Calcul d'une politique déterministe dans un MDP avec récompenses imprécises

Pegah Alizadeh **, Aomar Osmani *, Emiliano Traversi*

*Léonard de Vinci Pôle Universitaire, Centre Recherche
** LIPN-UMR CNRS 7030, PRES Sorbonne Paris Cité

Résumé. Pour beaucoup d'applications réelles nécessitant une prise de décision séquentielle dans un cadre incertain, on utilise un processus de décision Markovien avec récompenses inconnues (IRMDP) en calculant naturellement des politiques stochastiques. Une politique stochastique n'est pas facilement interprétable pour l'utilisateur final. Celui-ci a souvent besoin d'une politique déterministe et compréhensible. Pour mieux motiver l'utilisation d'une procédure exacte pour trouver une politique déterministe, nous montrons quelques cas où l'idée intuitive d'utiliser une politique déterministe obtenue après une «déterminisation» (arrondi) de la politique stochastique optimale donne une politique déterministe différente de la politique optimale.

1 Introduction

Les processus de décision de Markov (MDP) se sont avérés être des modèles efficaces pour représenter et résoudre des problèmes de décision séquentiels avec incertitude. Les applications dans divers domaines (robotique, navigation, composition de services, etc.) nécessitant une prise de décision dans un environnement dynamique utilisent naturellement des modèles MDP. Dans le cas de la conduite autonome, par exemple, à chaque étape, les effets probabilistes de l'exécution d'une action (déplacement, rotation) conduit aux étapes suivantes, chacune avec une récompense ou pénalité différente. Ceci motive l'utilisation des MDP qui rendent compte de ces incertitudes dans les paramètres du modèle. Dans ce cas, l'environnement (ici le trafic routier) sera modélisé par un ensemble d'actions probabilistes (par exemple, les probabilités d'accidents et leur coûts) et un ensemble d'états. L'objectif peut être défini comme un problème de maximisation de la somme des récompenses attendue.

En dépit de connaître l'objectif final, spécifier des récompenses ou des punitions pour choisir des actions à partir des états n'est jamais évident. De plus, comme l'a montré Mannor et al. (2007), les stratégies trouvées via un processus d'optimisation sous MDPs avec des paramètres numériques, peuvent parfois être bien pire que la politique anticipée. Parmi les raisons qui peuvent expliquer cette situation, on peut citer : (1) l'insuffisance des données pour estimer les récompenses, (2) la complexité de la construction des modèles [1] et (3) la présence d'élicitations contradictoires entre les utilisateurs. Dans le cas du véhicule autonome, si le modèle est

1. Dans le cas du véhicule autonome, par exemple, la définition des récompenses exactes pour toutes les actions prend du temps et est compliquée à faire. De plus, ces récompenses varient au cours du processus de conduite.