

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

Dans cet exposé, on présente des **liens étroits** entre ces deux cadres, autour du problème de la régression parcimonieuse.

# Plan de l'exposé

Notre exposé s'articulera comme suit :

- ① Cadre statistique : quelques rappels
- ② Cadre séquentiel déterministe
- ③ Liens étroits autour de la régression parcimonieuse

- 1 Cadre statistique : quelques rappels
  - Modèle de régression
  - Quelques estimateurs classiques
- 2 Régression linéaire séquentielle
  - Cadre
  - Exemple d'algorithme séquentiel
  - Quels liens avec le cadre statistique classique ?
- 3 Liens autour de la régression parcimonieuse en grande dimension
  - Grande dimension : même problème qu'en statistique
  - Algorithme séquentiel et bornes associées
  - Application au cadre statistique classique

# Modèle de régression

Le statisticien **observe**  $T$  couples aléatoires  $(X_1, Y_1), \dots, (X_T, Y_T) \in \mathbb{R}^d \times \mathbb{R}$  indépendants et donnés par :

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

où :

- la fonction de régression  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  est **inconnue** ;
- les v.a. de bruit  $\varepsilon_1, \dots, \varepsilon_T$  sont i.i.d., centrées et de carré intégrable.

L'objectif du statisticien est d'**estimer la fonction  $f$**  à partir de la seule donnée de l'échantillon  $(X_1, Y_1), \dots, (X_T, Y_T)$ . On distingue ci-après les cas où le *design*  $(X_1, \dots, X_T)$  est déterministe ou aléatoire.

## Design fixe versus design aléatoire

Le statisticien observe  $T$  couples aléatoires  $(X_1, Y_1), \dots, (X_T, Y_T) \in \mathbb{R}^d \times \mathbb{R}$  indépendants et donnés par :

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

- ① *Design fixe* : les vecteurs  $X_1, \dots, X_T \in \mathbb{R}^d$  sont **déterministes**.

Objectif de **débruitage** : estimer le vecteur  $(f(X_1), \dots, f(X_T))$  en norme euclidienne.

- ② *Design aléatoire* : les vecteurs  $X_1, \dots, X_T \in \mathbb{R}^d$  sont **aléatoires i.i.d.**

Objectif de **généralisation** : estimer la fonction  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  en norme  $\mathbb{L}^2(P_X)$ , i.e., estimer  $f(X)$  sur une nouvelle réalisation  $X \sim X_1$ .

## Exemple : modèle linéaire gaussien

Le *design*  $X_1, \dots, X_T \in \mathbb{R}^d$  est déterministe, le bruit  $\varepsilon_t \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$  gaussien et on postule une **relation linéaire** de la forme  $f(\mathbf{x}) = \sum_{j=1}^d u_j^* x_j$ :

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

Le vecteur  $\mathbf{u}^* \in \mathbb{R}^d$  est **inconnu**. A partir de l'**observation de**  $(X_t, Y_t)_{1 \leq t \leq T}$ , le statisticien peut viser l'un des deux objectifs suivants :

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

## Exemple : modèle linéaire gaussien

Le *design*  $X_1, \dots, X_T \in \mathbb{R}^d$  est déterministe, le bruit  $\varepsilon_t \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$  gaussien et on postule une **relation linéaire** de la forme  $f(\mathbf{x}) = \sum_{j=1}^d u_j^* x_j$ :

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

Le vecteur  $\mathbf{u}^* \in \mathbb{R}^d$  est **inconnu**. A partir de l'**observation** de  $(X_t, Y_t)_{1 \leq t \leq T}$ , le statisticien peut viser l'un des deux objectifs suivants :

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

Remarque en passant : comment étendre le modèle à un cadre non linéaire ?

On peut remplacer les  $X_{t,j}$  par  $\varphi_j(X_t)$  pour des fonctions non linéaires  $\varphi_j : \mathbb{R}^d \rightarrow \mathbb{R}$  (éléments d'une base de fonctions par ex).



# Réécriture sous forme matricielle

Le modèle linéaire gaussien

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

peut être réécrit matriciellement sous la forme  $Y = X\mathbf{u}^* + \varepsilon$ , avec

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_T \end{pmatrix}, \quad X = \begin{pmatrix} X_{1,1} & \dots & X_{1,d} \\ \vdots & & \vdots \\ X_{T,1} & \dots & X_{T,d} \end{pmatrix} \quad \text{et} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_T \end{pmatrix}.$$

L'objectif est d'estimer le vecteur inconnu  $X\mathbf{u}^* \in \mathbb{R}^T$ , i.e., de construire un estimateur  $\hat{\mathbf{u}} \in \mathbb{R}^d$  qui minimise l'**erreur quadratique moyenne** (EQM):

$$\frac{1}{T} \|X\mathbf{u}^* - X\hat{\mathbf{u}}\|^2,$$

où  $\|\cdot\|$  désigne la norme euclidienne standard sur  $\mathbb{R}^T$ .

# Estimateur des moindres carrés

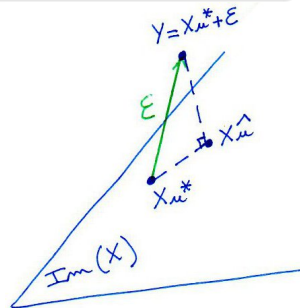
Objectif : construire un estimateur  $\hat{\mathbf{u}} \in \mathbb{R}^d$  qui minimise l'erreur quadratique moyenne  $\frac{1}{T} \|\mathbf{X}\mathbf{u}^* - \mathbf{X}\hat{\mathbf{u}}\|^2$ .

Un estimateur classique est l'estimateur des moindres carrés :

$$\hat{\mathbf{u}} \in \operatorname{argmin}_{\mathbf{u} \in \mathbb{R}^d} \frac{1}{T} \|\mathbf{Y} - \mathbf{X}\mathbf{u}\|^2.$$

Par définition,

$$\mathbf{X}\hat{\mathbf{u}} = \operatorname{proj}_{\operatorname{Im}(\mathbf{X})}(\mathbf{Y}) = \mathbf{X}\mathbf{u}^* + \operatorname{proj}_{\operatorname{Im}(\mathbf{X})}(\boldsymbol{\varepsilon})$$



## Proposition

L'estimateur des moindres carrés vérifie  $\mathbb{E}[\mathbf{X}\hat{\mathbf{u}}] = \mathbf{X}\mathbf{u}^*$  et

$$\mathbb{E} \left[ \frac{1}{T} \|\mathbf{X}\mathbf{u}^* - \mathbf{X}\hat{\mathbf{u}}\|^2 \right] = \frac{\operatorname{rg}(\mathbf{X}) \sigma^2}{T}.$$

# EQM de l'estimateur des moindres carrés

## Proposition

L'estimateur des moindres carrés vérifie  $\mathbb{E}[\mathbf{X}\hat{\mathbf{u}}] = \mathbf{X}\mathbf{u}^*$  et

$$\mathbb{E}\left[\frac{1}{T} \|\mathbf{X}\mathbf{u}^* - \mathbf{X}\hat{\mathbf{u}}\|^2\right] = \frac{\text{rg}(\mathbf{X}) \sigma^2}{T}.$$

On en déduit la majoration  $\mathbb{E}\left[\frac{1}{T} \|\mathbf{X}\mathbf{u}^* - \mathbf{X}\hat{\mathbf{u}}\|^2\right] \leq d\sigma^2/T$ . Cette erreur est faible en petite dimension  $d \ll T$ .

## EQM de l'estimateur des moindres carrés

## Proposition

L'estimateur des moindres carrés vérifie  $\mathbb{E}[\mathbf{X}\hat{\mathbf{u}}] = \mathbf{X}\mathbf{u}^*$  et

$$\mathbb{E}\left[\frac{1}{T} \|\mathbf{X}\mathbf{u}^* - \mathbf{X}\hat{\mathbf{u}}\|^2\right] = \frac{\text{rg}(\mathbf{X}) \sigma^2}{T}.$$

On en déduit la majoration  $\mathbb{E}\left[\frac{1}{T} \|\mathbf{X}\mathbf{u}^* - \mathbf{X}\hat{\mathbf{u}}\|^2\right] \leq d\sigma^2/T$ . Cette erreur est faible en petite dimension  $d \ll T$ .

Que se passe-t-il en grande dimension  $d > T$  ?

- Si la matrice  $\mathbf{X} \in \mathbb{R}^{T \times d}$  est de rang maximal, alors l'EQM est égale au niveau de bruit :  $\mathbb{E}\left[\frac{1}{T} \|\mathbf{X}\mathbf{u}^* - \mathbf{X}\hat{\mathbf{u}}\|^2\right] = \sigma^2$  (sur-apprentissage).
- La tâche de prévision est néanmoins possible sous des hypothèses supplémentaires.

# Grande dimension et parcimonie

$$\mathbb{E} \left[ \frac{1}{T} \|X\mathbf{u}^* - X\hat{\mathbf{u}}\|^2 \right] \leq d\sigma^2/T : \text{erreur faible en petite dimension } d \ll T.$$

En grande dimension  $d > T$ , pas de prévision intéressante en général.

→ Des hypothèses supplémentaires sont nécessaires.

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

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

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

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

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  $\mathbf{u}^*$  est inconnu, mais on peut imiter l'estimateur précédent en **régularisant** les moindres carrés :

$$\hat{\mathbf{u}} \in \operatorname{argmin}_{\mathbf{u} \in \mathbb{R}^d} \left\{ \frac{1}{T} \|Y - X\mathbf{u}\|^2 + \operatorname{reg}(\mathbf{u}) \right\} .$$

$\operatorname{reg}(\mathbf{u})$	$\mathbb{E}[\operatorname{EQM}]$	Hypothèses sur $X$	Coût algorithmique
$\ \mathbf{u}\ _0$	$\frac{s \ln(d/s)}{T}$	aucune	combinatoire
$\ \mathbf{u}\ _1$	$\frac{s \ln d}{T}$	$X_{\bullet,j}$ presque orthogonaux	minimisation convexe

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

- 1 Cadre statistique : quelques rappels
  - Modèle de régression
  - Quelques estimateurs classiques
- 2 Régression linéaire séquentielle
  - Cadre
  - Exemple d'algorithme séquentiel
  - Quels liens avec le cadre statistique classique ?
- 3 Liens autour de la régression parcimonieuse en grande dimension
  - Grande dimension : même problème qu'en statistique
  - Algorithme séquentiel et bornes associées
  - Application au cadre statistique classique

# Cadre séquentiel déterministe

## ① Cadre déterministe

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

**Tâche de prévision** : à chaque date  $t \geq 1$ , prévoir l'observation  $y_t \in \mathbb{R}$ . Pour ce faire, le statisticien dispose d'un vecteur de prévisions élémentaires  $\mathbf{x}_t = (x_{t,j})_{1 \leq j \leq d} \in \mathbb{R}^d$  qu'il peut combiner pour formuler sa prévision  $\hat{y}_t \in \mathbb{R}$ .



# Cadre séquentiel déterministe

## ① Cadre déterministe

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

**Tâche de prévision** : à chaque date  $t \geq 1$ , prévoir l'observation  $y_t \in \mathbb{R}$ . Pour ce faire, le statisticien dispose d'un vecteur de prévisions élémentaires  $\mathbf{x}_t = (x_{t,j})_{1 \leq j \leq d} \in \mathbb{R}^d$  qu'il peut combiner pour formuler sa prévision  $\hat{y}_t \in \mathbb{R}$ .

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

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

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

$$\sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \mathbf{x}_t)^2 + \Delta_{T,d}(\mathbf{u}) \right\},$$

où le terme de **regret**  $\Delta_{T,d}(\mathbf{u})$  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  $\mathbf{x}_t \in \mathbb{R}^d$ .
- 2 Le statisticien formule sa prévision  $\hat{y}_t \in \mathbb{R}$  à l'aide des prévisions élémentaires  $x_{t,j}$  et des données passées  $(\mathbf{x}_s, y_s)$ ,  $1 \leq s \leq t-1$ .
- 3 L'environnement révèle l'observation  $y_t \in \mathbb{R}$  et le statisticien encourt la perte carrée  $(y_t - \hat{y}_t)^2$ .

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

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

où le terme de **regret**  $\Delta_{T,d}(\mathbf{u})$  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  $\hat{y}_t = \hat{\mathbf{u}}_t \cdot \mathbf{x}_t$ , où

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

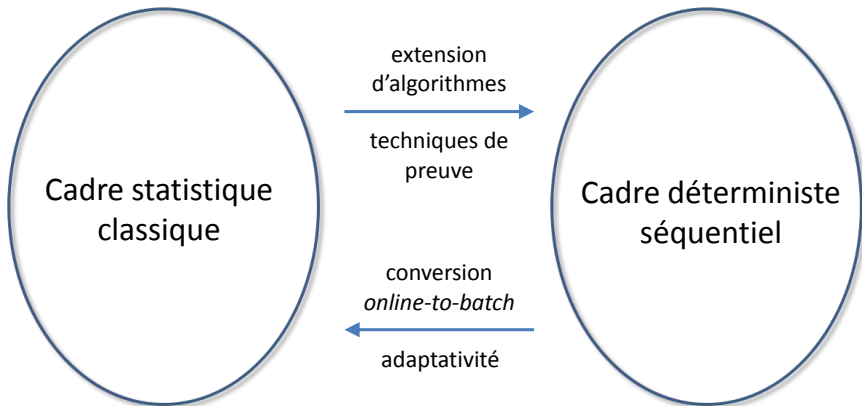
Cet algorithme vérifie, pour toute suite  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_T, y_T) \in \mathbb{R}^d \times \mathbb{R}$ ,

$$\sum_{t=1}^T (y_t - \hat{y}_t)^2 \leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (y_t - \mathbf{u} \cdot \mathbf{x}_t)^2 + \lambda \|\mathbf{u}\|_2^2 + d C_y \ln T \right\} + \dots$$

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

# Liens entre les cadres statistique et déterministe

Malgré des protocoles et hypothèses de prévision radicalement différents, les cadres statistique et déterministe entretiennent des **liens étroits**.



On illustre ces liens pour la régression parcimonieuse en grande dimension.

- 1 Cadre statistique : quelques rappels
  - Modèle de régression
  - Quelques estimateurs classiques
- 2 Régression linéaire séquentielle
  - Cadre
  - Exemple d'algorithme séquentiel
  - Quels liens avec le cadre statistique classique ?
- 3 Liens autour de la régression parcimonieuse en grande dimension
  - Grande dimension : même problème qu'en statistique
  - Algorithme séquentiel et bornes associées
  - Application au cadre statistique classique

# 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  $\mathbf{u}^* \in \mathbb{R}^d$  **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  $\mathbf{u}^*$ , i.e.,

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

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

## Bornes de parcimonie

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

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

où  $g$  croît au plus logarithmiquement en  $T$ ,  $d$  et  $\|\mathbf{u}\|_1 \triangleq \sum_{j=1}^d |u_j|$ .  
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

$$\mathbb{E}[R(\hat{\mathbf{u}})] \leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ R(\mathbf{u}) + C \frac{\|\mathbf{u}\|_0 \ln d}{T} \right\}.$$



# Algorithme SeqSEW (*Sequential Sparse Exponential Weighting*)

**Paramètres:** seuil  $B$ , température inverse  $\eta$  et paramètre d'échelle  $\tau$ .

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

$$\hat{y}_t \triangleq \int_{\mathbb{R}^d} [\mathbf{u} \cdot \mathbf{x}_t]_B p_t(d\mathbf{u}),$$

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

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

pour une constante de renormalisation  $W_t$ .

La probabilité a priori  $\pi_{\tau}$  sur  $\mathbb{R}^d$ , introduite par [DT08] dans le cadre stochastique, favorise la **parcimonie** :

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

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  $\pi_\tau$  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  $B_x$  :

$$y_1, \dots, y_T \in [-B_y, B_y] \quad \text{et} \quad \|\mathbf{x}_1\|_\infty, \dots, \|\mathbf{x}_T\|_\infty \leq B_x.$$

## Théorème (G.)

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

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

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  $\pi_\tau$  [DT08].

## Borne de parcimonie

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

$$y_1, \dots, y_T \in [-B_y, B_y] \quad \text{et} \quad \|\mathbf{x}_1\|_\infty, \dots, \|\mathbf{x}_T\|_\infty \leq B_x .$$

## Théorème (G.)

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

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

**Calibration automatique** : on peut prouver une borne similaire à l'aide de paramètres  $B_t$ ,  $\eta_t$  et  $\tau_t$  calibrés uniquement en fonction des données, i.e.,

$$\max_{1 \leq s \leq t-1} |y_s| \quad \text{et} \quad \max_{1 \leq s \leq t-1} \|\mathbf{x}_s\|_\infty .$$

# Application au cadre statistique classique

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

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

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  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  inconnue.

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

$$\sum_{t=1}^T (Y_t - \underbrace{\tilde{f}_t(X_t)}_{=\hat{y}_t})^2 \leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T (Y_t - \mathbf{u} \cdot X_t)^2 + 64 \max_{1 \leq t \leq T} Y_t^2 \|\mathbf{u}\|_0 \ln(\dots) \right\} + \dots$$

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

$$\tilde{f}_t(\mathbf{x}) \triangleq \int_{\mathbb{R}^d} [\mathbf{u} \cdot \mathbf{x}]_{B_t} \rho_t(d\mathbf{u}).$$

# Conversion *online-to-batch*

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

Rappel : on a la borne **déterministe**

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

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

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

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**

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

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

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

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**

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

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

$$\mathbb{E} \left[ \underbrace{(f(X) - \hat{f}_T(X))^2}_{+\sigma^2} \right] \leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \mathbb{E} \left[ \underbrace{(f(X) - \mathbf{u} \cdot X)^2}_{+\sigma^2} \right] + 64 \mathbb{E} \left[ \max_{1 \leq t \leq T} Y_t^2 \right] \frac{\|\mathbf{u}\|_0}{T} \ln(\cdot) \right\} + \dots$$

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

$$\begin{aligned} & \mathbb{E} \left[ (f(X) - \widehat{f}_T(X))^2 \right] \\ & \leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \mathbb{E} \left[ (f(X) - \mathbf{u} \cdot X)^2 \right] + 64 \mathbb{E} \left[ \max_{1 \leq t \leq T} Y_t^2 \right] \frac{\|\mathbf{u}\|_0}{T} \ln \left( 1 + \frac{\sqrt{dT} \|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) \right\} + \dots \end{aligned}$$

On peut majorer  $\mathbb{E} \left[ \max_{1 \leq t \leq T} Y_t^2 \right]$  sous diverses hypothèses : par ex., si  $\|f\|_\infty < +\infty$  et si  $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$  indép. de  $X_t$ , alors

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

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  $\sigma^2$ .

## 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,

$$\begin{aligned} & \mathbb{E} \left[ (f(X) - \widehat{f}_T(X))^2 \right] \\ & \leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ \mathbb{E} \left[ (f(X) - \mathbf{u} \cdot X)^2 \right] + 64 \mathbb{E} \left[ \max_{1 \leq t \leq T} Y_t^2 \right] \frac{\|\mathbf{u}\|_0}{T} \ln \left( 1 + \frac{\sqrt{dT} \|\mathbf{u}\|_1}{\|\mathbf{u}\|_0} \right) \right\} + \dots \end{aligned}$$

On peut majorer  $\mathbb{E} \left[ \max_{1 \leq t \leq T} Y_t^2 \right]$  sous diverses hypothèses : par ex., si  $\|f\|_\infty < +\infty$  et si  $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$  indép. de  $X_t$ , alors

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

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  $\sigma^2$** .

# 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.