Prédiction de suites individuelles
Cours 2 : Prédiction avec experts (suite)

Gilles Stoltz (CNRS – Ecole normale supérieure – HEC Paris)

Jeudi 5 mai 2011

RAPPELS : Jeu répété de prédiction.

Paramètres : N experts $j = 1, \ldots N$

Ensemble d'observations $\mathcal{Y}$ (arbitraire), ensemble de prédictions $\mathcal{X}$ (convexe)

Pour tous les tours $t = 1, 2, \ldots$

1) les experts procurent les conseils $f_{1t}, \ldots f_{Nt} \in \mathcal{X}$, révélés au statisticien;
2) l'environnement choisit l'observation $y_t \in \mathcal{Y}$, sans la révéler au statisticien;
3) le statisticien forme une prédiction $\hat{p}_t \in \mathcal{X}$, généralement de la forme $\hat{p}_t = \sum_{j=1}^{N} p_{jt} f_{jt}$ où $(p_{jt})_{j=1,\ldots N}$ est une combinaison convexe;
4) y_t est alors révélée au statisticien.

Evaluation : Une fonction de perte $\ell : \mathcal{X} \times \mathcal{Y} \longrightarrow [0, M]$ est donnée.

On veut minimiser la différence entre la perte cumulée du statisticien $\hat{L}_n = \sum_{t=1}^{n} \ell(\hat{p}_t, y_t)$ et celles des experts

$$L_{jn} = \sum_{t=1}^{n} \ell(f_{jt}, y_t).$$

Cette différence est appelée le regret

$$R_n = \hat{L}_n - \min_{j=1,\ldots N} L_{jn}.$$

On veut des stratégies telles que

$$\sup R_n \leqslant o(n)$$

où le supremum est pris sur l'ensemble des suites $y_1, \ldots y_n$ (cas d'un environnement ne réagissant pas aux prédictions du statisticien) ou des stratégies de l'environnement (cas du jeu contre le diable).

Algorithme : L'algorithme de pondération par poids exponentiels des performances propose $\mu_1 = (1/N, \dots 1/N)$

puis pour $t \geq 2$, μ_t définie par $\mu_{j,t} = \dfrac{\exp(-\eta L_{j,t-1})}{\sum_{k=1}^{N} \exp(-\eta L_{k,t-1})}$

<u>Théorème</u>: Si pour tout $y \in \mathcal{Y}$, $\ell(\cdot, y)$ est convexe, alors le choix $\eta = \dfrac{1}{M}\sqrt{\dfrac{8 \ln N}{n}}$ assure que

$$\sup R_n \leq M \sqrt{\frac{n}{2} \ln N}.$$

<u>Remarque</u>: Nous avons vu comment nous passer de la connaissance de n pour calibrer η.

<u>Application</u>: Le lemme de Sion.

Soient $\mathcal{X}$ et $\mathcal{Y}$ deux ensembles convexes et $f : \mathcal{X} \times \mathcal{Y} \longrightarrow [0, M]$ une fonction telle que

- $\forall x \in \mathcal{X}$, $f(x, \cdot)$ est concave,
- $\forall y \in \mathcal{Y}$, $f(\cdot, y)$ est convexe.

Alors, sous des hypothèses de régularité (p.ex. $\mathcal{X}, \mathcal{Y}$ métriques, et f uniformément continue sur $\mathcal{X} \times \mathcal{Y}$, et $\mathcal{X}$ compact),

$$\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} f(x,y) = \sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} f(x,y).$$

<u>Optimalité</u> (asymptotique), y compris de la constante :

<u>Théorème</u>: Pour $\mathcal{X} = [0,1]$, $\mathcal{Y} = \{0,1\}$ et $\ell(x,y) = |x-y|$, toute stratégie de prédiction encourt un regret R_n tel que

$$\liminf_{N \to +\infty} \liminf_{n \to +\infty} \frac{\sup R_n}{\sqrt{\frac{n}{2} \ln N}} \geq 1$$

où il suffit de prendre le sup sur des suites individuelles d'observations y_t et de conseils d'experts $f_{j,t} \in \{0,1\}$.

<u>Preuve</u>: Randomisation des Y_t et des $F_{j,t}$ iid $\sim \mathrm{Ber}(1/2)$...

$$\sup_{y_t,\, f_{j,t}} \left\{ \sum_{t=1}^{n} \left| \hat{p}_t - y_t \right| - \min_{j=1,\dots,N} \sum_{t=1}^{n} \left| f_{j,t} - y_t \right| \right\} = \sup R_n.$$

$$\geqslant \mathbb{E}\left[\sum_{t=1}^{n} \left| \hat{p}_t - Y_t \right| - \min_{j=1\dots N} \sum_{t=1}^{n} \left| F_{j,t} - Y_t \right| \right] = \mathbb{E}[R_n]$$

tq. Y_t et $F_{j,t}$ tous iid $\sim \mathcal{B}(1/2)$

$$= \underbrace{\dots}_{\text{cf. calculs ci-dessous}} = \frac{n}{2} - \mathbb{E}\left[\min_{j=1\dots N} B_{j,n} \right]$$

où les $B_{j,n}$ sont iid $\sim \mathrm{Bin}(n, 1/2)$

On fait apparaître une quantité du type de celle du TCL :

HEURISTIQUE

$$Z_{j,n} = 2\left(\frac{n/2 - B_{j,n}}{\sqrt{n}} \right) \longrightarrow \mathcal{N}(0,1)$$

$$\mathbb{E}[R_n] = \frac{\sqrt{n}}{2}\, \mathbb{E}\left[\max_{j \leq N} 2\left(\frac{n/2 - B_{j,n}}{\sqrt{n}} \right) \right] \underset{n\to+\infty}{\sim} \frac{\sqrt{n}}{2}\, \mathbb{E}\left[\max_{j \leq N} G_j \right]$$

où G_j iid $\sim \mathcal{N}(0,1)$

puis on utilisera que :

$$\mathbb{E}\left[\max_{j \leq N} G_j \right] \sim \sqrt{2 \ln N}.$$

1) Puisque les Y_t sont indépendantes de tout le reste et $\sim \mathrm{Ber}(1/2)$, notant

$$\mathbb{E}_t = \mathbb{E}\left[\,\cdot \mid F_{j,s} \text{ où } j = 1,\dots N \text{ et } s = 1,\dots t \text{ et } Y_s \text{ où } s = 1\dots t-1 \right]$$

l'espérance conditionnelle par rapport à l'information disponible au statisticien au début du tour t (pour former $\hat{p}_t$), il vient, $\hat{p}_t$ étant ainsi fixé :

$$\mathbb{E}_t\left[\ell(\hat{p}_t, Y_t) \right] = \mathbb{E}\left[\left| \hat{p}_t - Y_t \right| \right] = \frac{1}{2} \hat{p}_t + \frac{1}{2}(1 - \hat{p}_t) = 1/2$$

soit au final : $\mathbb{E}\left[\hat{L}_n \right] = n/2$

(Morale : c'est dû à la contrainte de prédiction séquentielle.)

Tandis que : $L_{jn} = \sum_{t=1}^{n} |F_{jt} - Y_t|$, dont on fait une évaluation rétrospective.

Conditionnellement à $Y_1 \dots Y_n$, les $|F_{jt} - Y_t|$ sont iid $\sim Ber(1/2)$ lorsque j et t varient, de sorte que les L_{jn} sont iid $\sim Bin(n, 1/2)$ lorsque j varie.

De plus, cette loi commune est indépendante des $Y_1, \dots Y_n$.

Ainsi :
$$\mathbb{E}[R_n] = \mathbb{E}[\hat{L}_n] - \mathbb{E}\left[\mathbb{E}\left[\min_{j=1\dots N} L_{jn} \mid Y_1 \dots Y_n\right]\right]$$
$$= n/2 - \mathbb{E}\left[\min_{j=1\dots N} B_{jn}\right]$$

où $B_{1n}, \dots B_{Nn}$ sont N variables aléatoires iid $\sim Bin(n, 1/2)$.

Soit :
$$\mathbb{E}[R_n] = \sqrt{n}\left(\mathbb{E}\left[\max_{j=1\dots N} (n/2 - B_{jn})/\sqrt{n}\right]\right)$$
$$= \sqrt{\frac{n}{2}}\ \mathbb{E}\left[\max_j Z_{jn}\right]$$

où :

2) Par TCL,
$$Z_{jn} = 2\left(\frac{n/2 - B_{jn}}{\sqrt{n}}\right) \longrightarrow \mathcal{N}(0,1)$$

Les Z_{jn} étant indépendants lorsque j varie, on tire de ces convergences marginales une convergence jointe :

$$\begin{pmatrix} Z_{1n} \\ \vdots \\ Z_{Nn} \end{pmatrix} \longrightarrow \mathcal{N}(\underline{0}, I_N).$$

On veut maintenant prouver la convergence des espérances (attention ! ce n'est pas impliqué en général par la convergence en loi).

En fait, vu l'énoncé du théorème, on veut juste prouver

$$\liminf_{n \to +\infty} \mathbb{E}\left[\max_{j=1,\dots,N} Z_{jn}\right] \geqslant \mathbb{E}\left[\max_{j=1,\dots,N} G_j\right].$$

Soit ϕ_L la fonction seuil à L : $\phi_L : x \in \mathbb{R} \longmapsto \begin{cases} x & \text{si } x \in [-L, L] \\ -L & \text{si } x \leqslant -L \\ L & \text{si } x \geqslant L \end{cases}$

$(x_1, \dots x_N) \mapsto \phi(\max_j x_j)$ étant continue bornée :

$$\forall L > 0, \qquad \mathbb{E}\left[\phi_L\left(\max_j Z_{jn}\right)\right] \longrightarrow \mathbb{E}\left[\phi_L\left(\max_j G_j\right)\right]$$

Or,

$$\mathbb{E}\left[\max_j Z_{jn}\right] \geqslant \mathbb{E}\left[\phi_L\left(\max_j Z_{jn}\right)\right] + \underbrace{\mathbb{E}\left[\left(L + \max_j Z_{jn}\right) \mathbb{1}_{\{\max_j Z_{jn} \leqslant -L\}}\right]}$$

est égal, par théorème de Fubini, à

$$-\int_{-\infty}^{0} \mathbb{P}\left\{\max_{j=1,\dots,N} Z_{jn} \leqslant u - L\right\} du.$$

En effet, pour X variable aléatoire, vu la "positivité" (négativité) :

$$\mathbb{E}\left[X \mathbb{1}_{\{X \leqslant 0\}}\right]$$
$$= \int_{-\infty}^{0} x \, d\mathbb{P}^X(x)$$
$$= \iint_{(\mathbb{R}_-)^2} \left(-\mathbb{1}_{\{x \leqslant u \leqslant 0\}}\right) d\mathbb{P}^X(x) du$$
$$= -\int_{-\infty}^{0} \mathbb{P}\{X \leqslant u\} \, du$$

et on applique cela à $X = \max_j Z_{jn} + L$

Or (pour $d \leq 0$) $\mathbb{P}\left\{ \max_{j \leq N} Z_{jn} \leq d \right\} = \left(\mathbb{P}\{Z_{1n} \leq d\} \right)^N \underset{\text{(Markov)}}{\leq} \left(\dfrac{\operatorname{Var} Z_{1n}}{d^2} \right)^N = \dfrac{1}{d^{2N}}$

Par construction, $\operatorname{Var} Z_{jn} = 1$

et donc
$$\mathbb{E}\left[\left(L + \max_j Z_{jn}\right) \mathbb{1}_{\{\max_j Z_{jn} \leq -L\}} \right] \geq - \int_{-\infty}^{0} \frac{1}{(u-L)^{2N}} \, du$$
$$= \underbrace{- \int_{-\infty}^{-L} 1/u^{2N} \, du}_{\hat{=}\, r_{L,N} \xrightarrow[L \to +\infty]{} 0}$$

Au final, $\forall L > 0$,
$$\liminf_{n \to +\infty} \mathbb{E}\left[\max_j Z_{jn} \right] \geq \liminf_{n \to +\infty} \mathbb{E}\left[\phi_L\left(\max_j Z_{jn}\right) \right] - r_{L,N}$$
$$\underset{\substack{\text{par} \\ \text{bornitude +} \\ \text{cv en loi}}}{=} \mathbb{E}\left[\phi_L\left(\max_j G_j \right) \right] - r_{L,N}$$

Ceci valant pour tout L, et vu $r_{L,N} \xrightarrow[L \to +\infty]{} 0$ (à N fixé)

et puisque (par convergence dominée par $|G_1| + \cdots |G_N|$)
$$\mathbb{E}\left[\phi_L\left(\max_j G_j \right) \right] \longrightarrow \mathbb{E}\left[\max_j G_j \right],$$

on a prouvé :
$$\liminf \mathbb{E}\left[\max_j Z_{jn} \right] \geq \mathbb{E}\left[\max_j G_j \right],$$

soit $\displaystyle \liminf_{n \to +\infty} 2\, \mathbb{E}[R_n] / \sqrt{n} \geq \mathbb{E}\left[\max_{j=1,\dots,N} G_j \right]$.

Pour conclure, il ne reste qu'à voir que $\displaystyle \liminf_{N \to +\infty} \mathbb{E}\left[\max_{j=1 \dots N} G_j \right] \Big/ \sqrt{2 \ln N} \geq 1$

$$\mathbb{E}\Big[\max_{j\leq N} G_j\Big] = \mathbb{E}\Big[\Big(\max_{j\leq N} G_j\Big)^+\Big] - \underbrace{\mathbb{E}\Big[\Big(\max_{j\leq N} G_j\Big)^-\Big]}$$

$$\geq \mathbb{E}\Big[\Big(\max_{j\leq N} G_j\Big)^+\Big] - 1$$

cette espérance est $\leq \mathbb{E}[G_1^-]$
$\leq \sqrt{\mathbb{E}[G_1^2]} = 1$

Soit Ψ_N la fonction de répartition de $\Big(\max_{j\leq N} G_j\Big)^+$

Φ celle de la $\mathcal{N}(0,1)$

$\Phi_N = \Phi^N$ celle de $\max_{j\leq N} G_j$

Alors $\Psi_N(0) = \Phi_N(0) = \big(\Phi(0)\big)^N = 2^{-N}$

et pour $t>0$,

$\Psi_N(t) = \big(\Phi(t)\big)^N$; en particulier, Ψ_N est inversible sur $\mathbb{R}_+^*$

D'où l'inverse généralisé :

$$\Psi_N^{(-1)}(u) = \begin{cases} 0 & \text{si } u < 2^{-N} \\ \Phi^{-1}(u^{1/N}) & \text{si } u > 2^{-N} \end{cases}$$

$:= \inf\{x : \Psi_N(x) \geq u\}$

Ainsi, $\mathbb{E}\Big[\Big(\max_{j\leq N} G_j\Big)^+\Big] = \displaystyle\int_0^1 \Psi_N^{(-1)}(u)\,du$

$$= \int_{2^{-N}}^1 \underbrace{\Phi^{-1}(u^{1/N})}_{\text{fonction croissante et } \geq 0}\,du \geq \int_\delta^1 \Phi^{-1}(u^{1/N})\,du$$

pour tout $\delta > 0$, au moins pour N assez grand

(Cf. si X v.a. de f.r. F, d'inverse généralisé $F^{(-1)}$, alors $\mathbb{E}[X] = \mathbb{E}[F^{(-1)}(U)]$ où $U \sim \mathcal{U}_{[0,1]}$)

$$\geq (1-\delta)\,\Phi^{-1}(\delta^{1/N})$$

On a donc prouvé :

$\forall \delta > 0$,

$$\lim_{N\to+\infty} \mathbb{E}\Big[\max_{j\leq N} G_j\Big]\Big/\sqrt{2\ln N} = \lim_{N\to+\infty} \mathbb{E}\Big[\Big(\max_{j\leq N} G_j\Big)^+\Big]\Big/\sqrt{2\ln N}$$

$$\geq \lim_{N\to+\infty} \frac{(1-\delta)\,\Phi^{-1}(\delta^{1/N})}{\sqrt{2\ln N}}$$

Or (exercice :)

$1 - \Phi(x) \sim e^{-x^2/2}\big/x\sqrt{2\pi}$ quand $x \to +\infty$ d'où $\dfrac{1}{1-\Phi(x)} \sim e^{x^2/2}\, x\sqrt{2\pi}$

Comme $\Phi^{-1}(v) \underset{v\to1}{\longrightarrow} +\infty$, il vient :

$$\frac{1}{1-v} = \frac{1}{1-\Phi(\Phi^{-1}(v))} \sim e^{+\Phi^{-1}(v)^2/2} \times \Phi^{-1}(v)\sqrt{2\pi}$$

et en passant aux logarithmes :

$$\ln\left(\frac{1}{1-v}\right) \sim +\Phi^{-1}(v)^2/2 \quad \text{puis} \quad \Phi^{-1}(v) \sim \sqrt{2\ln(1/_{1-v})}$$

quand $v \to 1$

On va réinjecter cet équivalent: $$(1-\delta)\,\Phi^{-1}(\delta^{1/N}) \underset{N\to+\infty}{\sim} (1-\delta)\sqrt{2\ln\left(\frac{1}{1-\delta^{1/N}}\right)}$$

$$\sim (1-\delta)\sqrt{2\ln N}$$

car $1-\delta^{1/N} = 1-\exp\left(\frac{1}{N}\ln\delta\right) \sim \frac{1}{N}\ln\delta$

Au final: $\forall \delta > 0$, $$\underset{N\to+\infty}{\underline{\lim}} \ \mathbb{E}\left[\max_{j\leq N} G_j\right]\Big/\sqrt{2\ln N}$$

$$\geq \underset{N\to+\infty}{\underline{\lim}} \ (1-\delta)\,\Phi^{-1}(\delta^{1/N})\Big/\sqrt{2\ln N} = 1-\delta$$

d'où suit le résultat désiré, en faisant $\delta \searrow 0$.

<u>Exercice:</u> $\Phi(x) = \int_{-\infty}^{x} e^{-t^2/2}/\sqrt{2\pi}\,dt$, alors $1-\Phi(x) \underset{x\to+\infty}{\sim} e^{-x^2/2}/\sqrt{2\pi}$.

<u>Preuve:</u> $$\frac{1}{\sqrt{2\pi}}\int_x^{+\infty} e^{-t^2/2}\,dt = \frac{1}{\sqrt{2\pi}}\int_x^{+\infty} t e^{-t^2/2}\cdot\frac{1}{t}\,dt$$

$$\underset{\text{IPP}}{=} \underbrace{\left[-\frac{1}{\sqrt{2\pi}}\,e^{-t^2/2}/t\right]_x^{+\infty}}_{=\ \frac{e^{-x^2/2}}{x\sqrt{2\pi}}} \underbrace{- \frac{1}{\sqrt{2\pi}}\int_x^{+\infty}\frac{e^{-t^2/2}}{t^2}\,dt}$$

or, à nouveau par IPP,

$$0 \leq \frac{1}{\sqrt{2\pi}}\int_x^{+\infty}\frac{t}{t}\frac{e^{-t^2/2}}{t^2}\,dt = \left[-\frac{1}{\sqrt{2\pi}}\,\frac{e^{-t^2/2}}{t^3}\right]_x^{+\infty} - \underbrace{\int \cdots}_{\geq 0}$$

de sorte que ce terme est dominant.

donc est $\leq \dfrac{e^{-x^2/2}}{x^3\sqrt{2\pi}}$

<u>Rappel:</u> Si $f(x) \sim g(x)$ et que $\begin{cases} f(x) \to \text{cste} \neq 1 \\ f(x) \overset{\text{ou}}{\to} +\infty \end{cases}$ alors $\ln f(x) \sim \ln g(x)$

En effet, $\ln f(x) = \ln f(x)/g(x) + \ln g(x) \sim \ln g(x)$

En fait, le sens difficile était la minoration. On peut montrer que $\mathbb{E}\left[\max_{j \leq N} G_j\right] \sim \sqrt{2 \ln N}$, via le résultat plus général suivant, qui indique que $\mathbb{E}\left[\max_{j \leq N} G_j\right] \leq \sqrt{2 \ln N}$.

<u>Lm</u>: Soit $\sigma > 0$ et $X_1 \ldots X_N$ (ni nécessairement indépendants ni nécessairement identiquement distribués) de fonctions génératrices sous-gaussiennes*:

$$\mathbb{E}\left[e^{\lambda X_i}\right] \leq e^{\lambda^2 \sigma^2 / 2} \qquad \forall \lambda > 0.$$

Alors $$\mathbb{E}\left[\max_{i=1,\ldots,N} X_i\right] \leq \sigma \sqrt{2 \ln N}.$$

<u>preuve</u>:

$$e^{\lambda \mathbb{E}\left[\max_{i=1..N} X_i\right]} \underset{[\text{Jensen}]}{\leq} \mathbb{E}\left[e^{\lambda \max_i X_i}\right] = \mathbb{E}\left[\max_{i=1..N} \underbrace{e^{\lambda X_i}}_{\geq 0}\right]$$

$$\underset{[\text{majoration dite « à la Pisier »}]}{\leq} \sum_{i=1}^{N} \mathbb{E}\left[e^{\lambda X_i}\right] \leq N e^{\lambda^2 \sigma^2 / 2}$$

et donc $$\mathbb{E}\left[\max X_i\right] \leq \frac{1}{\lambda} \ln\left(N e^{\lambda^2 \sigma^2 / 2}\right) = \underbrace{\frac{\ln N}{\lambda} + \lambda \sigma^2 / 2}$$

est minimal pour λ tq. $\lambda^2 = \dfrac{2 \ln N}{\sigma^2}$,

i.e., $\lambda = \sqrt{\dfrac{2 \ln N}{\sigma^2}}$, ce qui conclut.

* Des exemples de telles variables aléatoires sont celles vérifiant les hypothèses du lemme de Hoeffding p.ex., ou, évidemment, les variables gaussiennes.

RESUME & NOUVEAUX OBJECTIFS.

[Dans le cas non randomisé, où $\ell(\cdot, y)$ est convexe $\forall y$.

→ Agrégation séquentielle convexe de conseils d'experts :

Regret $R_n = \hat{L}_n - \min_{j=1..N} L_{j,n}$

$$\hat{L}_n = \sum_{t=1}^{n} \ell(\hat{p}_t, y_t) \qquad \sum_{t=1}^{n} \ell(f_{j,t}, y_t) \overset{\text{not.}}{=} L_{j,n}$$

où $\hat{p}_t = \sum_{j=1}^{N} p_{j,t} f_{j,t}$

Algorithme : $\quad p_1 = (1/N, \dots 1/N)$ et pour $t \geq 2$

$$p_{j,t} = \exp(-\eta_t L_{j,t-1}) \Big/ \sum_{k=1}^{N} \exp(-\eta_t L_{k,t-1})$$

Th : Pour un choix adaptatif des η_t,

$$\sup R_n \leq \square \, \|\ell\|_\infty \sqrt{n \ln N}$$

où les ordres de grandeur en n et N sont optimaux en général. (On vient de prouver cette optimalité.)

→ Objectifs suivants :

(1) Améliorer la vitesse avec des hypothèses sur ℓ (exp-concavité / convexité forte)

(2) Se comparer non pas seulement au meilleur expert j mais à la meilleure combinaison convexe q d'experts

↳ si $q = (q_1 \dots q_N)$ est une telle combinaison,

$$L_n(q) = \sum_{t=1}^{n} \ell\Big(\sum_{j=1}^{N} q_j f_{j,t}, y_t\Big)$$

et on veut que

$$\sup R_n^{cvx} \hat{=} \sup \Big\{ \hat{L}_n - \inf_q L_n(q) \Big\}$$

soit petit.

"Souvent" cet inf sera un min

EXP- CONCAVITÉ.

Déf : ℓ est η- exp. concave (ou η- fortement convexe) si pour tout $y \in \mathcal{Y}$, $F_{\eta,y} = e^{-\eta \ell(\cdot, y)}$ est concave.

Ex1 : Perte log #1 (marché boursier) :

$\mathcal{X} = \mathcal{P}$ simplexe de $\mathbb{R}^N$ (les experts proposent chacun une allocation des capitaux)

$\mathcal{Y} = (\mathbb{R}_+)^N$

$\ell(\underline{x}, \underline{y}) = -\log \underline{x} \cdot \underline{y}$ (éventuellement infini)

ℓ est 1- exp. concave ($\eta_0 = 1$)

On y revient en détails plus loin

Ex2 : Perte log #2 (liens avec le codage séquentiel → hélas, on n'aura pas le temps de développer ces liens)

$\mathcal{X} = \mathcal{P}$ simplexe de $\mathbb{R}^{\mathcal{A}}$

les experts proposent chacun une probabilité sur $\mathcal{A}$

$\mathcal{Y} = \mathcal{A} = \{1, \dots m\}$ un alphabet fini

$\ell(\underline{p}, y) = -\log p_y$ (éventuellement infini)

ℓ est 1- exp. concave ($\eta_0 = 1$)

Ex3 : Perte quadratique (prédictions statistiques)

$\mathcal{X} = \mathcal{Y} = [0, B]$

$\ell(x, y) = (x - y)^2$

Les $F_{\eta,y}$ sont deux fois dérivables, avec

$F'_{\eta,y}(x) = -2\eta(y - x) e^{-\eta(x-y)^2}$ $\quad F''_{\eta,y}(x) = \left[(2\eta(x-y))^2 - 2\eta \right] e^{-\eta(x-y)^2}$

et la condition de concavité est que

$F''_{\eta,y} \leqslant 0$ soit $\eta \leqslant \dfrac{1}{2(x-y)^2}$ $\forall x, y$

Par exemple, $\eta_0 = 1/2B^2$ est tel que ℓ est η_0 - exp. concave.

Rq: L' exp. concavité entraîne la convexité :

$$\ell(\cdot, y) = -\frac{1}{\eta} \log F_{\eta, y}$$

où $-\frac{1}{\eta}\log$ convexe et décroissante.

Stratégie 1: Mélange fini sur les experts

EWA(η) : $\left[\begin{array}{l} t \geqslant 2, \qquad \mu_{jt} = \exp\left(-\eta L_{j\,t-1}\right) \Big/ \sum_{k=1}^{N} \exp\left(-\eta L_{k\,t-1}\right) \\ \text{et } \hat{p}_t = \sum_{j=1} \mu_{jt} f_{jt}. \end{array}\right.$

Th1: Si ℓ est η-exp. concave, alors le regret de EWA(η) est borné par

$$\sup R_n \leqslant \frac{\ln N}{\eta}.$$

Rq: Pas d'hyp. sur le caractère borné ou non de ℓ!

Preuve : $w_{jt} = \exp\left(-\eta L_{jt}\right)$ et $W_t = w_{1t} + \dots w_{Nt}$

$$\ln \frac{W_n}{W_0} \geqslant -\eta L_{jn} - \ln N$$

$t = 1 \dots n$:

$$\ln \frac{W_t}{W_{t-1}} = \ln \sum_{j=1}^{N} \mu_{jt}\, e^{-\eta\, \ell(f_{jt}, y_t)}$$

$$\underset{[\text{exp. concavité}]}{\leqslant} \ln e^{-\eta\, \ell\left(\sum_j \mu_{jt} f_{jt}, y_t\right)}$$

$$= -\eta\, \ell(\hat{p}_t, y_t).$$

Stratégie 2: Mélange uniforme : $\mu_1 = (1/N, \dots 1/N)$ et pour $t \geqslant 2$,

Unif(η):

$$\mu_t = \int_{\mathcal{P}} q\, e^{-\eta L_{t-1}(q)}\, d\mu(q) \Big/ \int_{\mathcal{P}} e^{-\eta L_{t-1}(q)}\, d\mu(q)$$

où μ est la mesure uniforme sur $\mathcal{P}$ (dérive de Lebesgue).

Note : Cette stratégie est reminiscente du mélange de Laplace en codage.

Rq: Comment calculer μ_t en pratique (avec MATLAB) ? Un argument de discrétisation (grille de pas ε) est trop coûteux : $\approx 1/\varepsilon^{N-1}$

Par méthode stochastique plutôt :

On utilise que par loi des grands nombres, si $Q_1, Q_2, \dots$ sont tirés au hasard (ie, selon μ) dans $\mathcal{P}$, alors pour toute fonction f borné, éventuellement à valeurs N-dimensionnelles,

lorsque m est grand $\displaystyle\int_{\mathcal{P}} f(q)\, d\mu(q) \simeq \frac{1}{m} \sum_{k=1}^{m} f(Q_k)$

(en fait, précisément : $\displaystyle\frac{1}{m} \sum_{k=1}^{m} f(Q_k) \xrightarrow[m \to +\infty]{ps} \int_{\mathcal{P}} f(q)\, d\mu(q)$).

Mais comment tirer iid selon μ ?

Une méthode est de tirer $X_1, \dots X_{N-1}$ iid $\sim \mathcal{U}_{[0,1]}$

de les réordonner $X_{(1)} \leqslant X_{(2)} \leqslant \dots \leqslant X_{(N-1)}$

puis de considérer les N segments créés :

$$Q = \left(X_{(1)},\; X_{(2)} - X_{(1)},\; \dots,\; X_{(N-1)} - X_{(N-2)},\; 1 - X_{(N-1)} \right).$$

Théorème. Si ℓ est η-exp. concave et prend ses valeurs dans $[0, M]$, alors la stratégie de mélange continu $\mathrm{Unif}(\eta)$ est telle que

$$\sup R_n^{cvx} = \sup \left\{ \hat{L}_n - \inf_{q \in \mathcal{P}} L_n(q) \right\} \leqslant \frac{N-1}{\eta} \max \left\{ 1, \log \frac{e \eta M n}{N-1} \right\}.$$

Rq: On atteint donc une vitesse $\log n \ll \sqrt{n}$ malgré la comparaison à la meilleure combinaison convexe constante. // Note : Il n'y a pas de problème majeur d'homogénéité en M dans la borne, puisque η varie normalement en $1/M$.

Preuve: On note que pour tout t, $\displaystyle \ell(\mu_t \cdot (f_{j,t})_j, y_t) = -\frac{1}{\eta} \log e^{-\eta \ell(\mu_t \cdot (f_{j,t})_j, y_t)}$

$$\leqslant -\frac{1}{\eta} \log \frac{\int_{\mathcal{P}} e^{-\eta \ell(q \cdot (f_{j,t})_j, y_t)} e^{-\eta L_{t-1}(q)} d\mu(q)}{\int_{\mathcal{P}} e^{-\eta L_{t-1}(q)} d\mu(q)}$$

où l'on a utilisé $\left\{ \begin{array}{l} \text{l'inégalité de Jensen} \\ \text{la convexité forte} \end{array} \right\}$

et la décroissance de $-\log$

Soit $\ell(\mu_t \cdot (f_t)_j, y_t) \leq -\frac{1}{\eta} \log \dfrac{\int_{\mathcal{P}} e^{-\eta L_t(q)} d\mu(q)}{\int_{\mathcal{P}} e^{-\eta L_{t-1}(q)} d\mu(q)}$

Puis en sommant sur $t = 1, 2, \ldots$ (et en se rappelant la convention $L_0 = 0$)

$$\hat{L}_n \leq -\frac{1}{\eta} \log \int_{\mathcal{P}} e^{-\eta L_n(q)} d\mu(q).$$

§ Si l'inf n'est pas atteint, on prend q^*_ε tq. $L_n(q^*_\varepsilon)$ soit proche à ε près de $\inf_{q \in \mathcal{P}} L_n(q)$

On note $q^* \in \underset{q \in \mathcal{P}}{\operatorname{argmin}} L_n(q)$; et on utilise que les q qui sont dans un voisinage de q^* ont des performances similaires à lui :

si $q = (1-\alpha) q^* + \alpha r$ pour $\alpha \in [0,1[$ et $r \in \mathcal{P}$, alors comme (par composition et somme) L_n est convexe :

$$L_n(q) \leq (1-\alpha) L_n(q^*) + \alpha L_n(r) \leq (1-\alpha) L_n(q^*) + \alpha M_n$$

de sorte que $\displaystyle\int_{\mathcal{P}} e^{-\eta L_n(q)} d\mu(q) \geqslant e^{-\eta\left((1-\alpha)L_n(q^*) + \alpha M_n\right)} \times \mu\left(V(q^*, \alpha)\right)$

où $V(q^*, \alpha)$ désigne le α-voisinage de q^* :

$$V(q^*, \alpha) = \{ q : \exists r \in \mathcal{P} \mid q = (1-\alpha) q^* + \alpha r \}$$

Or, μ étant uniforme, $\mu\left(V(q^*, \alpha)\right) = \dfrac{\text{mesure de Lebesgue de } \alpha\mathcal{P} \text{ dans l'hyperplan affine engendré par } \mathcal{P}}{} = \alpha^{N-1} \mu(\mathcal{P}) = \alpha^{N-1}$

En réinjectant :

$$\hat{L}_n \leq -\frac{1}{\eta} \log \left(e^{-\eta\left((1-\alpha) L_n(q^*) + \alpha M_n\right)} \alpha^{N-1} \right)$$

$$= \underbrace{(1-\alpha)}_{\leq 1} L_n(q^*) + \alpha M_n + \frac{N-1}{\eta} \log 1/\alpha$$

$$\hat{L}_n - L_n(q^*) \leq \inf_{\alpha \in [0,1]} \left\{ \alpha M_n + \frac{N-1}{\eta} \log \frac{1}{\alpha} \right\} \overset{\text{not.}}{=} \Psi_{M,N,\eta}$$

(Si on avait q^*_ε, alors on a prouvé $\hat{L}_n - \inf_{q \in \mathcal{P}} L_n(q) \leq \varepsilon + \hat{L}_n - L_n(q^*_\varepsilon)$
$\leq \varepsilon + \Psi_{M,N,\eta}$ et

il suffit de faire $\varepsilon \to 0$.)

Reste à calculer $\varphi_{M,N,\eta}$:

$$\varphi : \underset{\in \mathbb{R}_+^*}{\alpha} \longmapsto \alpha M n + \frac{N-1}{\eta} \log \frac{1}{\alpha}$$

$$\varphi'(\alpha) = Mn - \frac{N-1}{\eta \alpha}, \qquad \varphi''(\alpha) = \frac{N-1}{\eta \alpha^2} > 0$$

donc φ atteint son minimum en $\alpha^* = \dfrac{N-1}{Mn\eta}$.

mais attention ! cette valeur de α n'est pas toujours <1 !

Ce n'est le cas que si $n > \dfrac{(N-1)}{M\eta}$; pour $n < \dfrac{(N-1)}{M\eta}$, on utilise que le regret est plus petit que $\widehat{L}_n$ et donc que $Mn < \dfrac{N-1}{\eta}$.

Pour $n > \dfrac{N-1}{M\eta}$, on a $\varphi(\alpha^*) = \dfrac{N-1}{\eta} + \dfrac{N-1}{\eta} \log \dfrac{Mn\eta}{N-1}$.

Ainsi, dans tous les cas : $\varphi_{M,N,\eta} = \dfrac{N-1}{\eta} + \dfrac{N-1}{\eta}\left(\log \dfrac{Mn\eta}{N-1}\right)^+$, ce qui donne la borne proposée.