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

Coordonnées : stoltz@math.ens.fr --- http://www.math.ens.fr/~stoltz/

Apprentissage par agrégation séquentielle d'experts.

Cadre : Ensembles : convexe $\mathcal{X}$ de prévisions, arbitraire $\mathcal{Y}$ d'observations } Fonction de perte $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$

N experts (= boîtes noires prédictives)

Pour $t = 1, 2, \dots$

1) L'environnement choisit en secret $y_t \in \mathcal{Y}$
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 $f_{jt} \in \mathcal{X}$, $j = 1 \dots N$
3) Le statisticien choisit sa prévision $\hat{y}_t \in \mathcal{X}$ 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 $\ell(\hat{y}_t, y_t)$

Objectif :

$$\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) = \inf_{q \in \Delta_{\{1..N\}}} \sum_{t=1}^{n} \ell\Big(\sum_j q_j f_{jt}, y_t\Big) + \Big(\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_q \sum_{t=1}^{n} \ell\Big(\sum_j q_j f_{jt}, y_t\Big) \Big)$$

- $\inf_{q \in \Delta_{\{1..N\}}} \sum_{t=1}^{n} \ell\Big(\sum_j q_j f_{jt}, y_t\Big)$: erreur d'approximation ↔ performance des experts
- $\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_q \sum_{t=1}^{n} \ell\Big(\sum_j q_j f_{jt}, y_t\Big)$: "regret" ↔ difficulté d'estimation séquentielle

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

$\sup_{y_1 \dots y_n, (f_{jt})_{t \leq n, j \leq N}} R_n$ → "suites individuelles" ↘ "agrégation robuste"

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

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

Exemples : Prévisions (méta-) statistiques de phénomènes quantitatifs

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

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

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

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

CAS DE LA SIMPLE COMPARAISON AUX EXPERTS.

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

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

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

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

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

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

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

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

et donc :

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

On applique alors Hoeffding :

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

$$\left(\leftrightarrow\ s\,\mathbb{E}X\right) \qquad \left(\leftrightarrow \log \mathbb{E}[e^{sX}]\right) \qquad \left(\leftrightarrow -\frac{s^2}{8}(b-a)^2\right)$$

D'où

$$\sum_{t=1}^{n}\sum_{j=1}^{N} p_{jt}\ell_{jt} \leq \sum_t -\frac{1}{\eta}\log\left(\sum_j p_{jt}e^{-\eta\ell_{jt}}\right) + \sum_t \frac{\eta(b-a)^2}{8}$$

$$\leq \frac{\eta n}{8}(b-a)^2 - \frac{1}{\eta}\sum_{t=1}^{n}\log\frac{\sum_j e^{-\eta\sum_{s=1}^{t}\ell_{js}}}{\sum_j e^{-\eta\sum_{s=1}^{t-1}\ell_{js}}}$$

$$\overset{\text{(télescopage)}}{=} \frac{\eta n}{8}(b-a)^2 - \frac{1}{\eta}\log\left(\frac{1}{N}\sum_{j=1}^{N} e^{-\eta\sum_{t=1}^{n}\ell_{jt}}\right)$$

$$\geq e^{-\eta\min_j\sum_{t=1}^{n}\ell_{jt}}$$

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 $\underline{\ell}_t = (\ell_{1t}, \dots \ell_{Nt}) \in [a;b]^N$
3) Le statisticien observe $\underline{\ell}_t$.

But : Le statisticien veut contrôler

$$R_n^* = \sup_{\text{sur } \underline{\ell}_1, \dots \underline{\ell}_m \text{ ou sur les stratégies de l'environnement}} \left\{ \sum_{t=1}^{n}\sum_{j=1}^{N} p_{jt}\ell_{jt} - \min_{j=1\dots N}\sum_{t=1}^{n}\ell_{jt} \right\}$$

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

pondération par poids exponentiels :

$$p_{jt} = e^{-\eta \sum_{s=1}^{t-1} \ell_{js}} \Big/ \sum_{k=1}^{N} e^{-\eta \sum_{s=1}^{t-1} \ell_{ks}}.$$

RETOUR À LA COMPARAISON À LA MEILLEURE COMBINAISON CONVEXE.

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

Alors (inégalité des pentes) :

$$\forall y, \quad \forall x, x' \in \mathcal{X}, \quad \ell(x,y) - \ell(x',y) \leq \nabla \ell(x,y) \cdot (x - x')$$

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

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

$$\sup_{\ell_t} \sup_{q \in \Delta_{\{1,\dots,N\}}} \sum_{t=1}^{n} \left(\ell\left(\sum_{j=1}^{N} p_{jt} f_{jt}, y_t \right) - \ell\left(\sum_j q_j f_{jt}, y_t \right) \right)$$

ou stratégies

$$\leq \sup_{\ell_t} \sup_{q} \sum_{t=1}^{n} \nabla \ell\left(\sum_j p_{jt} f_{jt}, y_t \right) \cdot \left(\sum_j p_{jt} f_{jt} - \sum_j q_j f_{jt} \right)$$

on note $\ell_{jt} = \nabla \ell\left(\sum_k p_{kt} f_{kt}, y_t \right) \cdot f_{jt}$ (↑ j-ième composante du gradient de la fonction ψ_t définie à la page suivante.)

$$= \sup_{\ell_t} \left\{ \sup_{q} \sum_{t=1}^{n} \sum_j p_{jt} \ell_{jt} - \sum_t \sum_j q_j \ell_{jt} \right\}$$

$$= \sup_{\ell_t} \left\{ \sum_{t=1}^{n} \sum_j p_{jt} \ell_{jt} - \min_j \sum_{t=1}^{n} \ell_{jt} \right\}$$ (car linéaire en q)

$$= \text{un certain regret générique}$$

Hyp : $(x,y,x') \mapsto \nabla \ell(x,y) \cdot x'$ est bornée, G sur $\|\cdot\|_\infty$

de sorte que :

$$\forall j,t, \quad \ell_{jt} \in [-G,G] = [a,b].$$

Stratégie = poids exponentiels sur les pseudo-pertes gradient

$$\ell_{jt} = \nabla \ell\Big(\sum_k p_{kt} f_{kt}, y_t\Big) \cdot f_{jt}$$

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

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

⚠ Extension à l'hypothèse $X \subset \mathbb{R}^d$ et $\forall y \in Y$, $x \mapsto \ell(x,y)$ convexe

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

Ex: $X = Y = [0,B]$ et $\ell(x,y) = |x-y|$ ↪ souci de différentiabilité en 0

MAis : inégalité des pertes ok : $x = y$ fixé,

$$\forall x', \quad \ell(x,y) - \ell(x',y) = 0 - |x'-y| \overset{?}{\leq} a(x-x') = a(y-x')$$

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

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

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

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

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

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 fallait choisir $\eta = \Delta\sqrt{\frac{\ln N}{n}}$ 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. $\longrightarrow$ Adaptation à n nécessaire.

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

Idée : Faire varier η au cours du temps,

pour $t \geq 1$, $\quad p_{jt} = e^{-\eta_t \sum_{s=1}^{t-1} \ell_{js}} \Big/ \sum_{k=1}^{N} e^{-\eta \sum_{s=1}^{t-1} \ell_{ks}}$.

Lem : Si (η_t) est décroissante, alors

$$\sup\left\{ \sum_{t=1}^{n} \sum_{j=1}^{N} p_{jt}\ell_{jt} - \min_{i=1\dots N} \sum_{t=1}^{n} \ell_{it} \right\} \leq \frac{\ln N}{\eta_{n+1}} + \sum_{t=1}^{n} \frac{\eta_t}{8}(b-a)^2$$

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

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

sait :

$$\sup R_n \leq (b-a)\sqrt{(n+1)\ln N}\left(\frac{1}{\gamma} + \frac{\gamma}{4}\right)$$

↳ $= 1$ pour $\gamma = 2$.

↳ $\leq \int_1^{n+1} \frac{1}{\sqrt{t}}\,dt \leq 2\sqrt{n+1}$

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

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

$$N \times \frac{1}{N}\sum_j p_{jt}\, e^{-\eta_t \ell_{jt}} = N \times \frac{1}{N}\sum_j \left(p_{jt}^{\eta_{t+1}/\eta_t} e^{-\eta_{t+1}\ell_{jt}}\right)^{\eta_t/\eta_{t+1}}$$

$$\overset{\text{Jensen}}{\geqslant} N \times \left(\frac{1}{N}\sum_j p_{jt}^{\eta_{t+1}/\eta_t} e^{-\eta_{t+1}\ell_{jt}}\right)^{\eta_t/\eta_{t+1}}$$

$$= N^{1-\eta_t/\eta_{t+1}} \times \frac{\left(\sum_{j=1}^N e^{-\eta_{t+1}\sum_{s=1}^t \ell_{js}}\right)^{\eta_t/\eta_{t+1}}}{\sum_{k=1}^N e^{-\eta_t \sum_{s=1}^{t-1}\ell_{ks}}}$$

De sorte que

$$\sum_{j=1}^N p_{jt}\ell_{jt} \leqslant \underbrace{(1/\eta_{t+1} - 1/\eta_t)}\log N - \frac{1}{\eta_t}\log N^{1-\eta_t/\eta_{t+1}} + \frac{\eta_t}{8}(b-a)^2$$

$$- \frac{1}{\eta_t} \times \frac{\eta_t}{\eta_{t+1}} \log\left(\sum_{j=1}^N e^{-\eta_{t+1}\sum_{s=1}^t \ell_{js}}\right) + \frac{1}{\eta_t}\log \sum_{j=1}^N e^{-\eta_t \sum_{s=1}^{t-1}\ell_{js}}$$

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

$$\sum_{t=1}^n \sum_{j=1}^N p_{jt}\ell_{jt} \leqslant \left(\frac{1}{\eta_{n+1}} - \frac{1}{\eta_1}\right)\ln N + \sum_{t=1}^n \frac{\eta_t}{8}(b-a)^2$$

$$- \frac{1}{\eta_{n+1}} \log\Big(\underbrace{\sum_j e^{-\eta_{n+1}\sum_{t=1}^n \ell_{jt}}}_{\geqslant\, e^{-\eta_{n+1}\min_{j=1\dots N}\sum_{t=1}^n \ell_{jt}}}\Big) + \frac{1}{\eta_1}\ln N$$

Ce qui conclut la preuve.