

# Prédiction de suites individuelles

Cours 3 et 4 : Agrégation convexe ; Prédiction randomisée

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

Jeudis 12 et 19 mai 2011

RÉSUMÉ & NOUVEAUX OBJECTIFS.

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

→ Agrégation séquentielle convexe de conseils d'experts :

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

$\hat{L}_n = \sum_{t=1}^n l(\hat{p}_t, y_t)$        $L_{j,n} = \sum_{t=1}^n l(f_{j,t}, y_t)$   
 où  $\hat{p}_t = \sum_{j=1}^N \mu_{j,t} f_{j,t}$

Algorithme :

$\mu_1 = (1/N, \dots, 1/N)$  et pour  $t \geq 2$ ,

$\mu_{j,t} = \frac{\exp(-\eta_t L_{j,t-1})}{\sum_{k=1}^N \exp(-\eta_t L_{k,t-1})}$

Th: Pour un choix adaptatif des  $\eta_t$ ,

$\sup R_n \leq \alpha \|l\|_\infty \sqrt{n \ln N}$

où les ordres de grandeur en  $n$  et  $N$  sont optimaux en général.

→ Objectifs :

- (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,

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

"souvent", cet inf sera un min.

et on veut que soit petit.

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

MÉTHODE 1:

EXP-CONCAVITÉ.

Def:  $l$  est  $\eta$ -exp-concave (ou  $\eta$ -fortement convexe) si pour tout  $y \in Y$ ,

$$F_{\eta,y} = e^{-\eta l(\cdot,y)}$$

est concave.

Ex1: Perte-log #1 (marché boursier):  
 $X = \mathcal{P}$  simplexe de  $\mathbb{R}^N$  (les experts proposent chacun une allocation de capitaux)  
 $Y = (\mathbb{R}_+)^N$   
 $l(x,y) = -\log x \cdot y$  (éventuellement infini)  
 $l$  est 1-exp-concave ( $\eta_0 = 1$ )

On y revient en détails plus tard

Ex2: Perte-log #2 (liens avec le codage séquentiel  $\rightarrow$  hélas on n'aura pas le temps de développer ces liens)  
 $X = \mathcal{P}$  simplexe de  $\mathbb{R}^A$   
 $Y = A = \{1, \dots, m\}$  un alphabet fini  
 $l(p,y) = -\log p_y$  (éventuellement infini)  
 $l$  est 1-exp-concave ( $\eta_0 = 1$ )

les experts proposent chacun une probabilité sur  $A$

Ex3: Perte quadratique (prédictions statistiques)

$$X = Y = [0, B]$$

$$l(x,y) = (x-y)^2$$

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

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

et la condition de concavité est que

$$F''_{\eta,y} \leq 0 \quad \text{soit} \quad \eta \leq \frac{1}{2(x-y)^2} \quad \forall x,y$$

Par exemple  $\eta_0 = \frac{1}{2B^2}$  est tel que

$l$  est  $\eta_0$ -exp-concave.

Rq: L'exp. concavité entraîne la convexité:

$$l(\cdot, y) = -\frac{1}{\eta} \log F_{\eta, y} \quad \text{où } -\frac{1}{\eta} \log$$

convexe et  
décroissante.

Stratégie 1: Mélange fini sur les experts

$$\text{EWA}(\eta): \begin{cases} t \geq 2, & \mu_{jt} = \exp(-\eta L_{jt-1}) / \sum_{k=1}^N \exp(-\eta L_{kt-1}) \\ \text{et } \hat{p}_t = \sum_{j=1}^N \mu_{jt} f_{jt}. \end{cases}$$

Th1: Si  $l$  est  $\eta$ -exp. concave, alors le regret de  $\text{EWA}(\eta)$  est borné par

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

$$\sup R_n \leq \frac{\ln N}{\eta}$$

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

$$\ln \frac{w_{jt}}{w_{j0}} \geq -\eta L_{jt} - \ln N$$

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

$$\leq \ln e^{-\eta l(\sum_{j=1}^N \mu_{jt} f_{jt}, y_t)}$$

[exp. concavité]

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

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

$$\text{Unif}(\eta): \quad \mu_t = \int_{\mathcal{F}} f e^{-\eta L_{t-1}(f)} d\mu(f) / \int_{\mathcal{F}} e^{-\eta L_{t-1}(f)} d\mu(f)$$

où  $\mu$  est la mesure uniforme sur  $\mathcal{F}$  (dérivée de Lebesgue).

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

Rq: Comment calculer  $\mu_t$  en pratique (avec MATLAB)? Un argument de discrétisation (grille de pas  $\varepsilon$ ) 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  $\mu$ ) dans  $\mathcal{P}$ , alors pour toute fonction  $f$  bornée, éventuellement à valeurs  $N$ -dimensionnelles,

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

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

Mais comment tirer iid selon  $\mu$  ?

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

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

puis de considérer les  $N$  segments créés :

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

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

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

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  $\eta$  varie moralement en  $1/M$ .

Preuve: On note que pour tout  $t$ ,  $l(\mu \cdot (\frac{f}{\|f\|}, \frac{f}{\|f\|})) = -\frac{1}{\eta} \log e^{-\eta l(\mu \cdot (\frac{f}{\|f\|}, \frac{f}{\|f\|}))}$

où l'on a utilisé  $\left\{ \begin{array}{l} \text{l'inégalité de Jensen} \\ \text{la convexité forte} \end{array} \right\}$   $\leq -\frac{1}{\eta} \log \frac{\int_{\mathcal{P}} e^{-\eta l(q \cdot (\frac{f}{\|f\|}, \frac{f}{\|f\|}))} e^{-\eta L_t(q)} d\mu(q)}{\int_{\mathcal{P}} e^{-\eta L_t(q)} d\mu(q)}$

et la décroissance de  $-\log$

Soit 
$$L(\mu, (\frac{f_t}{\sum_{j=1}^n f_j}, y_t)) \leq -\frac{1}{\eta} \log \frac{\int_{\mathcal{P}} e^{-\eta L_t(q)} d\mu(q)}{\int_{\mathcal{P}} e^{-\eta L_{t-1}(q)} d\mu(q)}$$

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

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

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

On note  $q^* \in \operatorname{argmin}_{q \in \mathcal{P}} L_n(q)$ ; 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 
$$\int_{\mathcal{P}} e^{-\eta L_n(q)} d\mu(q) \geq e^{-\eta((1-\alpha)L_n(q^*) + \alpha M_n)} \times \mu(V(q^*, \alpha))$$

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

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

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

En réinjectant :

$$\begin{aligned} \hat{L}_n &\leq -\frac{1}{\eta} \log \left( e^{-\eta((1-\alpha)L_n(q^*) + \alpha M_n)} \alpha^{N-1} \right) \\ &= \underbrace{(1-\alpha)}_{\leq 1} L_n(q^*) + \alpha M_n + \frac{N-1}{\eta} \log \frac{1}{\alpha} \end{aligned}$$

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

( Si on avait  $q^*_\varepsilon$ , 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  $\varepsilon \rightarrow 0$ .)

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

$$\varphi: \begin{matrix} \alpha \\ \in \mathbb{R}_+^* \end{matrix} \mapsto \alpha M_n + \frac{N-1}{\eta} \log \frac{1}{\alpha}$$

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

donc  $\varphi$  atteint son minimum en  $\alpha^* = \frac{N-1}{M_n \eta}$ .

mais attention! cette valeur de  $\alpha$  n'est pas toujours  $< 1$ !

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

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

Ainsi,  $\varphi_{M, N, \eta} = \frac{N-1}{\eta} + \frac{N-1}{\eta} \left( \log \frac{M_n \eta}{N-1} \right)^+$ , ce qui donne la borne proposée.

MÉTHODE 2:

EXPONENTIELLE DES GRADIENTS.

Rappel (inégalité des pentes en dimension  $d$ ):

$\mathcal{D}$  un domaine convexe de  $\mathbb{R}^d$ ,  $f: \mathcal{D} \rightarrow \mathbb{R}$  convexe et différentiable,

alors  $\forall \underline{u}, \underline{v} \in \mathcal{D}$ ,  $f(\underline{u}) - f(\underline{v}) \leq \nabla f(\underline{u}) \cdot (\underline{u} - \underline{v})$

(où  $\cdot$  désigne le produit scalaire euclidien et  $\nabla f$  la différentielle de  $f$ ).

Remarque de culture mathématique: \* Si  $f: \mathcal{D} \rightarrow \mathbb{R}$  est convexe, alors

sur  $\mathcal{D}$ ,  $f$  est continue et l'ensemble  $\mathcal{G}(\underline{u})$  des vecteurs  $\underline{g}$

tels que  $\forall \underline{v} \in \mathcal{D}$ ,  $f(\underline{u}) - f(\underline{v}) \leq \underline{g} \cdot (\underline{u} - \underline{v})$  est non-vide.

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

\* Dans ce qui suit, on travaille effectivement sur un domaine

l'intérieur  $\rightarrow$  ouvert, on n'aura donc pas besoin de la différentiabilité, même si dans

des exemples que nous avons en tête, les fonctions de perte le sont sur l'intérieur de leur domaine de définition.

\* Lorsque  $f$  est différentiable en  $\underline{u}$ ,

$$\mathcal{G}(\underline{u}) = \{ \nabla f(\underline{u}) \}.$$

Application: Majoration quasi-linéaire du regret: Pour des  $\ell(\cdot; y)$  convexes,

$$R_n = \hat{L}_n - \min_{q \in \mathcal{P}} L_n(q) = \max_{q \in \mathcal{P}} \sum_{t=1}^n \ell(\sum_{k=1}^N \mu_k f_{kt}, y_t) - \ell(\sum_{k=1}^N q_k f_{kt}, y_t)$$

(où  $g(\mu_t)$  sous-gradient de la fonction convexe  $q \mapsto \ell(\sum_{k=1}^N q_k f_{kt}, y_t)$ )

$$\leq \max_{q \in \mathcal{P}} \sum_{t=1}^n g(\mu_t) \cdot (\mu_t - q)$$

en  $\mu_t \in \mathcal{J}_j$ ,  $\mu_t \in \mathcal{J}$  étant assuré

$$= \max_{j=1, \dots, N} \sum_{t=1}^n g(\mu_t) \cdot (\mu_t - \delta_j)$$

par l'algorithme EG décrit à la page suivante)

$$= \tilde{R}_n \stackrel{\text{def.}}{=} \sum_{t=1}^n \sum_{k=1}^N \mu_{kt} \tilde{\ell}_{kt} = \min_{j=1, \dots, N} \sum_{t=1}^n \tilde{\ell}_{jt}$$

(où les  $\delta_j$  sont les Dirac en  $j$ )

(en notant  $g(\mu_t) = (\tilde{\ell}_{kt})_{k=1, \dots, N}$ ).

On considère l'algorithme de pondération exponentielle sur les sous-gradients appelé EG ("exponentiated gradient") et pour  $t \geq 2$ ,  $\mu_t$  définie par

$$\mu_t^j = \frac{\exp(-\eta \sum_{s=1}^{t-1} (g_s(\mu_t))_j)}{\sum_{k=1}^N \exp(-\eta \sum_{s=1}^{t-1} (g_s(\mu_t))_k)}$$

où  $(g_t(\mu_t))_k$  désigne la  $k$ -ième composante de  $g_t(\mu_t)$ .

pseudo-perte de  $j$

Les résultats des cours précédents entraînent (attention au fait que l'étendue des  $\mathcal{I}_{k,t}$  est  $[-M, M]$ ) :

Théorème :

On note  $M$  une borne sur la norme  $\|\cdot\|_\infty$  des sous-gradients.

Alors pour le choix  $\eta = \frac{1}{2M} \sqrt{\frac{8 \ln N}{m}}$ ,

on a : 
$$\sup R_n^{\text{Cvx}} = \sup \hat{L}_n - \min_{g \in \mathcal{P}} L_n(g) \leq M \sqrt{2n \ln N}.$$

Rq: - En choisissant adaptativement des  $\eta_t$ , on peut évidemment se passer de la connaissance de  $M$  et  $n$ .

- Au 1<sup>er</sup> cours, on a traité le cas des pertes  $\geq 0$ , mais cette hypothèse n'est pas cruciale (notamment : lorsque  $M$  est connu, on peut translater les pseudo-pertes.)

Exemples

\* Perte quadratique (prévisions statistiques : régression séquentielle).

Les prédictions des experts sont formées dans l'intervalle  $X = [a, b]$ , et tant qu'à faire, on suppose que notre connaissance du problème est suffisamment bonne pour que l'on sache que  $Y = [a, b]$  également.

Les pertes sont

$$l_t(\mu_t) \stackrel{\text{def.}}{=} l\left(\sum_{k=1}^N \mu_{k,t} f_{k,t}, y_t\right) = \left(\sum_{k=1}^N \mu_{k,t} f_{k,t} - y_t\right)^2.$$

On a que les  $l_t$  sont convexes différentiables, avec

$$(g_t(\mu_t))_j = (\nabla l_t(\mu_t))_j = 2 \left( \underbrace{\sum_{k=1}^N \mu_{k,t} f_{k,t} - y_t}_{\in [-B, B]} \right) f_{j,t}$$

soit  $M = 2B^2$ .

Application pratique à

- la prédiction de pics d'ozone
- la prédiction de consommation électrique

\* Perte logarithmique (investissement boursier).

Les experts sont p.ex.  $N$  valeurs boursières, on note  $x_{jt}$  le facteur multiplicatif par lequel évolue  $j$  au jour  $t$  (si  $j$  valait 1 € à l'ouverture au jour  $t$ , elle en vaut  $x_{jt}$  à la fermeture), et  $x_t = (x_{jt})_j$ .

Une action au jour  $t$  pour le statisticien est de choisir un portefeuille  $\mu_t \in \mathcal{P}$  qui indique comment diviser son capital entre les  $N$  valeurs.   
 [  $\hookrightarrow$  Et l'expert  $j$  est la  $j$ ème valeur boursière ]

On prend: 
$$l_t(\mu_t) = -\log(\mu_t \cdot x_t) = -\log \sum_{j=1}^N \mu_{jt} x_{jt}$$

En effet, avoir de bons résultats, c'est assurer que le gain final

$$\hat{G}_n = \prod_{t=1}^n \mu_t \cdot x_t$$

est grand, soit que  
est petit.

$$\hat{L}_n = -\sum_{t=1}^n \log \mu_t \cdot x_t$$

$R_n$ : Classe optimale pour un marché iid  
 $\downarrow$   
...

La  $(-\log)$ -fortune obtenue par les éléments de la classe de comparaison est :

$$G_n(q) = \prod_{t=1}^n q_t \cdot x_t$$

$$\text{et } L_n(q) = -\log G_n(q)$$

$\uparrow$  Chaque jour, on achète et vend pour retomber sur une distribution  $q$  donnée.

On peut assurer que 
$$R_n^{cvx} = \hat{L}_n - \min_{q \in \mathcal{P}} L_n(q) = \log \frac{\max_{q \in \mathcal{P}} G_n(q)}{\hat{G}_n}$$

soit 
$$\hat{G}_n = e^{-R_n^{cvx}} \max_{q \in \mathcal{P}} G_n(q)$$

où certes,  $R_n^{cvx} = o(n)$  mais  $e^{-R_n^{cvx}}$  peut être très petit !

Compléments :

→ "Optimalité" de la classe de compensation :

Lm: si  $X_1, \dots, X_n$  iid

alors

$$\max_{\varphi} E[\log G_n(\varphi)] = \max_{\text{stratégies}} E[\log \hat{G}_n] \quad (\text{preuve omise})$$

→ Algorithme par mélange uniforme (unif(1)) introduit par Cover '91.

En fait on peut améliorer dans ce cas la borne sup générale, il se trouve que l'on n'a pas besoin d'hypothèse de marché borné.

→ Algorithme EG par exponentielle des gradients :

Nécessite, lui, des hypothèses de marché borné, au moins dans sa version la plus simple :

$$0 < b < B \quad \text{tels que} \quad x_{jt} \in [b, B] \quad \forall j, t$$

alors

$$(\nabla_{\mu} \ell(\mu))_j = -x_{jt} / \mu \cdot x_t \in [-B/b, 0]$$

et  $M = B/b$  soit un regret de l'ordre de  $\frac{B}{b} \sqrt{m \ln N}$

→ Rq: on a négligé tout frais de transactions (achat/vente) ici.

## INÉGALITÉS ORACLE

Code: Soit  $(Y_1, Y_2, \dots)$  une suite iid de variables aléatoires, à valeurs dans  $\mathcal{Y}$ , et  $Q: \mathcal{P} \times \mathcal{Y} \rightarrow \mathbb{R}$  une fonction de perte bornée (où  $\mathcal{P}$  désigne le simplexe de  $\mathbb{R}^N$ ).

But: Etant données  $Y_1, \dots, Y_n$ , trouver un estimateur  $\hat{\theta}_n = \hat{\theta}_n(Y_1, \dots, Y_n)$  tel que

$$E[Q(\hat{\theta}_n, Y)] \leq \inf_{\theta \in \mathcal{P}} E[Q(\theta, Y)] + \Delta_n$$

où  $Y$  est indépendant des  $Y_t$

et distribué selon la même loi, l'espérance  $E$  est par rapport à  $Y$  et aux  $Y_t$ , et  $\Delta_n = o(1)$ .

Méthode: Faire faussement comme si les  $Y_t$  n'étaient disponibles que séquentiellement, construire des

$$\tilde{\theta}_t = \tilde{\theta}_t(Y_1, \dots, Y_{t-1})$$

et considérer

$$\hat{\theta}_n = \frac{1}{n} \sum_{t=1}^n \tilde{\theta}_t$$

(Hyp:  $Q$  convexe en son premier argument.)

P. ex.:  $\tilde{\theta}_t = (\tilde{\theta}_{jt})_{j=1, \dots, N}$  définis par pondération exponentielle des gradients

$$\tilde{\theta}_{jt} = \frac{\exp(-\eta \sum_{s=1}^{t-1} (\nabla Q(\tilde{\theta}_s, Y_s))_j)}{\sum_{k=1, \dots, N} \exp(-\eta \sum_{s=1}^{t-1} (\nabla Q(\tilde{\theta}_s, Y_s))_k)}$$

(Correspond à des experts Dirac en la  $j$ :  $\delta_j$ )

assure que, pour  $\eta$  bien choisi (ce qui est facile, parce qu'en fait on travaille en réalité off-line!)

$$\stackrel{\text{ps}}{\neq} R_n = \sum_{t=1}^n Q(\tilde{\theta}_t, Y_t) - \min_{\theta \in \mathcal{P}} \sum_{t=1}^n Q(\theta, Y_t) \leq M \sqrt{2n \ln N}$$

$$\text{où } M = \sup_{y \in \mathcal{Y}} \|\nabla Q(\cdot, y)\|_{\infty}$$

L'inégalité valant ps, on peut en prendre l'espérance :

On utilise que  $(\tilde{\theta}_t, \mathcal{Y}_t) \stackrel{(d)}{=} (\tilde{\theta}_t, Y)$

pour écrire  $E[Q(\tilde{\theta}_t, \mathcal{Y}_t)] = E[Q(\tilde{\theta}_t, Y)]$

$$\begin{aligned} \text{puis } E[R_n] &= E\left[\sum_{t=1}^n Q(\tilde{\theta}_t, \mathcal{Y}_t)\right] - \underbrace{E\left[\inf_{\theta \in \mathcal{P}} \sum_{t=1}^n Q(\theta, \mathcal{Y}_t)\right]}_{\substack{\uparrow \\ \text{Jensen.}}} \\ &\geq n \left( E\left[\underbrace{\frac{1}{n} \sum_{t=1}^n Q(\tilde{\theta}_t, Y)}_{\substack{\uparrow \\ \text{Jensen.}}} \right] - \inf_{\theta \in \mathcal{P}} E[Q(\theta, Y)] \right) \\ &\geq n \left( E[Q(\hat{\theta}_n, Y)] - \inf_{\theta \in \mathcal{P}} E[Q(\theta, Y)] \right). \end{aligned}$$

On a ainsi :

$$\begin{aligned} E[Q(\hat{\theta}_n, Y)] &\leq \inf_{\theta \in \mathcal{P}} E[Q(\theta, Y)] + \frac{1}{n} E[R_n] \\ &\leq \inf_{\theta \in \mathcal{P}} E[Q(\theta, Y)] + \underbrace{M \sqrt{\frac{2}{n} \ln N}}_{\hat{=} \Delta_n = o(1)} \end{aligned}$$

### Conclusion:

Nos bornes déterministes valent en particulier à des bornes stochastiques!  
Ce n'est évidemment pas une surprise...

PREDICTION RANDOMISÉE.

← Que faire si  $X$  n'est pas convexe ?

Ex: Déplacement dans  $Z^2$ ,  $X = \{ \text{haut, bas, droite, gauche} \}$   
 il faut vraiment choisir une direction, on ne peut pas agréger!

Solution: Randomiser permet de convexifier (et même, linéariser):  
 on choisit une probabilité  $\mu_t$  sur  $\{1, \dots, N\}$  puis  
 on tire l'indice  $I_t$  d'un expert selon  $\mu_t$ , pour  
 au final prédire  $\hat{f}_t = f_{I_t, t}$ .

Attention: Si le statisticien a une stratégie aléatoire, alors les observations  $y_t$  et les prédictions  $\hat{f}_t$  sont elles aussi aléatoires en général (dans le cas d'un jeu contre le diable).

On va considérer la version suivante ;  $X, Y$  et  $l$  sont arbitraires.

Déroulement: A chaque tour  $t = 1, 2, \dots$

- 1) L'environnement choisit les prédictions  $f_{1t}, \dots, f_{Nt} \in X$ ;
- 2) Le statisticien forme une probabilité  $\mu_t$  sur  $\{1, \dots, N\}$ ;
- 3) L'environnement (pouvant observer  $\mu_t$ ) choisit  $y_t \in Y$ ;
- 4) Le statisticien tire  $I_t \sim \mu_t$  et prédit  $\hat{f}_t = f_{I_t, t}$ ;
- 5) Les deux observent  $I_t, \mu_t$  et  $y_t$ .

↳ de pouvoir "remettre" la précision entre plusieurs experts

Note: L'environnement est ici bien plus puissant qu'avant, puisqu'il peut observer (part 3)  $\mu_t$  avant de choisir  $y_t$ . Il peut donc lire dans les pensées du statisticien et connaître sa stratégie.  
 Le fait de s'en remettre à une randomisation externe (part 4) sauve le statisticien. Le fait que l'environnement puisse lire les pensées du statisticien ne doit pas vers surprendre. Le regret

est en effet évalué dans le cas le pire.

But: Le regret est toujours défini par

$$R_n = \sum_{t=1}^n \ell(\hat{p}_t, y_t) - \min_{j=1, \dots, N} \sum_{t=1}^n \ell(f_{j,t}, y_t) = \sum_{t=1}^n \ell(f_{I_t, t}, y_t) - \min_{j=1, \dots, N} \sum_{t=1}^n \ell(f_{j,t}, y_t)$$

et on veut: Pour toute stratégie du diable (ou pour toute suite  $y_1, \dots, y_n$  contre la nature)  $\lim_m \frac{R_n}{m} \leq 0$  ps  $\uparrow$  par rapport aux randomisations auxiliaires du point 4).  
 $\hookrightarrow \bar{\alpha}$ , il n'y aura plus vraiment d'uniformité possible (ou alors, en un sens plus faible).

Méthode:

$$R_n = \bar{R}_n + \Delta_n$$

$$\text{ou } \bar{R}_n = \sum_{t=1}^n \ell_t(\bar{I}_t) - \min_{j=1, \dots, N} \sum_{t=1}^n \ell_t(j)$$

$$\text{et } \Delta_n = \sum_{t=1}^n \ell_t(I_t) - \sum_{t=1}^n \ell_t(\bar{I}_t)$$

ou  $\ell_t: j \in \{1, \dots, N\} \mapsto \ell(f_{j,t}, y_t)$

est étendue linéairement au simplexe:

$$j \in \mathcal{P}(\{1, \dots, N\}) \mapsto \sum_{j=1}^N \mu_j \ell(f_{j,t}, y_t).$$

L'algorithme des poids exponentiels s'applique et donne

$$\sup \bar{R}_n \leq M \sqrt{\frac{n}{2} \ln N}$$

si  $\ell$  est bornée dans  $[0, M]$  et  $n$  est connue.

Comment contrôler  $\Delta_n$  ?

Lemme [inégalité de Hoeffding-Azuma]: Si  $X_1, \dots, X_n$  sont  $n$  variables aléatoires  $(\mathcal{F}_n)$ -adaptées telles que  $a_j \leq X_j \leq b_j$  ps pour des réels  $a_1, \dots, a_n$  et  $b_1, \dots, b_n$ , alors

$$\mathbb{P} \left\{ (X_1 + \dots + X_n) - (E_0(X_1) + E_1(X_2) + \dots + E_{n-1}(X_n)) \geq \varepsilon \right\} \leq \exp \left( - \frac{2\varepsilon^2}{\sum_{t=1, \dots, n} (b_t - a_t)^2} \right)$$

où  $E_s = E[\cdot | \mathcal{F}_s]$  pour  $s \geq 1$  et  $E_0 = E$ .

De manière équivalente, avec probabilité au moins  $1-\delta$ ,

$$X_1 + \dots + X_n \leq E_0[X_1] + \dots + E_{t-1}[X_t] + \sqrt{\frac{\sum_{t=1}^n (b_t - a_t)^2}{2} \ln \frac{1}{\delta}}$$

Remarques: \* Evidemment, ce résultat en entraîne un plus fort:

il suffit d'avoir les encadrements

$$a'_j + E_{j-1}[X_j] \leq X_j \leq E_{j-1}[X_j] + b'_j$$

et en général, on le obtient avec des constantes  $a'_j, b'_j$  telles que  $b'_j - a'_j \leq b_j - a_j$ .

\* L'inégalité de Hoeffding est le cas particulier où les variables  $X_j$  sont indépendantes.

Preuve de l'inégalité de Hoeffding-Azuma:

Similaire à celle d'Hoeffding.

Le lemme de Hoeffding admet une version conditionnelle:

$$\ln E[e^{sX_j} | \mathcal{F}_{j-1}] \leq s E[X_j | \mathcal{F}_{j-1}] + \frac{s^2}{8} (b_j - a_j)^2$$

(voir des éléments de preuve plus bas)

On étudie  $\Psi_{X_1, \dots, X_n}(s) = E\left[\exp\left(s(X_1 + \dots + X_n - \sum_{t=1}^n E_{t-1}[X_t])\right)\right]$

$$E = E[E[\cdot | \mathcal{F}_{n-1}]] \rightarrow = E\left[\exp\left(s(X_1 + \dots + X_{n-1} - \sum_{t=1}^{n-1} E_{t-1}[X_t])\right) \times e^{-sE_{n-1}[X_n]} \times E[e^{sX_n} | \mathcal{F}_{n-1}]\right]$$

$$\leq \text{lemme de H.} \quad \Psi_{X_1, \dots, X_{n-1}}(s) \times e^{\frac{s^2}{8}(b_n - a_n)^2}$$

$$\leq \dots \leq \exp\left(\frac{s^2}{8} \sum_{t=1}^n (b_t - a_t)^2\right)$$

en itérant

On applique ensuite une borne de Chernoff

$$\mathbb{P}\left\{ X_1 + \dots + X_n - \sum_{t=1}^n \mathbb{E}_{t-1}(X_t) \geq \varepsilon \right\}$$

$$= \mathbb{P}\left\{ \exp\left(s\left(X_1 + \dots + X_n - \sum_{t=1}^n \mathbb{E}_{t-1}(X_t)\right)\right) \geq e^{s\varepsilon} \right\}$$

$$\stackrel{\text{(Markov)}}{\leq} e^{-s\varepsilon} \mathbb{E}\left[\exp\left(s(\cdot)\right)\right] = e^{-s\varepsilon} \varphi_{X_1, \dots, X_n}(s)$$

$$\leq \exp\left(-s\varepsilon + \frac{s^2}{8} \sum_{t=1}^n (b_t - a_t)^2\right)$$

$$\stackrel{\text{choix (optimal) de } s}{=} \exp\left(-\frac{2\varepsilon^2}{\sum_t (b_t - a_t)^2}\right)$$

$$s = \frac{4\varepsilon}{\sum_t (b_t - a_t)^2}$$

Application (revenons à nos montons) :

$$R_n \leq \bar{R}_n + \Delta_n$$

$$\leq M \sqrt{\frac{n}{2} \ln N} + \Delta_n$$

$$\text{ou } \Delta_n = \sum_{t=1}^n \ell(\mathbb{F}_{t+1}, y_t) - \sum_{t=1}^n \sum_{k=1}^N \mu_{kt} \ell(\mathbb{F}_{kt}, y_t)$$

On prend  $\mathcal{F}_0 = \{\emptyset, \Omega\}$

et  $\mathcal{F}_t = \sigma(I_1, \dots, I_t)$ ;

alors si  $X_t = \ell(\mathbb{F}_{t+1}, y_t)$ , on a bien que  $(X_t)$  est  $(\mathcal{F}_t)$ -adaptée;

$$\text{de plus, } \mathbb{E}[X_t | \mathcal{F}_{t-1}] = \sum_{k=1}^N \mu_{kt} \ell(\mathbb{F}_{kt}, y_t)$$

↑  
l'espérance conditionnelle  
fixe  $\mu$  et  $y_t$ , seul le  
tirage  $I_t \sim \mu_t$  est aléatoire

Ainsi, par l'inégalité d'Hoeffding,

Avec probab.  $1-\delta$

$$\Delta_n \leq M \sqrt{\frac{n}{2} \ln \frac{1}{\delta}}$$

(par rapport à la randomisation)

Conclusion : Avec proba.  $1-\delta$

$$R_n \leq M \sqrt{\frac{n}{2}} \left( \sqrt{\ln N} + \sqrt{\ln \frac{1}{\delta}} \right) \\ \leq M \sqrt{m \ln \frac{N}{\delta}}$$

Par Borel-Cantelli (avec  $\delta = \delta_n = \frac{1}{n^2}$  p.ex.) :

Pour toute stratégie de l'adversaire (pour toute suite individuelle)

$$\lim_{n \rightarrow +\infty} \frac{R_n}{M \sqrt{2m \ln n}} \leq 1 \quad \text{ps}$$

(soit aussi :  $\lim_{n \rightarrow +\infty} \frac{R_n}{n} \leq 0$  ps.)

Remarque : En utilisant une version maximale de Hoeffding-Azuma :

$$\begin{cases} \text{Avec proba } 1-\delta \\ \max_{t \leq n} \{ X_1 + \dots + X_t \} \leq M \sqrt{\frac{n}{2} \ln \frac{1}{\delta}} \end{cases}$$

(preuve : appliquer l'inégalité de Doob au moment de la majoration par Chernoff)

et en considérant des régimes constitués des temps  $[2^r - 1, \dots, 2^{r+1}]$  pour  $r = 1, 2, \dots$

et en prenant  $\delta_r = \frac{1}{r^2}$

il vient :

Pour toute stratégie de l'adversaire,

$$\lim_{n \rightarrow +\infty} \frac{R_n}{M \sqrt{m \ln \ln n}} < 1 \quad \text{ps}$$

↳ Un résultat à rapprocher de la bi du logarithme itéré.

Preuve(s) du lemme de Hoeffding conditionnel.

Énoncé:  $X$  une variable aléatoire tq.  $a \leq X \leq b$  ps.  
 Alors, pour toute tribu  $\mathcal{F}$ , pour tout  $s \in \mathbb{R}$ ,

$$\ln \mathbb{E}[e^{sX} | \mathcal{F}] \leq s \mathbb{E}[X | \mathcal{F}] + \frac{s^2}{8} (b-a)^2.$$

Rappel: [Lemme de Hoeffding inconditionnel.]

$X$  une variable aléatoire tq.  $a \leq X \leq b$  ps.

Alors, pour tout  $s \in \mathbb{R}$ ,

$$\ln \mathbb{E}[e^{sX}] \leq s \mathbb{E}[X] + \frac{s^2}{8} (b-a)^2.$$

Premier type de preuve: en reprenant la preuve du cas inconditionnel.

- Cela ne pose pas de problème d'étendre celle reposant sur la convexité d'exponentielle, puisqu'elle commence par une inégalité ps, qu'il suffit d'intégrer via  $\mathbb{E}[\cdot | \mathcal{F}]$ .

- C'est moins clair pour la preuve par changement de probabilité de la semaine dernière, plus élégante en version inconditionnelle mais pour qui la version conditionnelle mettrait en jeu un changement de lois conditionnelles par rapport à  $\mathcal{F}$ ...

Autre solution: renforcement "automatique" du lemme de Hoeffding

On exploite le fait qu'il vaut pour toute bi.

En particulier, pour tout  $A \in \mathcal{F}$ , notant  $\mathbb{E}_A = \mathbb{E}[\cdot | A]$  l'espérance sachant  $A$  lorsque  $P(A) > 0$ , on a:

$$\forall A \in \mathcal{F} \text{ tq. } P(A) > 0, \quad \ln \mathbb{E}_A[e^{sX}] \leq s \mathbb{E}_A[X] + \frac{s^2}{8} (b-a)^2$$

On écrit cela parce que l'on se souvient que  $E[X|F]$  est caractérisé par le fait que

$$\forall A \in \mathcal{F}, \quad E[X \mathbb{1}_A] = E[E[X|F] \mathbb{1}_A]$$

(en particulier, lorsque  $P(A) > 0$ , on a :

$$E_A[X] = E_A[E[X|F]].$$

Puis on introduit  $A_r = \left\{ E[e^{sX} | F] - e^{s E[X|F]} \cdot e^{s^2/8(b-a)^2} \geq r \right\}$   
pour  $r \in \mathbb{Q}^{+*}$

et il s'agit de montrer que  $P(A_r) = 0 \quad \forall r \in \mathbb{Q}^{+*}$

Si ce n'était pas le cas, on aurait  $P(A_r) > 0$

puis :

$$E_{A_r}[e^{sX}] = E_{A_r}[E[e^{sX} | F]]$$

$$\geq r + E_{A_r}[e^{s E[X|F]}] \cdot e^{s^2/8(b-a)^2} \quad \left. \begin{array}{l} \text{par def.} \\ \text{de } A_r \end{array} \right\}$$

$$\stackrel{[\text{Jensen}]}{\geq} r + e^{s E_{A_r}[E[X|F]]} e^{s^2/8(b-a)^2}$$

$$= r + e^{s E_{A_r}[X]} \cdot e^{s^2/8(b-a)^2}$$

ce qui contredirait Hoeffding inconditionnel appliqué à  $E_{A_r}[\cdot]$ .

Cela conclut que  $P(A_r) = 0 \quad \forall r \in \mathbb{Q}^{+*}$ , et donne le résultat attendu.