Introduction à la prédiction de suites individuelles :

(1) Inégalités d'oracle

(2) Prédiction avec information imparfaite.

Par : Gilles Stoltz (CNRS - ENS - HEC)

Au : Groupe de travail statistiques, Chevaleret

Les : 12 et 26 novembre 2007.

---

0. Cadre (suites individuelles, information parfaite).

$\mathcal{D}$ un ensemble convexe = l'espace des décisions du statisticien

égal au simplexe $\mathcal{X}$ de $\mathbb{R}^N$ dans les cas les plus simples

Jeu de prédiction : Pour tout $t = 1, 2, \dots$,

- le statisticien choisit $\hat{p}_t \in \mathcal{D}$ (en fonction des infos des tours passés)
- la nature (le diable) choisit $\ell_t : \mathcal{D} \to \mathbb{R}$, fonction de perte du tour t (en fonction des infos des tours passés et même en fonction de $\hat{p}_t$)
- le statisticien et la nature observent $\hat{p}_t$ et ℓ_t ; le statisticien subit une perte $\ell_t(\hat{p}_t)$.

Ex 1 : [Agrégation séquentielle de prédicteurs pour la prévision des pics d'ozone]

Au début du tour t, le statisticien dispose de $f_{1t}^s, \dots f_{Nt}^s \in [0,B]$ les prédictions de N experts pour le jour t à la station s ($s \in \mathcal{Y}$).

Il choisit $\hat{p}_t \in \mathcal{X}$ (ou $\hat{p}_t \in \mathbb{R}^N$, selon qu'on autorise $\mathcal{D} = \mathbb{R}^N$ pour éventuellement débiaiser, ou qu'on se restreint à $\mathcal{D} = \mathcal{X}$) et prédit, au point s, $\hat{y}_t^s = \hat{p}_t \cdot (f_{jt}^s)_{j=1,\dots N} = \sum_{j=1}^{N} \hat{p}_{jt} f_{jt}^s$, qu'on compare au vrai pic y_t^s observé en fin de journée.

Au total,

$$\ell_t(\hat{p}_t) = \sum_{s\in\mathcal{S}} \left(\hat{p}_t \cdot (f_{jt}^s)_{j=1\dots N} - y_t^s \right)^2.$$

Note : pour les besoins d'interprétabilité, on prend le même $\hat{p}_t$ en toutes les stations $s\in\mathcal{S}$.

Cf. stage M2 Boris Mauricette pour les performances pratiques.

Ici, les experts sont N différents modèles physico-chimiques & numériques de propagation des polluants (& statistiques, si besoin).

Df. [regret] $$R_n = \sum_{t=1}^n \ell_t(\hat{p}_t) - \inf_{p\in\mathcal{X}} \sum_{t=1}^n \ell_t(p)$$

On veut $R_n = o(n)$, soit $\lim R_n/n = 0$;

cf. si les ℓ_t sont uniformément bornées, $R_n = O(n)$ est acquis.

Ex 1 [suite] : Alors, la RMSE du statisticien est proche de la RMSE de la meilleure combinaison convexe / linéaire constante.

Ex 2 : [Quand on doit choisir 1 action ou 1 expert parmi un nombre fini d'entre eux, et que l'on ne peut agréger] $\rightarrow$ déplacement dans un graphe.

Randomisation auxiliaire

$\{1,2,\dots N\}$ les N actions possibles ou $f_{1t},\dots f_{Nt}$ les N avis d'experts

$\mathcal{D} = \mathcal{X}$, on choisit $\hat{p}_t \in \mathcal{X}$ et on tire $I_t \sim \hat{p}_t$; on joue alors $I_t \in \{1,2,\dots N\}$ ou prédit $f_{I_t,t}$.

La perte est $\ell_t(I_t)$ ou $\ell_t(f_{I_t,t}) = \ell'_t(I_t)$,

d'espérance conditionnelle (par rapport au choix aléatoire de I_t selon $\hat{p}_t$) :

$$E_t\left[\ell_t(I_t)\right] = \sum_{j=1}^N \hat{p}_{jt}\, \ell_t(j)$$

Note : le diable peut observer $\hat{p}_t$ mais pas I_t lorsqu'il choisit ℓ_t.

Regret en espérance conditionnelle :

$$\overline{R}_n = \sum_{t=1}^n \underbrace{\sum_{j=1}^N \hat{p}_{jt}\, \ell_t(j)}_{\text{noté } \ell_t(\hat{p}_t)} - \underbrace{\inf_{p\in\mathcal{X}} \sum_{t=1}^n \sum_{j=1}^N p_j\, \ell_t(j)}_{= \min_{i=1,\dots N} \sum_{t=1}^n \ell_t(i)}$$

Est proche (inégalité de Hoeffding - Azuma) du regret inconditionnel

$$R_n = \sum_{t=1}^{n} \ell_t(I_t) - \min_{i=1,\dots N} \sum_{t=1}^{n} \ell_t(i)$$

à un facteur $M\sqrt{\frac{n}{2}\ln 1/\delta}$ près, avec proba. $1-\delta$ (où M est l'étendue max. des ℓ_t : $M \geq \max_{t=1,\dots n} \left(\max_j \ell_t(j) - \min_i \ell_t(i)\right)$).

<u>Note</u> : si $\ell_t = \ell(\cdot, y_t)$, les bornes sur $\bar{R}_n$ donnent aussi des bornes sur l'espérance du regret $\mathbb{E}R_n$ (si Y_t aléatoires), voir § sur inégalités oracle.

Lorsque $\bar{R}_n = o(n)$, par H-Az + Borel-Cantelli, on a également $R_n = o(n)$ p.s.

C'est donc pour ces 2 raisons que l'on s'intéressera à $\bar{R}_n$.

1. Premiers résultats (information parfaite, suites individuelles, $\mathcal{D} = \mathcal{X}$ simplexe).

1.1. Cas de l'Ex2 ($\mathcal{D} = \mathcal{X}$ et ℓ_t linéaire $\mathcal{X} \to \mathbb{R}$): <u>LINÉAIRE.</u>

$$\ell_t(p) = \sum_{j=1}^{N} p_j \ell_t(j) \quad \text{où } p \in \mathcal{D} = \mathcal{X}$$

$$M \geq \max_{t=1,\dots n} \left(\max_j \ell_t(j) - \min_i \ell_t(i)\right) \quad (*)$$

mais pas d'hyp. de signe (pour l'instant)

horizon n

1.1.1. Cas où M et n sont connus.

Stratégie : $\hat{p}_1 = (1/N, \dots 1/N)$ probabilité uniforme

et
$$\hat{p}_{j,t} = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \ell_s(j)\right)}{\sum_{i=1,\dots N} \exp\left(-\eta \sum_{s=1}^{t-1} \ell_s(i)\right)} = \hat{p}_{j,t-1} \frac{e^{-\eta \ell_{t-1}(j)}}{\text{normalisation}}$$

implémentation aisée !

$\eta > 0$ donné par le Th. (Sources: Littlestone & Warmuth '94, Vovk '90 puis Cesa-Bianchi et al. '97, etc.)

Th: Pour $\eta = \frac{1}{M}\sqrt{8 \ln N / n}$, on a
$$R_n = \sum_{t=1}^{n} \ell_t(\hat{p}_t) - \min_{i=1,\dots,N} \sum_t \ell_t(i)$$
$$\leq M\sqrt{\frac{n}{2}\ln N}$$

et par conséquent, avec proba $\geq 1-\delta$,
$$R_n \leq M\sqrt{\frac{n}{2}\ln N} + M\sqrt{\frac{n}{2}\ln 1/\delta}.$$

preuve: $W_t = \sum_{i=1} w_{it}$ où $\begin{cases} w_{it} = \exp\left(-\eta \sum_{s=1}^{t} \ell_s(i)\right), & t \geq 1 \\ w_{i0} = 1 \end{cases}$

de sorte que
$$\hat{p}_{jt} = \frac{w_{jt-1}}{W_{t-1}}.$$

D'une part,
$$\ln \frac{W_n}{W_0} = \ln\left(\sum_j \exp\left(-\eta \sum_{s=1}^{n} \ell_s(j)\right)\right) - \ln N$$
$$\geq \ln\left(\max_j \dots\right) - \ln N$$
$$= -\eta \min_{j=1,\dots,N} \sum_{s=1}^{n} \ell_s(j) - \ln N$$

D'autre part, pour tout $t \geq 1$,
$$\ln \frac{W_t}{W_{t-1}} = \ln \sum_{j=1}^{N} w_{jt-1} \cdot e^{-\eta \ell_t(j)} / W_{t-1}$$
$$= \ln \sum_{j=1}^{N} \hat{p}_{jt}\, e^{-\eta \ell_t(j)} = \ln \mathbb{E}\left[e^{-\eta \dots}\right]$$
$$\underset{\text{(lemme Hoeffding)}}{\leq} -\eta\, \mathbb{E}[\dots] + \frac{\eta^2}{8} M^2$$
$$= -\eta \underbrace{\sum_{j=1}^{N} \hat{p}_{jt}\, \ell_t(j)}_{\ell_t(\hat{p}_t)} + \frac{\eta^2}{8} M^2$$

Soit, en sommant sur $t = 1, \dots n$ et en combinant avec la borne inf:
$$-\eta \min_{j=1,\dots,N} \sum_{s=1}^{n} \ell_s(j) - \ln N \leq -\eta \sum_{t=1}^{n} \ell_t(\hat{p}_t) + \frac{\eta^2}{8} M^2$$

soit $$\sum_{t=1}^{n} \ell_t(\hat{p}_t) - \min_{j=1,\dots N} \sum_{t=1}^{n} \ell_t(j) \leq \frac{\ln N}{\eta} + \eta \frac{M^2 n}{8}$$

(forme à retenir)

et il suffit d'optimiser en η.

Rq1 [optimalité minimax]

On a : $$\sup \overline{R}_n \leq M\sqrt{\frac{n}{2}\ln N}$$

sur les stratégies de choix séquentiel des ℓ_t tq. $(*)$

Et on peut montrer (cf. Cesa-Bianchi et al. '97) que

$$\lim_{N\to+\infty} \lim_{n\to+\infty} \left(\inf_{\substack{\text{stratégies} \\ \text{statisticien}}} \sup_{\substack{\text{strat. diable} \\ \text{tq. } (*)}} \overline{R}_n \right) \Big/ \sqrt{\frac{n}{2}\ln N} \geq 1,$$

donc $= 1$.

Eléments de preuve : on le fait pour $\mathbb{E}[\overline{R}_n]$ en considérant un diable qui choisit $\ell_t(j) = j^{\text{ème}}$ composante du développement dyadique de y_t et sélectionne ensuite $y_1, \dots y_n, \dots$ iid $\sim \mathcal{U}_{[0,1]}$. Or,

$$\mathbb{E}[\overline{R}_n] = \underbrace{\mathbb{E}\left[\sum_{t=1}^{n} \ell_t(\hat{p}_t)\right]}_{= n/2} - \mathbb{E}\Big[\min_j \underbrace{\sum_{t=1}^{n} \ell_t(j)}\Big]$$

N v.a. iid. $\sim Bin(n, 1/2)$, donc

$$\mathbb{E}[\ \ ] = n/2 - (1+o(1))\sqrt{\frac{n}{2}\ln N}$$

où $\sqrt{n}$ vient du TCL, et $\sqrt{\ln N}$ du fait que $\mathbb{E}\left[\min_{j=1,\dots N} Z_j\right] \sim -\sqrt{2\ln N}$ lorsque Z_j iid $\sim \mathcal{N}(0,1)$

On fera une autre preuve (par FANO) donnant un résultat non-asymptotique (valant pour tous n, N) dans le cadre plus général de la prédiction économe en observations.

Rq2 : on n'est pas obligé d'utiliser des poids exp., des poids polynômiaux marchent bien également ...

1.1.2. Adaptation en η, M connu.

$$\hat{p}_{jt} = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} \ell_s(j)\right)}{\sum_{i=1}^{N} \exp\left(-\eta_t \sum_{s=1}^{t-1} \ell_s(i)\right)}$$

ie, $\eta = \eta_t$ dépend du passé.

(Première occurrence de ce genre de mise à jour des poids : Auer, Cesa-Bianchi & Gentile '02).

$$\bar{\Phi}(\hat{p}_t, \eta_t, \ell_t)$$

$$= \frac{1}{\eta_t} \ln\left(\sum_{i=1}^{N} \hat{p}_{it}\, e^{\eta_t(\ell_t(\hat{p}_t) - \ell_t(i))}\right)$$

Cesa-Bianchi & Lugosi '06, Cesa-Bianchi, Mansour & Stoltz '05, '07

Lm : Pour toute stratégie de choix des ℓ_t et toute stratégie de choix des η_t tq. $\eta_t \downarrow$, la pondération par poids exponentiels sur les η_t est telle que le regret conditionnel est borné par

$$\bar{R}_n \leq \left(\frac{2}{\eta_{n+1}} - \frac{1}{\eta_1}\right) \ln N + \sum_{t=1}^{n} \bar{\Phi}(\hat{p}_t, \eta_t, \ell_t).$$

(où $\eta_1 \geq \eta_2$ est quelconque).

Note : aucune hypothèse sur le signe des pertes pour l'instant !

Lm : $\bar{\Phi}(\hat{p}_t, \eta_t, \ell_t) \leq \frac{\eta_t}{8} M^2$ (par lemme de Hoeffding).

Th : Pour $\eta_t = \frac{1}{M}\sqrt{\frac{8 \ln N}{t-1}}$ $(t \geq 2)$, on a :

$$\bar{R}_n \leq M\left(2\sqrt{\frac{n}{2}\ln N} + 1\right).$$

1.1.3. Adaptation en η et M ; bornes plus fines ("data - dependent").

Lm : si $\ell_t \geq 0$, alors $\bar{\Phi}(\hat{p}_t, \eta_t, \ell_t) \leq \eta_t \sum_{i=1}^{N} \hat{p}_{it} \frac{\ell_t(i)^2}{2}$.

preuve : $e^{-x} \leq 1 - x + \frac{x^2}{2}$ pour $x \geq 0$, et $\ln(1+u) \leq u$, soit

$$\ln\left(\sum_{i=1}^N \hat{p}_t\, e^{-\eta_t \ell_t(i)}\right) \leq \ln\left(\sum_{i=1}^N \hat{p}_t\left(1-\eta_t \ell_t(i) + (\eta_t \ell_t(i))^2/2\right)\right)$$

$$\leq -\eta_t\, \ell_t(\hat{p}_t) + \eta_t^2 \sum_{i=1}^n \hat{p}_t\, \ell_t(i)^2/2$$

et $\Phi(\quad) = \frac{1}{\eta_t} \ln\left(\sum_{i=1}^N \hat{p}_t\, e^{-\eta_t(\ell_t(i)-\ell_t(\hat{p}_t))}\right) \leq \eta_t \sum_{i=1}^N \hat{p}_t\, (\ell_t(i))^2/2$.

En appliquant un lemme de la page 6: si les $\ell_t \geq 0$,

$$\bar{R}_n \leq \frac{2}{\eta_{n+1}} \ln N + \sum_{t=1}^n \eta_t\, \sigma_t^2$$

où $\sigma_t^2 \triangleq \sum_{i=1}^N \hat{p}_t\, (\ell_t(i))^2/2$

et on a alors le théorème:

<u>Th</u>: Pour $\eta_t = \square \sqrt{\frac{\ln N}{\sum_{s=1}^{t-1} \sigma_s^2}}$, et sous la seule hyp de pertes $\ell_t \geq 0$,

$$\bar{R}_n \leq \square \sqrt{\left(\sum_{t=1}^n \sigma_t^2\right) \ln N} + \dots$$

et par conséquent,

$$\bar{R}_n \leq \square \sqrt{M_n \left(\min_{j=1,\dots,N} \sum_{t=1}^n \ell_t(j)\right) \ln N} + \dots$$

où $M_n = \max_{t \leq n} \max \ell_t$.

<u>Rq1</u>: Cette borne implique en particulier des vitesses de convergence plus rapides que $\sqrt{n}$ (p.ex., quand (M_n) bornée et $\min_{j=1,\dots,N} \sum_{t=1}^n \ell_t(j) \ll n$, quand au moins 1 expert est bon).

<u>preuve</u>: * suffit de voir que $\sum_t \eta_t \sigma_t \leq \square \sqrt{\left(\sum_{t=1}^n \sigma_t^2\right) \ln N}$:

$S_t = \sum_{s=1}^t \sigma_s^2$: $\quad \sum_{t=1}^n \eta_t\, \sigma_t^2 = \square \sum_{t=1}^n \frac{\sqrt{\ln N}}{\sqrt{S_{t-1}}} \left(S_t - S_{t-1}\right)$

$$= \square \sum_{t=1}^n \frac{\sqrt{\ln N}}{\sqrt{S_{t-1}}} \left(\sqrt{S_t} - \sqrt{S_{t-1}}\right)\left(\sqrt{S_t}+\sqrt{S_{t-1}}\right)$$

$$\leq \square \left(\sum_{t=1}^n \sqrt{S_t} - \sqrt{S_{t-1}}\right) \sqrt{\ln N} + \dots = \square \sqrt{S_n \ln N} + \dots$$

$*$ Puis, $\displaystyle \sum_{t=1}^{n} \sigma_t^2 \underset{\substack{\uparrow \\ \text{chaque } \ell_t(i)^2 \le M_n}}{\le} \frac{M_n}{2} \sum_{t=1}^{n} \hat{p}_{it}\, \ell_t(i) = \frac{M_n}{2} \sum_{t=1}^{n} \ell_t(\hat{p}_t)$

Soit
$$\overline{R}_n = \sum_{t=1}^{n} \ell_t(\hat{p}_t) - \min_{j=1,\dots,N} \sum_{t=1}^{n} \ell_t(j)$$
$$\le \square \sqrt{M_n \left(\sum_{t=1}^{n} \ell_t(\hat{p}_t) \right) \ln N}$$

et le résultat en résolvant l'inéquation en
$$\sum_{t=1}^{n} \ell_t(\hat{p}_t).$$

<u>Rq2</u>: Si ℓ_t ont un signe arbitraire, on a le même résultat ou genre de résultats totalement adaptatifs et raffinés $\left(M_n = \max\limits_{t \le n} \max \ell_t - \min \ell_t \right)$ voir Cesa-Bianchi, Mansour & Stoltz '05, '07.

<u>1.2. Cas plus général</u> $(\mathcal{D} = \mathcal{X}$ et ℓ_t convexe $\mathcal{D} \to \mathbb{R})$: <u>CONVEXE</u>.

On peut se ramener au cas du § 1.1 en linéarisant par l'inégalité des pentes :
$$\ell_t(\hat{p}_t) - \ell_t(p) \le \nabla \ell_t(\hat{p}_t) \cdot (\hat{p}_t - p)$$
où $\nabla \ell_t$ est un sous-gradient de ℓ_t au point (intérieur, cf. ci-dessous, mise à jour par poids exponentiels) $\hat{p}_t$.

En particulier,
$$R_n = \sum_{t=1}^{n} \ell_t(\hat{p}_t) - \inf_{p \in \mathcal{X}} \sum_{t=1}^{n} \ell_t(p)$$
$$\le \sup_{p \in \mathcal{X}} \underbrace{\sum_{t=1}^{n} \nabla \ell_t(\hat{p}_t) \cdot (\hat{p}_t - p)}_{\text{linéaire en } p \,:\, \sup \text{ atteint en 1 coin}}$$
$$= \sum_{t=1}^{n} \sum_{j=1}^{N} \hat{p}_{jt}\, \tilde{\ell}_t(j) - \min_{i=1,\dots,N} \sum_{t=1}^{n} \tilde{\ell}_t(i)$$
(où $\tilde{\ell}_t(j) = \left(\nabla \ell_t(\hat{p}_t) \right)_j$, j-ème composante du sous-gradient)
$$= \sum_{t=1}^{n} \tilde{\ell}_t(\hat{p}_t) - \min_{i=1,\dots,N} \sum_{t=1}^{n} \tilde{\ell}_t(i),$$
ce qui est exactement le cadre du § précédent : on utilise des poids exponentiels, $\hat{p}_{jt} > 0 \ \forall j, t$.

On retrouve donc les mêmes bornes (à peu de choses près : p.ex., l'étendue M_n est calculée sur les sous-gradients et pas directement sur les poids).

Réf. Kivinen & Warmuth '97, Cesa-Bianchi '99

+ un peu de travail / simplifications de ma part.

1.3. Autre cas de vitesses rapides ($\Theta = \mathcal{X}$ et $\ell_t : \mathcal{X} \to \mathbb{R}$ exp-concave) :

EXP-CONCAVE.

Déf : [Haussler, Kivinen & Warmuth '95, '98, p.ex. ; mais des références antérieures semblent exister.]

ℓ_t est η-exp concave ($\eta > 0$) si $e^{-\eta \ell_t(\cdot)}$ est concave.

Ex : $\ell_t(p) = \sum_{s \in \mathcal{Y}} \left(p \cdot \left(f_{jt}^s \right)_{j=1,\dots,N} - y_t^s \right)^2$

Ex : $\ell_t(p) = -\log p \cdot r_t$ marché boursier.

sous des hypothèses de bornitude des y_t^s et f_{jt}^s, les ℓ_t sont toutes η_0-exp concaves pour η_0 suffisamment petit (et qu'il n'est pas difficile d'expliciter).

Th : *(1) Si ℓ_t est η_0-exp concave, alors le prédicteur par poids exponentiels utilisant $\eta = \eta_0$ indépendamment de t,

$$\hat{p}_{jt} = \frac{\exp\left(-\eta_0 \sum_{s=1}^{t-1} \ell_s(j)\right)}{\sum_{i=1}^{N} \exp\left(-\eta_0 \sum_{s=1}^{t-1} \ell_s(i)\right)} ,$$

vérifie :

$$\sum_{t=1}^{n} \ell_t(\hat{p}_t) - \min_{j=1,\dots,N} \sum_{t=1}^{n} \ell_t(j) \leq \frac{\ln N}{\eta_0} .$$

*(2) Si ℓ_t est η_0-exp concave, alors le prédicteur effectuant un mélange exponentiel sur le simplexe,

$$\hat{p}_t = \frac{\int_{\mathcal{X}} p \exp\left(-\eta_0 \sum_{s=1}^{t-1} \ell_s(p)\right) d\mu(p)}{\int_{\mathcal{X}} \exp\left(-\eta_0 \sum_{s=1}^{t-1} \ell_s(p)\right) d\mu(p)}$$

où μ est la mesure uniforme sur le simplexe, vérifie :

$$R_n = \sum_{t=1}^{n} \ell_t(\hat{p}_t) - \inf_{p \in \mathcal{X}} \sum_{t=1}^{n} \ell_t(p) \leqslant \square \, N \ln n .$$

Rq : - problème d'implémentation pratique (Cover et co. pour marché boursier)

- réf. : Cover '91 pour (1) et $\ell_t = -\log(\cdot \, r_t)$
  Blum & Kalai '97 pour (2) et $\ell_t = -\log(\cdot \, r_t)$
  Stoltz (thèse) '05 pour (2) en général.

- optimalité : <u>oui</u> au sens des suites individuelles, cf. Ordentlich & Cover '98.

<u>Preuve</u> (juste de (1)) : Par récurrence, pour tout j,

celle de (2) est un peu du même acabit.

$$\sum_{t=1}^{n} \ell_t(\hat{p}_t) - \sum_{t=1}^{n} \ell_t(j)$$

$$= -\frac{1}{\eta_0} \ln \frac{\exp\left(-\eta_0 \sum_{t=1}^{n-1} \ell_t(\hat{p}_t)\right) \cdot \exp\left(-\eta_0 \ell_n(\hat{p}_n)\right)}{\exp\left(-\eta_0 \sum_{t=1}^{n} \ell_t(j)\right)}$$

$$\leqslant -\frac{1}{\eta_0} \ln \frac{\exp\left(-\eta_0 \sum_{t=1}^{n-1} \ell_t(\hat{p}_t)\right) \sum_{i=1}^{N} \hat{p}_{in} \exp\left(-\eta_0 \ell_n(i)\right)}{\sum_{i=1}^{N} \exp\left(-\eta_0 \sum_{t=1}^{n} \ell_t(i)\right)}$$

au num. : exp. conc.
au dénom : majoration par Σ

$$= -\frac{1}{\eta_0} \ln \underbrace{\frac{\exp\left(-\eta_0 \sum_{t=1}^{n-1} \ell_t(\hat{p}_t)\right)}{\sum_{i=1}^{N} \exp\left(-\eta_0 \sum_{t=1}^{n-1} \ell_t(i)\right)}}_{\text{et on continue la réc.}} \cdot \underbrace{\frac{\sum_{i=1}^{N} \exp\left(-\eta_0 \sum_{t=1}^{n-1} \ell_t(i)\right) \sum_{i=1}^{N} \hat{p}_{in} e^{-\eta_0 \ell_n(i)}}{\sum_{i=1}^{N} \exp\left(-\eta_0 \sum_{t=1}^{n} \ell_t(i)\right)}}_{= 1 \text{ par déf. des } \hat{p}_{in}}$$

$$\leqslant \ldots \leqslant -{}^{1}/_{\eta_0} \ln \frac{1}{N} = \frac{\ln N}{\eta_0} .$$

2. Obtention d'inégalités oracle.

Réf: Littlestone & Warmuth '89
Cesa-Bianchi, Conconi & Gentile '04 et '05
Juditzky, Nazin, Tsybakov, Vayatis '05, etc.

Cadre : $Z_1, \dots Z_n$ iid $\sim \nu$ (sur un espace $\mathcal{Z}$ arbitraire)

Ensemble de paramètres $\theta \in \Theta$, disons $\Theta = \mathcal{X}$ (voir §3 pour des Θ plus gros)

Fonction de perte fixée $Q : \Theta \times \mathcal{Z} \to \mathbb{R}$

But : Trouver $\bar{\theta}_n = \bar{\theta}_n(Z_1^n)$ tq.

$$\mathbb{E}_\nu\left[Q(\bar{\theta}_n, Z) \right] \leq \inf_{\theta \in \Theta} \mathbb{E}_\nu\left[Q(\theta, Z) \right] + \Delta_n$$

où $\Delta_n = o(n)$

et $Z \sim \nu$ indép. des Z_1^n (l'espérance $\mathbb{E}_\nu$ est prise par rapport à Z et aux Z_1^n).

Méthode : Faire comme si les Z_1^n étaient disponibles uniquement séquentiellement, $\hat{\theta}_t = \hat{\theta}_t(Z_1^{t-1})$ avec les expressions en termes de poids exp. des pertes cumulées, puis $\bar{\theta}_n = \frac{1}{n}(\hat{\theta}_1 + \dots + \hat{\theta}_n)$.

Ex: $\ell_t = Q(\cdot, Z_t)$ hyp : $Q(\cdot, z)$ convexe sur $\mathcal{X}$ pour tout $z \in \mathcal{Z}$.

$$\hat{\theta}_{j,t} = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} \left(\nabla Q(\hat{\theta}_s, Z_s)\right)_j\right)}{\text{normalisation}}$$

où (η_t) bien choisie de manière adaptative

est une suite tq.

p.s.
$$R_n = \sum_{t=1}^n Q(\hat{\theta}_t, Z_t) - \inf_{\theta \in \mathcal{X}} \sum_{t=1}^n Q(\theta, Z_t)$$
$$\leq \Box \sqrt{\left(\sum_{t=1}^n \hat{\theta}_{it} \left(\nabla Q(\hat{\theta}_t, Z_t)_i\right)^2 \right) \ln N} + \dots$$
$$\leq \Box \sqrt{\left(\sum_{t=1}^n \|\nabla Q(\hat{\theta}_t, Z_t)\|_\infty^2 \right) \ln N} + \dots$$

cf. th. page 7

Cesa-Bianchi, Mansour & Stoltz '05, '07 montrent qu'il est OK même en cas de fonctions de perte à signe quelconque.

L'inégalité valant p.s., on peut l'intégrer : on utilise que $(\hat{\theta}_t, Z_t) \overset{(d)}{=} (\hat{\theta}_t, Z)$

d'où : $\mathbb{E}\left[Q(\hat{\theta}_t, Z_t) \right] = \mathbb{E}\left[Q(\hat{\theta}_t, Z) \right]$

$$\mathbb{E}[R_n] = \mathbb{E}\Big[\sum_t \dots\Big] - \mathbb{E}\Big[\inf_{\theta\in\mathcal{X}} \sum_t \dots\Big]$$

$$\geqslant \mathbb{E}\Big[\sum_{t=1}^{n} Q(\hat\theta_t, Z)\Big] - \inf_{\theta\in\mathcal{X}} \mathbb{E}\Big[\sum_{t=1}^{n} Q(\theta, Z_t)\Big]$$

$$\underset{\substack{Q \text{ convexe,} \\ \text{Jensen}}}{\geqslant} n\Big(\mathbb{E}[Q(\bar\theta_n, Z)] - \inf_{\theta\in\mathcal{X}} \mathbb{E}[Q(\theta, Z)]\Big)$$

et $$\mathbb{E}[R_n] \leqslant \square\, \mathbb{E}\left[\sqrt{\Big(\sum_{t=1}^{n} \|\nabla Q(\hat\theta_t, Z_t)\|_\infty^2\Big)\ln N}\,\right] + \dots$$

$$\underset{\text{Jensen}}{\leqslant} \square \sqrt{\Big(\sum_{t=1}^{n} \underbrace{\mathbb{E}\big[\|\nabla Q(\hat\theta_t, Z_t)\|_\infty^2\big]}\Big)\ln N} + \dots$$

$$= \mathbb{E}\big[\|\nabla Q(\hat\theta_t, Z)\|_\infty^2\big]$$

$$\underset{\substack{\text{passer} \\ \text{par esp.} \\ \text{cond.} \\ /t \text{ à } Z_1^{t-1}}}{\leqslant} \sup_{\theta\in\mathcal{X}} \mathbb{E}\big[\|\nabla Q(\theta, Z)\|_\infty^2\big]$$

Soit au final,

$$\mathbb{E}[Q(\bar\theta_n, Z)] - \inf_{\theta\in\mathcal{X}} \mathbb{E}[Q(\theta, Z)]$$

$$\leqslant \square \sqrt{\frac{\ln N}{n} \sup_{\theta\in\mathcal{X}} \mathbb{E}\big[\|\nabla Q(\theta, Z)\|_\infty^2\big]} + \dots \qquad (\ast\ast\ast)$$

où la borne est obtenue pour une procédure <u>totalement adaptative</u>.
À comparer au Th2 de Juditzky et al. '05 : c'est le même résultat à la constante $\square$ près. Je crois me souvenir qu'elle est meilleure chez Juditzky et al. '05, qui fait une preuve directe (pour le même algo, au choix des η_t près) du résultat final et n'a pas de borne ps. intermédiaire, qui notons-le, est <u>beaucoup</u> plus forte.

<u>Q</u>: Dans Juditzky et al. '05, voir si $(\ast\ast\ast)$ était obtenue de manière totalement adaptative ou pas? (Je ne me souviens plus, poser la question pendant l'exposé.)

3. Brève introduction au cas où $\mathcal{D}$ est plus gros que le simplexe $\mathcal{X}$.

Rq: si perte $(\ )^2$, il existe une technique ad hoc pour être compétitif face à $\mathcal{D} = B_{\|\cdot\|_1}(0, U)$ où $U > 0$ fixé, voir Kivinen & Warmuth '97.

Dans le cas général, on peut s'inspirer de Stoltz & Lugosi '07, Stoltz (thèse '05) ; voir aussi Vovk '07 :

hyp (si on veut des vitesses de convergence uniformes) :

- $\mathcal{D}$ convexe et compact ;
- perte de la forme $Q : \mathcal{D} \times \mathcal{Z} \to \mathbb{R}$ où les $Q(\cdot, z)$ sont convexes et uniformément Lipschitziennes (constante δ).

* À horizon n fixé, étant donné $\varepsilon > 0$ paramètre de couverture :

$d_1, \dots d_{N(\varepsilon)}$ un "ε-covering" de $\mathcal{D}$

ie, tout point de $\mathcal{D}$ est à ε près d'un d_j pour la métrique de $\mathcal{D}$

ainsi,

$$\inf_{d \in \mathcal{D}} \sum_{t=1}^{n} Q(d, z_t) \geq \min_{j=1,\dots N(\varepsilon)} \sum_{t=1}^{n} Q(d_j, z_t) - \delta n \varepsilon$$

il suffit de se comparer à $\min_{j=1,\dots N(\varepsilon)} \sum_{t=1}^{n} Q(d_j, z_t)$, ce que l'on peut faire en employant les techniques du §1 (on se ramène au simplexe de $\mathbb{R}^{N(\varepsilon)}$) :

$$\hat{d}_t = \sum_{j=1\dots N(\varepsilon)} \hat{p}_{jt}\, d_j$$

où $\hat{p}_{jt} = \exp\left(-\eta_t \sum_{s=1}^{t-1} Q(d_j, z_s)\right) / \text{renormalisation}$ de sorte que, pour

$$\sum_{t=1}^{n} Q(\hat{d}_t, z_t) - \min_{i=1,\dots,N(\varepsilon)} \sum_{t=1}^{n} Q(d_i, z_t) \underset{\text{(Jensen)}}{\leq} \sum_{t=1}^{n} \sum_{j=1}^{N(\varepsilon)} \hat{p}_{jt}\, Q(d_j, z_t) - \min_{i=1\dots N(\varepsilon)} \sum_{t=1}^{n} Q(d_i, z_t)$$

$$\leq \|\nabla Q\|_\infty \sqrt{n \ln N(\varepsilon)}$$

soit au final,

$$\sum_{t=1}^{n} Q(\hat{d}_t, z_t) - \inf_{d\in\mathcal{D}} \sum_{t=1}^{n} Q(d, z_t) \leq \underbrace{\|\nabla Q\|_\infty \sqrt{n \ln N(\varepsilon)}}_{\text{terme regret}} + \underbrace{\delta n \varepsilon}_{\text{terme approx.}}$$

* On en déduit une borne uniforme en partitionnant le temps

$1, 2, \ldots n, \ldots$

en régimes $\mathcal{R}_r$, $r = 1, 2, \ldots$. Au début de chaque régime, on fait repartir l'algo précédent avec ε_r et pour une durée L_r.

Soit r_n le régime dans lequel se situe n :

$$R_n = \sum_{t=1}^{n} Q(\hat{d}_t, z_t) - \inf_{d\in\mathcal{D}} \sum_{t=1}^{n} Q(d, z_t)$$

$$\leq \sum_{r=1,\ldots r_n} \underbrace{\sum_{t\in\mathcal{R}_r} Q(\hat{d}_t, z_t) - \inf_{d\in\mathcal{D}} \sum_{t\in\mathcal{R}_r} Q(d, z_t)}_{\leq \|\nabla Q\|_\infty \sqrt{L_r \ln N(\varepsilon_r)} + \delta L_r \varepsilon_r}$$

$$\leq \underbrace{\sum_{r=1,\ldots r_n} \|\nabla Q\|_\infty \sqrt{L_r \ln N(\varepsilon_r)} + \delta L_r \varepsilon_r}.$$

le choix précis de L_r, ε_n etc. est à faire dans chaque cas particulier, voir 1 ex. dans Stoltz & Lugosi '07 ($\mathcal{D}$ = ensemble des fonctions Lipsch. de cste $\leq C$ fixé).

c'est typiquement $o(n)$:

- $\sum L_r = n$ mais comme $\varepsilon_r \to 0$, $\sum L_r \varepsilon_r = o(n)$
- $\sum \sqrt{L_r} \sim \sqrt{n}$ pour $L_r = 2^r$ p.ex.

et donc souvent, $\sum \sqrt{L_r \ln N(\varepsilon_r)} = o(n)$ si $\mathcal{D}$ n'est pas trop massif.

<u>Rq</u> : là, on a uniformité en $d \in \mathcal{D}$. Si l'on veut juste

$\forall d \in \mathcal{D}$, $$\sum_{t=1}^{n} Q(\hat{d}_t, z_t) - \sum_{t=1}^{n} Q(d, z_t) = o(n),$$

alors des hyp. d'absolue continuité de Q et de séparabilité de $\mathcal{D}$ suffisent.

Partie 2 : INFORMATION IMPARFAITE.

0. Résumé de l'épisode précédent (de la partie 1), pour le cas de l'agrégation linéaire.

Le statisticien a un nombre fini d'actions $\{1, \dots N\}$ et il est autorisé à utiliser des actions randomisées : il choisit à chaque tour $\hat{p}_t \in \mathcal{D}$ (qui est le simplexe $\mathcal{X}$ de $\mathbb{R}^N$) puis tire son action $I_t \in \{1, \dots N\}$ selon $\hat{p}_t$.

Jeu de prédiction : Pour tout $t = 1, 2, \dots$

- le statisticien choisit $\hat{p}_t \in \mathcal{X}$ (en fonction des infos des tours passés);
- le diable choisit $\underline{\ell}_t = (\ell_{jt})_{j=1, \dots N} \in \mathbb{R}^N$, vecteur des pertes (en fonction des infos des tours passés et même en fonction de $\hat{p}_t$);
- le statisticien tire $I_t \sim \hat{p}_t$ et subit la perte $\ell_{I_t, t}$;
- le statisticien et le diable observent I_t et $\underline{\ell}_t$.

But : Avoir un regret (inconditionnel) faible :
$$R_n = \sum_{t=1}^{n} \ell_{I_t, t} - \min_{j=1, \dots N} \sum_{t=1}^{n} \ell_{jt}$$
$$= \sum_{t=1}^{n} \ell_{I_t, t} - \min_{p \in \mathcal{X}} \sum_{t=1}^{n} p \cdot \underline{\ell}_t$$

Moyen : Considérer le regret conditionnel : $\overline{R}_n = \sum_{t=1}^{n} \hat{p}_t \cdot \underline{\ell}_t - \min_{p \in \mathcal{X}} \sum_{t=1}^{m} p \cdot \underline{\ell}_t$

En effet : Soit $\mathbb{E}_t$ l'espérance conditionnelle par rapport à l'info disponible au début du tour t (ie, par