Gilles Stoltz

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

Professeur affilié à HEC Paris

Apprentissage statistique

Série de cours sur l'apprentissage séquentiel

Applications à la théorie des jeux répétés

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

Apprentissage séquentiel et théorie des jeux.

Notion de jeu (à somme nulle):

Joueur 1 a un ensemble fini d'actions $A = \{1, \dots N\}$

2 $B = \{1, \dots M\}$

Une matrice de paiements est donnée: $\underline{R} = [R(i,j)]_{(i,j) \in A \times B}$ (comme des joueurs)

Interprétation: Si $(I_t, J_t) \in A \times B$ est joué, alors:

* le joueur 1 obtient le paiement $R(I_t, J_t)$
* 2 $-R(I_t, J_t)$

Objectif: Chaque joueur essaie de maximiser la somme / moyenne de ses paiements

L'objectif de chaque joueur est donc antagoniste de celui de l'autre!

→ Ce cadre correspond au cadre d'un environnement qui réagit.

Déroulements On considérera deux cadres.

* Jeu en un coup ("one-shot"): les joueurs choisissent respectivement (et simultanément) des lois μ et ν sur A et B (on note $\mu \in \Delta(A)$ et $\nu \in \Delta(B)$) et reçoivent les paiements respectifs

$$\mu^T \underline{R} \nu = \sum_{i \in A} \sum_{j \in B} \mu_i \nu_j R(i,j) \quad \text{et} \quad -\mu^T \underline{R} \nu$$

$$= \mathbb{E}[R(I,J)] \quad \text{où} \quad \left.\begin{matrix} I \sim \mu \\ J \sim \nu \end{matrix}\right\} \text{tirés indépendamment}$$

* Jeu répété: à chaque tour $t = 1, 2, \dots$, chaque joueur choisit en secret son action, éventuellement en randomisant ($I_t \sim \mu_t$ et $J_t \sim \nu_t$, respectivement); ce choix d'actions est fondé sur le passé, i.e., sur les actions $(I_s, J_s)_{s \le t-1}$. A la fin du tour t, I_t et J_t sont révélés.

Application des bornes de regret : (Cas du jeu répété)

* Du point de vue du joueur 1, tout se passe comme si le joueur 2 choisissait le vecteur de pertes

$$\ell_t = -R(\cdot, J_t)$$

au tour t et qu'il devait simultanément en choisir une composante $I_t \in A$.
C'est exactement le cadre que nous avons étudié la dernière fois.

On note $[-\|R\|_\infty, \|R\|_\infty]$ l'étendue des éléments de R : on a exhibé une stratégie pour le joueur 1 (pondération par poids exponentiels) telle que :

$\forall n$, avec probabilité $\geq 1-\delta$,

$$\underbrace{\sum_{t=1}^{n} -R(I_t, J_t)}_{\text{perte cumulée joueur 1}} - \min_{i \in A} \underbrace{\sum_{t=1}^{n} R(i, J_t)}_{\text{perte cumulée de la meilleure composante}} \leq 2\|R\|_\infty \left(\underbrace{\sqrt{(n+1)\ln N}}_{\text{vient du contrôle déterministe sur l'esp. conditionnelle}} + \underbrace{\sqrt{\frac{n}{2}\ln\frac{1}{\delta}}}_{\text{cf. inég. Hoeffding-Azuma}} \right)$$

ici est :

$$\frac{1}{n}\sum_{t=1}^{n} R(I_t, J_t) \geq \max_{i \in A} \frac{1}{n}\sum_{t=1}^{n} R(i, J_t) - \frac{2\|R\|_\infty}{n}\left(\sqrt{(n+1)\ln N} + \sqrt{\frac{n}{2}\ln\frac{1}{\delta}}\right)$$

En particulier, nous avions vu qu'une application de Borel-Cantelli menait à :

$$(*) \quad \frac{1}{n}\sum_{t=1}^{n} R(I_t, J_t) \geq \max_{i \in A} \frac{1}{n}\sum_{t=1}^{n} R(i, J_t) - \varepsilon_n$$

où ε_n est une suite de v.a. positives avec $\varepsilon_n \to 0$ ps

* Symétriquement, le joueur 2 dispose d'une stratégie telle que :

$\forall$ stratégie du joueur 1,

$$\frac{1}{n}\sum_{t=1}^{n} \underbrace{(-R(I_t, J_t))}_{\text{paiement moyen du joueur 2}} \geq \max_{j \in B} \frac{1}{n}\sum_{t=1}^{n} \underbrace{(-R(I_t, j))}_{\text{paiement moyen de la meilleure composante}} - \delta_n$$

où $\delta_n \geq 0$ et $\delta_n \to 0$ ps

ici est :

$$(**) \quad \frac{1}{n}\sum_{t=1}^{n} R(I_t, J_t) \leq \min_{j \in B} \frac{1}{n}\sum_{t=1}^{n} R(I_t, j) + \delta_n$$

Notion de valeur du jeu

Lorsque (*) et (**) sont simultanément vraies : notant $\hat{\nu}_n = \frac{1}{n}\sum_{t=1}^{n} \delta_{J_t}$ et $\hat{\mu}_n = \frac{1}{n}\sum_{t=1}^{n} \delta_{I_t}$, il vient :

$$\varliminf_{n\to+\infty} \frac{1}{n}\sum_{t=1}^{n} R(I_t, J_t) \underset{\text{par (*)}}{\geqslant} \varliminf_{n\to+\infty} \max_{i=1\dots N} \frac{1}{n}\sum_{t=1}^{n} R(i, J_t)$$

$$\overset{\text{not.}}{=} \varliminf_{n\to+\infty} \max_{i=1\dots N} \delta_i^T \underline{\underline{R}}\, \hat{\nu}_n$$

$$= \underset{\text{par linéarité}}{} \varliminf_{n\to+\infty} \max_{\mu\in\Delta(A)} \mu \underline{\underline{R}}\, \hat{\nu}_n$$

(inf = min ici car $\nu \mapsto \max_\mu \mu^T \underline{\underline{R}}\, \nu$ est Lipschitzienne et définie sur un compact) $\longrightarrow$

$$\geqslant \inf_{\nu\in\Delta(B)} \max_{\mu\in\Delta(A)} \mu^T \underline{\underline{R}}\, \nu = \min_{\nu\in\Delta(B)} \max_{\mu\in\Delta(A)} \mu^T \underline{\underline{R}}\, \nu$$

et de même, par (**) :

$$\varlimsup_{n\to+\infty} \frac{1}{n}\sum_{t=1}^{n} R(I_t, J_t) \leqslant \max_{\mu\in\Delta(A)} \min_{\nu\in\Delta(B)} \mu^T \underline{\underline{R}}\, \nu$$

Or, comme l'on a toujours $\max_{\mu\in\Delta(A)} \min_{\nu\in\Delta(B)} \mu^T \underline{\underline{R}}\, \nu \leqslant \min_{\nu} \max_{\mu} \mu^T \underline{\underline{R}}\, \nu$

on a donc que (*) et (**) entraînent :

- $\max_\mu \min_\nu \mu^T \underline{\underline{R}}\, \nu = \min_\nu \max_\mu \mu^T \underline{\underline{R}}\, \nu \overset{\text{not.}}{=} v$, la valeur du jeu ; cette égalité est appelée le théorème minmax de von Neumann
- $\frac{1}{n}\sum_{t=1}^{n} R(I_t, J_t)$ admet v pour limite, et par théorème des gendarmes :

(*) indique : $v \leqslant \max_\mu \mu^T \underline{\underline{R}}\, \hat{\nu}_n \leqslant \frac{1}{n}\sum_{t=1}^{n} R(I_t, J_t) + \varepsilon_n \longrightarrow v$

donc également : $(**)$ $\max_\mu \mu^T \underline{\underline{R}}\, \hat{\nu}_m \longrightarrow v$

Symétriquement, par (**), vient : $(**)$ $\min_\nu \hat{\mu}_n \underline{\underline{R}}\, \nu \longrightarrow v$

Interprétation de la valeur d'un jeu

* Jeu en un coup :

- $v = \max_{\mu \in \Delta(A)} \min_{\nu \in \Delta(B)} \mu^T \underline{\underline{R}} \nu$: $\forall \mu \in \Delta(A), \exists \nu \in \Delta(B) \mid \mu^T \underline{\underline{R}} \nu \leq v$

  ↳ le joueur 1 ne peut pas gagner, dans le cas le pire, plus que v ; il peut se garantir v en jouant $\mu^* \in \operatorname{argmax}_{\mu \in \Delta(A)} \min_{\nu \in \Delta(B)} \mu^T \underline{\underline{R}} \nu$

- $v = \min_{\nu \in \Delta(B)} \max_{\mu \in \Delta(A)} \mu^T \underline{\underline{R}} \nu$

  ↳ le joueur 2 ne peut pas gagner, dans le cas le pire, plus que $-v$; il peut se garantir $-v$ en jouant $\nu^* \in \operatorname{argmin}_{\nu \in \Delta(B)} \max_{\mu \in \Delta(A)} \mu^T \underline{\underline{R}} \nu$

* Jeu répété :

- Soit $\nu^* \in \Delta(B)$ telle que $v = \max_{\mu \in \Delta(A)} \mu^T \underline{\underline{R}} \nu^*$; i.e. :

  alors, si le joueur 2 tire ses actions J_t iid $\sim \nu^*$, pour toute stratégie du joueur 1,

  $$\overline{\lim_{n \to +\infty}} \frac{1}{n} \sum_{t=1}^{n} R(I_t, J_t) \overset{\text{Hoeffding-Azuma + Borel-Cantelli}}{=} \overline{\lim_{n \to +\infty}} \frac{1}{n} \sum_{t=1}^{n} \mu_t \underline{\underline{R}} \nu^* \leq \max_{\mu} \mu^T \underline{\underline{R}} \nu^* = v$$

  ou encore $\underline{\lim_{n \to +\infty}} \frac{1}{n} \sum_{t=1}^{n} \left(- R(I_t, J_t) \right) \geq -v$

  c'est-à-dire :
  - que le joueur 1 ne peut avoir plus de v comme paiement moyen asymptotique
  - que le joueur 2 se garantit $-v$ comme paiement moyen asymptotique

- Symétriquement, en tirant I_t iid $\sim \mu^*$ où $\mu^* \in \operatorname{argmax}_{\mu \in \Delta(A)} \min_{\nu \in \Delta(B)} \mu^T \underline{\underline{R}} \nu$ le joueur 1 assure que

$\underline{\lim}_{n\to+\infty} \frac{1}{n} \sum_{t=1}^{n} R(I_t, J_t) \geq v$ (le joueur 1 se garantit v comme paiement moyen asymptotique)

et $\overline{\lim}_{n\to+\infty} \frac{1}{n} \sum_{t=1}^{n} (-R(I_t, J_t)) \leq -v$ (le joueur 2 ne peut obtenir plus de $-v$ comme paiement moyen asymptotique)

* Conclusion : $\exists v$ est le paiement (du jeu en un coup ou le paiement moyen asymptotique) obtenu quand chaque joueur joue optimalement dans le cas le pire.

En particulier, on observe également ce comportement quand chaque joueur minimise son regret, ie, quand (*) et (**) sont réalisés.

↳ Situation d'équilibre, en un certain sens ...

Notion d'équilibre minmax (= cas particulier de l'équilibre de Nash).

Déf : (μ^*, ν^*) est un équilibre si les contraintes suivantes sont respectées dans le jeu en un coup :

- pas de déviations profitable pour le joueur 1, ie, $\forall \mu \in \Delta(A)$, $\mu^T \underline{\underline{R}} \nu^* \leq \mu^{*T} \underline{\underline{R}} \nu^*$

  ou encore : (1) $\mu^* \underline{\underline{R}} \nu^* = \max_{\mu \in \Delta(A)} \mu^T \underline{\underline{R}} \nu^*$

- pas de déviations profitable pour le joueur 2, i.e., $\forall \nu \in \Delta(B)$, $-\mu^{*T} \underline{\underline{R}} \nu \leq -\mu^{*T} \underline{\underline{R}} \nu^*$

  ou encore : (2) $\mu^* \underline{\underline{R}} \nu^* = \min_{\nu \in \Delta(B)} \mu^{*T} \underline{\underline{R}} \nu$.

Question : Existe-t-il toujours un équilibre (quelle que soit $\underline{\underline{R}}$) ?

Oui : il n'est pas difficile de montrer que

$$\operatorname{argmax}_{\mu} \{ \min_{\nu} \mu^T \underline{\underline{R}} \nu \} \times \operatorname{argmin}_{\nu} \{ \max_{\mu} \mu^T \underline{\underline{R}} \nu \} \subset \mathcal{E}$$

où $\mathcal{E}$ est l'ensemble des équilibres.

(Vous pouvez le prouver en exercice.)

Nous prouverons ce fait au détour d'une preuve : nous allons montrer que

toutes les valeurs d'adhérence de $(\hat{\mu}_n, \hat{\nu}_n)$ – et il en existe, cf. théorème de Bolzano–Weierstrass – sont dans $\mathcal{G}$ lorsque (*) et (**) sont vérifiés.

Lemme (justifiant le nom d'équilibre minimax) : Si (μ^*, ν^*) est un équilibre alors $\mu^{*T} \underline{\underline{R}} \nu^* = v$.

Preuve : (1) indique que $\mu^{*T} \underline{\underline{R}} \nu^* = \max_{\mu \in \Delta(A)} \mu^T \underline{\underline{R}} \nu^* \geq \min_{\nu} \max_{\mu} \mu^T \underline{\underline{R}} \nu = v$

tandis que (2) conduit à $\mu^{*T} \underline{\underline{R}} \nu^* = \min_{\nu \in \Delta(B)} \mu^{*T} \underline{\underline{R}} \nu \leq \max_{\mu} \min_{\nu} \mu^T \underline{\underline{R}} \nu = v$.

Contre-exemple : La réciproque est fausse : il n'est pas vrai que si un couple (μ, ν) vérifie $\mu^T \underline{\underline{R}} \nu = v$ alors ce soit un équilibre.

Considérons en effet le jeu de "matching pennies" : $\underline{\underline{R}} = \begin{matrix} \\ H \\ B \end{matrix} \begin{matrix} G & D \\ \end{matrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

La valeur du jeu est bien $1/2$:

- avec $\mu = \frac{1}{2}\delta_H + \frac{1}{2}\delta_B$: $\forall \nu$, $\mu^T \underline{\underline{R}} \nu = 1/2$ donc $v \geq 1/2$
- avec $\nu = \frac{1}{2}\delta_G + \frac{1}{2}\delta_D$: $\forall \mu$, $\mu^T \underline{\underline{R}} \nu = 1/2$ donc $v \leq 1/2$

Or, le couple $\mu = \frac{1}{2}\delta_H + \frac{1}{2}\delta_B$ et $\nu = \delta_G$ vérifie :

- $\mu^T \underline{\underline{R}} \nu = 1/2$ (cf. ci-dessus)
- que le joueur 1 a une déviation profitable : $1 = \delta_B \underline{\underline{R}} \nu \geq \mu \underline{\underline{R}} \nu = 1/2$.

Convergence vers l'ensemble des équilibres minimax.

On rappelle que l'on a noté $\hat{\mu}_n = \frac{1}{n}\sum_{t=1}^{n} \delta_{I_t}$ et $\hat{\nu}_n = \frac{1}{n}\sum_{t=1}^{n} \delta_{J_t}$.
On s'intéresse ici à la convergence de $(\hat{\mu}_n, \hat{\nu}_n)$.

On va montrer le résultat (surprenant) suivant :

- la notion d'équilibre met vraiment en jeu un <u>couple</u> de stratégies
- les minimisations des regrets (*) et (**) se font chacune indépendamment et de manière dite "myope" (sans trop penser à la stratégie de l'autre joueur)
- et pourtant :

Théorème : Si les minimisations du regret (*) et (**) sont réalisées, alors la suite des $(\hat{\mu}_n, \hat{\nu}_n)$ converge vers le sous-ensemble $\mathcal{E}$ de $\Delta(A) \times \Delta(B)$ formé par les équilibres minmax :

$$\mathcal{E} = \{ (\mu^*, \nu^*) : \quad \mu^* \text{ et } \nu^* \text{ vérifient (1) et (2)} \}$$

Rq : Il s'agit bien d'une convergence vers $\mathcal{E}$, ie,

$$\inf_{(\mu,\nu)\in\mathcal{E}} \| (\hat{\mu}_n, \hat{\nu}_n) - (\mu, \nu) \| \longrightarrow 0 \quad ps$$

(norme sur $\mathbb{R}^{A\times B}$)

et non pas en général d'une convergence vers un point de $\mathcal{E}$.

Preuve : On a vu en page 3 que (*) et (**) entraînaient

$$(**) \quad \max_{\mu} \mu^T \underline{\underline{R}}\, \hat{\nu}_n \longrightarrow v \qquad \text{et} \qquad (***) \quad \min_{\nu} \hat{\mu}_n^T \underline{\underline{R}}\, \nu \longrightarrow v$$

En particulier, par théorème des gendarmes : $\hat{\mu}_n \underline{\underline{R}}\, \hat{\nu}_n \longrightarrow v$

de sorte que

$$\begin{cases} (3) & \lim_{n\to+\infty} \hat{\mu}_n^T \underline{\underline{R}}\, \hat{\nu}_n - \max_{\mu} \mu^T \underline{\underline{R}}\, \hat{\nu}_n = 0 \\ (4) & \lim_{n\to+\infty} \hat{\mu}_n^T \underline{\underline{R}}\, \hat{\nu}_n - \min_{\nu} \hat{\mu}_n^T \underline{\underline{R}}\, \nu = 0 \end{cases}$$

(3) et (4) sont les pendants "jeu répété" des conditions (1) et (2) du jeu en un coup. On raisonne par l'absurde. Si $((\hat{\mu}_n, \hat{\nu}_n))$ ne convergeait pas vers Σ, il existerait $\varepsilon > 0$ tel qu'une suite extraite est incluse dans le complémentaire de l'ε-voisinage ouvert de Σ, un ensemble fermé du compact $\Delta(A) \times \Delta(B)$, donc lui-même compact. Par théorème de Bolzano-Weierstrass, on pourrait en ré-extraire une sous-sous-suite $((\hat{\mu}_{\psi(u)}, \hat{\nu}_{\psi(u)}))$ convergeant vers (μ^*, ν^*) dans cet ensemble, en particulier : $\notin \Sigma$.
Dès lors, (1) ou (2) — disons, (1) — n'est pas vérifiée par (μ^*, ν^*) : on aurait $\exists \mu \in \Delta(A) \mid \mu^T \underline{\underline{R}} \nu^* > \mu^{*T} \underline{\underline{R}} \nu^*$. Mais par ailleurs (3) implique sur la suite extraite que $\mu^{*T} \underline{\underline{R}} \nu^* = \max_{\mu} \mu^T \underline{\underline{R}} \nu^*$, ce qui forme une contradiction et conclut la preuve.

Rq : Cette preuve montre au passage que Σ n'est pas vide : en reprenant les arguments, on voit que toute valeur d'adhérence de $((\hat{\mu}_n, \hat{\nu}_n))$ vérifie (1) et (2) lorsque (3) et (4) sont vérifiés, donc en particulier lorsque (*) et (**) sont vérifiés. Or, par théorème de Bolzano-Weierstrass, il existe toujours une telle valeur d'adhérence.