ANALYSE PRÉLIMINAIRE – APPRENTISSAGE PAR RENFORCEMENT SÉCURITAIRE

A PREPRINT

Mouhamed Gando Diallo
Apprentissage par renforcement – IFT-7201
Université Laval
mouhamed-gando.diallo.1@ulaval.ca

Renaud Djekornonde Raoudel
Apprentissage par renforcement – IFT-7201
Université Laval
renaud.djekornonde-raoudel.1@ulaval.ca

31 mars 2025

Revue de la littérature

L'apprentissage par renforcement sécuritaire (Safe Reinforcement Learning, SRL) vise à apprendre des politiques efficaces tout en respectant des contraintes de sécurité, particulièrement critiques dans des contextes industriels où des actions peuvent entraîner des conséquences catastrophiques. Dans cette section, nous présentons trois variantes distinctes du problème de SRL explorées dans la littérature récente.

Méthodes, théories et applications du SRL (vue d'ensemble) Gu et al. [1] proposent une revue structurée des approches modernes en apprentissage par renforcement sécuritaire. Les auteurs formalisent cinq grandes dimensions du SRL (résumées par l'acronyme "2H3W") : *Why* SRL est nécessaire, *What* à sécuriser, *How* sécuriser l'apprentissage, *When* sécuriser (apprentissage vs exécution), et *Where* appliquer ces méthodes. L'article couvre à la fois les fondements théoriques (notamment les garanties de sécurité et la complexité échantillonnale) et les avancées algorithmiques, tout en offrant un panorama des domaines d'application et des environnements de test couramment utilisés. Ce travail constitue une référence importante pour comprendre la diversité des approches, ainsi que les compromis entre performance et sécurité. Il permet également d'identifier les lacunes actuelles et les pistes de recherche futures dans le domaine du SRL.

SRL appliqué au contrôle optimal industriel Lu et al. [3] présentent une approche de SRL conçue pour le contrôle optimal dans un procédé industriel de grillage de minerai d'or. L'algorithme proposé satisfait des contraintes de chance conjointes avec une haute probabilité, assurant ainsi que les politiques apprises respectent des seuils critiques de sécurité dans un contexte industriel réel. Cette étude met en évidence l'importance d'intégrer des contraintes opérationnelles spécifiques dans le processus d'apprentissage. En se basant sur un cas concret, les auteurs démontrent que des approches de RL peuvent être viables dans des systèmes à risques élevés si des garanties probabilistes sont imposées. Leur méthodologie fournit un exemple solide d'intégration entre théorie du SRL et applications industrielles réelles.

Transfert sécurisé de politiques simulées vers des systèmes réels Hsu et al. [2] s'attaquent à la problématique du transfert de politique, en introduisant une approche "Sim-to-Lab-to-Real" combinant apprentissage par renforcement avec un mécanisme de surveillance actif (shielding). Leur méthode repose sur une architecture à double politique : une politique primaire optimise la performance, tandis qu'une politique secondaire agit comme garde-fou pour prévenir les actions dangereuses. Ce cadre permet un transfert fiable entre simulation et déploiement dans des environnements réels, tout en maintenant des garanties de sécurité. Cette approche est particulièrement pertinente pour notre étude, car elle s'attaque explicitement aux risques liés au déploiement de politiques apprises en simulation dans des environnements dont les dynamiques sont imparfaitement connues. Elle combine élégamment robustesse, sécurité et généralisation, trois axes clés pour des systèmes réels critiques.

Formulation du problème sélectionné

Dans cette étude, nous nous intéressons à la problématique du *transfert de politiques sécuritaires* apprises en simulation vers des environnements réels, dans le cadre de l'apprentissage par renforcement. Le défi principal réside dans le fait que des politiques efficaces en simulation peuvent engendrer des comportements dangereux lorsqu'elles sont appliquées directement à des environnements réels dont les dynamiques diffèrent légèrement.

Nous formulons ce problème comme un processus de décision markovien (MDP) défini par un tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$, où :

- $\mathcal{S}$ est l'espace des états,
- $\mathcal{A}$ est l'espace des actions,
- $P(s' \mid s, a)$ est la dynamique de transition (différente entre la simulation et la réalité),
- $R(s, a)$ est la fonction de récompense, contenant des *pénalités sévères pour les états considérés comme dangereux*,
- $\gamma \in [0, 1]$ est le facteur d'actualisation.

Le but est d'apprendre une politique $\pi : \mathcal{S} \to \mathcal{A}$ qui maximise la récompense attendue tout en *minimisant le risque de transitions vers des états à récompenses catastrophiques*.

La difficulté provient du fait que l'agent s'entraîne uniquement dans un environnement de simulation $\mathcal{M}_{\text{sim}}$, et que la politique apprise doit être transférée vers un environnement légèrement différent $\mathcal{M}_{\text{real}}$, sans exploration directe de celui-ci. Nous considérons une formulation de type *constrained RL* avec une contrainte sur la probabilité de visiter un sous-ensemble d'états dangereux $\mathcal{S}_{\text{bad}}$:

$$\max_{\pi} \mathbb{E}_{\pi}[G] \quad \text{sous la contrainte} \quad \mathbb{P}_{\pi}(s_t \in \mathcal{S}_{\text{bad}}) \leq \delta, \ \forall t$$

ou δ est un seuil de tolérance au risque.

Présentation des méthodes à étudier

Nous avons sélectionné deux approches complémentaires pour étudier notre problématique de transfert sécurisé de politiques. La première est une méthode classique de renforcement profond, largement utilisée et non spécifique à la sécurité. La seconde est une méthode spécialisée conçue pour traiter explicitement les risques liés au transfert de politiques.

Méthode classique : Deep Q-Network (DQN) Deep Q-Network (DQN) [4] est une méthode de RL classique qui combine les Q-learning avec des réseaux de neurones profonds pour approximer la fonction de valeur d'action $Q(s, a)$. DQN est capable d'apprendre des politiques efficaces dans des environnements à grand espace d'états, ce qui le rend adapté à notre contexte de simulation. Toutefois, DQN ne tient pas compte des aspects de sécurité : il peut générer des politiques performantes mais potentiellement risquées lorsqu'elles sont transférées dans un environnement réel.

Nous utiliserons l'implémentation de DQN disponible dans la bibliothèque `Stable-Baselines3` [1].

Méthode sécuritaire : Safe Reinforcement Learning with Shielding (Sim-to-Lab-to-Real) Nous adoptons la méthode proposée par Hsu et al. [2], qui combine apprentissage profond et mécanisme de *shielding* (surveillance active). Le principe est d'entraîner une politique principale dans la simulation, tout en la couplant à une politique de sécurité chargée de surveiller les actions proposées. Si une action risque de conduire à un état dangereux, la politique de sécurité intervient pour bloquer ou remplacer cette action. Ce cadre permet de garantir des limites de sécurité probabilistes lors du transfert vers l'environnement réel.

Ce type de méthode est compatible avec les environnements de type OpenAI Gym et peut être implémenté à l'aide de la bibliothèque `SafeRL` ou via une adaptation personnalisée dans `Stable-Baselines3`.

Conclusion L'utilisation combinée de DQN (méthode standard) et du cadre Sim-to-Lab-to-Real (méthode sécuritaire) nous permettra de comparer des politiques optimisées selon des critères classiques avec des politiques explicitement conçues pour minimiser les risques de comportements catastrophiques en environnement réel.

1. `https://github.com/DLR-RM/stable-baselines3`

Références

[1] Shangding Gu, Lin Yang, Yiming Du, Guannan Chen, Florian Walter, Jun Wang, and Alois Knoll. A review of safe reinforcement learning : Methods, theory and applications. *arXiv preprint arXiv :2205.10330*, 2022.

[2] Kai-Chieh Hsu, Yujia Luo, Rohan Kumar, Claire J Tomlin, and Dorsa Sadigh. Sim-to-lab-to-real : Safe reinforcement learning with shielding and generalization guarantees. *arXiv preprint arXiv :2201.08355*, 2022.

[3] Yuchen Lu, Chenguang Song, Jinfeng Liu, Lihua Xie, and Tao Liu. Safe reinforcement learning for industrial optimal control : A case study in gold ore roasting. *Information Sciences*, 631 :110–125, 2023.

[4] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540) :529–533, 2015.