Régression linéaire séquentielle pour des suites déterministes arbitraires

Liens avec le cadre statistique classique

Sébastien Gerchinovitz

DMA, École Normale Supérieure / Université Paris-Sud 11

Introduction

On va aborder un problème statistique usuel – la régression linéaire – dans deux cadres différents :

- un cadre statistique classique, où les données $(Y_t)_{t \geqslant 1}$ sont modélisées de façon stochastique ;
- un cadre séquentiel déterministe, où l'on ne fait aucune hypothèse stochastique sur la suite $(y_t)_{t \geqslant 1}$ des données à prévoir ; cela conduit à des algorithmes de prévision très robustes.

Ces deux cadres entretiennent des liens étroits qu'on présentera après quelques rappels.

Plan de l'exposé

Modèle linéaire gaussien (*design* fixe)

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

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

En grande dimension $d > T$:

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

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

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

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

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

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 :

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

$\mathrm{reg}(\boldsymbol{u})$	$\mathbb{E}\left[R(\widehat{\boldsymbol{u}})\right]$	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].

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 $y_t \in \mathbb{R}$.

Ouvrage de référence : [CBL06].

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

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

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

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

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

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

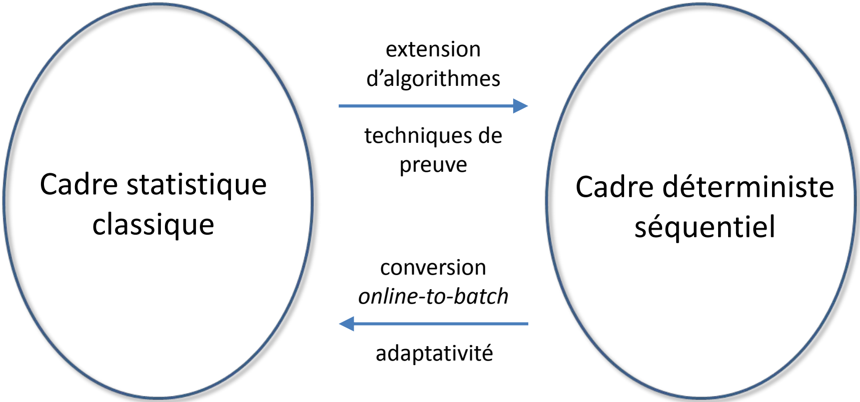
Cet algorithme vérifie, pour toute suite $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_T, y_T) \in \mathbb{R}^d \times \mathbb{R}$,

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

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 en grande dimension.

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 $d \ll T / \ln T$.

En grande dimension $d > T / \ln T$, 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 [DT08] dans le cadre stochastique, favorise la parcimonie :

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

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*

$$\sum_{t=1}^T (y_t - \widehat{y}_t)^2$$

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

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 [DT08].

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*

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

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 $\mathrm{SeqSEW}_\tau^{B,\eta}$ *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

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

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

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

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

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

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

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

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

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

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

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

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 $(X_1, Y_1, \dots, X_T, Y_T)$. Alors,

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

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

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

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 $(X_1, Y_1, \dots, X_T, Y_T)$. Alors,

$$\mathbb{E}\left[\left(f(X) - \widehat{f}_T(X)\right)^2\right]$$

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

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

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

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.

Conclusion

Synthèse et contributions principales :

- La prévision de suites individuelles et le cadre statistique classique entretiennent des liens étroits.
- Nous avons importé la notion de borne de parcimonie dans le cadre déterministe.
- En retour, notre algorithme séquentiel automatique peut être utilisé sur des données i.i.d. pour s'adapter à la variance inconnue du bruit.

Perspectives directes (en cours) :

- Obtention de bornes de parcimonie pour des algorithmes eux-mêmes parcimonieux (Lasso séquentiel par ex.).

Conclusion

Synthèse et contributions principales :

- La prévision de suites individuelles et le cadre statistique classique entretiennent des liens étroits.
- Nous avons importé la notion de borne de parcimonie dans le cadre déterministe.
- En retour, notre algorithme séquentiel automatique peut être utilisé sur des données i.i.d. pour s'adapter à la variance inconnue du bruit.

Perspectives directes (en cours) :

- Obtention de bornes de parcimonie pour des algorithmes eux-mêmes parcimonieux (Lasso séquentiel par ex.).

Les résultats de parcimonie sont tirés du papier [Ger11]. Cet exposé est disponible sur ma page web : http://www.math.ens.fr/∼gerchinovitz

Merci pour votre attention !

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.
Mach. Learn., 43(3):211–246, 2001.

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

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 and G. Lugosi.
Prediction, Learning, and Games.
Cambridge University Press, 2006.

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, sharp PAC-Bayesian bounds and sparsity.
Mach. Learn., 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. Gerchinovitz.
Sparsity regret bounds for individual sequences in online linear regression.
JMLR Workshop and Conference Proceedings, 19 (COLT 2011 Proceedings):377–396, 2011.

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.

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. 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.

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.