Algorithmes efficaces pour le contrôle et l'apprentissage par renforcement

E. Berthier,
ENS et Inria Paris,

Email : eloise.berthier@inria.fr

Mots Clés : contrôle optimal, apprentissage par renforcement, méthodes numériques

Biographie – Ancienne élève de l'École polytechnique puis du master MVA de l'ENS Paris-Saclay, Eloïse Berthier est étudiante en 2e année de thèse à l'ENS et Inria Paris, sous la direction de Francis Bach, au sein de l'équipe Sierra, équipe spécialisée dans l'apprentissage automatique et l'optimisation. Cette thèse est intégralement financée par la Direction Générale de l'Armement.

Resumé :

L'apprentissage par renforcement désigne pour un agent le fait d'apprendre les actions qu'il doit effectuer dans un environnement incertain, de façon à maximiser sa récompense sur le long terme [4]. Il trouve son origine dans le domaine du contrôle optimal et dans certains travaux en psychologie. L'augmentation des capacités de calcul, l'utilisation d'algorithmes efficaces et de méthodes d'approximation comme les réseaux de neurones ont permis des succès récents dans le domaine des jeux (Go, Starcraft...), sans pour autant systématiquement fournir des garanties théoriques. Quant au domaine du contrôle optimal, pour lequel un modèle de l'environnement est fourni, il a connu des développements théoriques solides dès les années 1960, avec des outils numériques centrés sur les systèmes linéaires, qui ont fait leurs preuves en aéronautique. Le résolution numérique de problèmes de contrôle non-linéaires de grande dimension est plus récente, et reste aujourd'hui relativement ouverte [5]. S'ils sont formulés différemment, les problèmes du contrôle et de l'apprentissage par renforcement sont néanmoins très proches, surtout si on s'intéresse au *model-based reinforcement learning* et au contrôle stochastique, et pourraient bénéficier d'échanges réciproques.

Nous cherchons à développer des algorithmes efficaces pour le contrôle et l'apprentissage par renforcement, en proposant des méthodes qui peuvent s'appliquer en particulier en robotique, et pour lesquelles on cherche à obtenir des garanties théoriques. L'application à la robotique présente certaines particularités. D'une part, les dimensions du système non-linéaire sont telles qu'il est impossible d'espérer résoudre les problèmes exactement, ce qui pousse à chercher des méthodes d'approximation. D'autre part, s'agissant de systèmes physiques, ils sont imparfaitement simulés par des modèles, ce qui limite encore l'intérêt d'une solution numérique exacte. Enfin, les calculs doivent être faits en temps réel, comme dans le *model-predictive control*, voire dans des systèmes embarqués, ce qui limite à la fois la puissance et le temps alloués aux calculs.

Dans un premier temps [2], on considère le problème de la discrétisation d'un processus de décision markovien déterministe à espace d'état continu. On propose d'utiliser une méthode d'approximation max-plus linéaire [1] pour approcher la fonction valeur avec une base de fonctions particulière. Cette méthode fournit une discrétisation du problème qui est plus parcimonieuse qu'une discrétisation naïve de l'espace d'état, dont la taille croît exponentiellement avec la dimension du problème. On propose également un stratégie pour adapter la discrétisation à une instance du problème, afin d'atténuer le fléau de la dimension. Enfin, on applique numériquement cette méthode à quelques problèmes de faible dimension.

Dans un second temps [3], on s'intéresse au problème de l'estimation de régions de stabilité pour des systèmes dynamiques non-linéaires. L'outil de choix pour stabiliser un système dynamique autour d'un point d'équilibre est le régulateur linéaire-quadratique (LQR). Pour des systèmes dynamiques non-linéaires, ce régulateur n'est valable que localement, et estimer cette région de

validité est un problème important en pratique. On propose plusieurs certificats qui garantissent que le contrôleur LQR stabilise le système dans une certaine région. Ces certificats sont rapides à calculer, et robustes sur une classe de systèmes dynamiques dont les dérivées premières ou secondes sont bornées. Associés à un oracle efficace pour borner les dérivées du système, ils fournissent un algorithme simple pour estimer des régions de stabilité. On compare cette méthode avec une approche classique basée sur l'optimisation polynomiale, sur des systèmes de dimensions variées, dont un système robotique.

Références

[1] Marianne Akian, Stéphane Gaubert, and Asma Lakhoua. The max-plus finite element method for solving deterministic optimal control problems : basic properties and convergence analysis. *SIAM Journal on Control and Optimization*, 47(2) :817–848, 2008.

[2] Eloïse Berthier and Francis Bach. Max-Plus Linear Approximations for Deterministic Continuous-State Markov Decision Processes. *IEEE Control Systems Letters*, 4(3) :767–772, July 2020.

[3] Eloïse Berthier, Justin Carpentier, and Francis Bach. Fast and Robust Stability Region Estimation for Nonlinear Dynamical Systems. *accepted to European Control Conference (ECC)*, 2021.

[4] Richard S Sutton and Andrew G Barto. *Reinforcement learning : An introduction*. MIT press, 2018.

[5] Emmanuel Trélat. Contrôle optimal : Théorie & applications. 01 2005.