Optimisation en ligne pour la décision distribuée dans l'incertain

Emmanuel Rachelson
ONERA-DCSD, Toulouse
Supaéro, Toulouse
emmanuel.rachelson@onera.fr

Sous la direction de : P. Fabiani, J.-L. Farges, F. Garcia

Résumé

Dans le cadre de problèmes de décision dans l'incertain impliquant plusieurs agents engagés dans des tâches nécessitant coordination et coopération, on s'intéresse à des algorithmes de recherche de politiques quasi-optimales. On part du formalisme des processus décisionnels de Markov et des outils associés afin de développer les algorithmes et modèles nécessaires au traitement de ce type de problème. Les deux besoins d'une prise en compte d'une variable temporelle et d'un algorithme d'optimisation de la politique conjointe des agents ont été en partie traitées en stage de M2R et leurs perfectionnement, approfondissement et test font l'objet des premiers travaux de thèse.

1 Introduction et problème mono-agent

Afin d'apporter une solution robuste aux problèmes de décision en environnement incertain, une méthode couramment utilisée est la recherche de politiques, associant à chaque état de l'univers de l'agent une action, considérée optimale selon un certain critère. En fonction de la complexité du problème cette politique peut être plus ou moins longue et difficile à déterminer. Les processus décisionnels de Markov (MDP) fournissent un cadre mathématique et théorique aux problèmes de décision dans l'incertain en associant à chaque triplet (état de départ, action, état d'arrivée) une probabilité de réalisation et une récompense. Les algorithmes de programmation dynamique permettent alors de rechercher la politique optimale selon un critère qui maximise une fonction des récompenses. Basés sur le principe de l'équation de Bellman, les algorithmes d'itération de la valeur et de la politique (et leurs variantes) parcourent l'espace d'état et d'action pour améliorer incrémentalement la valeur du critère espérée en appliquant la politique π en chaque état s de l'espace d'état. Le problème de la décision mono-agent s'avère particulièrement ardu pour des systèmes complexes comportant une modélisation riche, en effet, le *curse of dimensionality* de Bellman illustre bien le fait que la complexité du problème croit exponentiellement avec les tailles des espaces d'état et d'action. De nombreux travaux se sont attachés à conserver au problème ses propriétés de structure afin de les exploiter dans la résolution, on peut notamment citer les MDP décomposés qui divisent le problème en sous-problèmes faiblement couplés de tailles réduite ([Hauskrecht *et al.*, 1998], [Parr, 1998], [Dean and Lin, 1995]), les principes de hiérarchisation des espaces d'état (issus des MDP décomposés) ou d'action (comme l'approche MAXQ, [Diettrich, 1998]), ou encore les approches factorisées symboliques utilisant les arbres de décision afin de représenter et de manipuler le problème sous une forme compacte et structurée ([Boutilier *et al.*, 1999]).

Le cadre de cette étude porte sur des problèmes de décision où la résolution est distribuée entre plusieurs agents. L'objectif à terme est de produire des réponses réactives (en ligne), utilisables et quasi-optimales à des problèmes de coordination, coopération ou missions communes pour des systèmes comportant moins d'une dizaine d'agents.

2 Problématique multi-agents

Les missions sur lesquelles s'appliquent les travaux de thèse sont des ensembles d'engins autonomes de type drone ou robot terrestre où les agents doivent délibérer pour combiner des tâches de navigation et de prise d'information en coopération pour réaliser un objectif. Les projets comme Robea AcroBate, Ressac, la coupe de France de robotique ou des problèmes plus théoriques comme l'exemple des "deux taxis" fournissent un grand nombre de cas de tests.

On cherche à traiter ce type de problèmes dans le cadre MDP. On souhaite également que les agents soient capables de réparer un plan qui s'avère être inefficace ou de corriger un plan si le modèle de l'univers change. Enfin, par souci de capacité de calcul et de réalisme, on adopte un formalisme où chaque agent résout un problème qui lui est propre sans nécessairement disposer de toute l'information à propos du système global. La problématique peut donc se reformuler de la façon suivante : "Comment construire une politique commune quasi-optimale en établissant des plans individuels et en communiquant le minimum d'information possible ?".

La recherche et l'établissement d'une politique individuelle optimale dans notre cadre implique immédiatement que l'action optimale pour l'agent A dépend de ce que fait l'agent B. Si, selon une approche naïve, on cherchait à rajouter tous les états de B au problème de A afin de résoudre ce dernier dans le cadre bi-agents (et réciproquement) on se retrouve avec un problème dont la taille est exponentielle en le nombre d'agents et qu'on ne sait pas résoudre pour des problèmes réalistes. On se retrouve d'ailleurs avec une approche quasi-centralisée ce qu'on a cherché à éviter. La clé semble donc être dans la résolution par chaque agent d'un MDP individuel mono-agent tenant compte de la déclaration du plan de l'autre agent. L'idée principale est de propager les

conséquences des plans de B dans le MDP de A pour que A résolve un problème mono-agent où il "se croit seul" et où l'univers intègre déjà les probabilités de changement issues du plan de B. On voit apparaître simultanément deux besoins : le premier concerne un algorithme d'échange d'information et d'optimisation de la politique commune. Le second découle des conséquences de la propagation du plan de B dans le MDP de A. Prenons par exemple un monde grille où évoluent deux agents. Une unique récompense est placée à un coin de la grille et l'objectif est de prendre la récompense le plus rapidement possible pour pouvoir couper le moteur. Initialement, aucun plan n'a été déclaré de part et d'autre donc, comme indiqué plus haut, chaque agent résout le problème individuellement et construit une politique, optimale pour lui, où il va chercher la récompense et s'arrête. Il est immédiat que l'union de ces deux politiques n'est pas la politique conjointe optimale. La propagation du plan de B dans le MDP de A revient à intégrer dans le MDP de A l'information quantitative : "étant donné le plan de B, la récompense a telle probabilité d'avoir disparu à telle date".

Cet exemple introduit naturellement la variable temporelle dans le problème. Elle est intrinsèquement présente dès que l'on considère la coordination des agents. Ainsi le second besoin pour la résolution de notre problème est l'intégration de la variable temporelle dans l'espace d'état.

3 Contributions et perspectives

Les deux contributions apportées en stage de M2R et en début de thèse concernent les deux besoins identifiés à la section précédente. Pour le besoin d'intégration de la variable temporelle dans l'état, deux modèles de prise en compte du temps ont été proposés et leurs caractéristiques étudiées. Dans un premier temps, une distinction a été faite entre la notion de durée et celle de date (qui implique une origine). Cette distinction justifie le fait que l'on s'éloigne des modèles stationnaires classiques comme le modèle SMDP.

Le premier modèle développé est une approche intégrale qui traite le temps comme une variable continue et associe à chaque état, non plus la valeur de l'état, mais une fonction de t qui donne la valeur de l'état à l'instant t. Cette approche est une extension du modèle SMDP. Sous certaines hypothèses, il est possible de calculer simplement les fonctions de valeur d'une politique, cependant, dans le cas général, ce modèle nécessite des calculs trop complexes pour être utilisable. Le second modèle est basé sur des distributions finies sur la durée de chaque action et sur une variable temporelle dont le pas est égal au pgcd des durées des transitions. Ce modèle permet d'obtenir une discrétisation temporelle à pas fixe le plus grand possible sans perte d'optimalité. Les problèmes sont alors similaires aux problèmes MDP classiques et on parvient ainsi à traiter des exemples (comme Robea).

La contribution algorithmique concernant l'échange d'information et l'optimisation de la politique commune a été désigné sous le nom d'itération de la chronologie. Cet algorithme décrit comment, à partir d'une politique déterminée, chaque agent construit un message minimal (une chronologie) qu'il émet. A partir de ce message, son homologue met à jour son propre MDP en introduisant les échéances déclarées par le premier agent, prolonge sa propre politique et résout le problème. Les agents itèrent ainsi jusqu'à convergence et trouvent une stratégie localement optimale (dans l'espace des stratégies). De plus, la comparaison entre la chronologie et le déroulement effectif de la mission peut donner lieu à des réparations du plan de chaque agent.

Les pistes de développement à partir de cette base sont nombreuses. Il faut d'une part tester les algorithmes et modèles développés afin d'acquérir une meilleure connaissance des problèmes et des difficultés. D'autre part, le développement des modèles de prise en compte du temps nécessite de nouveaux approfondissements, notamment en intégrant des notions POMDP ([Cassandra *et al.*, 1994]) ou un traitement type "variables continues" ([Munos and Moore, 2002]). Une approche de politique robuste a également été envisagée ([Nillim and El Ghaoui, 2005]). Par ailleurs, l'intérêt pour un système en ligne des notions comme l'apprentissage par renforcement est immédiat ([Kaelbling *et al.*, 1996]). Enfin, l'adaptation des algorithmes heuristiques comme sfDP ([Teichteil-Königsbuch, 2005]) aux problèmes multi-agents est une piste intéressante.

Références

[Boutilier *et al.*, 1999] C. Boutilier, R. Dearden, and M. Goldszmidt. Stochastic dynamic programming with factored representations. In *Journal of Artificial Intelligence*, 1999.

[Cassandra *et al.*, 1994] A.R. Cassandra, L.P. Kaelbling, and M.L. Littman. Acting optimaly in partially observable stochastic domains. In *AAAI*, 1994.

[Dean and Lin, 1995] T. Dean and S.-H. Lin. Decomposition techniques for planning in stochastic domains. In *IJCAI*, 1995.

[Diettrich, 1998] T. Diettrich. The MAXQ method for hierarchical reinforcement learning. In *15th international Conf. on Machine Learning*, 1998.

[Hauskrecht *et al.*, 1998] M. Hauskrecht, N. Meuleau, L.P. Kaelbling, T. Dean, and C. Boutilier. Hierarchical solution of Markov decision processes using macro-actions. In *Uncertainty in Artificial Intelligence*, 1998.

[Kaelbling *et al.*, 1996] L.P. Kaelbling, M. Littman, and A. Moore. Reinforcement learning, a survey. In *Journal of Artificial Intelligence*, 1996.

[Munos and Moore, 2002] R. Munos and A. Moore. Variable resolution discretization in optimal control. In *Machine learning*, 2002.

[Nillim and El Ghaoui, 2005] A. Nillim and L. El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. In *Operations Research*, 2005.

[Parr, 1998] R. Parr. Flexible decomposition algorithms for weakly coupled Markov decision problems. In *Uncertainty in Artificial Intelligence*, 1998.

[Teichteil-Königsbuch, 2005] F. Teichteil-Königsbuch. Approche Symbolique et Heuristique de la Planification en Environnement Incertain. In *PhD Thesis*, 2005.