Echantillonnage adaptatif pour l'identification de la politique optimale dans les PDMs

Aymen Al Marjani, Alexandre Proutiere
ENS Lyon, KTH Royal Institute of Technology

Email : aymen.al_marjani@ens-lyon.fr

Mots Clés : Processus de Décision Markoviens, Exploration Pure, Modèle Génératif.

Biographie – Aymen Al Marjani est doctorant en première année à l'ENS Lyon sur les thématiques de la décision séquentielle, co-encadré par Aurélien Garivier et Emilie Kaufmann. Avant de commencer sa thèse, il a poursuivi le cycle ingénieur à l'Ecole Polytechnique avec un double diplôme à l'ENS Paris-Saclay (master MVA).

Resumé :

Les algorithmes d'apprentissage par renforcement (RL) sont conçus pour interagir avec un système dynamique stochastique inconnu et, grâce à cette interaction, identifier aussi rapidement que possible une politique de contrôle optimale. L'efficacité de ces algorithmes peut-être mesurée (entre autres) par leur *complexité d'échantillonnage*, définie comme le nombre d'échantillons (le nombre de fois que l'algorithme interagit avec le système) requis pour identifier une politique optimale avec un certain niveau de précision et de certitude. Ce travail[1], comme la plupart des travaux connexes dans ce domaine, se concentre sur les systèmes et les objectifs de contrôle qui sont modélisés comme des processus de décision de Markov (PDM) escomptés standard avec des espaces d'état et d'action finis. Différents modèles d'interaction ont été étudiés, mais les analyses de complexité d'échantillonnage ont été principalement menées dans le cadre du modèle dit "génératif", où à chaque étape, l'algorithme dispose d'un simulateur qui lui permet d'observer une transition et une récompense à partir de n'importe quelle paire (état, action) donnée. Nous limitons également notre attention à ce modèle.

Nous étudions la conception d'algorithmes RL avec une complexité d'échantillonnage minimale. Ce problème a attiré beaucoup d'attention au cours des deux dernières décennies. La plupart des études suivent une approche minimax. Par exemple, on sait [1] que pour le pire PDM possible, l'identification d'une politique ε-optimale avec une probabilité $1-\delta$ nécessite au moins $\frac{SA}{\varepsilon^2(1-\gamma)^3}\log(\frac{SA}{\delta})$ échantillons, où S et A sont le nombre d'états et d'actions, respectivement, et γ est le facteur d'escomptage. Depuis l'apparition de la borne inférieure minimax mentionnée ci-dessus, la plupart des chercheurs ont cherché à concevoir des algorithmes correspondant à cette borne. En revanche, nous nous intéressons à l'analyse de la complexité minimale d'échantillon *spécifique-au-problème*. Plus précisément, nous cherchons à comprendre la dépendance de la complexité d'échantillon vis-à-vis du PDM qui doit être appris. Les mesures de performance spécifiques au problème sont beaucoup plus informatives que leurs équivalents minimax, car elles codent et expriment la dureté inhérente du PDM. Les métriques minimax ne représentent que la dureté du pire PDM. En particulier, établir que la complexité d'échantillon d'un algorithme ne dépasse pas la limite inférieure minimax révèle simplement que l'algorithme est performant pour ce pire PDM. Cependant, cela n'indique pas si l'algorithme s'adapte à la dureté du PDM, c'est-à-dire si la politique optimale d'un PDM très facile serait apprise très rapidement. En fait, un algorithme dont la complexité d'échantillonnage correspond à la limite inférieure minimax consiste simplement à échantillonner des paires (état, action) uniformément au hasard, et ne s'adapte pas au PDM.

La complexité d'échantillonnage spécifique-au-problème pour l'identification du meilleur bras dans les problèmes de bandit stochastiques à plusieurs bras (MAB) est maintenant bien comprise [2, 4].

[1] Présentation basée sur le travail [3].

Dans cet article, les auteurs déduisent une limite inférieure de complexité d'échantillon spécifique-au-problème et conçoivent un algorithme de suivi et d'arrêt qui atteint cette limite inférieure. Un premier pas vers l'extension de ces résultats à l'apprentissage dans les PDMs a été fait dans [5]. Les auteurs y présentent BESPOKE, un algorithme qui exploite la structure du PDM en fonction du problème et qui, à son tour, offre des garanties de complexité d'échantillon spécifiques au problème. En effet, la limite supérieure de la complexité de l'échantillon de BESPOKE dépend explicitement des écarts de sous-optimalité des actions sous-optimales et des variances des récompenses et de la fonction de valeur de l'état suivant. Dans ce travail, nous explorons si la méthodologie utilisée dans [2] pour les problèmes MAB peut être étendue aux problèmes RL. Cette méthodologie consiste à prouver d'abord une borne inférieure de complexité d'échantillon spécifique au problème. Cette dernière devrait révéler l'allocation d'échantillon menant à la complexité d'échantillon minimale. On peut ensuite concevoir un algorithme de suivi et d'arrêt qui (i) suit l'allocation optimale de l'échantillon identifiée dans la borne inférieure, et (ii) s'arrête lorsque l'information recueillie est jugée suffisante pour obtenir les garanties PAC souhaitées. Il s'avère que l'extension de cette méthodologie aux problèmes de RL soulève des questions fondamentales, principalement en raison de la difficulté de calculer l'allocation d'échantillon menant à la complexité minimale d'échantillon spécifique au problème. Nous proposons un ensemble d'outils pour résoudre ces problèmes. Nos contributions sont les suivantes :

1. Nous prouvons une borne inférieure de complexité d'échantillon spécifique au problème pour identifier une politique optimale dans un PDM ϕ donné. Cette limite est exprimée par $T(\phi)\log(1/\delta)$, où le *temps caractéristique* $T(\phi)$ encode la dureté du PDM ϕ. $T(\phi)$ est la valeur d'un problème d'optimisation non convexe. Cette complexité rend difficile la conception d'un algorithme de suivi et d'arrêt similaire à celui proposé dans [2] et l'obtention de la borne inférieure de complexité d'échantillon. Pour contourner cette difficulté, nous prouvons une borne supérieure explicite $U(\phi)$ de $T(\phi)$. L'avantage de $U(\phi)$ est double : (i) $U(\phi)$ reste spécifique au problème, et dépend explicitement des fonctionnelles du PDM caractérisant sa dureté. (ii) $U(\phi)$ correspond à une allocation d'échantillon explicite et simple. Cela nous permet de concevoir une procédure qui suit cette allocation.

2. Sur la base de notre analyse de la borne supérieure, nous concevons KLB-TS (KL Ball Track-and-Stop), un algorithme dont la complexité de l'échantillon est au maximum $U(\phi)\log(1/\delta)$. Notre algorithme repose sur une procédure de suivi de l'allocation d'échantillon conduisant à $U(\phi)$, et sur une règle d'arrêt que nous appelons règle d'arrêt KL-Ball en raison de son analogie avec la manière dont nous prouvons la borne supérieure $U(\phi)$.

Références

[1] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349, 2013.

[2] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 998–1027, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

[3] Aymen Al Marjani and Alexandre Proutiere. Adaptive sampling for best policy identification in markov decision processes, 2021.

[4] D. Russo. Simple bayesian algorithms for best arm identification. In *Proceedings of the 29th Conference On Learning Theory*, 2016.

[5] Andrea Zanette, Mykel J Kochenderfer, and Emma Brunskill. Almost horizon-free structure-aware best policy identification with a generative model. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5625–5634. Curran Associates, Inc., 2019.