



Gilles Stoltz

Chargé de recherche CNRS
à l'Ecole normale supérieure, Paris

Professeur affilié à HEC Paris

Apprentissage statistique

*Série de cours sur la prévision séquentielle
avec avis d'experts*

Cours 1 / 2

Agrégation convexe des avis d'experts

Apprentissage par agrégation séquentielle d'experts.

Cadre: Ensembles: Convexe X de prévisions
 arbitraire Y d'observations } Fonction de perte $l: X \times Y \rightarrow \mathbb{R}$
 N experts (= boîtes noires prédictives)

Pour $t = 1, 2, \dots$

- 1) L'environnement choisit en secret $y_t \in Y$
- 2) Les experts (en se fondant sur y_1, \dots, y_{t-1} et éventuellement sur d'autres informations à leur disposition) déterminent des prévisions $f_{jt} \in X, j = 1, \dots, N$

- 3) Le statisticien choisit sa prévision $\hat{y}_t \in X$ sous la forme d'une agrégation convexe
 (où $p_{jt} \geq 0, \sum_j p_{jt} = 1$) $\hat{y}_t = \sum_{j=1}^N p_{jt} f_{jt}$

on notera $p_t \in \Delta\{1, \dots, N\}$ ←

- 4) La vraie observation y_t est révélée et le statisticien encourt la perte $l(\hat{y}_t, y_t)$

Objectif: $\sum_{t=1}^n l(\hat{y}_t, y_t) = \inf_{q \in \Delta\{1, \dots, N\}} \sum_{t=1}^n l(\sum_j q_j f_{jt}, y_t) + \left(\sum_{t=1}^n l(\hat{y}_t, y_t) - \inf_q \sum_{t=1}^n l(\sum_j q_j f_{jt}, y_t) \right)$

↑
 erreur d'approximation
 ↔ performance des experts

↑
 "regret" ↔ difficulté d'extrapolation séquentielle

Note: On aura même des contrôles uniformes sur le regret:

$$\sup_{y_1, \dots, y_n} \sup_{(f_{jt})_{t \leq n, j \leq N}} R_n$$

→ "suites individuelles"
 ↘ "agrégation robuste"

avec pour objectif:
 $\sup R_n = o(n)$

Hyp (cette semaine): $l(\cdot, y)$ convexe $\forall y$

Exemples:

Prévisions (méta-) statistiques de phénomènes quantitatifs

$$X = Y = [b, B] \quad \text{et} \quad \ell(x, y) = (x - y)^2$$

Pics d'ozone: $b = 0$, $B = 300 \mu\text{g m}^{-3}$ 4 experts = modèles numériques

Consommation élec: $b = 0$, $B = 110 \text{ MW}$ 8 experts = méthodes stats EDF

Taux de change: $b = 0.5$, $B = 2$ (euros) 8 experts = méthodes économétriques
EUR / USD

CAS DE LA SIMPLE COMPARAISON AUX EXPERTS.

Algorithme : $\hat{y}_t = \sum_{j=1}^N p_{jt} f_{jt}$ où $f_t \in \Delta\{1, \dots, N\}$ est définie comme

$$p_{jt} = \frac{e^{-\eta \sum_{s=1}^{t-1} \ell(f_{js}, y_s)}}{\sum_{k=1}^N e^{-\eta \sum_{s=1}^{t-1} \ell(f_{ks}, y_s)}}$$

Théorème : Si ℓ est bornée à valeurs dans $[a, b] \subset \mathbb{R}$ et convexe en son premier argument ($\forall y : x \mapsto \ell(x, y)$ est convexe) alors la stratégie ci-dessus est telle que

$$\begin{aligned} \sup R_n &= \sup \left\{ \sum_{t=1}^n \ell(\hat{y}_t, y_t) - \min_j \sum_{t=1}^n \ell(f_{jt}, y_t) \right\} \\ &\leq \frac{\ln N}{\eta} + \eta \frac{(b-a)^2}{8} n \end{aligned}$$

Corollaire : $\sup R_n \leq (b-a) \sqrt{\frac{n}{2} \ln N}$ lorsque l'on connaît a, b et n et que l'on avait pris $\eta = \frac{1}{b-a} \sqrt{\frac{8 \ln N}{n}}$

↳ Savci : Evidemment, n et parfois a et b sont en général inconnus, et la calibration de η est un savci. On y reviendra en détails.

Preuve : On commence par utiliser la convexité des $\ell(\cdot, y_t)$:

$$\ell(\hat{y}_t, y_t) \leq \sum_{j=1}^N p_{jt} \underbrace{\ell(f_{jt}, y_t)}_{=: \ell_{jt}} \quad [\text{Jensen}]$$

et donc :

$$R_n \leq \sum_{t=1}^n \sum_{j=1}^N p_{jt} \ell_{jt} - \min_j \sum_{t=1}^n \ell_{jt}$$

On applique alors Hoeffding :

$$-\eta \sum_{j=1}^N p_{jt} \ell_{jt} \geq \log \mathbb{E} \left[\sum_j p_{jt} e^{-\eta \ell_{jt}} \right] - \frac{\eta^2}{8} (b-a)^2$$

$$\left(\Leftrightarrow \text{s. ex} \right) \quad \left(\Leftrightarrow \log \mathbb{E} [e^{SX}] \right) \quad \left(\Leftrightarrow -\frac{S^2}{8} (b-a)^2 \right)$$

Donc

$$\begin{aligned} \sum_{t=1}^n \sum_{j=1}^N p_{jt} l_{jt} &\leq \sum_t -\frac{1}{\eta} \log \left(\sum_j p_{jt} e^{-\eta l_{jt}} \right) + \sum_t \frac{\eta}{8} (b-a)^2 \\ &\leq \frac{\eta}{8} (b-a)^2 - \frac{1}{\eta} \sum_{t=1}^n \log \frac{\sum_j e^{-\eta \sum_{s=1}^t l_{js}}}{\sum_j e^{-\eta \sum_{s=1}^n l_{js}}} \\ &\stackrel{(\text{téléscopage})}{=} \frac{\eta}{8} (b-a)^2 - \frac{1}{\eta} \log \left(\frac{1}{N} \sum_{j=1}^N e^{-\eta \sum_{t=1}^n l_{jt}} \right) \\ &\geq e^{-\eta \min_j \sum_{t=1}^n l_{jt}} \end{aligned}$$

Ce qui conclut la preuve.

CADRE GÉNÉRIQUE

Si l'on regarde comment fonctionne la preuve et l'algorithme précédents, on voit qu'ils valent également pour le cadre dit générique suivant :

Déroulement :

Pour $t=1, 2, \dots$:

- 1) Le statisticien choisit $p_t \in \Delta\{1, \dots, N\}$
- 2) L'environnement, éventuellement au vu de p_t , choisit un vecteur $\underline{l}_t = (l_{t1}, \dots, l_{tN}) \in [a, b]^N$
- 3) Le statisticien observe \underline{l}_t .

But :

Le statisticien veut contrôler

$$R_n^* = \sup_{\substack{\text{sur } \underline{l}_1, \dots, \underline{l}_n \\ \text{ou sur } \mathcal{L} \\ \text{stratégies de l'environnement}}} \left\{ \sum_{t=1}^n \sum_{j=1}^N p_{jt} l_{jt} - \min_{j=1, \dots, N} \sum_{t=1}^n l_{jt} \right\}$$

Résultat :

$$R_n^* \leq \frac{\ln N}{\eta} + \frac{\eta}{8} (b-a)^2 \quad \text{pour l'algorithme de}$$

pondération par poids exponentiels :

$$p_j^t = e^{-\eta \sum_{s=1}^{t-1} l_{js}} / \sum_{k=1}^N e^{-\eta \sum_{s=1}^{t-1} l_{ks}}$$

RETOUR À LA COMPARAISON À LA MEILLEURE COMBINAISON CONVEXE.

Hypothèses : $X \subset \mathbb{R}^d$ et $\forall y \in Y, x \mapsto l(x, y)$ convexe et différentiable sur X

Alors (inégalité des pentes) :
 $\forall y, \forall x, x' \in X, l(x, y) - l(x', y) \leq \nabla l(x, y) \cdot (x - x')$

Ex : $X = Y = [a, b]$ et $l(x, y) = (x - y)^2$ [cas de la prévision statistique]

Application : On s'intéresse au regret face à la meilleure combinaison convexe

$$\sup_{\substack{\underline{L}_t \\ \text{ou stratégies}}} \sup_{q \in \Delta\{1, \dots, N\}} \sum_{t=1}^n \left(l\left(\sum_{j=1}^N p_j^t f_{jt}, y_t\right) - l\left(\sum_{j=1}^N q_j f_{jt}, y_t\right) \right)$$

on note $\leq \sup_q \sup_{\underline{L}_t} \sum_{t=1}^n \nabla l\left(\sum_{j=1}^N p_j^t f_{jt}, y_t\right) \cdot \left(\sum_{j=1}^N p_j^t f_{jt} - \sum_{j=1}^N q_j f_{jt}\right)$

$$= \sup_q \left\{ \sup_{\underline{L}_t} \sum_{t=1}^n \sum_{j=1}^N p_j^t l_{jt} - \sum_{t=1}^n \sum_{j=1}^N q_j l_{jt} \right\}$$

$$= \sup_q \left\{ \sum_{t=1}^n \sum_{j=1}^N p_j^t l_{jt} - \min_j \sum_{t=1}^n l_{jt} \right\}$$
 Car linéaire en q

= un certain regret générique

↑
 l_{jt} = $\nabla l\left(\sum_{k=1}^N p_k^t f_{kt}, y_t\right) \cdot f_{jt}$
 j-ème composante du gradient de la fonction Ψ_t définie à la page suivante.

Hyp : $(x, y, x') \mapsto \nabla l(x, y) \cdot x'$ est bornée, G sur $\|\cdot\|_\infty$
 de sorte que : $\forall j, t, l_{jt} \in [-GG] = [a, b]$

Stratégie = poids exponentiels sur les pseudo-pertes gradient

$$l_{jt} = \alpha \left(\sum_k p_{kt} f_{kt}(y_t) \right) \cdot f_{jt}$$

Boone uniforme = $\frac{\ln N}{\eta} + \frac{\eta^m}{2} G^2$, optimisable en $G \sqrt{2m \ln N}$.

Ex: perte $(\cdot)^2$: $G = 2B^2$.

! Extension à l'hypothèse $X \subset \mathbb{R}^d$ et $Y \subset \mathbb{Y}$, $z \mapsto l(z, y)$ convexe

[mais pas d'hyp. suppl. sur la différentiabilité]

Ex: $X = Y = [0, B]$ et $l(z, y) = |z - y|$ \hookrightarrow sans de différentiabilité en 0

MAIS: inégalité des pertes car: $x = y$ fixé,

$$\forall x', \quad l(x, y) - l(x', y) = 0 - |x' - y|$$

$$\stackrel{?}{\leq} a(x - x') = a(y - x')$$

oui \rightarrow pour tout $a \in [-1, 1]$: tout un choix possible de points...

Th: Si $g: \mathcal{D} \rightarrow \mathbb{R}$ est convexe, avec $\mathcal{D} \subset \mathbb{R}^d$, alors sur \mathcal{D} , g est continue et pour tout $\underline{u} \in \mathcal{D}$, l'ensemble $\mathcal{G}(\underline{u})$ de vecteurs $\underline{\kappa}$ tels que

$$\forall \underline{v} \in \mathcal{D}, \quad g(\underline{u}) - g(\underline{v}) \leq \underline{\kappa} \cdot (\underline{u} - \underline{v})$$

est non vide.

On appelle $\mathcal{G}(\underline{u})$ le sous-gradient de g en \underline{u} .

Lorsque g est différentiable en \underline{u} , $\mathcal{G}(\underline{u}) = \{ \nabla g(\underline{u}) \}$.

Ici: on considère les $\Psi_t: q \in \Delta\{1, \dots, N\} \mapsto l\left(\sum_j q_j f_{jt}, y_t\right)$ où $\mathcal{D} = \Delta\{1, \dots, N\}$ et on utilise le fait que vu notre algorithme, les points p_t retenus sont toujours tels que $p_t \in \mathcal{D}$.

C'est un des sous-gradients de Ψ_t qui donne le vecteur de pseudo-pertes $\underline{l}_t = (l_{1t}, \dots, l_{Nt})$.

CALIBRATION DE η :

ADAPTATION AU TEMPS n

Dans le théorème de performance de l'algorithme de pondération par poids exponentiels, pour avoir la borne $\Delta \sqrt{n \ln n}$ sur le regret, il faut choisir $\eta = \Delta \sqrt{\frac{\ln n}{m}}$ pour optimiser la borne théorique.

Si η est mal choisi ou si on joue pour une infinité de tours, alors potentiellement la borne peut être linéaire.

→ Adaptation à n nécessaire.

(Ici, on ne traitera pas de l'adaptation à l'étendue des pertes.)

Ide: Faire varier η au cours du temps, pour $t \geq 1$,

$$P_t^j = e^{-\eta t \sum_{s=1}^{t-1} l_{js}} / \sum_{k=1}^N e^{-\eta \sum_{s=1}^{t-1} l_{ks}}$$

Lm: Si (η_t) est décroissante, alors

$$\sup \left\{ \sum_{t=1}^n \sum_{j=1}^N P_t^j - \min_{i=1, \dots, N} \sum_{t=1}^n l_{it} \right\} \leq \frac{\ln n}{\eta_m} + \sum_{t=1}^n \frac{\eta_t}{8} (b-a)^2$$

Cor: Avec $\eta_t = \frac{1}{b-a} \sqrt{\frac{4 \ln n}{t}}$, $\sup R_n \leq (b-a) \sqrt{(\ln n) \ln n}$.

↳ Le prix de l'adaptation est donc (essentiellement) un facteur multiplicatif $\sqrt{2}$.

PR (Cor): $\eta_t = \frac{\gamma}{b-a} \sqrt{\frac{\ln n}{t}}$ (même forme que η constant optimal)

Alors $\sup R_n \leq (b-a) \sqrt{\ln n} \left(\frac{\sqrt{\ln n}}{\gamma} + \frac{\gamma}{8} \sum_{t=1}^n \frac{1}{\sqrt{t}} \right)$

soit: $\sup R_n \leq (b-a) \sqrt{(\ln n) \ln n} \left(\frac{1}{\gamma} + \frac{\gamma}{4} \right) \leq \int_1^{\ln n} \frac{1}{\sqrt{t}} dt \leq 2\sqrt{\ln n}$
↳ = 1 pour $\gamma=2$.

PR (Lm): Hoeffding donne: $\sum_{j=1}^N P_t^j l_{jt} \leq -\frac{1}{\eta_t} \log \left(\sum_{j=1}^N P_t^j e^{-\eta_t l_{jt}} \right) + \frac{\eta_t}{8} (b-a)^2$

Or, puisque $\eta_{t+1} \leq \eta_t$, $x \mapsto x^{\eta_t / \eta_{t+1}}$ est convexe, de sorte que

$$\begin{aligned}
N \times \frac{1}{N} \sum_j P_j^t e^{-\eta_t l_j^t} &= N \times \frac{1}{N^2} \sum_j^2 \left(P_j^t \eta_{t+1} / \eta_t e^{-\eta_{t+1} l_j^t} \right)^{\eta_t / \eta_{t+1}} \\
&\stackrel{\text{Jensen}}{\geq} N \times \left(\frac{1}{N} \sum_j P_j^t \eta_{t+1} / \eta_t e^{-\eta_{t+1} l_j^t} \right)^{\eta_t / \eta_{t+1}} \\
&= N^{1 - \eta_t / \eta_{t+1}} \times \frac{\left(\sum_j^2 e^{-\eta_{t+1} \frac{t}{s_j} l_j^t} \right)^{\eta_t / \eta_{t+1}}}{\sum_{k=1}^N e^{-\eta_t \sum_{s=1}^{t-1} l_{j_s}^k}}
\end{aligned}$$

De sorte que

$$\begin{aligned}
\sum_{j=1}^N P_j^t l_j^t &\leq \underbrace{\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \log N}_{-\frac{1}{\eta_t} \log N^{1 - \eta_t / \eta_{t+1}}} + \frac{\eta_t}{8} (b-a)^2 \\
&\quad - \frac{1}{\eta_t} \times \frac{\eta_t}{\eta_{t+1}} \log \left(\sum_j^2 e^{-\eta_{t+1} \frac{t}{s_j} l_j^t} \right) + \frac{1}{\eta_t} \log \sum_{j=1}^N e^{-\eta_t \sum_{s=1}^{t-1} l_{j_s}^k}
\end{aligned}$$

En sommant sur $t=1, \dots, n$, des télescopes ont lieu:

$$\begin{aligned}
\sum_{t=1}^n \sum_{j=1}^N P_j^t l_j^t &\leq \left(\frac{1}{\eta_n} - \frac{1}{\eta_1} \right) \log N + \sum_{t=1}^n \frac{\eta_t}{8} (b-a)^2 \\
&\quad - \frac{1}{\eta_n} \log \left(\underbrace{\sum_j e^{-\eta_n \sum_{t=1}^n l_{j_t}^t}}_{\geq e^{-\eta_n \min_{j=1 \dots N} \sum_{t=1}^n l_{j_t}^t}} \right) + \frac{1}{\eta_1} \log N
\end{aligned}$$

Ce qui conclut la preuve.