
Un formalisme pour la construction automatique d'interactions dans les SMA réactifs

Vincent Thomas — Christine Bourjot — Vincent Chevrier

LORIA, UMR 7503
BP239
54506 Villers-les -Nancy CEDEX
{vthomas,bourjot,chevrier}@loria.fr

RÉSUMÉ. Nous proposons un nouveau formalisme de représentation des actions et des interactions dans les SMA réactifs inspiré des processus de décision Markovien décentralisés (DEC-MDP). Ce formalisme appelé Interac-DEC-MDP permet de représenter dans un même cadre homogène les actions individuelles et les interactions directes entre agents. Ainsi il permet de calculer automatiquement les prises de décisions des agents relativement aux actions et aux déclenchements et résolution des interactions par l'introduction de la rationalité au travers de la notion de récompense. Un premier problème simple de partage de ressources impliquant 2 agents a été modélisé selon le formalisme proposé et les comportements des agents ont été construits automatiquement par *Q*-learning. Les premiers résultats bien qu'obtenus avec des hypothèses limitatives montrent qu'il est possible à partir d'apprentissages simples de construire automatiquement des comportements collectifs pertinents.

ABSTRACT. In this paper, we propose an original formalism inspired by Markov Decision Process in order to represent homogeneously actions and direct interactions between agents. Thanks to the introduction of individual rewards characterising the problem to be solved, this formalism is a first step towards the automatic computation of policies and use of available interactions. A simple problem of distributing resources among 2 agents has been modelled and solved by a *Q*-learning based approach. First results show that it is possible to benefit from simple individual learnings to produce efficient collective behaviour.

MOTS-CLÉS : Interaction, Apprentissage, SMA réactifs, Processus Décisionnels de Markov

KEYWORDS: Interaction, Learning, Reactive MAS, Markov Decision process

1. Introduction

Le contexte de notre travail est celui des systèmes multi-agents réactifs, sans mémoire et sans représentation globale du système, de la tâche à effectuer ou des autres agents. Dans ce cadre, nous nous focalisons sur la résolution collective de problèmes. Il s'agit de définir les comportements individuels et les interactions du système pour que le système parvienne à résoudre de manière décentralisée le problème posé à la collectivité. Nous nous concentrons sur des formalismes mathématiques pour décrire un problème et cherchons à développer des algorithmes permettant de construire automatiquement le comportement des agents du système. Cette résolution peut être la conséquence d'interactions médiées par l'environnement comme dans les approches stigmergiques [LUM 94] [DOR 99] ou d'interactions liées à la position des agents dans l'environnement comme [REY 87] [GEC 03]. Elle peut aussi être la conséquence d'interactions ponctuelles lorsqu'un agent essaie d'influencer directement le comportement d'un autre agent [SIM 00] [BUR 91].

Les modèles markoviens et leurs extensions [BER 00] en introduisant la rationalité au niveau individuel permettent à des agents d'apprendre automatiquement les comportements à adopter pour résoudre une tâche donnée [BUF 03] mais ne représentent pas de manière explicite ces interactions.

Dans cet article, nous proposons le formalisme Interac-DEC-MDP dans lequel la notion d'interaction directe entre agents a été ajoutée. Les agents peuvent alors mettre à jour leurs comportements vis à vis des interactions et prendre des décisions collectives au bénéfice de l'avancement de la tâche collective en cours. Ces soucis sont relativement proches de ceux d'autres approches centrées sur l'interaction comme [SIM 00] et [GEO 03]. Cet article se décompose en quatre parties. Dans la première partie, nous présentons le formalisme Interac-DEC-MDP. La partie suivante propose un processus d'apprentissage fondé sur des Q-learning. Nous présentons enfin un exemple simple d'utilisation des Interac-DEC-MDP et discutons du formalisme.

2. Interac-DEC-MDP

2.1. Le formalisme Interac-DEC-MDP

Nous nous intéressons à des problèmes de **coopération** (à savoir que l'objectif est de maximiser une fonction globale) dans un système **perçu globalement** par chaque agent, pour lequel les récompenses sont **partiellement observées** et dont la résolution est la conséquence de décisions individuelles. Nous proposons le formalisme Interac-DEC-MDP étendant le cadre des DEC-MDP par l'ajout du concept d'interaction.

Celui-ci est constitué de deux modules :

- Un "module d'action" équivalent à un DEC-MDP $\langle S, A, T, R, n \rangle$ ([BER 00]) dans lequel la fonction de récompense globale du système R peut se décomposer en récompenses locales additives ($R = \sum_i r_i$). Chaque agent i n'accède qu'à r_i . Cette

décomposition fait le lien entre le niveau global (la tâche caractérisée par R) et le niveau local de l'agent (r_i).

– Un "module d'interaction" utilisant des interactions directes réactives définies comme "des envois de signaux impliquant deux agents, l'agent émetteur à l'origine de l'interaction et l'agent receveur, et conduisant à une prise de décision collective d'action jointe coordonnée".

L'intérêt des interactions réside dans la prise de décision collective permettant à deux agents de prendre en compte des considérations plus globales alors qu'ils n'ont accès qu'à des récompenses individuelles.

2.2. Module d'interaction

Les interactions considérées sont ponctuelles. Une interaction I_k admet l résultats $RI_{k,l}$. Chacun correspond à une action jointe coordonnée et est défini par une matrice de transition $T_{RI_{k,l}} : S \times Agent_{[0..n]} \times Agent_{[0..n]} \rightarrow P(S)$ ($P(S)$ désigne une distribution de probabilité sur S) faisant évoluer le système en fonction des indices des agents impliqués dans l'interaction.

Chaque agent est tour à tour émetteur d'interaction selon trois étapes :

– **Une étape de déclenchement** consistant à choisir l'interaction que l'agent va déclencher et l'agent receveur. Cette prise de décision est représentée par la politique $\pi_{trig,i} : S \rightarrow P(I, Agent_{[0..n]})$.

– **Une étape de résolution des interactions** consistant en une prise de décision impliquant les deux agents au sujet du résultat de l'interaction I_i qui a été déclenchée. Elle est représentée par une politique $\Pi_{I_k} : S \times Agent_{[0..n]} \times Agent_{[0..n]} \rightarrow P(RI_k)$.

– **Une transition** faisant évoluer l'état du système. Elle correspond à l'exécution de l'action jointe coordonnée décidée $s \leftarrow T_{RI_{i,l}}(s, i, n_i)$

3. Processus de résolution

Les Interac-DEC-MDP permettent de représenter des actions individuelles et des interactions dans un même formalisme. Ils définissent une nouvelle classe de problèmes dans laquelle les agents peuvent interagir directement. Résoudre un problème avec ce formalisme consiste à trouver pour chaque agent i ses politiques individuelles π_i et $\pi_{trig,i}$, et les politiques d'interaction Π_{i,j,I_k} pour chaque couple. Comme dans [BUF 03], nous cherchons en outre à construire ces politiques de manière **décentralisée** par Q-learning [SUT 98] sans représenter au sein d'un agent les comportements des autres.

3.1. Représentations internes des politiques

– **Les politiques individuelles** π_i peuvent être représentées comme dans un Q-learning classique où chaque agent dispose d'une table de Q – valeurs. $Q_i(s, a_i)$ représente le gain espéré de l'agent i à effectuer l'action a_i dans l'état s .

– **Les politiques de déclenchement d'interaction** $\pi_{i, trig}$ font appel à des prises de décision individuelles. Elles peuvent être représentées de manière analogue à l'aide de Q – valeurs individuelles de déclenchement : $Q_{trig,i} : S \times I \times Agent_{[0..n]} \rightarrow \mathfrak{R}$.

– **Les politiques de résolution d'interaction** Π sont représentées de manière distribuée : Nous faisons l'hypothèse qu'un échange d'information concernant les Q – valeurs des deux agents impliqués dans l'interaction (l'émetteur E et le receveur R) suffit pour prendre une décision collective efficace. Chaque agent dispose alors de Q – valeurs d'interaction $Q_{I_k,i} : S \times RI_{I_k} \times \{R, E\} \rightarrow \mathfrak{R}$. Quand deux agents interagissent, la politique d'interaction est reconstruite à partir de ces deux valeurs. Une politique gloutonne consiste alors à choisir le résultat d'interaction maximisant la somme des espérances individuelles. Elle peut conduire un agent à émettre un comportement altruiste pour améliorer les performances globales du système au détriment de la sienne.

3.2. Construction des politiques

Dans cette partie, nous présentons un processus de résolution consistant à apprendre les politiques π , π_{trig} et Π . L'apprentissage des interactions est relativement simple puisqu'il consiste simplement à tirer parti de politiques individuelles déjà apprises par Q-learning pour améliorer le comportement collectif du système. Cette construction des politiques se fait en trois étapes consécutives :

1) **Un apprentissage des politiques individuelles** : Les agents apprennent simultanément et indépendamment leurs politiques individuelles π_i en fonction des récompenses individuelles r_i reçues, sans qu'aucune interaction ne soit émise.

2) **Un apprentissage des politiques jointes** $\Pi_{i,j}$: Les politiques individuelles π_i sont ensuite figées et les agents apprennent à résoudre conjointement les interactions déclenchées. Après chaque interaction, un résultat d'interaction est choisi. Ce résultat RI_i est connu par les deux agents concernés et fait évoluer le système de l'état s vers l'état s' . Chaque agent sait alors quelle est sa récompense escomptée en fonction de l'état s' : $max_a(Q(s', a))$. Les agents peuvent donc mettre à jour leurs Q – valeurs d'interaction Q_I à l'aide des Q – valeurs précédentes : $Q_{I_k,i}(s, RI_{k,l}, E) \leftarrow (1 - \alpha) \cdot Q_{I_k,i}(s, RI_{k,l}, E) + \alpha \cdot (r_i + \gamma \cdot max_{a'}(Q_i(s', a'))$ (α désigne un coefficient d'apprentissage et γ le discount factor [SUT 98]).

3) **Un apprentissage des déclenchements** : Les politiques individuelles π_i et les politiques collectives $\Pi_{i,j}$ sont figées et les agents apprennent à déclencher les interactions. Chaque agent peut savoir a posteriori la récompense escomptée lorsqu'il déclenche une certaine interaction dans un certain état et mettre à jour ses Q -valeurs de déclenchement : $Q_{trig,i}(s, I, n) = (1 - \alpha) \cdot Q_{trig,i}(s, I, n) + \alpha \cdot (\gamma \cdot max_{a'}(Q_i(s, a'))$

4. Exemple

Afin de mettre en évidence l'intérêt du concept d'interaction dans ce cadre, nous avons envisagé un problème simple impliquant deux agents où l'objectif est de montrer que l'ajout d'interactions permet d'améliorer les performances globales du système. Deux agents peuvent accéder à la nourriture en traversant un couloir immergé. Ils sont caractérisés par leur faim et leur possession de nourriture. Chaque agent peut soit **plonger** pour avoir de la nourriture (avec une récompense négative dépendant de ses capacités), soit **attendre** (sa faim augmente de 1 et l'agent reçoit une récompense négative) ou **manger** la nourriture qu'il possède (avec une récompense positive). Les actions des agents sont indépendantes les unes des autres. Enfin, les agents peuvent échanger de la nourriture. Il s'agit d'une interaction directe (I_1) avec deux résultats possibles : soit l'échange a effectivement lieu entre les deux agents ($RI_{1,1}$), soit l'échange n'est pas effectif ($RI_{1,2}$).

Deux scénarios sont envisagés. Dans le premier, les deux agents sont des agents bons plongeurs. Dans le second, le coût de plongée associé à un agent, l'agent qualifié d'agent mauvais plongeur, est plus important. Pour chacun de ces scénarios, les trois apprentissages successifs présentés précédemment ont été conduits. Enfin, nous avons comparés ces résultats à ceux optimaux obtenus pour un MDP centralisé équivalent (avec prise en compte d'interactions de manière centralisée). Vu le faible nombre d'agents, il est en effet encore possible d'utiliser une approche centralisée. Cependant, celle-ci ne pourra plus être envisagée si le nombre d'agents croît.

Les résultats obtenus, détaillés dans [THO 04], montrent que dans le cas homogène, l'apprentissage effectué conduit au comportement optimal calculé à partir du MDP centralisé pour lequel les agents ne déclenchent pas d'interactions. Dans le cas hétérogène, l'ajout de l'interaction améliore les performances globales : l'agent bon plongeur, en donnant sa nourriture, réalise une interaction "altruiste" intéressante pour le système. Enfin, l'apprentissage limite les interactions émises dans le système aux interactions utiles. Dans ce dernier cas, il est à noter que le comportement optimal n'est pas atteint car les comportements individuels ne sont pas remis en cause lors de l'ajout d'interactions. D'autres apprentissages plus élaborés sont alors nécessaires.

Cet exemple montre qu'il est possible de construire des comportements collectifs pertinents à partir de Q-learning sans représentation explicite des agents. Un simple échange d'information concernant les Q-valeurs des agents est suffisant pour décider de manière efficace de l'issue d'une interaction. De plus, l'exemple met en évidence l'intérêt de la notion d'interaction dans les MDPs par rapport à celle d'action classique qui se limite à un point de vue purement local et égoïste.

5. Conclusion

Dans cet article, nous avons proposé un nouveau formalisme de représentation des actions et des interactions dans les SMA réactifs inspiré des processus de décision Markoviens décentralisés (DEC-MDP). Ce formalisme original, l'Interac-DEC-MDP, est constitué de deux modules, un module d'action permettant aux agents de choisir

simultanément des influences à exercer sur le système et un module d'interaction permettant des prises de décisions collectives. Ce dernier module autorise d'autres considérations que celles simplement égoïstes et peut conduire de ce fait à une amélioration globale du système par rapport à des apprentissages purement individuels.

Un problème simple a été modélisé selon le formalisme proposé et les comportements des agents ont été construits automatiquement par Q-learning. Les premiers résultats montrent qu'il est possible à partir d'apprentissages simples de générer des comportements collectifs pertinents, en limitant les interactions utilisées à celles utiles pour le système. L'exemple présenté reste relativement simple : il n'implique que deux agents et les actions d'un agent n'influencent pas le résultat des actions de l'autre agent. Ce travail reste une première étape vers la construction automatique de comportements individuels et des interactions pour des agents réactifs. Nous envisageons par la suite d'autres problèmes plus complexes impliquant un plus grand nombre d'agents. Nous envisageons de même un apprentissage cyclique plutôt que séquentiel des différentes politiques pour obtenir des comportements collectifs plus efficaces.

6. Bibliographie

- [BER 00] BERNSTEIN D., ZILBERSTEIN S., IMMERMANN N., « The Complexity of Decentralized Control of Markov Decision Processes », *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI'00), Stanford, California, 2000*, p. 32-37.
- [BUF 03] BUFFET O., « Une double approche modulaire de l'apprentissage par renforcement pour des agents intelligents adaptatifs », PhD thesis, Univ. H. Poincaré - Nancy 1, 2003.
- [BUR 91] BURA S., DROGOUL A., FERBER J., JACOPIN E., « Eco-résolution : un modèle de résolution par interactions », *8ème congrès RFIA, 1991*, p. 1299-1308.
- [DOR 99] DORIGO M., CARO G. D., GAMBARDELLA L. M., « Ant algorithms for discrete optimization », *Artificial Life*, vol. 5, n° 2, 1999, p. 137-172, MIT Press.
- [GEC 03] GECHTER F., CHEVRIER V., CHARPILLET F., « Une architecture réactive pour la localisation en robotique mobile », *JFSMA 2003, 2003*, p. 345-358.
- [GEO 03] GEORGÉ J.-P., GLEIZES M.-P., GLIZE P., « Conception de systèmes adaptatifs à fonctionnalité émergente : la théorie Amas », *Revue d'Intelligence Artificielle RIA*, vol. 17, n° 4, 2003, p. 591-626.
- [LUM 94] LUMER E. D., FAIETA B., « Diversity and adaptation in populations of clustering ants », *3rd conference of Simulation of adaptive behavior : from animals to animats 3*, MIT Press, 1994, p. 501-508.
- [REY 87] REYNOLDS C., « Flocks, Herds, and Schools : A Distributed Behavioral Model », *SIGGRAPH, Computer Graphics*, vol. 21, 1987, p. 25-34.
- [SIM 00] SIMONIN O., FERBER J., « Modeling Self Satisfaction and Altruism to handle Action Selection and Reactive Cooperation », *6th conference of simulation of adaptive behaviour : from animals to animats 6*, MIT Press, 2000, p. 314-323.
- [SUT 98] SUTTON R., BARTO G., *Reinforcement Learning : an introduction*, Bradford Book, MIT Press, Cambridge, MA, 1998.
- [THO 04] THOMAS V., BOURJOT C., CHEVRIER V., « Un formalisme pour la construction automatique d'interactions dans les SMA réactifs - étendu », rapport interne, 2004, INRIA.