META APPRENTISSAGE POUR L'APPRENTISSAGE PAR RENFORCEMENT

Company: Laboratoire LIPN, Devinci Research Center

Supervisors: Aomar OSMANI, Pegah ALIZADEH

PHD Student :

Keywords: Meta Learning, Deep Reinforcement Learning, Sample complexity of Reinforcement Learning

Date: 01.09.2021

---

1 Contexte

Le principe de fonctionnement de tous les objets autonomes, qu'ils soient vivants ou artificiels, consiste à prendre des décisions séquentielles en réaction aux informations captées de l'environnement [28]. Ces réactions génèrent des récompenses ou des punitions par rapport à un objectif final à atteindre. L'apprentissage par renforcement permet d'apprendre une stratégie optimale (séquence de décisions) en maximisant la somme attendue de récompenses [4, 28].

C'est notamment le cas dans les systèmes de l'internet des objets [23] et de la conduite autonome où chaque objet connecté est représenté par un véhicule qui doit se déplacer dans un environnement dynamique en optimisant à la fois des objectifs locaux et des objectifs globaux [9, 16]. Dans le cas des êtres vivants, les mécanismes d'apprentissage par renforcement (AR) sont plus élaborés. Ils permettent notamment de structurer des objectifs élémentaires en vue de l'accomplissement de plusieurs objectifs globaux, parfois antagonistes et surtout d'utiliser de la mémoire et des modes de représentation des connaissances et de raisonnement complémentaires au service d'une meilleure définition des objectifs à optimiser.

Habituellement, pour utiliser les approches AR, il est nécessaire de définir plusieurs paramètres principaux du système comme les états (ensemble de situations finies ou infinies S du système), des actions (ensemble de décisions finies ou infinies A), des fonctions de transition (probabilité de changer d'états en fonction des actions $P(s'|s,a)$) ainsi qu'une récompense (ou une punition) $r(s,a)$) associé au choix de la décision a en étant dans l'état s [28].

L'objectif de cette thèse est d'étendre les travaux en AR en intégrant des métaconnaissances liées à la fois au changement de l'environnement, au changement des probabilités de transition et au changement des récompenses en fonction du contexte [11, 20]. Dans le cas de l'internet des objets, la pertinence d'émettre des données par un capteur ainsi que la récompense associée dépend de l'importance des données collectées et de la situation. Il est naturel, dans ce cas, de pouvoir changer, à la fois, les probabilités de transition associées à un état ainsi que les récompenses associées au couple (état, transition).

Cette thèse investiguera l'apport du métaapprentissage [13] pour la généralisation de l'apprentissage AR à des contextes plus riche en intégrant à la fois le contexte et l'impact des décisions antérieurs en utilisant des mécanismes de mémorisation [19, 25, 29]. Le métaapprentissage vise aussi à concevoir des modèles capables d'acquérir de nouvelles compétences ou de s'adapter rapidement à de nouveaux environnements avec très peu

d'exemples d'apprentissage. Elle traitera aussi de la recherche de modèles de structuration de l'espace des hypothèses permettant de réduire le nombre d'exemples nécessaires pour apprendre.

2 Challenges scientifiques

Le premier challenge scientifique qui sera abordé dans cette thèse sera la prise en compte du contexte : comment apprendre la stratégie optimale pour les systèmes AR avec différents contextes (ou de manière similaire pour différentes fonctions de transition) en utilisant des approches de métaapprentissage ? Par exemple, en supposant que les états et les actions sont fixés pour les familles générales de problèmes, mais que les matrices de transition dépendent du contexte qu'on nommera σ et chaque système avec un contexte particulier sera noté S_{RL} : (S, A, P_σ, R). Dans ce cas, le modèle général de l'environnement peut être modélisé par un ensemble fixé d'états et d'actions, les probabilités contextuelles P_σ permettent ainsi d'éliminer, en fonction du contexte, les états inaccessibles et les actions associées et d'éviter le traitement des probabilités à valeurs nulles.

Nous pourrons ainsi définir une famille d'environnements avec un ensemble fini ou infini de contextes : $\{(S, A, P_\sigma, R_\sigma)\}_{\sigma\in\Sigma}$. Parmi les extensions possibles du modèle AR standard, nous nous focaliseront sur ceux utilisant le métaapprentissage [13, 23] ainsi que l'utilisation d'invariants pour réduire les espaces de recherche. Dans [19], les auteurs combinent un modèle dynamique appris a priori avec les nouvelles données collectées pour l'adapter rapidement au contexte courant. Une autre approche consiste à regrouper les contextes (les probabilités de transitions par catégories et d'apprendre une stratégie optimale par des méthodes classiques AR pour chaque catégorie [8, 10].
Une autre manière de faire est de définir plusieurs hyper-paramètres qui gèrent plusieurs hypothèses du modèle ($\{(S, A, P_\sigma, R_\sigma)\}_{\sigma\in\Sigma}$). Parmi ces hypothèses, on peut citer : les fonctions de transition en tant que fonction linéaire ou quadratique d'une fonction multiple [7, 30] ($\forall s \in S, a \in A, \sigma \in \Sigma s, P_\sigma(s'|s, a) = \alpha_1 P_{\alpha_1}(s'|s, a) + \cdots \alpha_k P_{\alpha_k}(s'|s, a)$), les fonctions de transition basées sur une fonction de confiance [2], les fonctions de transition basées sur une distribution donnée [1]: ($\forall\sigma \in \Sigma,\ P_\sigma(s'|s, a) \sim d(\sigma)$) ou encore l'apprentissage du modèle de fonction de transition [15, 24].

Le deuxième challenge de cette thèse est de savoir comment utiliser le métaapprentissage pour réduire la complexité du modèle d'apprentissage en ajoutant des éléments de structure qui réduisent l'espace d'hypothèses. Ces approches se combinent avec les développements actuels de l'AR profond [10, 12, 22, 27] et l'utilisation du cadre du métaapprentissage basé sur des gradients. Ces méthodes mettent à jour les paramètres du modèle afin d'obtenir de bonnes performances de généralisation sur les nouvelles tâches. Ils sont généralement applicables à tout modèle tant que le gradient peut être estimé [10, 21]. Le métaapprentissage est aussi utilisé pour optimiser une procédure d'optimisation de la région de confiance pour le choix de la politique et des fonctions de valeur représentées par des réseaux de neurones [26] ou encore l'utilisation d'une mémoire épisodique à la fois pour intégrer les événements passés avec leur contexte et accélérer la vitesse de convergence [5, 6].

Le troisième défi est d'étudier la complexité théorique des méthodes proposées et de travailler également sur les bornes des politiques ϵ-optimales en s'inspirant de nombreux travaux de la littérature comme ceux basés sur la taille de l'espace de recherche pour calculer les politiques optimales, les bornes de la complexité de l'espace de recherche des politiques ϵ-optimales [14]. Ce qui permet de travailler sur les approches de réduction des espaces de recherche des politiques optimales en utilisant des modèles génératifs [3], une combinaison linéaire de modèles [17] ou une généralisation de l'espace des politiques [18].

References

[1] Yasin Abbasi-Yadkori, Peter L Bartlett, and Csaba Szepesvári. Online learning in markov decision processes with adversarially chosen transition probability distributions. *arXiv preprint arXiv:1303.3055*, 2013.

[2] Mahsa Asadi, Mohammad Sadegh Talebi, Hippolyte Bourel, and Odalric-Ambrym Maillard. Model-based reinforcement learning exploiting state-action equivalence. In *Asian Conference on Machine Learning*, pages 204–219. PMLR, 2019.

[3] Mohammad Gheshlaghi Azar, Rémi Munos, and Bert Kappen. On the sample complexity of reinforcement learning with a generative model. *arXiv preprint arXiv:1206.6461*, 2012.

[4] Alan W Beggs. On the convergence of reinforcement learning. *Journal of economic theory*, 122(1):1–36, 2005.

[5] Charles Blundell, Benigno Uria, Alexander Pritzel, Yazhe Li, Avraham Ruderman, Joel Z Leibo, Jack Rae, Daan Wierstra, and Demis Hassabis. Model-free episodic control. *arXiv preprint arXiv:1606.04460*, 2016.

[6] Matthew Botvinick, Sam Ritter, Jane X Wang, Zeb Kurth-Nelson, Charles Blundell, and Demis Hassabis. Reinforcement learning, fast and slow. *Trends in cognitive sciences*, 23(5):408–422, 2019.

[7] Sayak Ray Chowdhury and Aditya Gopalan. Online learning in kernelized markov decision processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3197–3205. PMLR, 2019.

[8] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

[9] Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. Going deeper: Autonomous steering with neural memory networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 214–221, 2017.

[10] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[11] Thomas Furmston and David Barber. Variational methods for reinforcement learning. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 241–248. JMLR Workshop and Conference Proceedings, 2010.

[12] Abhishek Gupta, Benjamin Eysenbach, Chelsea Finn, and Sergey Levine. Unsupervised meta-learning for reinforcement learning. *arXiv preprint arXiv:1806.04640*, 2018.

[13] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*, 2020.

[14] Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, UCL (University College London), 2003.

[15] Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020.

[16] Lars Kröger, Tobias Kuhnimhof, and Stefan Trommer. Does context matter? a comparative study modelling autonomous vehicle impact on travel behaviour for germany and the usa. *Transportation research part A: policy and practice*, 122:146–161, 2019.

[17] Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR, 2020.

[18] Wenlong Mou, Zheng Wen, and Xi Chen. On the sample complexity of reinforcement learning with policy space generalization. *arXiv preprint arXiv:2008.07353*, 2020.

[19] Anusha Nagabandi, Ignasi Clavera, Simin Liu, Ronald S Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. *arXiv preprint arXiv:1803.11347*, 2018.

[20] Gergely Neu, Andras Gyorgy, and Csaba Szepesvári. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Artificial Intelligence and Statistics*, pages 805–813. PMLR, 2012.

[21] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *CoRR*, abs/1803.02999, 2018.

[22] Georgios Papoudakis, Filippos Christianos, Arrasy Rahman, and Stefano V Albrecht. Dealing with non-stationarity in multi-agent deep reinforcement learning. *arXiv preprint arXiv:1906.04737*, 2019.

[23] Guanjin Qu and Huaming Wu. Dmro: A deep meta reinforcement learning-based task offloading framework for edge-cloud computing. *arXiv preprint arXiv:2008.09930*, 2020.

[24] Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, pages 5478–5486. PMLR, 2019.

[25] Steindór Sæmundsson, Katja Hofmann, and Marc Peter Deisenroth. Meta reinforcement learning with latent variable gaussian processes. *arXiv preprint arXiv:1803.07551*, 2018.

[26] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

[27] Akshay Smit, Damir Vrabac, Yujie He, Andrew Y Ng, Andrew L Beam, and Pranav Rajpurkar. Medselect: Selective labeling for medical image classification combining meta-learning with deep reinforcement learning. *arXiv preprint arXiv:2103.14339*, 2021.

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

[29] Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

[30] Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.