Apprentissage de la coordination multiagent : Q-learning par jeu adaptatif

O. Gies
gies@damas.ift.ulaval.ca

B. Chaib-draa
Chaib@ift.ulaval.ca

Équipe DAMAS
Département informatique-génie logiciel
Faculté des Sciences-Génie
Université Laval
Québec– Canada

Résumé :
Dans le cadre de l'apprentissage multiagent, de nombreux travaux ont cherché jusqu'à présent à établir des algorithmes convergents vers un équilibre de Nash en jeux stochastiques. De tels algorithmes sont cependant limités dans la mesure où ils sont incapables de gérer la multiplicité des équilibres de Nash et de converger vers l'équilibre Pareto-optimal si celui-ci existe. Ces algorithmes utilisent généralement une convention pour la sélection de l'équilibre de Nash le plus approprié en cas d'équilibres multiples. Pour palier à cela, nous proposons un algorithme d'apprentissage étendant le Q-learning aux jeux stochastiques non-coopératifs, qui converge en jeux uniformes (en anglais "self-play", ce sont des jeux où tous les agents utilisent le même algorithme d'apprentissage) vers l'équilibre de Nash Pareto-optimal. Nous présentons des résultats expérimentaux montrant la convergence d'un tel algorithme en jeux homogènes vers un équilibre de Nash, en tant qu'équilibre de meilleure réponse mutuelle (donc vers un équilibre de Nash Pareto-optimal), sans besoin de convention de coordination explicite.

1 Introduction

Récemment, un intérêt significatif a été porté à l'extension de l'apprentissage par renforcement monoagent ([10]) aux jeux stochastiques[1]([13]). Les jeux stochastiques sont présentés comme un cadre pertinent pour l'apprentissage multiagent par certains chercheurs comme Littman [11]. De tels jeux étendent à la fois le cadre formel des processus décisionnels de Markov, dans lequel est défini l'algorithme d'apprentissage par renforcement associant action-état et appelé l'algorithme du Q-learning ([16]), et de la théorie des jeux ([5]).

L'un des problèmes majeurs dans l'extension de l'apprentissage monoagent aux systèmes multiagent est l'interaction entre agents : les actions individuelles ne peuvent plus être considérées indépendamment des actions des autres agents, car leurs conséquences sont interdépendantes. Dans le cas où les autres agents seraient stationnaires, un agent seul utilisant l'apprentissage par renforcement peut converger vers une politique optimale face, car la stationnarité des agents adverses peut être incluse dans le modèle de l'environnement, auquel cas le problème revient à un environnement monoagent. Cependant, en présence de non-stationnarité induite par d'autres agents (parce qu'ils apprennent ou simplement qu'ils suivent une politique non-stationnaire inconnue) l'apprentissage par renforcement monoagent ne permet pas de prendre en compte la présence des autres agents.

Beaucoup d'efforts sont de nos jours portés sur les apports possibles de la théorie des jeux en apprentissage multiagent. Parmi les chercheurs les plus actifs, il convient de citer Littman [11] qui a proposé l'algorithme de minimax-Q learning, dont il a prouvé la convergence pour des jeux purement compétitifs (i.e. récompenses opposées). Pour leur part, Claus et Boutilier [2] ont introduit le concept d'agent *joint-action learner* qui apprend la valeur des actions conjointes plutôt que la seule valeur de ses propres actions, et ont prouvé la convergence vers un équilibre de Nash pour les jeux purement coopératifs (récompenses identiques). Les auteurs Hu et Wellman [9] ont, quant à eux, introduit l'algorithme de NashQ-learning dans le cadre des jeux stochastiques non-coopératifs, avec récompenses décorrélées. Dans le même contexte, Greenwald et Hall [7] ont proposé une version similaire au NashQ-learning, appelées CE-Q learning (*Correlated Equilibria*) qui apprend en utilisant la valeur des équilibres corrélés (en cas de récompenses corrélées) plutôt que les équilibres de Nash. Littman [12] a pour sa part ré-interprété le NashQ-learning, dans l'algorithme *Friend-or-Foe Q-learning*, comme la combinaison d'un algorithme coopératif et d'un algorithme compétitif, et a prouvé la convergence vers un équilibre de Nash pour différentes classes de jeux stochastiques. Tesaura [14] a, pour sa part, proposé l'algorithme de *Hyper-Q learning*, qui apprend la valeur des stratégies conjointes mixtes plutôt que celles des actions conjointes (i.e. stratégies conjointes pures).

À l'exception de deux approches (celles de [15]

[1] Également appelés *jeux de Markov* dans la littérature.

et de [17]) la plupart des approches relatives à l'apprentissage multiagent sont limitées dans la mesure où les algorithmes qu'elles proposent, sont incapables de gérer la multiplicité des équilibres de Nash et de converger vers l'équilibre Pareto-optimal si celui-ci existe. De tels algorithmes utilisent en fait une convention pour la sélection de l'équilibre de Nash le plus approprié en cas d'équilibres multiples. Pour palier à cela, nous proposons un algorithme qui s'inspire du jeu adaptatif, proposé par Young [18] comme méthode de coordination en cas d'équilibres multiples. Nous le désignons dans la suite par le terme *Q-learning par jeu adaptatif*. Il s'agit d'une variante du jeu fictif [1] dans laquelle les agents ont une mémoire limitée et utilisent une règle de décision bruitée. Young [18] a prouvé que cette règle de décision peut permettre aux joueurs de se coordonner sur l'équilibre de Nash Pareto-optimal.

2 Concepts préalables

2.1 Apprentissage monoagent

En environnement inconnu, l'algorithme de programmation dynamique adaptative (PDA) alterne entre calcul du modèle (récompenses et transitions) et évaluation d'une politique donnée pour le modèle appris. Les algorithmes de la classe "différences temporelles" TD(λ) (*Temporal Difference*), sont une variante de la PDA calculant les fonctions de valeurs de leur politique donnée *indépendamment du modèle*. Plus précisément, étant donnée la politique fixée π, l'apprentissage utilisant les différences temporelles (appelé par la suite TD-learning) parcourt le processus de décision markovien (PDM ou MDP en anglais) en mettant à jour les fonctions de valeurs des différents états. Pour chaque transition observée entre les états s et s' suite à l'action a, les fonctions de valeurs U sont mises à jour selon la règle suivante :

$$U^{\pi}(s) \leftarrow U^{\pi}(s) + \alpha(R(s,\pi(s)) + \gamma U^{\pi}(s') - U^{\pi}(s)) \quad (1)$$

où α est le coefficient d'apprentissage et γ est le coefficient d'actualisation. Notons que le terme pondéré par α correspond à l'égalité dans l'équation générale suivante :

$$U^{\pi}(s) = R(s,\pi(s)) + \gamma \sum_{s' \in S} T(s,\pi(s),s') U^{\pi}(s') \quad (2)$$

dans laquelle on a enlevé le terme de probabilité de transition entre les états s et s' selon la politique π, noté $T(s,\pi(s),s')$. Cette simplification permet de s'abstraire du modèle lors de l'apprentissage en considérant chaque transition comme une approximation locale du modèle de transition. Si le modèle de transition n'est pas déterministe, les probabilités sont implicitement prises en compte dans la proportion de transition observées pour un grand nombre de visite de l'état s. Une telle approche est appelée *apprentissage par renforcement sans modèle*.

L'apprentissage par Q valeurs appelé plus communément "Q-learning" est une méthode de résolution dynamique qui dérive du TD-learning. Elle consiste à apprendre la valeur des actions selon les états, ce qui permet de calculer les utilités *et* la politique optimale dynamiquement. En Q-learning précisément, l'agent possède une fonction $Q : S \times A \mapsto \Re$ qui attribue à chaque couple *état-action* (s,a) une *Q-valeur* $Q(s,a)$, correspondant à la récompense espérée obtenue en effectuant l'action a dans l'état s et en suivant une politique optimale à partir de l'état suivant :

$$Q(s,a) = R(s,a) + \gamma \sum_{s' \in S} T(s,a,s') \max_{a \in A} Q(a,s') \quad (3)$$

Le Q-learning consiste à évaluer cette *Q-fonction* dynamiquement par l'expérience de l'agent dans l'environnement. Étant donnés l'état précédent s, l'action a effectuée en s et l'état courant s', l'agent utilise la règle de mise à jour suivante, qui dérive de la règle de mise à jour du TD learning (équation 1) en remplaçant les utilités pour une politique donnée par les Q-valeurs des paires *(état, action)* :

$$Q(s,a) \leftarrow Q(s,a) + \alpha(s)\left(R(s,a) + \gamma \max_{a \in A} Q(s',a) - Q(s,a)\right) \quad (4)$$

où $\alpha(s) = 1/n(s)$ est un taux d'apprentissage inversement proportionnel au nombre de fois que l'état s a été visité.

La convergence des Q-valeurs vers les Q-valeurs optimales, et par conséquent vers la politique optimale, a été démontrée par Watkin [16] sous l'hypothèse que les couples état-action sont visités une infinité de fois et d'autres conditions restrictives sur les paramètres d'apprentissage.

Le Q-learning est un algorithme d'apprentissage monoagent, qui peut être utilisé en environnement multiagent, mais sans prendre en compte explicitement la présence des autres agents. La théorie des jeux est un cadre formel qui permet de prendre en compte la présence des autres agents et d'agir en conséquence.

2.2 Théorie des jeux

Dans un jeu où le mouvement simultané d'un ensemble de joueurs est constitué d'un seul coup, chaque joueur i choisit simultanément une stratégie[2] $m^i \in M^i$. Le vecteur de stratégies des joueurs est appelé profil de stratégies et il est noté par $m \in M \equiv \times_{i=1}^{N} M^i$, avec N le nombre d'agents. Chaque joueur reçoit alors une utilité (appelée aussi payement ou récompense et pouvant, dans certains cas, rejoindre la fonction valeur d'un MDP). La combinaison de l'ensemble des joueurs, de l'espace des stratégies et les fonctions "utilités" est appelée forme normale ou stratégique d'un jeu. Les stratégies peuvent être "pures" ou "mixtes". Dans le cas des stratégies mixtes, chaque joueur i utilise les stratégies pures de manière aléatoire avec des probabilités notées $\sigma^i \in \Sigma^i \equiv \Delta(M^i)$, et où l'espace de distribution de probabilités est noté par $\Delta(.)$. Les profils de stratégies mixtes sont alors notées $\sigma \in \Sigma = \times_{i=1}^{N} \Sigma^i$.

Partant de là, chaque joueur i a comme utilité espérée : $u^i(\sigma) = \sum_m u^i(m) \Pi_{j=1}^{N} \sigma^j(m^j)$ Là aussi, chaque joueur va tenter de maximiser sa propre utilité espérée. Bien entendu, cela va dépendre de comment il anticipe les jeux des autres agents. La question soulevée par l'apprentissage via les jeux tente de répondre à cela en formant justement de telles *anticipations*. Supposons pour le moment que i croit que la distribution de probabilités de ses opposants est σ^{-i}. Dans ce cas, i doit jouer la meilleure réponse, c'est à dire une stratégie $\hat{\sigma}^i$ telle que :

$$u^i(\hat{\sigma}^i, \sigma^{-i}) \geq u^i(\sigma^i, \sigma^{-i}) \quad \forall \sigma^i$$

L'ensemble des meilleures réponses (BR pour best responses) à σ^{-i} est noté $BR^i(\sigma^{-i})$, avec bien entendu, $\hat{\sigma}^i \in BR^i(\sigma^{-i})$. La notion de solution en théorie des jeux repose sur le concept d'*équilibre stratégique*, qui correspond à une stratégie conjointe, optimale selon un certain critère. On parle de *Pareto*-optimalité pour une action conjointe telle que jouer une autre action conjointe réduit l'utilité d'au moins l'un des joueurs. Une telle action conjointe est appelée un équilibre Pareto-optimal. Formellement, un profil stratégique $\overline{\sigma} = \left(\overline{m}^1, \ldots, \overline{m}^N\right)$ est Pareto-optimal si :

$$\forall \sigma \neq \overline{\sigma},\ \exists j \text{ tel que } u^j(\sigma) < u^j(\overline{\sigma})$$

[2]En théorie des jeux, une stratégie (au sens tactique ou manœuvre) s^i, au sens de la manœuvre, est une règle qui indique à l'agent i quelle action il convient de choisir à chaque instant t du jeu. C'est une notion spécifique à théorie des jeux, et qu'il convient de ne pas confondre avec la "stratégie" au sens des processus de Markov.

Une notion d'équilibre plus largement utilisée est celle d'*équilibre de Nash*. Un équilibre de Nash est une action conjointe telle que dévier *individuellement* de son action pour chaque joueur i réduit son utilité propre. En d'autres termes, un équilibre de Nash est une stratégie conjointe où la stratégie de chaque agent est une meilleure réponse au profil stratégique adverse. Formellement, un profil stratégique $\hat{\sigma}$ est un équilibre de Nash si :

$$\hat{\sigma}^i \in BR^i(\hat{\sigma}^{-i}) \quad \forall i$$

Un équilibre de Nash peut être Pareto-optimal, mais bien que les équilibres Pareto-optimaux semblent meilleurs pour tous les joueurs, les équilibres de Nash sont plus fréquents et plus faciles à déterminer, tout en proposant une notion d'équilibre de meilleure réponse mutuelle. Particulièrement, il existe *toujours* un équilibre de Nash en stratégies mixtes. Le jeu fictif est une méthode de coordination en théorie des jeux, dont la convergence vers un équilibre de Nash (en stratégies pures ou mixtes) a été prouvée pour plusieurs classes de jeux. Nous présentons ce processus dans la section suivante.

2.3 Jeu fictif

Le jeu fictif[3] est un processus d'apprentissage de la théorie des jeux, établi par Brown [1]. En jeu fictif, les joueurs entretiennent des croyances empiriques individuelles sur les stratégies suivies par les autres joueurs. Pour fixer les idées, considérons seulement 2 joueurs ayant des espaces de stratégies finies M^1 et M^2 et les utilités u^1 et u^2. Le modèle de jeux fictifs suppose que les joueurs choisissent leurs actions à chaque période pour maximiser leur utilité espérée, étant donnée leur évaluation des distributions des actions d'autrui durant cette période. Cette évaluation prend la forme suivante. On suppose que i a une fonction de poids initiale qui serait $c_0^i : M^{-i} \rightarrow \Re_+$. Ce poids est mis à jours en ajoutant 1 au poids de chacune des stratégies de son opposent lorsque cette stratégie est jouée :

$$c_t^i\left(m^{-i}\right) = c_{t-1}^i\left(m^{-i}\right) + \begin{cases} 1 & \text{si} \quad m_{t-1}^{-i} = m^{-i} \\ 0 & \text{sinon} \end{cases}$$

Dan ces conditions la probabilité que le joueur i assigne au joueur $-i$ jouant m^{-i} à la date t est donnée par :

$$\gamma_t^i(m^{-i}) = \frac{c_t^i\left(m^{-i}\right)}{\sum_{\tilde{m}^i \in S^i} c_t^i(\tilde{m}^{-i})}$$

[3]Ce concept est essentiellement traité en littérature anglophone sous le terme *fictitious play*.

On peut maintenant associer une règle par les moyens de $\Gamma(.)$ telle que $\Gamma_t^i(\gamma_t^i)$ fait partie des meilleures réponses de i, soit : $\Gamma_t^i(\gamma_t^i) \in BR^i(\gamma_t^i)$. Dès lors $\Gamma(.)$ indique la meilleure réponse de i quand $-i$ joue m^{-i} avec la probabilité $\gamma_t^i(m^{-i})$. Cependant il n'y a pas qu'une unique règle puisque il peut y avoir plusieurs meilleures réponses. Étant données ses croyances sur le profil stratégique adverse γ_t^i, chaque joueur i choisit alors sa réponse aléatoirement dans l'ensemble des meilleures réponses, soit : $m_{t+1}^i = random(BR^i(\gamma_t^i))$

La convergence vers un équilibre de Nash des stratégies des joueurs et des croyances empiriques sur ces stratégies a été prouvée pour plusieurs classes de jeux. Nous référons à Hofbauer et Sandholm [8] pour un aperçu couvrant l'essentiel de ces travaux, et une étude générale de la convergence du jeu fictif stochastique introduit par Fudenberg et Levine [4].

L'inconvénient principal du jeu fictif est sa forte dépendance aux croyances initiales. Cette forte dépendance aux croyances initiales se traduit également dans l'apparition de cycles stratégiques dégénérés.

2.4 Jeu adaptatif

Pour surmonter les défauts du jeu fictif, Young [18] a introduit une variante du jeu fictif appelée *jeu adaptatif*. L'idée principale du jeu adaptatif est, d'une part, d'introduire une exploration stochastique des actions et d'autre part, de supprimer l'influence des croyances initiales en utilisant une mémoire limitée. De plus, pour limiter l'influence du bruit introduit par l'exploration des joueurs adverses, seul un échantillon de la mémoire est utilisé pour construire les croyances empiriques sur les stratégies.

Formellement, chaque joueur i garde en mémoire un historique $H_t = \{\sigma_{t-p}, \ldots, \sigma_t\}$ des p dernières stratégies au temps t. Pour chaque joueur adverse j, le joueur i tire aléatoirement et sans remise un échantillon $\hat{H}_t^j = \left\{\sigma_{k_1}^j, \ldots, \sigma_{k_l}^j\right\}$ de l stratégies passées du joueur j prises dans l'historique H_t. La croyance du joueur i sur la stratégie future du joueur j est alors calculée sur l'échantillon $\hat{H}_t^j$ de taille $\left|\hat{H}_t^j\right| = l$ de la manière suivante :

$$\eta_t^i(\hat{H}_t^j) = \frac{n_{\hat{H}_t^j}\left(\sigma_{k_j}^j\right)}{l} \quad (5)$$

où $n_{\hat{H}_t^j}\left(\sigma_{k_j}^j\right)$ est le nombre de fois qu'apparaît la stratégie $\sigma_{k_j}^j$ dans l'échantillon $\hat{H}_t^j$. Avec $1-\epsilon$, l'agent i joue alors une meilleure réponse à la distribution statistique qu'il a échantillonnée η_t^i et ; avec une probabilité de ϵ, il choisit au hasard son action. Notons que cette méthode d'exploration se retrouve largement au delà du cadre de la théorie des jeux sous le terme ϵ-*glouton*. Le coefficient d'exploration ϵ peut être fixé (exploration stationnaire) ou variable (par exemple, décroissant si l'on recherche une convergence).

Young [18] a démontré que le jeu adaptatif permet de sélectionner l'équilibre de Nash Paretoopimal - s'il existe. Notons qu'il permet également aux agents de se coordonner sans convention en cas d'équilibres de Nash multiples.

2.5 Jeux stochastiques

Les jeux stochastiques sont un cadre formel permettant de modéliser un environnement multiagent non-coopératif. Le fait d'être non-coopératif signifie que les agents ne poursuivent pas a priori de but commun, mais des objectifs individuels. Ils peuvent cependant être amenés à se coordonner, voire à coopérer, pour atteindre leurs buts individuels. Le cadre formel des jeux stochastiques forme un modèle qui étend les MDP à un cadre multiagent. Il peut aussi être considéré comme une extension à plusieurs états des *jeux en forme normale* (voir section 2.2 plus haut) de la théorie des jeux. Notons particulièrement que chaque état d'un jeu stochastique peut être vu comme un jeu en forme normale, comme représenté sur la figure 1. Dans la suite, on emploie donc indifféremment les termes *agent* et *joueur*.

Formellement, un jeu stochastique est un tuple $\left\langle N, S, \{A^1, \ldots, A^N\}, \{R^1, \ldots, R^N\}, T\right\rangle$ où N est le nombre d'agents modélisés, S l'ensemble fini d'états du jeu, $A^i = \left\{a_1^i, \ldots, a_{|A^i|}^i\right\}$ l'ensemble d'actions de l'agent i, $R^i : S \times A^1 \times \ldots \times A^N \mapsto \Re$ est la fonction de récompense de l'agent i et $T : S \times A^1 \times \ldots \times A^N \times S \mapsto \Re$ est le modèle de transition entre états, dépendant de l'action conjointe des agents.

A chaque tour de jeu, étant donné l'état

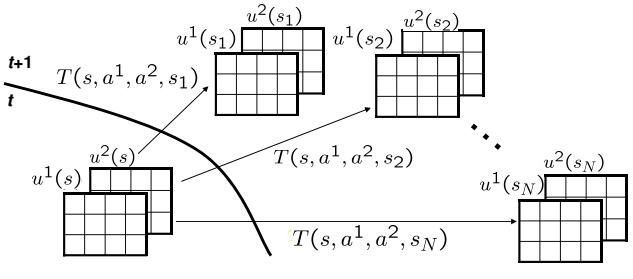

FIG. 1 – Jeu stochastique à deux joueurs : transitions possibles entre le tour t et le tour $t+1$ lorsque les joueurs jouent les actions a_1 et a_2 ; chaque état peut être vu comme un jeu en forme normale.

courant s, les agents choisissent les actions $a^1, \ldots, a^N$. Ils obtiennent alors les récompenses $\{R^i\left(s, a^1, \ldots, a^N\right)\}_{i\in[1,n]}$ et le système passe dans l'état s' en suivant le modèle de transition T, qui vérifie $\sum_{s'\in S} T\left(s, a^1, \ldots, a^N, s'\right) = 1$. Une politique $\pi^i : S \mapsto [0,1]^{|A^i|}$ pour l'agent i définit une stratégie locale en chaque état au sens de la théorie des jeux. Autrement dit, $\pi^i(s)$ est un vecteur dont les éléments sont des masses de probabilité sur les actions du joueur i, spécifiques au jeu en forme normale défini par l'état s.

Le terme d'utilité espérée d'un joueur en théorie des jeux désigne l'espérance de "récompense" sur les *stratégies des joueurs adverses*, alors que la fonction de valeur en MDP est l'espérance *temporelle* de la récompense. Nous emploierons donc le concept d'utilité U^i en jeu stochastique comme l'espérance temporelle des utilités espérées u^i_s de l'agent i définies pour chaque état s de manière similaire aux utilités espérées des jeux en forme normale. Les utilités U^i des états pour chaque joueur i, associée à la politique conjointe $\breve{\pi}(s) \equiv \times_{i=1}^{N}\pi^i$, sont donc définies comme l'utilité espérée par l'agent i à partir de l'état s si tous les agents suivent cette politique conjointe :

$$\begin{aligned} &U^i_{\pi^1,\ldots,\pi^N}(s) \\ &\quad = E\left[\sum_{t=0}^{\infty}\gamma^t u^i_{s_t}\left(\pi^1(s_t), \ldots, \pi^N(s_t)\right) | s_0 = s\right] \\ &\quad = u^i_s\left(\pi^1(s), \ldots, \pi^N(s)\right) \\ &\qquad + \gamma \sum_{s'\in S} T\left(s, \pi^1(s), \ldots, \pi^N(s), s'\right) U^i_{\pi^1,\ldots,\pi^N}(s') \end{aligned} \quad (6)$$

En jeu stochastique, une politique conjointe $\pi^1_\circ, \ldots, \pi^N_\circ$ est un équilibre de Nash si les stratégies qu'elle définit en chaque état forment un équilibre de Nash (EqNash) pour cet état au sens de la théorie des jeux. Formellement :

$$\begin{array}{c} \pi^1_\circ, \ldots, \pi^N_\circ \text{ est un EqNash} \\ \Updownarrow \\ \forall s \in S, \pi^1_\circ(s), \ldots, \pi^N_\circ(s) \text{ est un EqNash pour l'état } s \end{array}$$

De la même manière, on définit pour l'agent i une *politique de meilleure réponse* π^i_{br} aux politiques adverses $\{\pi^j\}_{j\neq i}$ comme une politique définissant en chaque état une stratégie de meilleure réponse aux stratégies définies par les politiques adverses pour cet état :

$$\begin{array}{c} \pi^i_{br} \text{ politique de meilleure réponse à } \{\pi^j\}_{j\neq i} \\ \Updownarrow \\ \forall s \in S,\ \pi^i_{br}(s) \in BR^i(\pi^j(s)) \end{array}$$

Soulignons qu'à l'instar des jeux en forme normale, un équilibre de Nash en jeux stochastiques est une politique conjointe où la politique de chaque agent est une politique de meilleure réponse aux politiques adverses. Le Q-learning est une méthode d'apprentissage *dynamique* qui permet d'apprendre simultanément les utilités des états ainsi que la politique optimale de l'agent en MDP. Le jeu adaptatif, quant à lui, permet à plusieurs joueurs de se coordonner sur un *équilibre optimal* dans un jeu en forme normale, i.e. un équilibre de meilleure réponse mutuelle. En utilisant la notion de solution d'équilibre de Nash présenté en théorie des jeux, ainsi que le concept de meilleure réponse, nous étendons donc le Q-learning en utilisant le jeu adaptatif pour prendre en compte la présence des autres agents, et ce dans le cadre formel multiagent formé par les jeux stochastiques. Notons qu'un autre avantage du jeu adaptatif est de permettre une adaptation rapide des croyances de l'agent sur les stratégies adverses, même si ces dernières sont non-stationnaires. Notre algorithme présente donc des capacités potentielles d'adaptation à des agents non-stationnaires. Cet aspect d'adaptation à la non stationnarité a été développé ailleurs [6].

3 Apprentissage de la coordination par la méthode Q-learning par jeu adaptatif

Nous proposons un algorithme qui étend le Q-learning monoagent aux jeux stochastiques, en utilisant le *jeu adaptatif* proposé par Young [18]. Dans cette section, nous décrivons l'algorithme du Q-learning par jeu adaptatif ainsi que les modes d'exploration utilisés dans

nos expériences. En Q-learning par jeu adaptatif, l'agent i apprend les stratégies adverses *en chaque état s* par jeu adaptatif. Plus précisément, il mémorise H_t un historique des actions conjointes en chaque état, et calcule les probabilités empiriques des actions de l'agent j en chaque état sur un échantillon $\hat{H}_t^j$ des actions de j prises dans l'historique H_t. Il sélectionne ensuite la meilleure réponse à sa croyance sur la stratégie conjointe adverse $\breve{\pi}^{-i}(s)$ dans l'état s. Il convient de préciser que le modèle de jeu stochastique étend les jeux en forme normale à plusieurs états et par conséquent l'utilité d'une action conjointe ne dépend plus du seul état courant, mais également des utilités des états futurs. Or le Q-learning permet de tenir compte de l'utilité espérée, au sens de l'espérance temporelle, d'une action conjointe dans un état donné. En Q-learning monoagent, l'utilité d'un état est calculée comme étant la Q-valeur *maximale* de cet état sur toutes les actions possibles. Comme expliqué précédemment, le critère d'optimalité en jeux stochastiques ne peut pas être une maximisation des Q-valeurs sur les actions individuelles des agents, car leurs conséquences dépendent de l'action conjointe. L'extension du Q-learning au cadre multiagent (avec N agents) prend en compte l'influence des actions des autres joueurs par la définition de Q-valeurs état-action conjointe $Q^i\left(s, a^1, \ldots a^N\right)$ pour chaque agent i. On définit formellement une Q-valeur comme suit :

$$Q^i\left(s, a^1, \ldots a^N\right) = R^i\left(s, a^1, \ldots, a^N\right) + \gamma \sum_{s' \in S} T\left(s, a^1, \ldots a^N, s'\right) \Gamma_{\pi^1,\ldots,\pi^N}\left[U^i_{\pi^1,\ldots,\pi^N}\left(s'\right)\right] \tag{7}$$

Dans cette équation 7, $\Gamma_{\pi^1,\ldots,\pi^N}$ est un opérateur générique qui définit la politique conjointe $\pi^1, \ldots, \pi^N$ qui est supposée suivie par les agents à partir de l'état suivant. Cet opérateur définit le critère d'optimalité recherché. En Q-learning monoagent, $\Gamma_\pi = \max_{a \in A}$. Hu et Wellman [9] ont par la suite introduit l'opérateur $\Gamma_\pi \equiv NashQ$, qui définit la politique conjointe $\pi^1_\circ, \ldots, \pi^N_\circ$ comme un équilibre de Nash pour le jeu stochastique considéré. De manière similaire, Weinberg et Rosenchein [17] ont proposé un opérateur $\Gamma(.)$ qui définit une politique conjointe constituée des *politiques stationnaires limites* des autres agents et de la politique de meilleure réponse pour l'agent i.

En ce qui nous concerne, nous adoptons la formulation des Q-valeurs suivante :

$$Q^i\left(s, a^1, \ldots a^N\right) = R^i\left(s, a^1, \ldots, a^N\right) + \gamma \sum_{s' \in S} T\left(s, a^1, \ldots a^N, s'\right) U^i_{\pi^1_\star,\ldots,\pi^i_{br},\ldots,\pi^N_\star}\left(s'\right) \tag{8}$$

où $\pi^j_\star$ est la politique *non-stationnaire réellement suivie* par l'agent j, et π^i_{br} une politique de meilleure réponse du joueur i aux politiques adverses $\{\pi^j_\star\}_{j \neq i}$. Nous faisons ici l'hypothèse qu'une "bonne approximation" de $\pi^j_\star(s')$ est $\hat{\pi}^j(s')$ qui représente la croyance empirique de l'agent i sur la stratégie de l'agent j dans l'état s', calculée selon la règle (5) du jeu adaptatif. Bien entendu, ceci n'est valable que si $\lim_{x \to \infty} |\pi^j_\star(s') - \hat{\pi}^j(s')| \to 0$, ce qui a été montré par ailleurs [18].

Dans ce cas, la formule de mise à jour des Q-valeurs devient alors :

$$\begin{aligned} &Q^i_{t+1}\left(s, a^1, \ldots a^N\right) \leftarrow (1-\alpha) Q^i_t\left(s, a^1, \ldots a^N\right) \\ &+\alpha[R^i\left(s, a^1, \ldots, a^N\right) \\ &+\gamma u^i_{s'}\left(\hat{\pi}^1(s'), \ldots, \pi^i_{br}(s'), \ldots, \hat{\pi}^N(s')\right)] \end{aligned} \tag{9}$$

où $\pi^i_{br}(s')$ est une action de meilleure réponse dans l'état s' au profil stratégique adverse reflété par $\hat{\pi}^j(s')_{j \neq i}$. Conformément à ce qu'on avait dit auparavant, l'agent i choisit son action dans l'état s' en suivant la règle de décision du jeu adaptatif, i.e. $\pi^i_{br}(s') = (1-\epsilon) BR^i(\hat{\pi}^j(s')) + \epsilon.random(A^i)$.

L'exploration stationnaire consiste à attribuer une valeur fixée à ϵ : quelque soit le stade d'apprentissage de l'agent, celui-ci choisit une action aléatoire avec une probabilité ϵ ***fixée***, et sa meilleure réponse avec la probabilité $1 - \epsilon$.

Pour ϵ on pourrait adopter GLIE qui est l'acronyme de *Greedy in the Limit with Infinite Exploration*. Cette approche consiste à faire tendre ϵ vers 0 à l'infini. L'idée est de permettre une exploration intensive au début de l'apprentissage, et d'utiliser exclusivement les données apprises à l'infini. Une manière intuitive de définir ϵ est de lui assigner l'inverse du nombres de tours de jeu depuis le début : $\epsilon = 1/t$.

Cette définition est plutôt simpliste, et peut présenter un inconvénient lorsque les espaces d'état et d'actions sont vastes. On utilise dans notre étude une version qui permet d'explorer les actions possibles *en chaque état* : $\epsilon = 1/n(s)$, où $n(s)$ est le nombre de fois que l'état s a été visité.

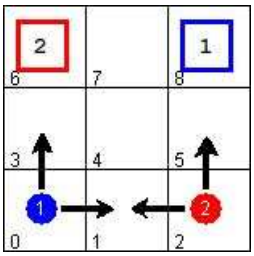

FIG. 2 – Jeu de coordination : les agents doivent se coordonner pour atteindre la case dans le coin supérieur opposé à leur position initiale sans se gêner.

4 Expérimentations

Les performances d'apprentissage du Q-learning par jeu adaptatif sont évaluées dans une grille de jeu où se déplacent deux agents, représentées sur la figure 2. Le jeux se déroule en temps discret où à chaque pas (tour de jeu), les agents choisissent leurs actions simultanément, et se déplacent dans la grille. Les conséquences des actions sont déterministes : un agent qui effectue l'action $haut$ se déplacera dans la case supérieure au tour suivant avec une probabilité 1 (sauf en cas de collision dans le jeu de coordination). Le jeu est un jeu de coordination inspiré de Hu et Wellman [9].

4.1 Un aperçu sur le jeu de coordination

Dans le jeu de coordination (figure 2), les deux agents sont initialement positionnés dans les coins inférieurs de la grille. Ils doivent atteindre leurs cases destinations, situées dans les coins supérieurs opposés à leurs positions initiales. S'ils tentent d'aller sur la même case au même tour de jeu, ils entrent en collision et sont replacés sur la case d'où ils viennent. Pour atteindre leur destination en un temps minimal, les agents doivent donc coordonner leurs actions pour ne pas se gêner.

Un agent reçoit une récompense (1) de $+80$ s'il atteint sa destination en ne suivant pas les pas de l'autre agent, (2) de $80 + 10$ fois le nombre de cases déjà visitées par l'autre agent et (3) de -1 lors d'une collision. Pour toute autre situation, il reçoit une récompense nulle. Les deux agents disposent des actions $\{haut, bas, gauche, droite\}$, dépendant de la case dans laquelle ils se trouvent. Par exemple, $A^1((0,*)) = \{haut, droite\}$, autrement dit l'agent 1 ne peut choisir que l'action $haut$ ou l'action $droite$ lorsqu'il est sur la case 0 (Précisons que $(0,*)$ représente ici tous les états où l'agent 1 est sur la case 0, quelque soit la position de l'agent 2).

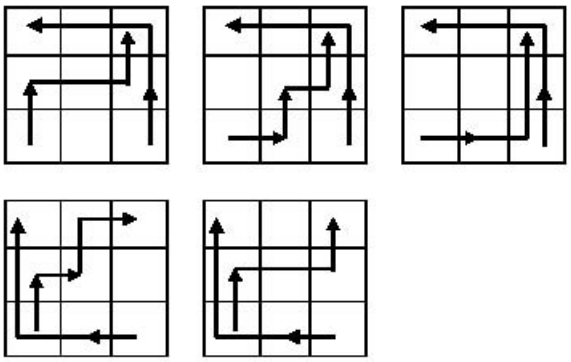

FIG. 3 – Équilibres de Nash pour le jeu de coordination.

L'environnement proposé ici peut être modélisé par un jeu stochastique $G = \langle S, \{A^i\}_{i\in\{1,2\}}, \{R^i\}_{i\in\{1,2\}}, T\rangle$ où les états sont les positions conjointes des agents, $S = \{(0,1),(0,2),\ldots\}$ et le modèle de transition T est déterministe, i.e.
$\forall\,(s,a^1,a^2)\,\exists!\boldsymbol{s'}$ tel que $T(s,a^1,a^2,s') = 1$.

Dans le jeu stochastique représentant cet environnement, un équilibre de Nash est une paire de "trajectoires optimales" où chaque agent atteint son but en un *minimum* de pas, sans détour ni collision. En effet, dans cette situation chaque trajectoire est une meilleure réponse à la trajectoire adverse. La figure 3 représente 5 des 10 équilibres de Nash du jeu de coordination, les 5 autres équilibres étant obtenus par symétrie. Dans cette configuration les utilités pour les joueurs sont de haut en bas : (80,80), (80,100), (90,80) et (90,80), le jeu le plus à droite de la figure 3 et son symétrique sont des configurations où le 2ème joueur suit les pas du 1er joueur dans un plus grand nombre de cases (soit ici 2 cases) et par conséquent ces deux jeux sont considérés comme Pareto-optimaux.

Nous avons étudié dans ce jeu la convergence empirique vers un équilibre de Nash pour différents algorithmes d'apprentissage (Q-learning classique, Q-learning par jeu fictif, Q-learning par jeu adaptatif) et différents modes d'exploration (exploitation pure, GLIE, exploration stochastique stationnaire).

Dans un premier temps, les agents passent par

TAB. 1 – Pourcentage de convergence vers un EN en self-play.

Apprentissage	Exploration	EqNash
Jeu fictif pur	sans	10%
Jeu fictif pur	GLIE	66%
Jeu fictif pur	ϵ fixé (0.15)	98%
Jeu adaptatif $(p = 32, l = 16)$	ϵ fixé (0.15)	100%

TAB. 2 – Pourcentage de convergence vers un EqNash Pareto-optimal.

Apprentissage	Nash Optimal
Jeu fictif pur (ϵ fixé à 0.5)	10%
Jeu adaptatif $(p = 32, l = 16)$ et ϵ fixé (0.15)	100%

une phase d'apprentissage *off-line* de 5000 épisodes. Au début de cette phase, les Q-valeurs des agents sont initialisées à 0, et les croyances sur les stratégies adverses sont initialisées à l'équiprobabilité. Au cours de l'apprentissage, les agents mettent à jour leurs Q-valeurs et leurs croyances, et sont replacés aléatoirement sur la grille après chaque épisode en conservant les données apprises. La durée moyenne d'un épisode durant l'apprentissage est de 4 pas (environ 20000 pas par apprentissage). Il y a 424 couples (s, a^1, a^2), donc après 5000 épisodes, chaque couple *(état, actions conjointes)* a été visité $20000/424 \approx 47$ fois. On alors $\alpha\,(s, a^1, a^2) \approx 1/47 \approx 0.02$. Les Q-valeurs et les probabilités ne sont donc quasiment plus modifiées. A la fin de l'apprentissage, on observe les politiques apprises par les agents partant de leurs positions initiales. Cette simulation est réitérée 50 fois, après quoi on évalue le pourcentage de simulations ayant mené à un équilibre de Nash.

4.2 Résultats sur la convergence dans le jeu de coordination

Le tableau 1 montre la performance en *self-play*[4] du Q-learning par jeu adaptatif par rapport au Q-learning par jeu fictif utilisant différents modes d'exploration. On voit que seule l'exploration stochastique stationnaire permet au Q-learning par jeu fictif pur de converger avec une très forte probabilité (98%), sans pour autant la garantir. Les simulations pour le Q-learning par jeu adaptatif en *self-play* convergent vers un équilibre de Nash dans 100% des expériences.

Le tableau 2 montre, quant à lui, la convergence vers l'équilibre de Nash Pareto-optimal. Comme on le voit, le fictif à exploration stochastique bien qu'il converge vers le Nash dans 98% des cas, il ne trouve le Nash Pareto-optimal que dans 20% des cas. Ceci serait sans aucun doute plus faible s'il n'y avait qu'un seul équilibre Pareto-optimal parmi les dix. L'adaptatif est assuré, quant à lui, de converger vers l'un des deux équilibres Pareto-optimal. Il faudra toutefois, pousser l'expérimentation plus loin et voir si une telle convergence se maintient même dans le cas d'un seul équilibre Pareto-optimal.

La figure 4 montre l'influence comparée du paramètre ϵ pour le Q-learning par jeu fictif (avec exploration stationnaire) et le Q-learning adaptatif avec une mémoire de 32 tours, et une taille d'échantillonnage de 16 actions. On constate qu'il converge dans plus de 98% des cas pour des valeurs de ϵ prises dans l'intervalle $[0.15, 0.55]$. Une valeur d'ϵ trop petite s'approche d'une exploitation pure des données apprises, ce qui explique la dégradation des performances pour $\epsilon < 0.15$. La convergence pour des valeurs de ϵ allant jusqu'à 0.55 s'explique par la structure spécifique du jeu. Le nombre d'actions disponibles pour un agent varie de 2 à 4 selon la position dans la grille. Ainsi, dans le pire des cas (4 actions disponibles pour la position 4), choisir l'action réalisant la meilleure réponse avec une probabilité $1 - \epsilon = 0.45$ et une action aléatoire avec une probabilité $\epsilon = 0.55$ revient à choisir l'action de meilleure réponse avec une probabilité de $0.45 + 0.55/4 \approx 0.59$. Pour 3 actions possibles, elle monte à $0.45 + 0.55/3 \approx 0.63$, et pour 2 actions possibles, à $0.45{+}0.55/2 \approx 0.73$. La meilleure réponse est donc choisie avec une probabilité moyenne sur toutes les positions de $(0.59 + 4 * 0.73 + 3 * 0.63)/(1 + 4 + 3) \approx 0.66$ malgré un paramètre d'exploration stochastique $\epsilon = 0.55$. Les probabilités entretenues sur les actions adverses gardent donc une bonne représentativité même si ϵ est élevé. Notons ici que lorsque le nombre d'actions augmente, la probabilité de choisir l'action de meilleure réponse tend vers ϵ.

La figure 5 indique l'influence de la taille mémoire m et de la taille d'échantillonnage s pour une valeur fixée de ϵ. On observe que la performance de coordination est de 100% à par-

[4]Les deux agents utilisent le même algorithme d'apprentissage

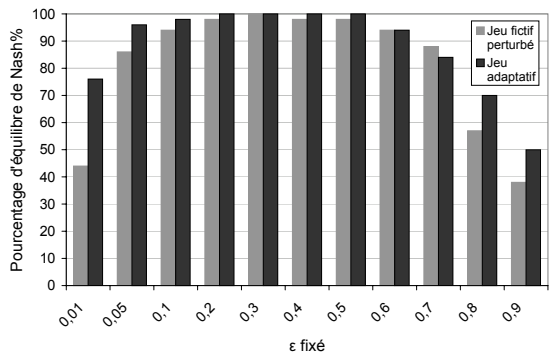

FIG. 4 – Influence du paramètre ϵ sur la convergence en self-play en Q-learning par jeu fictif et Q-learning par jeu adaptatif.

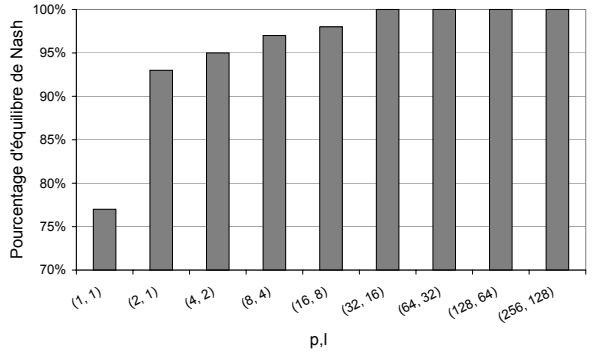

FIG. 5 – Influence de de la taille mémoire et de la taille d'échantillonnage sur la convergence du Q-learning par jeu adaptatif en self-play.

tir de $(m, s) = (32, 16)$. Pour un échantillonnage limité $(s < 16)$, les échecs de convergence s'expliquent par la nature très approximative des croyances sur les actions adverses. A l'inverse, une grande taille mémoire permet d'avoir une statistique plus fine de ces actions. Notons que le Q-learning par jeu fictif correspond au Q-learning par jeu adaptatif avec mémoire et taille d'échantillonnage infinies. L'avantage fondamental de la mémoire limitée du jeu adaptatif, *quelque soit sa taille*, est de supprimer l'influence des premières actions dans les croyances sur les stratégies adverses. En effet, celles-ci sont effectuées lors des toutes premières étapes d'apprentissage alors que les agents ont principalement une attitude d'exploration de l'environnement, et ne sont donc pas représentatives des politiques courantes des agents.

5 Conclusion

Nous avons présenté dans cet article un algorithme d'apprentissage multiagent appelé Q-learning par jeu adaptatif. Cet algorithme part du jeu fictif en faisant des hypothèses qui ne sont autres que des applications respectives des principes de base : information, évaluation et décision qui sont ici applicables à chaque instant. On pourrait voir ces trois principes ici de la façon suivante :

Information : Chaque agent à une profondeur de mémoire p, c'est à dire qu'il n'a pas accès aux événements antérieurs à $t - p, \ldots, t - 1$;

Évaluation : Chaque agent tire au sort un échantillon de taille l d'actions utilisées par ses partenaires voire adversaires, parmi celles utilisées en $t - p, \ldots, t - 1$.

Décision : L'agent joue alors une meilleure réponse avec une probabilité de $1 - \epsilon$ à la distribution qu'il a échantillonnée ; et joue au hasard une action aléatoire avec la probabilité ϵ qui reflète un *taux d'erreur*.

Selon Young et les autres chercheurs travaillant sur le jeu adaptatif, Les trois paramètres p, l, ϵ définissent un processus dynamique dit "processus adaptatif de mémoire p, de taille d'échantillonnage l et de taux d'erreur ϵ". Dans ce cas, le jeu fictif pur peut être assimilé à un processus adaptatif à mémoire infinie, échantillonnage exhaustif et sans erreur.

Les résultats expérimentaux que nous avons présentés pour un tel algorithme adaptatif, montrent sa convergence vers un équilibre de Nash Pareto-optimal où les deux joueurs utilisent le même algorithme d'apprentissage. Nous pouvons postulé dès maintenant bien que nous visons à le faire dans le cadre de nos futurs travaux, que la convergence peut être prouvée formellement, notamment sur la base des travaux de l'équipe de Sandholm [3, 15] en jeux stochastiques coopératifs.

Nous avons argumenté pourquoi à l'inverse des autres approches, notre algorithme converge vers l'équilibre de Nash optimal si celui existe et dès lors, il n'a nullement besoin d'être guidé par le concepteur pour y parvenir.

6 Remerciements

Nous aimerions remercier tout particulièrement Junling Hu, pour ses conseils et suggestions,

ainsi que tous les membres du DAMAS pour les discussions constructives et l'aide technique. Merci aussi aux relecteurs pour leur lecture perspicace. à laquelle cette version finale doit beaucoup. Cette recherche est supportée par le conseil de recherche en sciences naturelles et en génie du Canada (CRSNG).

Références

[1] G. W. Brown. Iterative solution of games by fictitious play. In T. C. Koopmans, editor, *Activity Analysis of Production and Allocation*, chapter XXIV. Wiley, New York, 1951.

[2] C. Claus and C. Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *AAAI/IAAI*, pages 746–752, 1998.

[3] V. Conitzer and T. Sandholm. AWESOME : a general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. In *Proceedings of the International Conference on Machine Learning (ICML 2003)*, pages 83–90, 2003.

[4] D. Fudenberg and D. K. Levine. *The Theory of Learning in Games*. The MIT Press, Cambridge, Massachussets, 1998.

[5] D. Fudenberg and J. Tirole. *Game Theory*. MIT Press, 1994.

[6] O. Gies and B. Chaib-draa. Jeux adaptatifs comme méthode d'apprentissage multiagent. *Revue d'intelligence artificielle (à paraître), numéro spécial "Décision et planification dans l'incertain"*, 2005.

[7] A. Greenwald and K. Hall. Correlated-Q-learning. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 242–249, 2003.

[8] J. Hofbauer and W. H. Sandholm. On the global convergence of stochastic fictitious play. *Econometrica*, 70(6) :2265–2294, November 2002.

[9] J. Hu and M. P. Wellman. Nash Q-learning for general-sum stochastic games. *J. Mach. Learn. Res.*, 4 :1039–1069, 2003.

[10] L. P. Kaelbling, M. L. Littman, and A. P. Moore. Reinforcement learning : A survey. *Journal of Artificial Intelligence Research*, 4 :237–285, 1996.

[11] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning (ML-94)*, pages 157–163, New Brunswick, NJ, 1994. Morgan Kaufmann.

[12] M. L. Littman. Friend-or-foe : Q-learning in general-sum stochastic games. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 322–328, 2001.

[13] L. S. Shapley. Stochastic games. In *Proceedings of the National Academy of Science*, volume 39, pages 327–332, 1953.

[14] G. Tesauro. Extending Q-learning to general adaptive multi-agent systems. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[15] X. Wang and T. Sandholm. Reinforcement learning to play an optimal nash equilibrium in team markov games. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 1571–1578. MIT Press, Cambridge, MA, 2003.

[16] C. J. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3/4) :279–292, 1992.

[17] M. Weinberg and J. S. Rosenschein. Best-response multiagent learning in non-stationary environments. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS-04)*, Columbia University, New York City, July 2004.

[18] H. P. Young. *Individual Strategy and Social Structure : An Evolutionary Theory of Instituions*. Princeton University Press, Princeton, New Jersey, 1998.