CATEGORICAL MULTI-OBJECTIVE DEEP Q-NETWORKS

Farès Chouaki, Aurélie Beynier, Nicolas Maudet, Paolo Viappiani

Available at

`https://gitlab.com/chouakifares/distributionnal-morl`

Sommaire

Exemples de Problèmes d'Apprentissage par Renforcement Multi-Objectif I

Les problèmes d'Apprentissage par Renforcement Multi-Objectif (MORL) apparaissent dans divers domaines :

- **Conduite Autonome :** Équilibrer la sécurité, l'efficacité énergétique et le temps de trajet.
- **Robotique :** Optimiser la vitesse, la précision et la consommation d'énergie des bras robotiques.
- **Santé :** Maximiser l'efficacité des traitements tout en minimisant les effets secondaires et les coûts.

Exemples de Problèmes d'Apprentissage par Renforcement Multi-Objectif II

- **Gestion des Ressources :** Allouer efficacement les ressources entre des besoins concurrents (par ex. distribution d'énergie, contrôle du trafic).
- **IA pour les Jeux :** Équilibrer les stratégies offensives et défensives dans les jeux multi-agents.

Ces problèmes nécessitent une prise de décision tenant compte de multiples objectifs souvent conflictuels.

Sommaire

Processus de décision Markovien

Soit $M = \{S, A, R, T, \mu\}$ le processus de décision Markovien défini par :

- S : Ensemble d'états formant l'environnement.
- A : Ensemble d'actions que l'agent peut effectuer.
- $R : S \times A \longrightarrow \mathbb{R}$: Fonction de récompense de l'agent, associant à chaque couple (état, action) **un réel**.
- $T : S \times A \times S' \longrightarrow [0, 1]$: Fonction de transition de l'environnement, telle que

$$\forall s \in S \quad \forall a \in A \quad \sum_{s' \in S} T(s, a, s') = 1$$

- $\mu : S \longrightarrow [0, 1]$: Distribution des états initiaux.

Processus de décision Markovien multi-objectifs

Soit $M = \{d, S, A, R, T, \mu\}$ le processus de décision Markovien multi-objectifs défini par :

- $d \geq 2$: Nombre d'objectifs à optimiser.
- S : Ensemble d'états formant l'environnement.
- A : Ensemble d'actions que l'agent peut effectuer.
- $R : S \times A \longrightarrow \mathbb{R}^d$: Fonction de récompense de l'agent, associant à chaque couple (état, action) **un vecteur de taille d**.
- $T : S \times A \times S' \longrightarrow [0, 1]$: Fonction de transition de l'environnement, telle que

$$\forall s \in S \quad \forall a \in A \quad \sum_{s' \in S} T(s, a, s') = 1$$

- $\mu : S \longrightarrow [0, 1]$: Distribution des états initiaux.

Critères d'optimisations pour algorithme à politique unique

Étant donnée une fonction d'agrégation des objectifs u, deux critères d'optimisation peuvent être définis

- Scalarized Expected Reward (SER)

$$V_u^{\pi} = u(\mathbb{E}(\sum_{i=0}^{\infty} \gamma^i \mathbf{r_i} | \pi, s_0)))$$

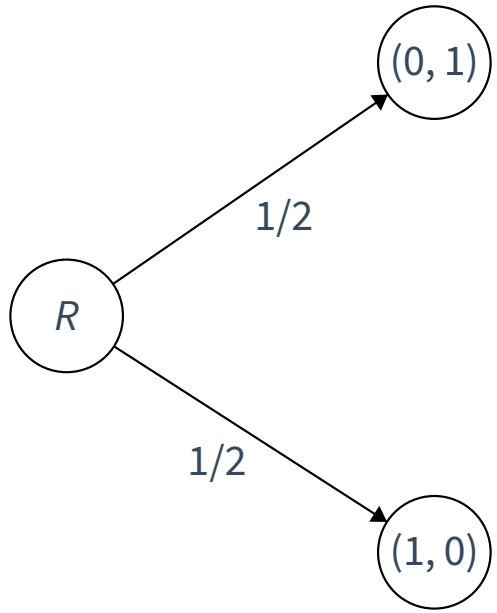
- Expected Scalarized Reward (ESR)

$$V_u^{\pi} = \mathbb{E}(u(\sum_{i=0}^{\infty} \gamma^i \mathbf{r_i}) | \pi, s_0))$$

Critère d'optimisation pour algorithme à politique unique – Exemple

Considérons la politique suivante π_1 :

La récompense obtenue par cette politique est donnée par la loterie ci-dessous :



En utilisant la fonction *min* comme agrégateur entre les objectifs, l'utilité associée à cette politique diffère selon le critère d'optimisation utilisé.

$ESR(\pi_1) = \frac{1}{2} \times min(0,1) + \frac{1}{2} \times min(1,0) = 0 \qquad SER(\pi_1) = min(\frac{1}{2} \times (0,1) + \frac{1}{2} \times (1,0)) = \frac{1}{2}$

Sommaire

Algorithmes existants pour l'apprentissage par renforcement multi-objectifs sous ESR

- Nash Social Welfare Q-Learning (NSW-QL) [Fan et al. 2023]
- Expected Utility Policy Gradient (EUPG) [Roijers et al. 2018]
- Multi-Objective Categorical Actor Critic (MOCAC) [Reymond et al. 2023]

Limites des algorithmes existants

- **Généralisation limitée des approches basées sur les politiques**
  EUPG et *MOCAC* exploitent avidement les expériences et ne peuvent pas se généraliser à d'autres fonctions de scalarisation.
- **Problèmes de passage à l'échelle dans les approches basées sur la valeur**
  NSW-Q-learning ne sont utilisables que pour des tâches avec des espaces d'états et d'actions de petite taille.
- **Nécessité d'adaptation pour les algorithmes de RL mono-objectif**
  Les algorithmes classiques de RL mono-objectif (comme *DQN, PER, C51-DQN*) ne peuvent pas traiter directement les problèmes multi-objectifs sans modification.

Sommaire

Objectifs I

Développer un algorithme de RL distributionnel basé sur la valeur pour les problèmes multi-objectifs

Introduction *Multi-Objective Categorical Deep Q-Networks (MO-CDQN)*, une extension de l'algorithme *C51-DQN* adaptée au cadre multi-objectif.

Améliorer l'efficacité d'échantillonnage par rapport aux méthodes basées sur les politiques

Démonstration que *MO-CDQN* apprend des politiques optimales sous ESR plus rapidement que les algorithmes existants.

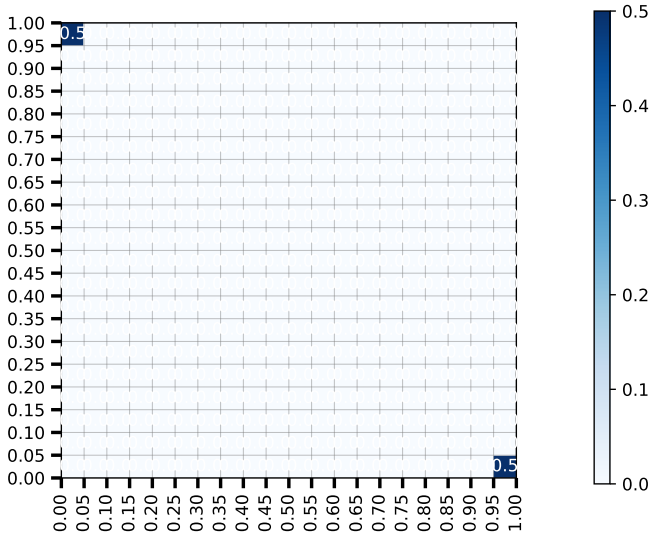
Méthode I

- **Représentation catégorielle des retours** pour capturer la distribution complète.
- **Sélection d'action** intégrant une scalarisation non linéaire.
- **Mise à jour par apprentissage temporel** (*TD Learning*), inspirée de Double DQN et C51. [Hasselt et al. 2016 ; Bellemare et al. 2017]

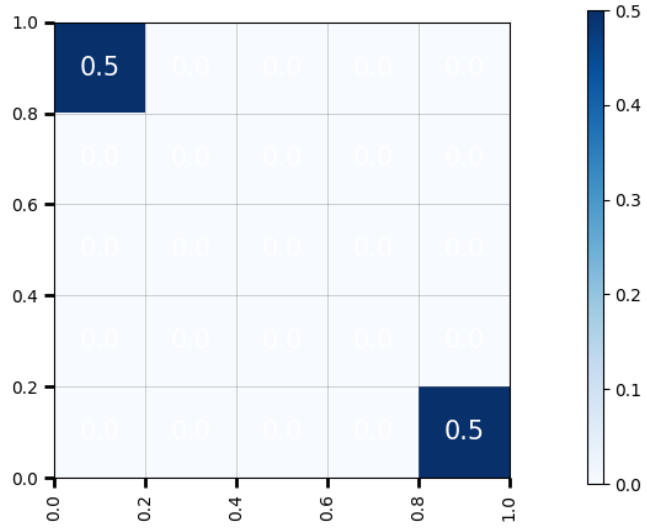
Représentation catégorielle des retours I

Problème : Les algorithmes classiques de RL n'estiment que l'espérance des retours, ce qui est insuffisant pour apprendre des politique optimale sous ESR.
Solution : MO-CDQN apprend la distribution complète des retours sous forme de **distribution catégorielle**.

Représentation catégorielle des retours II



Support à 20 atoms



Support à 5 atoms

Exemple de distribution représentant les retours obtenus sur l'exemple précedant

Représentation catégorielle des retours III

Avantage : Permet une meilleure estimation des retours futurs et permet d'appliquer la fonction de scalariazation sur le support de la distribution avant de calculer son espérance.

Séléction d'action I

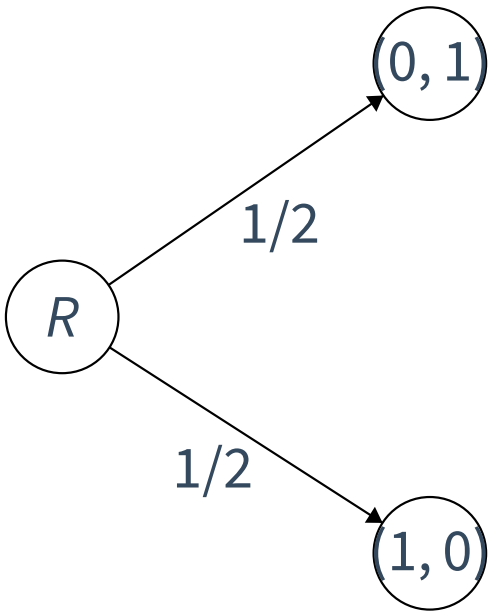
Problème : Comment choisir une action dans un état s étant données les distribution de retour multivariée apprises.

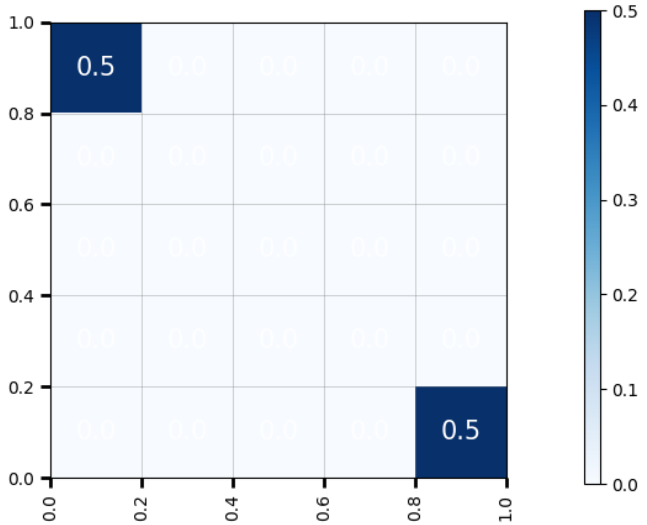
Solution : MO-CDQN ajuste sa sélection d'action en :

- Appliquant la scalarisation directement sur chaque atom du support de la distributions estimées.
- La valeur de chaque action est ensuite calculée en ponderant chacun des atoms scalarisé par sa probabilité.
- l'action est choisie ensuite via une politique **softmax** pondérée par les valeurs des actions.

Séléction d'action II

Exemple : Etant donnée la distribution montrée dans la Figure 1b.

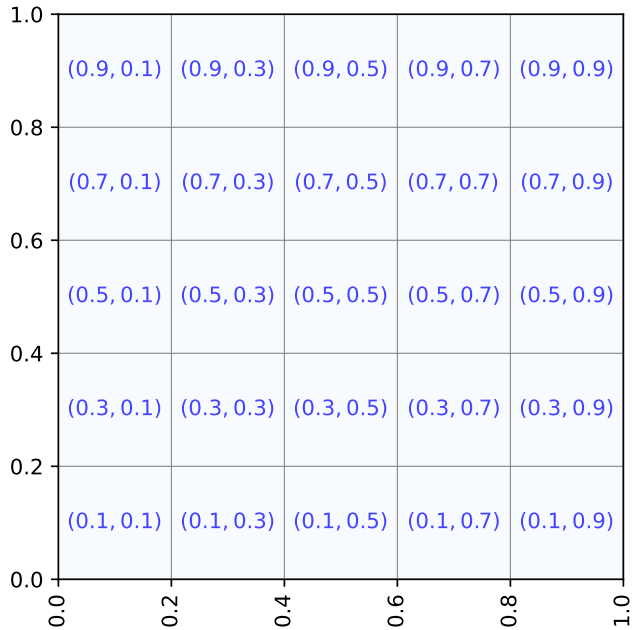




La Q valeur d'une action pareille, en utilisant le *min* comme fonction de scalarization, serait calculée comme suit :

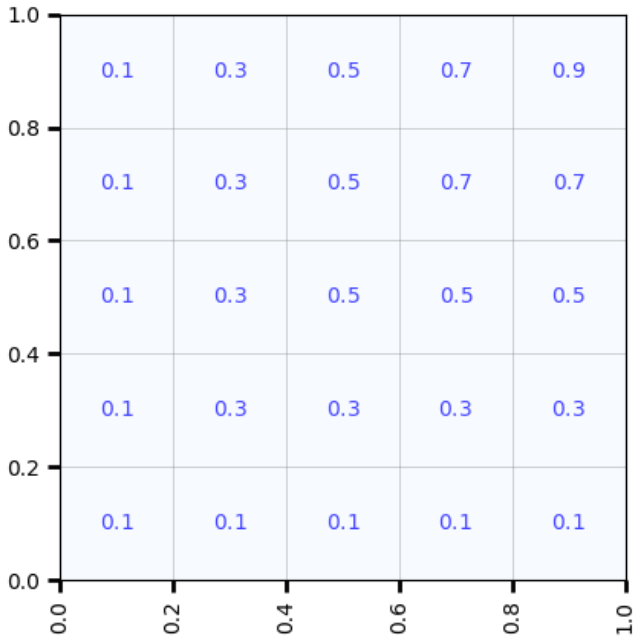
Séléction d'action III

1. Le support de la distribution est montré dans la figure suivante



Séléction d'action IV

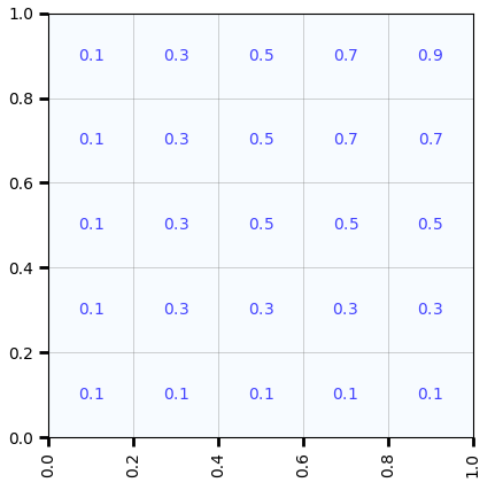
2. Les atoms sont scalarizés en utilisant la fonction *min* pour obtenir les atoms suivants

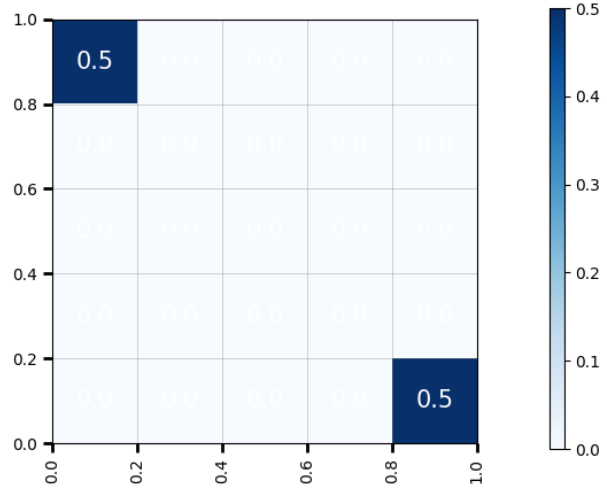


Séléction d'action V

3. On calcule ensuite l'esperance de la distribution,

$$Q = 0.5 \times 0.1 + 0.5 \times 0.1 = 0.1$$





Objectif de l'évaluation

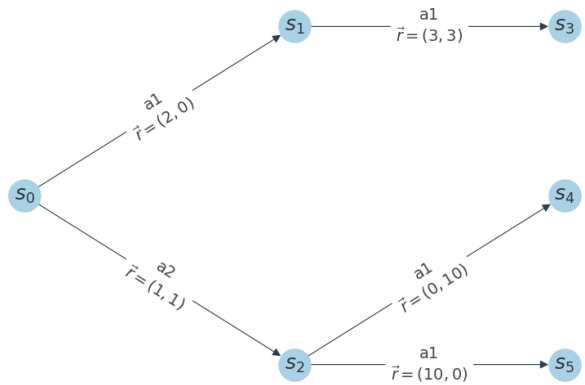
Valider l'algorithme sur des benchmarks multi-objectifs

Tester *MO-CDQN* sur un petit *MO-MDP* et sur l'environnement *Fruit-Tree Navigation* du benchmark *MO-gym* pour démontrer sa supériorité par rapport aux algorithmes optimisant ESR existants.

Évaluation et Résultats I

Validation et Comparaison :

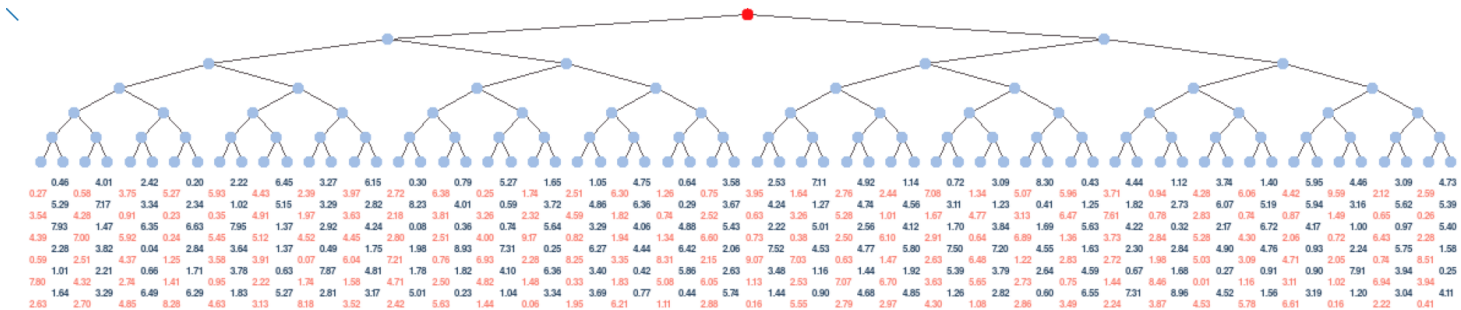
- **MO-MDP jouet :** Vérification de la convergence des distributions de retour.



MO-MDP jouet utilisé pour valider l'approche

Évaluation et Résultats II

- **Environnement FTN (MO-Gym)[Felten et al. 2023] :**
  Comparaison avec *EUPG* et *MOCAC*.



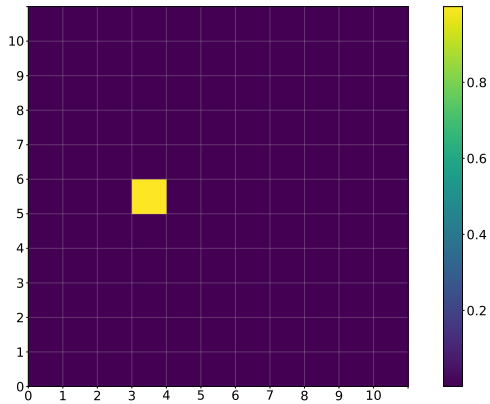
Évaluation et Résultats III

Protocole expérimental :

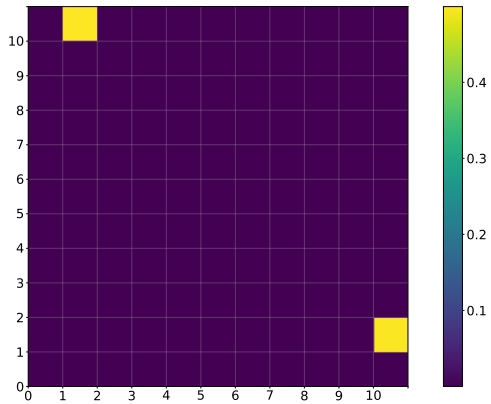
- Entraînement sur MO-MDP jouet (**2500 étapes**).
- Entraînement sur FTN : **500,000 interactions** pour MO-CDQN, **1,000,000** pour les autres.
- Évaluation périodique sur 100 épisodes.

Résultats expérimentaux

Résultats de Validation :



Distribution des retours prédite pour (s_0, a_0)

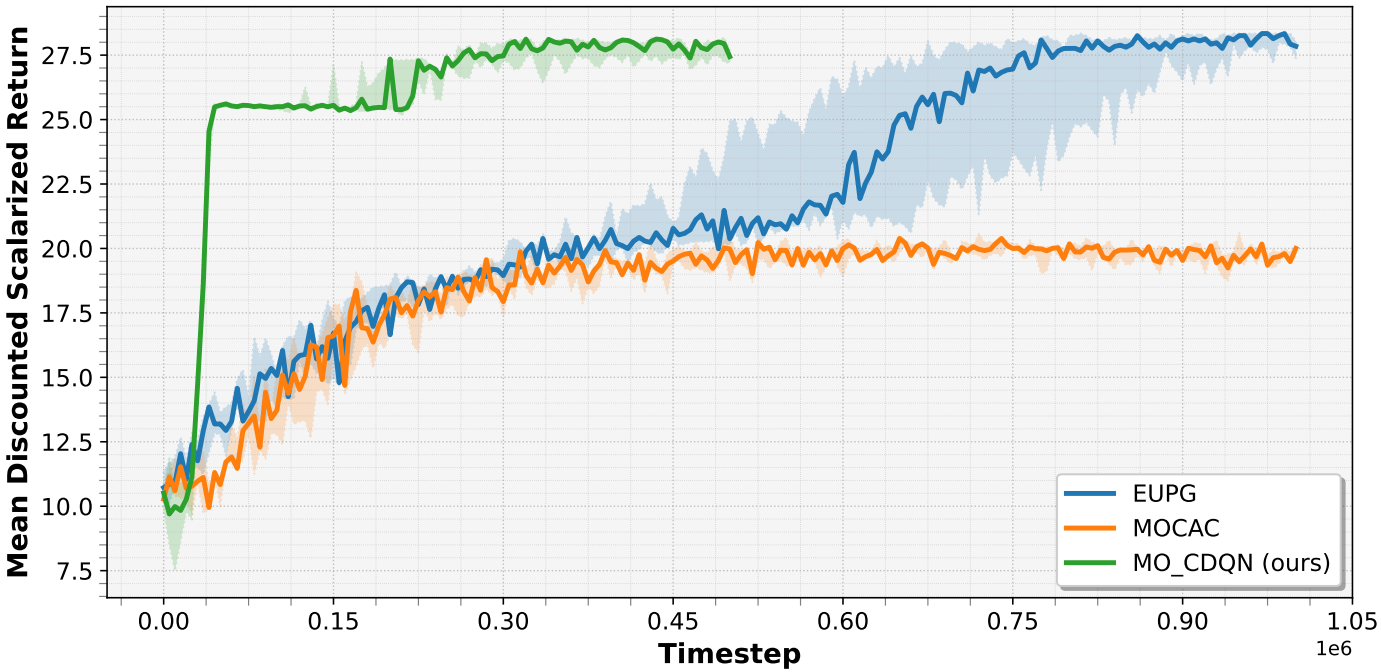


Distribution des retours prédite pour (s_0, a_1)

Distribution des retours apprise sur le MO-MDP jouet

Comparaison avec les baselines I

Comparaison des courbes d'apprentissage sur FTN :



Évolution du retour scalaire moyen sur FTN

Comparaison avec les baselines II

Performance de MO-CDQN :

- Convergence en **300 000 étapes**, contre **750 000** pour *EUPG*.
- *MOCAC* ne coverge pas vers la politique optimale dans le budget imparti.

MO-CDQN atteint une meilleure performance plus rapidement que les approches existantes, prouvant son efficacité pour l'optimisation ESR.

Sommaire

Perspectives

- Proposer des approches multi-politiques pour résoudre les problèmes de MORL selon le critère ESR.
- Adapter MO-CDQN au cas multi-agents et expérimenter sur davantage d'environnements.
- Proposer un moyen de réduire la complexité spatiale de l'algorithme.

https://gitlab.com/chouakifares/distributionnal-morl

Thank You

Questions?

Reference I

Hasselt, Hado van, Arthur Guez et David Silver (mars 2016). “Deep Reinforcement Learning with Double Q-Learning”. In : *Proceedings of the AAAI Conference on Artificial Intelligence* 30.1. doi : 10.1609/aaai.v30i1.10295. url : https://ojs.aaai.org/index.php/AAAI/article/view/10295.

Bellemare, Marc G., Will Dabney et Rémi Munos (2017). “A Distributional Perspective on Reinforcement Learning”. In : *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML’17. Sydney, NSW, Australia : JMLR.org, p. 449-458.

Roijers, Diederik, Denis Steckelmacher et Ann Nowe (juill. 2018). “Multi-objective Reinforcement Learning for the Expected Utility of the Return”. In.

Fan, Ziming, Nianli Peng, Muhang Tian et Brandon Fain (2023). “Welfare and Fairness in Multi-objective Reinforcement Learning”. In : *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. AAMAS ’23. London, United Kingdom : International Foundation for Autonomous Agents et Multiagent Systems, p. 1991-1999. isbn : 9781450394321.

Felten, Florian, Lucas N. Alegre, Ann Nowé, Ana L. C. Bazzan, El Ghazali Talbi, Grégoire Danoy et Bruno C. da Silva (2023). “A Toolkit for Reliable Benchmarking and Research in Multi-Objective Reinforcement Learning”. In : *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*.

Reymond, Mathieu, Conor Hayes, Denis Steckelmacher, Diederik Roijers et Ann Nowe (avr. 2023). “Actor-critic multi-objective reinforcement learning for non-linear utility functions”. In : *Autonomous Agents and Multi-Agent Systems* 37. doi : 10.1007/s10458-023-09604-x.