbol{u}) \ .$$

Conversion *online-to-batch*

On emploie la conversion *online-to-batch* [Lit89, CBCG04].

Rappel : on a la borne déterministe

$$\sum_{t=1}^{T}\big(Y_t - \widetilde{f}_t(X_t)\big)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}(Y_t - \boldsymbol{u}\cdot X_t)^2 + 64 \max_{1\leqslant t\leqslant T} Y_t^2 \, \|\boldsymbol{u}\|_0 \, \ln(\cdot)\right\} + \dots$$

En prenant l'espérance de la borne précédente et en appliquant l'inégalité de Jensen deux fois, on obtient, en posant $\widehat{f}_T \triangleq \frac{1}{T}\sum_{t=1}^{T}\widetilde{f}_t$,

$$\mathbb{E}\Big[\big(Y - \widehat{f}_T(X)\big)^2\Big] \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\mathbb{E}\Big[(Y - \boldsymbol{u}\cdot X)^2\Big] + 64\,\mathbb{E}\Big[\max_{1\leqslant t\leqslant T} Y_t^2\Big]\frac{\|\boldsymbol{u}\|_0}{T}\ln(\cdot)\right\} + \dots$$

où (X, Y) est une copie de (X_1, Y_1) indépendante de $(X_t, Y_t)_{t=1}^{T}$.

Conversion *online-to-batch*

On emploie la conversion *online-to-batch* [Lit89, CBCG04].

Rappel : on a la borne déterministe

$$\sum_{t=1}^{T}\left(Y_t - \widetilde{f}_t(X_t)\right)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\left(Y_t - \boldsymbol{u}\cdot X_t\right)^2 + 64 \max_{1\leqslant t\leqslant T} Y_t^2 \,\|\boldsymbol{u}\|_0 \,\ln(\cdot)\right\} + \dots$$

En prenant l'espérance de la borne précédente et en appliquant l'inégalité de Jensen deux fois, on obtient, en posant $\widehat{f}_T \triangleq \frac{1}{T}\sum_{t=1}^{T}\widetilde{f}_t$,

$$\mathbb{E}\left[\left(Y - \widehat{f}_T(X)\right)^2\right] \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\mathbb{E}\left[\left(Y - \boldsymbol{u}\cdot X\right)^2\right] + 64\,\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Y_t^2\right]\frac{\|\boldsymbol{u}\|_0}{T}\ln(\cdot)\right\} + \dots$$

où (X, Y) est une copie de (X_1, Y_1) indépendante de $(X_t, Y_t)_{t=1}^{T}$.

Conversion *online-to-batch*

On emploie la conversion *online-to-batch* [Lit89, CBCG04].

Rappel : on a la borne déterministe

$$\sum_{t=1}^{T}\left(Y_t - \widetilde{f}_t(X_t)\right)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\left(Y_t - \boldsymbol{u}\cdot X_t\right)^2 + 64 \max_{1\leqslant t\leqslant T} Y_t^2 \, \|\boldsymbol{u}\|_0 \, \ln(\cdot)\right\} + \dots$$

En prenant l'espérance de la borne précédente et en appliquant l'inégalité de Jensen deux fois, on obtient, en posant $\widehat{f}_T \triangleq \frac{1}{T}\sum_{t=1}^{T}\widetilde{f}_t$,

$$\mathbb{E}\left[\left(f(X) - \underset{\cancel{+\sigma^2}}{\widehat{f}_T(X)}\right)^2\right] \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\mathbb{E}\left[\left(f(X) - \underset{\cancel{+\sigma^2}}{\boldsymbol{u}}\cdot X\right)^2\right] + 64\,\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Y_t^2\right]\frac{\|\boldsymbol{u}\|_0}{T}\ln(\cdot)\right\} + \dots$$

où (X, Y) est une copie de (X_1, Y_1) indépendante de $(X_t, Y_t)_{t=1}^{T}$.

Adaptation en la variance inconnue du bruit

Théorème (Une inégalité oracle de parcimonie, G.)

Soit X une v.a. de même loi que X_1 et indépendante de $(X_1, Y_1, \ldots, X_T, Y_T)$. Alors,

$$\mathbb{E}\Big[\big(f(X) - \widehat{f}_T(X)\big)^2\Big]$$
$$\leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d} \left\{ \mathbb{E}\Big[\big(f(X) - \boldsymbol{u}\cdot X\big)^2\Big] + 64\,\mathbb{E}\Big[\max_{1\leqslant t\leqslant T} Y_t^2\Big] \frac{\|\boldsymbol{u}\|_0}{T} \ln\left(1 + \frac{\sqrt{dT}\,\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right)\right\} + \ldots$$

On peut majorer $\mathbb{E}\big[\max_{1\leqslant t\leqslant T} Y_t^2\big]$ sous diverses hypothèses : par ex., si $\|f\|_\infty < +\infty$ et $\mathbb{E}\big[\exp(\lambda\varepsilon_1) \,\big|\, X_1\big] \leqslant e^{\lambda^2\sigma^2/2}$ pour tout $\lambda \in \mathbb{R}$, alors

$$\mathbb{E}\Big[\max_{1\leqslant t\leqslant T} Y_t^2\Big] \leqslant 2\left(\|f\|_\infty^2 + 2\,\sigma^2 \ln(2e\,T)\right) .$$

On en déduit une borne de risque similaire à [DT11, Prop.1], mais de façon adaptative : l'estimateur $\widehat{f}_T$ n'utilise pas la connaissance de σ^2.

Adaptation en la variance inconnue du bruit

Théorème (Une inégalité oracle de parcimonie, G.)

Soit X une v.a. de même loi que X_1 et indépendante de $(X_1, Y_1, \ldots, X_T, Y_T)$. Alors,

$$\mathbb{E}\Big[\big(f(X) - \widehat{f}_T(X)\big)^2\Big]$$
$$\leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d} \left\{ \mathbb{E}\Big[\big(f(X) - \boldsymbol{u}\cdot X\big)^2\Big] + 64\,\mathbb{E}\Big[\max_{1\leqslant t\leqslant T} Y_t^2\Big] \frac{\|\boldsymbol{u}\|_0}{T} \ln\left(1 + \frac{\sqrt{dT}\,\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right)\right\} + \ldots$$

On peut majorer $\mathbb{E}\big[\max_{1\leqslant t\leqslant T} Y_t^2\big]$ sous diverses hypothèses : par ex., si $\|f\|_\infty < +\infty$ et $\mathbb{E}\big[\exp(\lambda\varepsilon_1)\,\big|\,X_1\big] \leqslant e^{\lambda^2\sigma^2/2}$ pour tout $\lambda \in \mathbb{R}$, alors

$$\mathbb{E}\Big[\max_{1\leqslant t\leqslant T} Y_t^2\Big] \leqslant 2\left(\|f\|_\infty^2 + 2\,\sigma^2 \ln(2eT)\right) .$$

On en déduit une borne de risque similaire à [DT11, Prop.1], mais de façon adaptative : l'estimateur $\widehat{f}_T$ n'utilise pas la connaissance de σ^2.

Rappel du cadre séquentiel déterministe

Tâche de prévision : à chaque date t, prévoir l'observation $y_t \in \mathbb{R}$ à partir du vecteur de prévisions élémentaires $\boldsymbol{x}_t \in \mathbb{R}^d$.

Étape initiale : l'environnement choisit deux suites déterministes arbitraires $(y_t)_{t \geqslant 1}$ dans $\mathbb{R}$ et $(\boldsymbol{x}_t)_{t \geqslant 1}$ dans $\mathbb{R}^d$ mais le statisticien n'y a pas accès.

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

1. L'environnement révèle le vecteur de prévisions élémentaires $\boldsymbol{x}_t \in \mathbb{R}^d$.
2. Le statisticien formule sa prévision $\widehat{y}_t \in \mathbb{R}$ à l'aide des prévisions élémentaires $x_{t,j}$ et des données passées $(\boldsymbol{x}_s, y_s)$, $1 \leqslant s \leqslant t-1$.
3. L'environnement révèle l'observation $y_t \in \mathbb{R}$ et le statisticien encourt la perte carrée $(y_t - \widehat{y}_t)^2$.

Objectif : comparaison à des boules ℓ^1

Objectif : 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., minimiser le regret

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

pour un rayon $U > 0$ et un horizon de temps $T \geqslant 1$ (fixés ou pas).

Remarques :

- Étend la tâche d'agrégation convexe (cf. [Nem00, Tsy03] en statistique).
- Une borne sur le regret précédent par $f_{T,d}(U)$ pour tout $U > 0$ (avec $f_{T,d} \nearrow$) implique une borne régularisée de la forme

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

De telles bornes ont été obtenues en statistique pour le Lasso [MM11].

Regret minimax

On s'intéresse à la vitesse optimale du regret minimax défini par

$$\inf_{(\widehat{y}_t)_{t\geqslant 1}} \sup_{\substack{\|\boldsymbol{x}_1\|_\infty,\ldots,\|\boldsymbol{x}_T\|_\infty \leqslant X \\ |y_1|,\ldots,|y_T|\leqslant Y}} \left\{ \sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 - \min_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 \right\} ,$$

où le sup est pris sur toutes les suites individuelles bornées par $X > 0$ et $Y > 0$ et où l'inf est pris sur tous les prédicteurs séquentiels $(\widehat{y}_t)_{t\geqslant 1}$.

On observe un phénomène de transition identique à celui observé en statistique par [BM01] en *design* fixe et par [Tsy03] en *design* aléatoire :

- sur un premier régime, le regret minimax croît en $\sqrt{T}$;
- sur un second régime, le regret minimax croît en $\ln T$.

Théorème (Majoration du regret minimax, G.)

Soit $d, T \geqslant 1$ et $U, X, Y > 0$. Le regret minimax vérifie, en fonction de U,

$$\inf_{(\widehat{y}_t)_{t \geqslant 1}} \sup_{\|x_t\|_\infty \leqslant X,\ |y_t| \leqslant Y} \left\{ \sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 - \min_{\|u\|_1 \leqslant U} \sum_{t=1}^{T} (y_t - u \cdot x_t)^2 \right\}$$

$$\lesssim \begin{cases} UXY\sqrt{T \ln(2d)} & \text{si} \quad U < U_1 \ , \\ UXY\sqrt{T \ln\left(1 + \frac{2dY}{\sqrt{T}UX}\right)} & \text{si} \quad U \in [U_1, U_2] \ , \\ dY^2 \ln\left(1 + \frac{\sqrt{T}UX}{dY}\right) & \text{si} \quad U > U_2 \ , \end{cases}$$

où U_1 et U_2 dépendent de d, T, X et Y.

Commentaires :

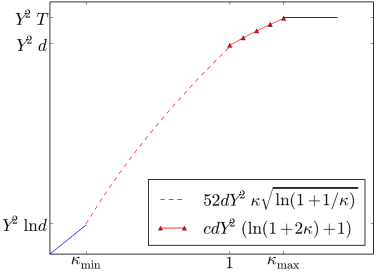
- 1ère borne : prouvée par [KW97] pour l'algorithme $\mathrm{EG}^{\pm}$;
- 2ème borne : technique inspirée de la littérature statistique, i.e., argument à la Mauray [Nem00, Tsy03, BN08] ; cette borne améliore légèrement celle de l'algorithme $\mathrm{EG}^{\pm}$;
- 3ème borne : conséquence directe des bornes de parcimonie. Cette borne améliore légèrement celle induite par l'algorithme *ridge* séquentiel.

Deux régimes distincts

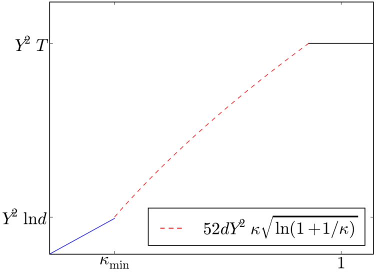
On peut réécrire la borne précédente en fonction de d, Y et d'une quantité intrinsèque $\kappa \triangleq \sqrt{T}UX/(2dY)$ qui relie d à $\sqrt{T}UX/(2Y)$:

$$\begin{cases} dY^2\kappa\sqrt{\ln(2d)} & \text{si} \quad \kappa < \kappa_{\min}(d) \ , \\ dY^2\kappa\sqrt{\ln(1+1/\kappa)} & \text{si} \quad \kappa \in \left[\kappa_{\min}(d), 1\right] \ , \\ dY^2 \ln(1+2\kappa) & \text{si} \quad \kappa > 1 \ . \end{cases}$$

En petite dimension $d \lesssim T$, on observe une transition en $\kappa = 1$:



(a) Transition si $d \lesssim T$.



(b) Pas de transition si $d \gtrsim T$.

Borne inférieure

La borne supérieure précédente

$$\begin{cases} dY^2\kappa\sqrt{\ln(2d)} & \text{si} \quad \kappa < \kappa_{\min}(d)\ , \\ dY^2\kappa\sqrt{\ln(1+1/\kappa)} & \text{si} \quad \kappa \in \left[\kappa_{\min}(d), 1\right]\ , \\ dY^2\ln(1+2\kappa) & \text{si} \quad \kappa > 1\ . \end{cases}$$

est optimale à un facteur logarithmique près pour tous $d \geqslant 1$, $Y > 0$ et $\kappa \geqslant \kappa_{\min}(d) \approx \sqrt{\ln d}/d$.

Pour le montrer, on a exhibé une borne inférieure de la façon suivante :

- réduction au cadre stochastique (via la conversion *online-to-batch*) ;
- emploi de la borne inférieure de [Tsy03] pour l'agrégation convexe.

Borne inférieure

La borne supérieure précédente

$$\begin{cases} dY^2\kappa\sqrt{\ln(2d)} & \text{si} \quad \kappa < \kappa_{\min}(d) \ , \\ dY^2\kappa\sqrt{\ln(1+1/\kappa)} & \text{si} \quad \kappa \in \big[\kappa_{\min}(d), 1\big] \ , \\ dY^2\ln(1+2\kappa) & \text{si} \quad \kappa > 1 \ . \end{cases}$$

est optimale à un facteur logarithmique près pour tous $d \geqslant 1$, $Y > 0$ et $\kappa \geqslant \kappa_{\min}(d) \approx \sqrt{\ln d}/d$.

Pour le montrer, on a exhibé une borne inférieure de la façon suivante :

- réduction au cadre stochastique (via la conversion *online-to-batch*) ;
- emploi de la borne inférieure de [Tsy03] pour l'agrégation convexe.

Conclusion : l'agrégation sur des boules ℓ^1 est approximativement de même complexité dans les cadres stochastique et déterministe.

Adaptativité en les paramètres du problème

Problème : certains algorithmes étudiés précédemment sont inefficaces en grande dimension d et requièrent la connaissance a priori de X, Y, T et U.

Un exemple de cheminement

1. Afin de résoudre ce problème, nous avons introduit une variante automatique de l'exponentielle des gradients (algorithme $\mathrm{EG}^{\pm}$) de [KW97], combinée à une pré-transformation lipschitzienne des pertes.

Adaptativité en les paramètres du problème

Problème : certains algorithmes étudiés précédemment sont inefficaces en grande dimension d et requièrent la connaissance a priori de X, Y, T et U.

Un exemple de cheminement

1. Afin de résoudre ce problème, nous avons introduit une variante automatique de l'exponentielle des gradients (algorithme $\mathrm{EG}^{\pm}$) de [KW97], combinée à une pré-transformation lipschitzienne des pertes.

Cet algorithme est quasi-optimal dans le régime $\kappa \leqslant 1$ et s'adapte en X, Y et T grâce à une calibration séquentielle.

On peut s'adapter à U en agrégeant des sous-algorithmes associés à une grille $\mathcal{U}_t = \left\{ U_r = a_t \, 2^r : r = 0, \ldots, R_t \right\}$ évoluant avec t.

Adaptativité en les paramètres du problème

Problème : certains algorithmes étudiés précédemment sont inefficaces en grande dimension d et requièrent la connaissance a priori de X, Y, T et U.

Un exemple de cheminement

1. Afin de résoudre ce problème, nous avons introduit une variante automatique de l'exponentielle des gradients (algorithme $\mathrm{EG}^{\pm}$) de [KW97], combinée à une pré-transformation lipschitzienne des pertes.

2. Comme régulièrement en recherche, nous nous sommes aperçus que le problème avait déjà été résolu ! (par [ACBG02] via un autre algo.)

Adaptativité en les paramètres du problème

Problème : certains algorithmes étudiés précédemment sont inefficaces en grande dimension d et requièrent la connaissance a priori de X, Y, T et U.

Un exemple de cheminement

1. Afin de résoudre ce problème, nous avons introduit une variante automatique de l'exponentielle des gradients (algorithme $\mathrm{EG}^{\pm}$) de [KW97], combinée à une pré-transformation lipschitzienne des pertes.
2. Comme régulièrement en recherche, nous nous sommes aperçus que le problème avait déjà été résolu ! (par [ACBG02] via un autre algo.)
3. Ainsi, notre algorithme fournit une deuxième solution pour s'adapter efficacement aux paramètres du problème.

Adaptativité en les paramètres du problème

Problème : certains algorithmes étudiés précédemment sont inefficaces en grande dimension d et requièrent la connaissance a priori de X, Y, T et U.

Un exemple de cheminement

1. Afin de résoudre ce problème, nous avons introduit une variante automatique de l'exponentielle des gradients (algorithme $\mathrm{EG}^{\pm}$) de [KW97], combinée à une pré-transformation lipschitzienne des pertes.
2. Comme régulièrement en recherche, nous nous sommes aperçus que le problème avait déjà été résolu ! (par [ACBG02] via un autre algo.)
3. Ainsi, notre algorithme fournit une deuxième solution pour s'adapter efficacement aux paramètres du problème.
4. Nous serions heureux que notre technique de pré-transformation lipschitzienne des pertes s'avère utile dans d'autres contextes !

Agrégation de modèles non linéaires

Cadre : modèle de régression gaussienne avec *design* fixe. Le statisticien observe le vecteur $(Y_1, \ldots, Y_n) \in \mathbb{R}^n$ donné par

$$Y_i = s_i + \sigma \xi_i \in \mathbb{R} \,, \quad 1 \leqslant i \leqslant n \,, \quad \xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

Objectif : étant donnée une collection au plus dénombrable $\left(\widehat{s}_m\right)_{m \in \mathcal{M}}$ d'estimateurs des moindres carrés associés à des modèles non-linéaires $S_m \subset \mathbb{R}^n$, estimer $(s_1, \ldots, s_n)$ presque aussi bien que le meilleur des $\widehat{s}_m$.

Résultats :

- Notre estimateur mélange les $\widehat{s}_m$ par pondération exponentielle, de même que [LB06] dans le cas de modèles linéaires.
- Cet estimateur vérifie une inégalité de type oracle avec grande probabilité pour des modèles S_m non-linéaires.
- On utilise des arguments de concentration de [BM01, Mas07].

Regrets interne et *swap*

Nous avons étudié d'autres formes de regret qui jouent un rôle important en théorie des jeux. Le cadre séquentiel est celui introduit par [FS97].

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

1. Le statisticien choisit un vecteur de poids $\boldsymbol{p}_t = (p_{i,t})_{1 \leqslant i \leqslant K}$ 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 $\boldsymbol{p}_t \cdot \boldsymbol{\ell}_t \triangleq \sum_{i=1}^K p_{i,t} \ell_{i,t}$.

Les regrets interne et *swap* sont de la forme

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

où l'infimum est pris sur un ensemble d'applications linéaires $\varphi : \mathcal{X}_K \to \mathcal{X}_K$ qui préservent le simplexe $\mathcal{X}_K \triangleq \left\{ \boldsymbol{x} \in \mathbb{R}_+^K, \quad \sum_{i=1}^K x_i = 1 \right\}$.

Vitesse minimax des regrets interne et *swap*

Nous avons étudié les vitesses minimax des regrets interne et *swap* en environnement stochastique ou déterministe.

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

Bornes inf et sup prouvées par [Sto05] et [BM07].

Vitesse minimax des regrets interne et *swap*

Nous avons étudié les vitesses minimax des regrets interne et *swap* en environnement stochastique ou déterministe.

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

Regret interne : agrégation exponentielle utile pour supprimer $\sqrt{\ln K}$.

Vitesse minimax des regrets interne et *swap*

Nous avons étudié les vitesses minimax des regrets interne et *swap* en environnement stochastique ou déterministe.

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

Regrets interne et *swap* : vitesses exactes en stochastique.

Vitesse minimax des regrets interne et *swap*

Nous avons étudié les vitesses minimax des regrets interne et *swap* en environnement stochastique ou déterministe.

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

Borne inférieure *swap* déterministe : preuve via l'inégalité de Pinsker.

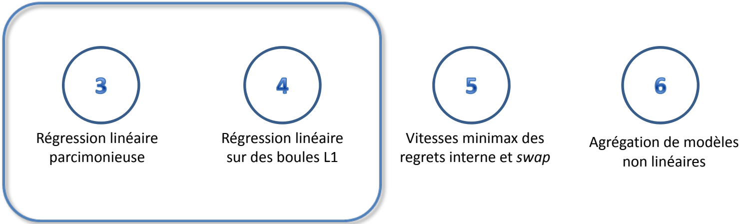
Vitesse minimax des regrets interne et *swap*

Nous avons étudié les vitesses minimax des regrets interne et *swap* en environnement stochastique ou déterministe.

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

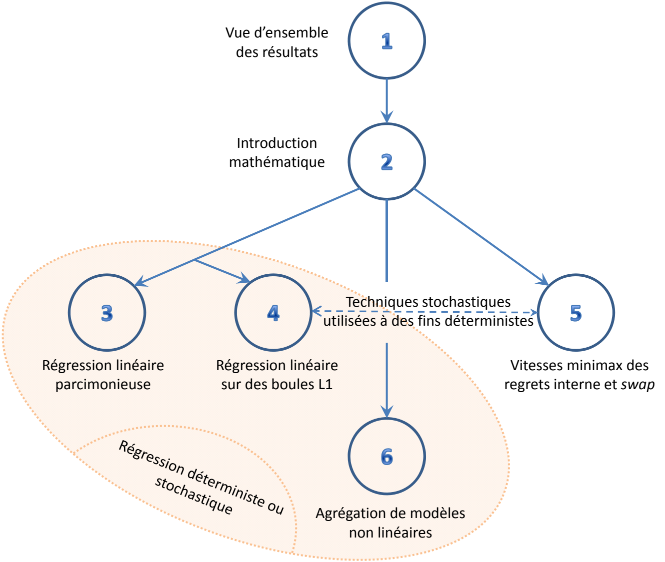
Problème ouvert : quid des facteurs $\sqrt{\ln K}$ en déterministe ?

Conclusion : structure de la thèse



Les liens entre ces chapitres peuvent être synthétisés comme suit.

Conclusion : structure de la thèse



Perspectives de recherche

Je souhaiterais aborder plusieurs problèmes au croisement des suites individuelles et du cadre statistique classique. En voici deux exemples.

1. Régression parcimonieuse : peut-on prouver des garanties déterministes pour une variante séquentielle de l'algorithme Lasso ?

2. Peut-on tisser des liens plus étroits entre les suites individuelles et la sélection de modèles en termes de ...
   - ... techniques de calibration ?
   - ... bornes ? (bornes de type oracle en suites individuelles ?)

Perspectives de recherche

Je souhaiterais aborder plusieurs problèmes au croisement des suites individuelles et du cadre statistique classique. En voici deux exemples.

1. Régression parcimonieuse : peut-on prouver des garanties déterministes pour une variante séquentielle de l'algorithme Lasso ?
2. Peut-on tisser des liens plus étroits entre les suites individuelles et la sélection de modèles en termes de ...
   - ... techniques de calibration ?
   - ... bornes ? (bornes de type oracle en suites individuelles ?)

Merci !

P. Auer, N. Cesa-Bianchi, and C. Gentile.
Adaptive and self-confident on-line learning algorithms.
64:48–75, 2002.

P. Alquier and K. Lounici.
PAC-Bayesian bounds for sparse regression estimation with exponential weights.
Electron. J. Stat., 5:127–145, 2011.

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

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

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

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

F. Bunea and A. Nobel.
Sequential procedures for aggregating arbitrary estimators of a conditional mean.
IEEE Trans. Inform. Theory, 54(4):1725–1735, 2008.

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

F. Bunea, A. B. Tsybakov, and M. H. Wegkamp.
Aggregation for Gaussian regression.
Ann. Statist., 35(4):1674–1697, 2007.

O. Catoni.
Universal aggregation rules with exact bias bounds.
Technical Report PMA-510, Laboratoire de Probabilités et Modèles Aléatoires, CNRS, Paris, 1999.

O. Catoni.
Statistical learning theory and stochastic optimization.
Springer, New York, 2004.

N. Cesa-Bianchi, A. Conconi, and C. Gentile.
On the generalization ability of on-line learning algorithms.
IEEE Trans. Inform. Theory, 50(9):2050–2057, 2004.

N. Cesa-Bianchi, P.M. Long, and M.K. Warmuth.
Worst-case quadratic loss bounds for prediction using linear functions and gradient descent.
IEEE Trans. Neural Networks, 7(3):604–619, 1996.

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

A. Dalalyan and A. B. Tsybakov.
Aggregation by exponential weighting and sharp oracle inequalities.
In *Proceedings of the 20th Annual Conference on Learning Theory (COLT'07)*, pages 97–111, 2007.

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

A. Dalalyan and A. B. Tsybakov.
Mirror averaging with sparsity priors.
Bernoulli, 2011.
To appear. Available at `http://hal.archives-ouvertes.fr/hal-00461580/`.

D. Foster.
Prediction in the worst-case.
Ann. Statist., 19:1084–1090, 1991.

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.

A. E. Hoerl and R. W. Kennard.
Ridge regression: biased estimation for nonorthogonal problems.
Technometrics, 12(1):55–67, 1970.

I. M. Johnstone and B. W. Silverman.
Empirical bayes selection of wavelet thresholds.
Ann. Statist., 33(4):1700–1752, 2005.

Jyrki Kivinen and Manfred K. Warmuth.
Exponentiated gradient versus gradient descent for linear predictors.
Inform. and Comput., 132(1):1–63, 1997.

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

N. Littlestone.
From on-line to batch learning.
In *Proceedings of the 2nd Annual Conference on Learning Theory (COLT'89)*, pages 269–284, 1989.

N. Littlestone and M. K. Warmuth.
The weighted majority algorithm.
Inform. and Comput., 108:212–261, 1994.

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

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

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

P. Rigollet and A. B. Tsybakov.
Exponential Screening and optimal rates of sparse estimation.
Ann. Statist., 39(2):731–771, 2011.

M. W. Seeger.
Bayesian inference and optimal design for the sparse linear model.
J. Mach. Learn. Res., 9:759–813, 2008.

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

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

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

V. Vovk.
Aggregating strategies.
In *Proceedings of the 3rd Annual Workshop on Computational Learning Theory (COLT'90)*, pages 371–383, 1990.

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