Learning in Multiagent Systems

Reinforcement learning and some issues

Stéphane Airiau

Université Paris Dauphine

Théorie des jeux

- On veut modéliser la décision d'agents rationnels
- La décision d'un agent peut avoir un effet sur d'autres agents

La Théorie des Jeux est un moyen de faire cette modélisation

Dans les exemples suivants, **deux** personnes doivent prendre une décision en **même temps**, **sans connaître** la décision de l'autre.

On représentera la situation à l'aide d'une table :

- la table a autant de ligne que le premier joueur à d'actions
- la table a autant de colonne que le second joueur à d'actions
- dans une case, on indique l'utilité du premier joueur (celui qui choisit la ligne) et du second joueur (celui qui choisit la colonne)
- plus l'utilité est grande, plus le joueur aime cet état.

Dilemme du prisonnier

Deux complices Ligne (L) et Colonne (C) sont arrêtés par la police et sont interrogés dans des pièces séparées.
Du point de vue d'un des complices, disons L, quatre situations sont possibles :

- C coopère avec la police et L refuse, la police réduit donc la peine de C à une peine légère.
- C ne coopère pas, mais L si, et donc C reçoit une lourde peine.
- C et L ne coopèrent pas avec la police, qui ne peut donc prouver la culpabilité totale. C et L reçoivent une peine "moyenne".
- C et L coopèrent, chacun reçoit une peine assez lourde.

L \ C	C est fidèle	C trahit
L est fidèle	3,3	1,4
L trahit	4,1	2,2

Le jeu du poulet

Dans *Rebel Without a Cause*, Buzz met au défi le personnage joué par James Dean, appelé Jim : ils doivent faire une course avec des voitures volée en se dirigeant vers une falaise. Le premier qui freine ou qui saute de la voiture a perdu.

	Jim continue	Jim freine
Buzz continue	-10,-10	5,0
Buzz freine	0,5	1,1

Cap ou pas cap ?

Deux amis font un pari, par exemple de venir en cours le lendemain dans une tenue ridicule. Si seul un des deux tient sa promesse, il sera ridicule. S'ils la tiennent tous les deux, cela sera un succès. Si personne ne tient sa promesse, personne ne sera embarassé, mais ce serait mauvais pour leur amitié.

	pas cap	cap
pas cap	1,1	2,0
cap	0,2	3,3

Problème du rendez-vous

	foot	opéra
opera	2,2	4,3
foot	3,4	1,1

- **Problème :** Match de foot ou Opéra avec son partenaire ?
- **Requirements :**
  - être ensemble !
  - profiter du spectacle qui vous convient le plus !

Questions

- Existe-t-il toujours un choix rationnel ?
- Si oui, comment le définir ?
- Si oui, comment le trouver ?

Représentation

Représentation générique du jeu est appelée jeu en forme normale.

Definition (Jeu en forme normale)

Un **jeu en forme normale (NFG)** est $(N, (S_i)_{i \in N}, (u)_{i \in N})$ où

- N est l'ensemble de n joueurs
- S_i est l'ensemble des stratégies/actions du joueur i.
- $u_i : S_1 \times \cdots \times S_n \to \mathbb{R}^n$ est l'utilité du joueur i : étant donné la stratégie de chacun des joueurs, la fonction retourne l'utilité du joueur i

Vocabulaire :

- un élément $s = \langle s_1, \ldots, s_n \rangle$ of $S_1 \times \cdots \times S_n$ est appelé profil des stratégies
- Soient $s \in S_1 \times \cdots \times S_n$ et $s'_i \in S_i$. On écrit (s'_i, s_{-i}) le profil des stratégie qui est le même que s sauf pour le joueur i qui joue la stratégie s'_i, c-a-d $(s'_i, s_{-i}) = \langle s_1, \ldots, s_{i-1}, s'_i, s_{i+1}, \ldots, s_n \rangle$

Que feriez-vous ?

- $N = \{\mathtt{L}, \mathtt{C}\}$
- $S_{Ligne} = S_{Colonne} = \{confiance, trahison\}$
- u_{Ligne} et $u_{Colonne}$ sont définis par.

L \ C	trahison	confiance
trahison	2,2	4,1
confiance	1,4	3,3

Ici, on ne veut pas utiliser le nom des actions (et ce qui est rattaché à ces noms), on veut seulement tenir compte des utilités

Peut-on utiliser un principe général pour choisir une action ?

Stratégie Dominante

Definition (Dominance forte)

Une stratégie $x \in S_i$ pour le joueur i **domine (fortement)** un stratégie $y \in S_i$ si le joueur i préfère (strictement) x à y indépendemment de la stratégie employée par les autres joueurs, i.e. $\forall s \in S_1 \times \cdots \times S_n,\ u_i(x, s_{-i}) > u_i(y, s_{-i})$

exemple : dilemme du prisonnier :

L \ C	trahison	confiance
trahison	2,2	4,1
confiance	1,4	3,3

Les deux joueurs ont des stratégies dominantes : trahir !

Du point de vue du joueur ligne L

- si C trahit, L a intérêt de trahir
- si C fait confiance, L a intérêt de trahir !

Stratégie Dominante

Problème du rendez-vous :

	G	D
H	2,2	4,3
B	3,4	1,1

Est-ce que les joueurs ont une stratégie dominante ?

Definition (Meilleurs réponse)

La stratégie s_i du joueur i est une **meilleure réponse** à un profil de stratégie s_{-i} ssi

$$\forall s_i' \in S_i,\ u_i(s_i, s_{-i}) \geqslant u_i(s_i', s_{-i}).$$

Definition (Equilibre de Nash)

Un profil de stratégie $s \in S_1 \times \cdots \times S_n$ est un **équilibre de Nash** si chaque s_i est une meilleure réponse à s_{-i}, c-a-d

$$(\forall i \in N)\left(\forall s_i' \in S_i\right)\ u_i(s_i, s_{-i}) \geqslant u_i(s_i', s_{-i})$$

Un équilibre de Nash est un profil de stratégie pour lequel aucun joueur ne peut améliorer son utilité en changeant seul sa stratégie.

Le problème du rendez-vous à deux équilibres de Nash : $\langle H, D\rangle$ and $\langle B, G\rangle$.

	G	D
H	2,2	4,3
B	3,4	1,1

Equilibre de Nash

L \ C	trahison	confiance
trahison	2,2	4,1
confiance	1,4	3,3

Il y a un unique équilibre de Nash : les deux joueurs trahissent !

Definition (Etat Pareto optimal)

Un profil de stratégies s est **Pareto optimal** s'il n'existe pas de profil de stratégies s'
$\forall i \in N\ u_i(s') \geqslant u_i(s)$ and $\exists i \in N u_i(s') > u_i(s)$

Un profil de stratégies est Pareto optimal quand on ne peut améliorer le bien-être d'un individu sans détériorer celui d'un autre

Avoir confiance (ne pas trahir) est Pareto optimal.

discussion : Il semble rationnel de trahir ! Cela semble contre-intuitif, car les deux joueurs ont intérêt à avoir confiance l'un en l'autre.

⇨Il y a donc un conflit entre une solution **stable** (l'équilibre de Nash dans lequel aucun joueur n'a intérêt à changer de stratégie) et une solution efficace (qui est Pareto optimale), ici, l'équilibre de Nash est dominé au sens de Pareto.

Existence d'un équilibre de Nash

L \ C	Pile	face
Pile	+1,-1	-1,+1
Face	-1,+1	+1,-1

Ce jeu n'admet pas d'équilibre de Nash.

Récapitulatif

- Quand il n'y a pas de stratégie dominante, un équilibre la meilleure chose qui reste.
- Un jeu n'a pas forcément un équilibre de Nash !
- Si un jeu possède un équilibre de Nash, il n'est pas forcément unique !
- Toute combinaison de stratégie dominante est un équilibre de Nash
- Un équilibre de Nash n'est pas forcément optimal au sens de Pareto
- Deux équilibres de Nash n'ont pas forcément les mêmes utilités

Definition (Stratégie Mixte)

Une stratégie mixte p_i pour un joueur i est une distribution de probabilité sur son espace de stratégie S_i.

exemple : $S_i = \{1,2,3\}$, le joueur décide de jouer la stratégie 1 avec une probabilité $\frac{1}{3}$, strategie 2 avec une probabilité de $\frac{1}{2}$ et la dernière action avec la probabilité $\frac{1}{6}$. Cette stratégie mixte sera notée $\left\langle \frac{1}{3}, \frac{1}{2}, \frac{1}{6} \right\rangle$.

Soit un profil de stratégies mixtes $p = \langle p_1, \ldots, p_n \rangle$, l'utilité espérée pour l'agent i est donc :

$$E_i(p) = \sum_{s \in S_1 \times \cdots \times S_n} \left(\left(\prod_{j \in N} p_j(s_j) \right) \times u_i(s) \right)$$

Problème du rendez vous

	G	D
H	2,2	4,3
B	3,4	1,1

On note

- $p_1 =< x, 1-x >$ la stratégie mixte du joueur ligne
- $p_2 =< y, 1-y >$ la stratégie mixte du joueur colonne

avec $x \in [0,1]$ et $y \in [0,1]$

L'utilité espérée du joueur ligne est donc :

$$xy \cdot 2 + x(1-y) \cdot 4 + (1-x)y \cdot 3 + (1-x)(1-y) \cdot 1 = -4xy + 3x + 2y + 1$$

Etant donnée un profil de stratégie mixte $p = \langle p_1, \ldots, p_n \rangle$, on écrit (p_i', p_{-i}) le même profil de stratégie mixte que p sauf pour le joueur i qui joue la stratégie mixte p_i', i.e., $(p_i', p_{-i}) = \langle p_1, \ldots, p_{i-1}, p_i', p_{i+1}, \ldots, p_n \rangle$.

Definition (Equilibre de Nash en stratégie mixte)

Un **équilibre de Nash en stratégie mixte** est un profil de stratégie mixte p tel que $E_i(p) \geqslant E_i(p_i', p_i)$ pour chaque joueur i et toute autre stratégie mixte p_i' pour le joueur i.

Problème du rendez-vous

	G	D
H	2,2	4,3
B	3,4	1,1

Soit la stratégie mixte $\langle \frac{3}{4}, \frac{1}{4} \rangle$.
aucun joueur n'a intérêt de changer de stratégie

$$E_{ligne}(T) = \frac{3}{4} \cdot 2 + \frac{1}{4} \cdot 4 = \frac{5}{2} \qquad E_{ligne}(B) = \frac{3}{4} \cdot 3 + \frac{1}{4} \cdot 1 = \frac{5}{2}$$

(les joueurs sont indifférents)

Théorème (J. Nash, 1950)

Tout jeu fini en forme normale possède au moins un équilibre de Nash en stratégie mixte.

note : La démonstration est non-constructive (i.e. elle ne donne pas le moyen de calculer un équilibre de Nash), elle utilise un théorème de point fixe (celui de Brouwer ou celui de Kakutani selon la version) pour garantir l'existence.

J.F. Nash. Equilibrium points in *n*-person games. in *Proc. National Academy of Sciences of the United States of America*, 36 :48-49, 1950.

Calcul d'un équilibre de Nash

Complexité : C'est un problème difficile, ce problème appartient à une classe appellée PPAD.

Daskalakis, Goldberg, Papadimitriou : **The complexity of computing a Nash equilibrium**, in *Proc. 38th Ann. ACM Symp. Theory of Computing (STOC)*, 2006

Il existe des algorithmes pour calculer des solutions pour certaines classes de jeux.

Y. Shoham & K. Leyton-Brown : **Multiagent Systems**, Cambridge University Press, 2009. (Chapter 4)

Nisan, Roughgarden, Tardos & Vazirani : **Algorithmic Game Theory**, Cambridge University Press, 2007. (chapters 2, 3)

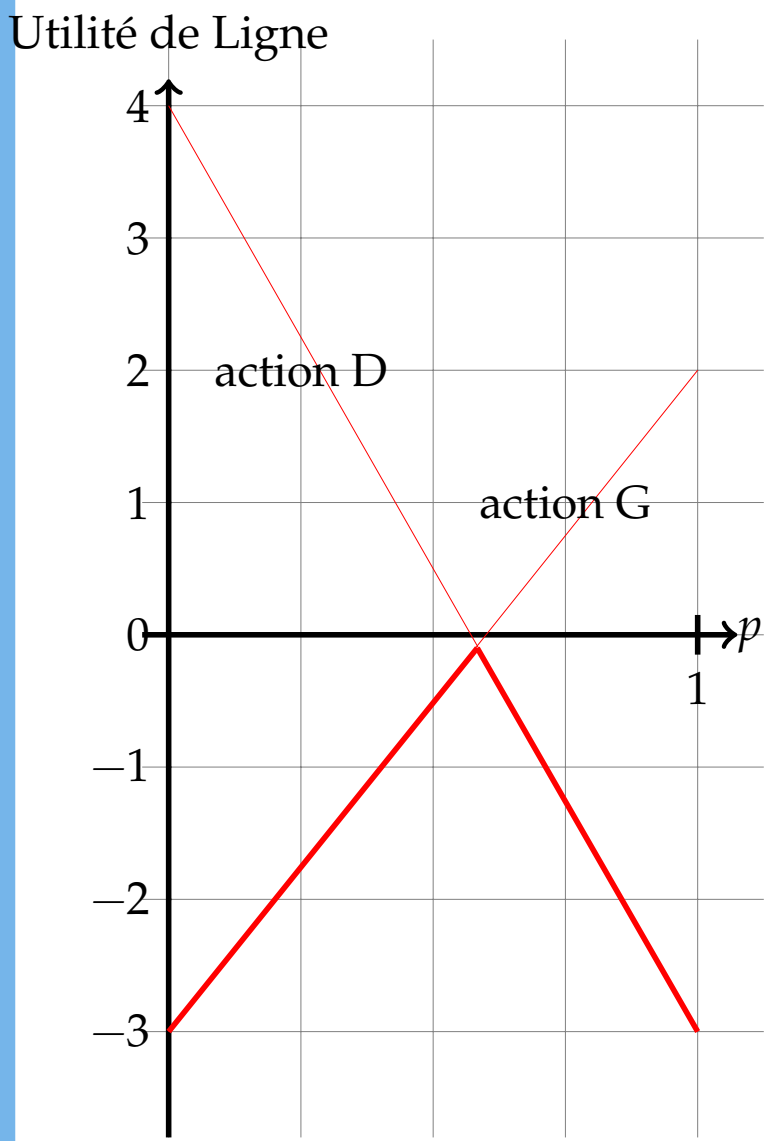
Jeu à somme nulle

La somme des utilités des agents pour une stratégie pure est égale à 0. On ne marque donc qu'un seul nombre (l'autre étant l'opposé !).

	G	M	D
H	30	-10	20
B	10	20	-20

Pour résoudre ce jeu, on peut calculer le MinMax en version probabiliste. On peut plus facilement travailler sur le jeu suivant

	G	D
H	2	-3
B	-3	4



	G	D
H	2	-3
B	-3	4

- Si C joue G
  L obtient $u(p) = 2p - 3 \times (1-p)$
- Si C joue D
  L obtient $u(p) = -3p + 4 \times (1-p)$

Pour maximiser son score, C cherche à minimiser l'utilité de L

- Il joue donc G avant que les droites se croisent
- il joue D après

Sachant cela, L va maximiser son utilité et jouer le point où les droites se croisent.

MaxMin (deux joueurs)

Chaque joueur peut donc jouer la stratégie

$$arg \max_{s_i \in \Delta(S)} \min_{s_{-i} \in \Delta(S)} u_i(s_i, s_{-i})$$

On reconnait donc bien la formulation maxmin pour jouer à des jeux à informations complètes à deux joueurs où chaque joueur joue une action en alternance (cours d'IA pour jouer aux dames, échecs, ...)

Definition (Maxmin)

La stratégie **maxmin** du joueur i est $\operatorname{argmax}_{s_i \in S_i} \min_{s_{-i} \in S_{-i}} u_i(s_i, s_{-i})$,

et sa valeur de **maxmin** est $\max_{s_i \in S_i} \min_{s_{-i} \in S_{-i}} u_i(s_i, s_{-i})$.

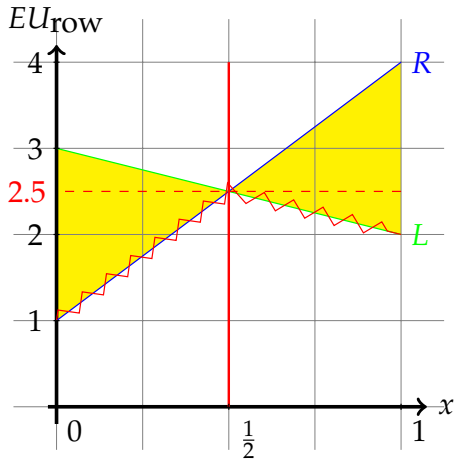
interprétation :

1) le joueur i choisit une stratégie (mixte ou non)

2) son adversaire $-i$ choisit une stratégie qui *minimise* les gains de i.

⇨ La stratégie *maximise* le pire gain du joueur i.

		y	$1-y$
		L	R
x	T	2,2	4,3
$1-x$	B	3,4	1,1



Gains de Ligne
lorsque colonne joue une stratégie pure (T ou R)
ou une stratégie mixte (en jaune)

Quel que soit ce que joue colonne, ligne peut garantir un gain de 2.5 en jouant $\langle \frac{1}{2}, \frac{1}{2} \rangle$.

MinMax (deux joueurs)

On peut étendre cette définition à n joueurs en faisant l'hypothèse que tous les autres agents coordonnent leurs actions – se liguent – pour minimiser le score de i.

⇨ stratégie *"conservative"* : l'agent cherche à maximiser son payoff quel que soit le comportement des autres agents (garantie minimale).

Bilan : On peut raisonner par rapport aux hypothèses sur l'adversaire

- adversaire rationel : stratégie de Nash
- adversaire imprévisible : stratégie maxmin nous donne une garantie sur les gains
- et si l'adversaire était "méchant" ou "mauvais" ?

Punir un autre agent avec maxmin

Stratégie duale de maxmin :

$$arg \min_{s_i \in \Delta(S)} \max_{s_{-i} \in \Delta(S)} u_{-i}(s_i, s_{-i})$$

Le joueur i cherche à minimiser le payoff de son adversaire même si son adversaire joue optimalement. ⇨ stratégie pour *"punir"* son adversaire

- Pour *deux joueurs*, la valeur "conservative" du maxmin est aussi sa valeur de punition (l'adversaire tente de punir le joueur, qui peut donc seulement obtenir sa valeur maxmin).
- Pour $n \geqslant 3$ *joueurs*, la valeur "conservative" est plus petite que la valeur du minmax.

Théorème (Théorème du minimax (von Neumann, 1928))

Pour un jeu fini à deux joueurs et à somme nulle, les stratégies minmax et maxmin sont les mêmes et sont en équilibre de Nash.

Choix de la stratégie

Bilan : On peut raisonner par rapport aux hypothèses sur l'adversaire

- adversaire rationel : stratégie de Nash
- adversaire imprévisible : stratégie maxmin nous donne une garantie sur les gains
- adversaire "méchant" : stratégie minmax pour le punir
- adversaire imprévisible : peut-on faire mieux ?
  maximiser le gain dans le pire des cas
  minimiser les pertes dans le pire des cas ?

Regret

	L	R
T	100,100	0,0
B	0,0	1,1

- (B,R) et (T,L) sont des équilibres de Nash
- Il n'y a pas de dominance forte.
- mais (T,L) Pareto domine (B,R)

⇨ Comment expliquer que (T,L) doit être préféré à (B,R) ?

Jeu "regret" r_i :
$r_i(s_i,s_{-i})$ est le **regret de i d'avoir choisi s_i et non $s_i^\star$**.

$$r_i(s_i,s_{-i}) = \left[\max_{a_i \in S} u_i(a_i,s_{-i})\right] - u_i(s_i,s_{-i}),$$

$r_i \backslash r_j$	L	R
T	**0,0**	1,100
B	100,1	0,0

On peut transformer la matrice des gains en une matrice qui traduit le regret.

Regret

On définit $regret_i(a_i)$ le regret maximal de i en choisissant l'action a_i

$$regret_i(a_i) = \max_{s_{-i} \in S} \left[\max_{a'_i \in S} u_i(a'_i, s_{-i}) \right] - u_i(a_i, s_{-i}),$$

$regret_i(a_i)$ est donc ce que i perd en jouant a_i plutôt que que sa meilleure réponse dans le pire des cas où l'adversaire joue l'action qui maximise sa perte.

Definition (Minimax regret)

$$arg\min_{a_i \in S} \left[\max_{s_{-i} \in S} \left(\max_{a'_i \in S} u_i(a'_i, s_{-i}) \right) - u_i(a_i, s_{-i}) \right]$$

Une stratégie minimisant le regret est une stratégie qui minimise le regret maximal.

Bilan

- adversaire rationel : stratégie de Nash
- adversaire imprévisible : stratégie maxmin nous donne une garantie sur les gains
- adversaire "méchant" : stratégie minmax pour le punir
- adversaire imprévisible : mininmiser le regret maximal

Il existe d'autres notions d'équilibres qu'on ne va pas présenter en détail.

- équilibres corrélés
  généralisation de l'équilibre de Nash. Les joueurs peuvent observer une variable aléatoire et peuvent coordonner leurs actions avec ce signal
- équilibre "trembling hand"
  Notion plus forte que l'équilibre de Nash, où une meilleure réponse doit être robuste à une pertubation de la stratégie de l'adversaire
- équilibre ϵ-Nash (la deviation doit excéder ϵ pour être effectuée)
  (attention, tout eq de Nash a dans son voisinage des éq de ϵ-Nash, mais pas l'inverse)

Variante de modèle de jeu : jeux répétés

One shot Vs Repeated

- Ce qu'on a vu jusqu'à présent : chaque joueur prend une décision et le jeu s'arrête. *"one shot game", there is no tomorrow*
- jeux répétés : il modélise la possibilité d'une nouvelle interaction avec le même adversaire
  - nombre de répétition fini : on peut représenter ce jeu comme un jeu avec alternance et on utilise une rétro induction pour résoudre le jeu.
  - nombre de répétition infini : l'arbre de jeu est infini, il faut d'autres techniques !
    ⇨ les PDMs ne sont peut être pas très loin !

Dilemme du prisonnier

	Trahison	Confiance
Trahison	2,2	4,1
Confiance	1,4	3,3

- "one shot" : stratégie dominante est de trahir !
- jeux répétés : si les mêmes agents jouent souvent ce jeu, ils auraient peut-être intérêt à toujours coopérer pour obtenir $\langle 3,3 \rangle$!

Stratégie

Définition la plus générale d'une **stratégie pure** : fonction qui dépend de l'historique des actions prises jusqu'ici.

$$h_0, h_1, \ldots, h_{t-1} \mapsto a_t,$$

où h_τ est un vecteur contenant les actions prises par chaque joueur à l'instant τ.

exemple : pour jouer au dilemme du prisonnier, une stratégie plutôt efficace est la suivante :

- $t = 0$: joue la confiance
- $t > 0$: joue l'action jouée par l'adversaire à $t-1$

Stratégie "Tit for Tat" ("coopération-réciprocité-pardon")
Tournoi d'Axelrod pour jouer au dilemme du prisonnier. Tit for Tat est la stratégie du vainqueur, Rapoport.

Axelrod, Robert
The Evolution of Cooperation, (1984), dernière édition 2006.

Objectif des joueurs

Comme pour les PDMs, la définition de l'objectif est importante puisque le jeu est appelé à être joué à l'infini.

- critère moyen : le joueur cherche à optimiser la moyenne des gains reçu durant toutes les parties

$$\lim_{t \to +\infty} \frac{\sum_{\tau=0}^{t} u_i(h_t)}{t},$$

où h_t est le vecteur des actions prises à l'instant t.

- critère somme dévaluée :

$$\sum_{t=0}^{\infty} \gamma^t u_i(h_t),$$

où γ est le facteur de dévaluation

interpretation : valeur à long terme ou le jeu s'arrête avec une probabilité $1-\gamma$.

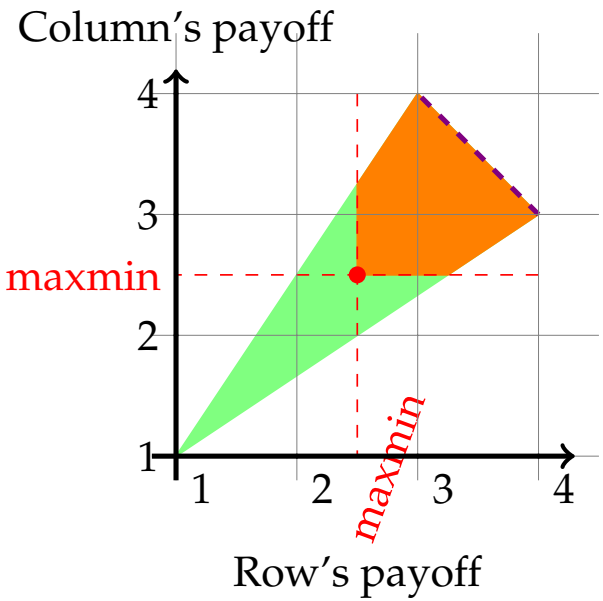
Théorème (A Folk theorem)

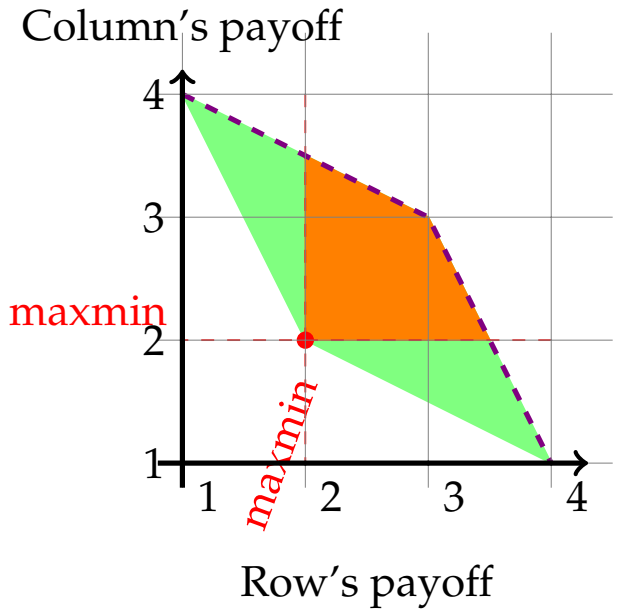
Avec le critère moyen, tout vecteur de gain v vérifiant

- v est **réalisable** : $\exists \lambda_a$ tel que $\sum_{a \in S^m} \lambda_a = 1$ et $v_i = \sum_{a \in S^m} \lambda_a u_i(a)$
  i.e. le gain moyen est bien dans l'enveloppe convexe des gains.
- v est **enforceable** $v_i \geqslant \max_{a \in S_i} \min_{s_{-i} \in S_{-i}} u_i(s_i, s_{-i})$

est le gain d'un équilibre de Nash.

Famille des équilibres de Nash





Problème du rendez-vous

	G	D
H	2,2	4,3
B	3,4	1,1

dilemme du prisonnier

	C est fidèle	C trahit
L est fidèle	2,2	4,1
L trahit	1,4	3,3

Famille des équilibres de Nash

Théorème (A Folk theorem)

Avec le critère moyen, tout vecteur de gain v vérifiant

- v est **réalisable** : $\exists \lambda_a$ tel que $\sum_{a \in S^m} \lambda_a = 1$ et $v_i = \sum_{a \in S^m} \lambda_a u_i(a)$
  i.e. le gain moyen est bien dans l'enveloppe convexe des gains.
- v est **enforceable** $v_i \geqslant \max_{a \in S_i} \min_{s_{-i} \in S_{-i}} u_i(s_i, s_{-i})$

est le gain d'un équilibre de Nash.

Idée de la démonstration :

- si v est plus petit que la valeur maxmin, l'agent peut dévier pour obtenir sa valeur de "sûreté". Donc être "enforceable" est nécessaire pour être un équilibre de Nash.
- on peut construire une stratégie qui donne exactement v aux agents car v est dans l'enveloppe convexe. Pour rendre cette stratégie un équilibre de Nash, on peut menacer de jouer la stratégie minmax pour toujours !

Une généralisation des jeux répétés : jeux stochastiques

- jeux répétés : c'est le même jeu qui est joué à chaque itération
- jeux stochastique : le jeu suivant dépend du jeu courant et du choix de l'action de chaque joueur

  ressemble à un des PDMs où l'action du PDM est le profil d'actions des joueurs et chaque agent possède sa fonction de récompense donné par le gain dans le jeu

Definition (Jeu stochastique)

un jeu stochastique est $(N,(S_i)_{i\in N},Q,P,(u_i)_{i\in N})$ où

- N est l'ensemble des joueurs
- S_i est l'ensemble des stratégies pour le joueur i
- Q est un ensemble de jeux en forme normale $q=(N,(S_i)_{i\in N},(v_i^q)_{i\in N})$
- $P:Q\times\prod_{i\in N}S_i\times Q\to[0,1]$ est la **fonction de transition**.
  $P(q,s,q')$ est la probabilité de jouer au jeu q' après avoir joué le profil d'action s dans le jeu q.
- $u_i:Q\times\prod_{i\in N}S_i$ est la fonction de **gains**
  $u_i(q,s)$ est le gain du joueur i quand le profil des action s a été joué dans le jeux q.

N.B. l'ensemble des actions est le même pour tous les jeux, mais ce n'est pas nécessaire.

Apprentissage Multiagent

Apprendre dans des jeux répétés

	Foot	Opéra
Foot	3,4	1,1
Opéra	2,2	3,4

rendez-vous

	Trahison	Confiance
Trahison	2,2	4,1
Confiance	1,4	3,3

Dilemme du prisonnier

Hypothèses

- information parfaite : chaque joueur peut observer les actions prises par son adversaire.
- information incomplète : chaque joueur ne connait pas forcémment les gains de son adversaire
- on répète le jeu ⇨ à chaque action, on reçoit un signal
  ⇨ l'apprentissage par renforcement semble très pertinent ici !

On peut se poser le problème dans un cadre plus complexe comme les jeux stochastiques.

Apprendre dans des jeux répétés

Si *un seul* agent apprend et les autres agents ont une stratégie stationnaire, on se retrouve dans un PDM.

⇨ **mais**, si d'autres agents apprennent en même temps, on n'a pas un PDM !

- hypothèse de Markov ?
- perte de stationarité ?

Que cherche-t-on vraiment à faire ?

- approche descriptive : comment l'apprentissage a lieu dans la *vraie* vie.
  - ⇨ similarité entre le modèle formel et la nature
  - ⇨ d'autant plus intéressant si le modèle formel possède des bonnes propriétés (convergence, etc)
    - Est-ce qu'on converge vers un équilibre de Nash ?
    - Est-ce que la fréquence observée converge vers un équilibre de Nash
    - Convergence vers un équilibre particulier (Pareto optimal ?)
- approche prescriptive : comment des agents (artificiels) devraient apprendre
  - un mécanisme d'apprentissage devrait garantir au moins le gain du maxmin (safety/rationalité individuelle)
  - si l'adversaire joue une stratégie stationnaire, le mécanisme devrait apprendre à jouer une meilleure réponse
  - le mécanisme ne devrait pas avoir un regret trop grand
  - si le mécanisme recontre un agent qui utilise le même mécanisme, il serait bon de garantir la convergence

quelques algorithmes

Fictitious Play (Brown 1951)

Hypothèse : l'adversaire joue une stratégie mixte fixée.
⇨ on apprend la distribution empirique des actions jouées par l'adversaire
⇨on joue une meilleure réponse à cette distribution !

```
1  initialise la fréquence p des actions jouées par l'adversaire
2  répète
3    joue une meilleure réponse à p
4    observe l'action jouée par l'adversaire
       et mise à jour des fréquences
```

Théorème

Si la fréquence de l'adversaire converge en utilisant fictitious play, alors elle converge vers un équilibre de Nash

- convergence pas toujours garantie (ex Pierre-feuille-ciseaux)

Joint Action Learning

(JAL) Claus et Bouilitlier 1998, (CJAL) Banerjee et Sen 2007

- apprendre une valeur q pour le profil des actions
- $$Q_{t+1}(a_t) \leftarrow Q_t(a_t) + \alpha\,(r_t - Q_t(a_t))$$
- r_t est la récompense obtenue après le profil d'actionsa_t
- α est le "taux" d'apprentissage

⇨ l'observation de sa récompense est suffisante

On utilise un modèle de l'adversaire pour calculer l'utilité espérée de ses actions (i.e. on estime $\mathbb{P}$)

$$\text{JAL}\quad \mathbb{E}(a_i) = \sum_{a_j \in S} \mathbb{P}(a_j) Q_{t+1}(a_i, a_j) \text{ où } \mathbb{P}(a_j) = \frac{|\{a_j^\tau = a_j | \tau \in [0..t]\}|}{t+1}$$

$$\text{CJAL}\quad \mathbb{E}(a_i) = \sum_{a_j \in S} \mathbb{P}(a_j|a_i) Q_{t+1}(a_i, a_j) \text{ où}$$

$$\mathbb{P}(a_j|a_i) = \frac{|\{a_j^\tau = a_j \wedge a_i^\tau = a_i | \tau \in [0..t]\}|}{|\{a_j^\tau = a_j | \tau \in [0..t]\}|}$$

Joint Action Learning

CJAL arrive à converger vers la coopération dans le dilemme du prisonnier !

Les trois méthodes que l'on vient de voir sont basées sur l'apprentissage d'un modèle de l'adversaire.

Il existe d'autres formes de modèles de l'adversaire (pas développées ici)

Méthodes où on peut observer le gain de l'adversaire

Idée générale :

- On construit un modèle pour chaque joueur
- On calcule une solution à partir du modèle
- le joueur joue un coup de la solution calculée

Solutions :

- hypothèse de jeux à somme nulle ⇨ minimax Minimax-Q (Littman 1994)
- équilibre de Nash Nash-Q (Hu & Wellman 2003)
- équilibre corrélé CE-Q (Greenwald & Hall 2003)

Apprentissage des valeurs de Q pour le profil des actions

$$Q(s,a_1,\ldots,a_n) \leftarrow (1-\alpha)Q(s,a_1,\ldots,a_n) + \alpha(r + \beta EQ_j(s')),$$

où EQ est une des solutions

Minimax-Q, Nash-Q, CE-Q

- Si les équilibres calculés ne sont pas uniques
- s'il y en a plusieurs, mais aucun ne se dégage (ex ils sont tous Pareto optimaux)

⇨ difficile de garantir la convergence !

Méthodes qui ne construisent pas un modèle de l'adversaire

- **idée 1** : chercher à monter dans l'espace des politiques
  la probabilité de jouer la meilleure action (selon la valeur de Q) augmente (légèrement) alors que la probabilité de jouer les autres décroit.
  ⇨on utilise Q-learning "classique"
- **idée 2** utiliser un taux d'apprentissage variable :
  - changement lent si on gagne (α_w petit)
  - changement plus rapide si on perf ($\alpha_l >> \alpha_w$)

- *Infinitesimal Gradient Ascent* (IGA) policy gradient ascent
  (convergence pas garantie pour tous les jeux)
- *Generalized* IGA $\rightarrow$ méthode basée sur le regret
  IGA converge vers un équilibre de Nash lorsque le jeu possède un équilibre de Nash en stratégie pure
- *Win or Lose Fast* IGA (WoLF-IGA)
  Converge vers un équilibre de Nash pour des jeux qvec deux actions
- Policy Hill Climber (PHC) and WoLF-PHC

Comparaison

Il est difficile de savoir quel algorithme utiliser.

- certains ont des garantie en "self-play"
- certains algorithmes s'en sortent mieux sur certains jeux, contre certains adversaire.
- Quel critère utiliser ? sur quel jeux ? Quel méthode de classement ?

Powers and Shoham 05, Airiau & Sen 05

Application to controlling a multiagent system

Scenario

- Collection of autonomous learning agents (e.g. robots, uavs, traffic controllers) works **for a system designer**
- The system designer wants to optimize a **collective criterion** (e.g. some objective function)

⇨ The utility function of the agents can be set up by the system designer.

- Agents cannot explicitly reason and communicate to reach the goal (system is too large, too difficult to compute).
- Agents only use their own experience

How to set up the individual utility functions so that, when each agents optimize its personal utility, the system converges to a good state ?

Difference Utility

- $N = \{1, \ldots, n\}$ is the set of agents
- $A = \{a_1, \ldots, a_k\}$ is the set of actions available to each agent
- $z \in A^N$ is the joint-action of the agents in the system
  (this may contain many entries)
  $\rightarrow z_i$ is the action of agent i
- $G : A^N \rightarrow \mathbb{R}$ is the collective utility function
  (set by the system designer).

The difference reward for agent i is of the form :

$$D_i = G(z) - G(z - z_i \cdot e_i + c_i \cdot e_i),$$

where $e_i \in A^n$ such that $e_i(j) = 0$ if $i \neq j$ and $e_i(i) = 1$.

$$D_i = G(z) - G(z - z_i \cdot e_i + c_i \cdot e_i),$$

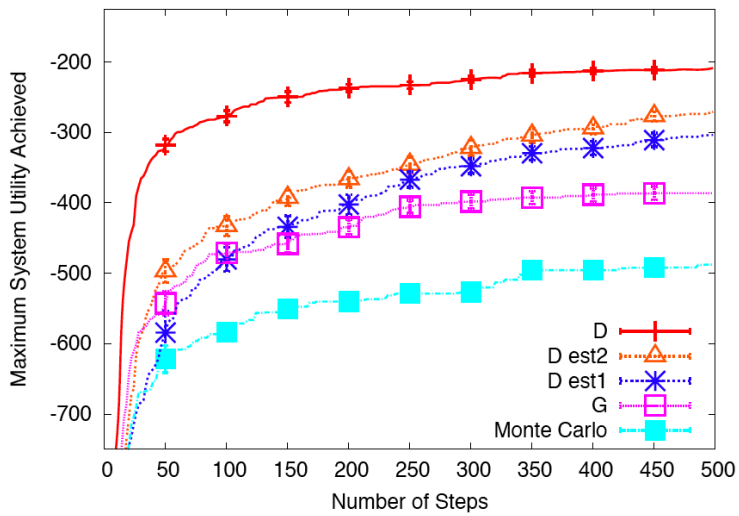
the action of agent i z_i is replaced by c_i

Sometimes, it is possible to choose c_i such that $z - z_i \cdot e_i + c_i \cdot e_i$ is **as if** i left the system.

⇨ D evaluates the contribution of agent i

- better signal ("learnability")
- As $G(z - z_i \cdot e_i + c_i \cdot e_i)$ does not depend on i, any action that improves D_i also improves G ! ("factoredness")

The form of G may be complex, but sometimes, each agent can "easily" approximate its D_i.



Tumer & A

Application to air traffic control

Conclusion

- L'apprentissage multiagent est toujours un domaine actif de recherche (même si les algorithmes que j'ai montrés commencent à dater)
- Le potentiel d'être vraiment utile
- Autres techniques à regarder
  - approches évolutionnaires (théorie des jeux évolutionnaire)
  - swarm intelligence (colonies de fourmis/abeilles)