M2 DAC

Rapport

Rapport de mini-projet 2

19 octobre 2024

Résumé

Dans ce mini-projet, nous avons étudié l'impact de la suppression des caractéristiques et des extensions temporelles des observations et actions dans l'environnement CartPoleContinuous-v1 sur la performance de l'algorithme DDPG. Nous avons analysé différentes configurations, incluant la suppression de $\dot{x}$, $\dot{\theta}$, ainsi que l'extension des observations et actions. Les performances ont été mesurées à l'aide de récompenses cumulées et comparées à travers des courbes d'apprentissage.

Table des matières

1 Méthologie

1.1 Environnement et Algorithme

Dans cette expérience, nous avons utilisé l'environnement CartPoleContinuous-v1, une version modifiée du classique CartPole, où les actions sont continues au lieu de discrètes, variant entre -1 et 1. Nous avons utilisé l'algorithme DDPG (Deep Deterministic Policy Gradient), qui est un algorithme d'apprentissage par renforcement off-policy, capable de traiter des actions continues. L'objectif principal de cette étude était d'évaluer l'impact de la suppression de certaines caractéristiques des observations, ainsi que l'ajout d'extensions temporelles aux observations et actions, sur les performances de DDPG.

Choix de l'algorithme : Nos essais ont montré que dans un environnement relativement simple comme celui-ci, DDPG trouve plus rapidement une stratégie appropriée grâce à son mécanisme de mise à jour plus direct et à une fréquence de mise à jour des réseaux plus élevée. Les améliorations apportées par TD3, bien qu'efficaces dans des environnements plus complexes, peuvent constituer un surcoût dans ce cas, voire entraîner une diminution des performances.

1.2 Wrappers Utilisés

Pour réaliser cette étude, nous avons conçu et appliqué plusieurs wrappers à l'environnement d'entraînement :

`FeatureFilterWrapper` : Ce wrapper supprime certaines caractéristiques de l'observation pour simuler un environnement partiellement observable. Cette approche permet d'évaluer la robustesse de l'algorithme face à des informations incomplètes.

`ObsTimeExtensionWrapper` : Ce wrapper étend l'espace d'observation en ajoutant l'état de l'instant précédent à l'état courant. Ainsi, au lieu de se baser uniquement sur l'état actuel, l'algorithme reçoit aussi une information temporelle qui pourrait potentiellement améliorer sa capacité à prédire les récompenses futures dans un environnement partiellement observable.

`ActionTimeExtensionWrapper` : Ce wrapper étend l'espace d'action en générant une séquence d'actions. Bien que seule la première action de cette séquence soit appliquée à l'environnement, l'algorithme critique évalue l'ensemble de la séquence, ce qui permet d'anticiper les conséquences à plus long terme d'une action.

1.3 Réglages des Hyperparamètres

Nous avons ajusté plusieurs hyperparamètres dans l'algorithme DDPG pour optimiser les performances :

`discount_factor` : Défini à 0.98. Pour l'environnement CartPoleContinuous-v1, qui est un problème de maintien de l'équilibre à long terme, un facteur de discount élevé est préférable. Cela aide l'agent à se concentrer sur des récompenses futures et non seulement sur des gains immédiats, ce qui est crucial pour maintenir l'équilibre à long terme.

`seed` : Pour garantir la robustesse des résultats, nous avons utilisé différents seeds aléatoires à chaque exécution de l'algorithme. Cela permet de réduire l'impact de l'aléatoire

d'un seul seed et de mieux évaluer la performance globale de l'algorithme. L'utilisation de multiples seeds nous a également permis d'effectuer des analyses statistiques plus précises, assurant ainsi la fiabilité des résultats obtenus.

1.4 Scénarios Testés

Nous avons testé plusieurs configurations d'environnement en combinant la suppression de caractéristiques et l'extension temporelle des observations et des actions. Notamment, nous nous sommes concentrés sur la combinaison des scénarios suivants :

Suppression de $\dot{x}$ ou $\dot{\theta}$: La vitesse du chariot ($\dot{x}$) ou la vitesse angulaire du pendule ($\dot{\theta}$) peut être retirée individuellement ou en combinaison, afin d'évaluer leur importance respective dans la prise de décision de l'agent.

Extension temporelle des observations : En ajoutant les états précédents aux observations actuelles, nous avons cherché à évaluer si les informations temporelles supplémentaires pouvaient compenser la perte des caractéristiques supprimées.

Extension temporelle des actions : L'extension de l'espace d'action permet à l'agent de prendre en compte plusieurs étapes d'action, facilitant ainsi la planification à plus long terme.

1.5 Outils et Librairies

Nous avons utilisé les outils et librairies suivants :

BBRL : Un framework basé sur PyTorch pour simplifier l'implémentation d'algorithmes de renforcement.

Gymnasium : Utilisé pour l'environnement de simulation CartPoleContinuous-v1.

Matplotlib : Utilisé pour tracer les courbes d'apprentissage et visualiser les performances des agents.

Omegaconf : Utilisé pour facilite la gestion des paramètres

typing : Utilisé pour les annotations.

2 Résultats Expérimentaux

2.1 L'impact de supprimer des caratéristiques

Nous avons répété plusieurs expériences avec des seeds différents pour tester respectivement l'effet de l'observation complète, de la suppression de $\dot{x}$ (la vitesse de la véhicule), de la suppression de $\dot{\theta}$ (la vitesse angulaire du bâton) et de la suppression des deux sur les performances d'apprentissage, et les résultats ont été illustrés dans le graphique ci-dessous.

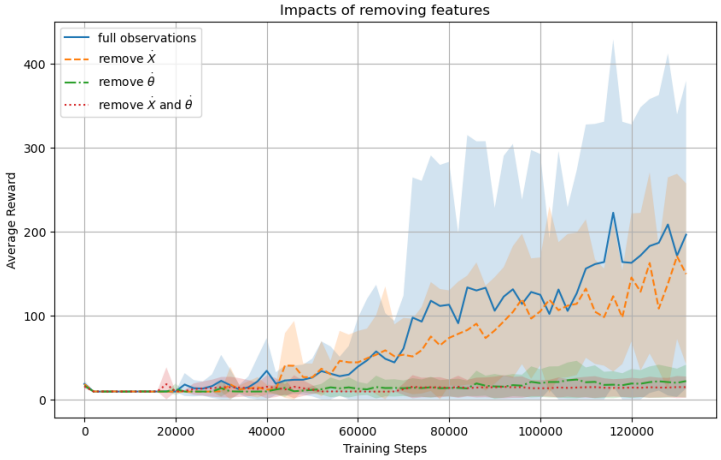


FIGURE 1 – L'impact de la suppression des caratéristiques

Nous pouvons constater à partir de ce graphique que la suppression de certaines fonctionnalités entraîne une diminution des performances d'apprentissage. Parmi celles-ci, la suppression de $\dot{\theta}$ (la vitesse angulaire du bâton) a l'impact le plus important. Si l'Agent ne connaît pas la vitesse actuelle du bâton et ne dispose que de la position de celui-ci, il sera incapable de prédire le mouvement du bâton à l'instant suivant. Par conséquent, il ne pourra pas proposer une action appropriée pour maintenir l'équilibre.

2.2 L'effet de l'observation extension

Nous avons créé un wrapper permettant de stocker l'historique des observations et avons réalisé plusieurs expériences en stockant l'observation actuelle ainsi que la précédente. Les résultats de ces expériences sont illustrés dans le graphique ci-dessous.

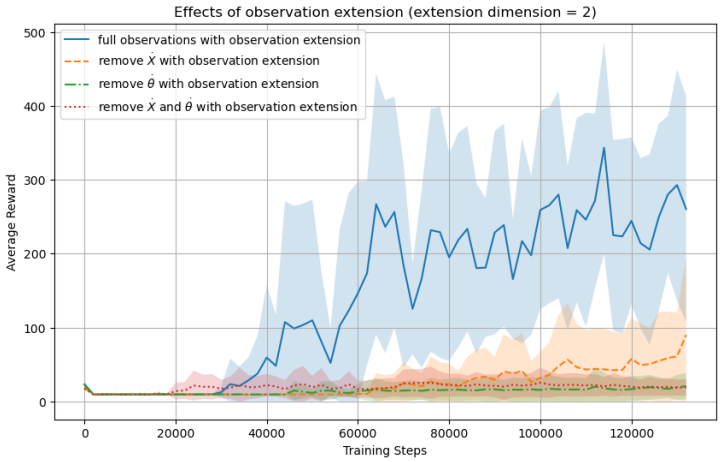


FIGURE 2 – L'impact de l'extension d'observation (dim=2)

Nous pouvons constater une amélioration significative des performances d'apprentissage avec l'observation complète. A priori, en combinant l'observation actuelle avec l'historique des observations et en les intégrant dans le modèle d'apprentissage, cela permet d'élargir la compréhension du modèle des variations des observations, ce qui explique l'amélioration des résultats d'apprentissage. Par la suite, nous comparerons l'impact des différents wrappers dans chaque situation.

2.3 L'effet de l'action extension

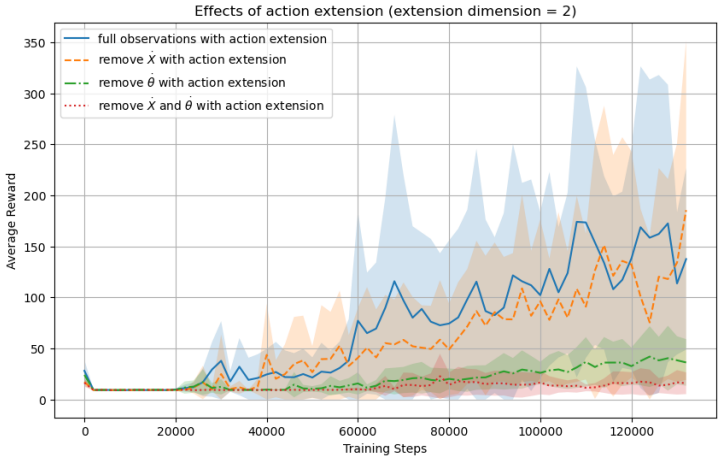


FIGURE 3 – L'impact de l'extension d'action(dim=2)

Pour le wrapper qui étend l'action, nous pouvons constater à partir du graphique ci-dessus qu'il contribue également à l'amélioration des performances d'apprentissage. Cela s'explique par le fait que le rôle du critic est d'estimer la valeur de $V(s)$ ou $Q(s,a)$ dans l'état actuel. Fournir un plus grand nombre d'actions équivaut à offrir au critic davantage d'échantillons, ce qui contribue à améliorer la précision de l'estimation.

2.4 L'effet de l'extension combinée

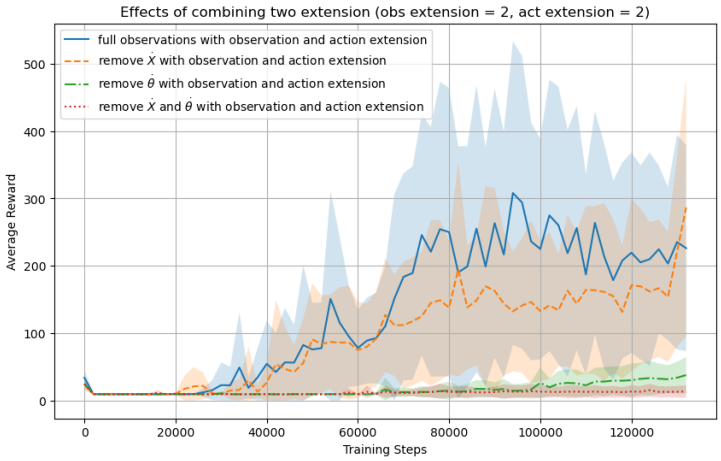
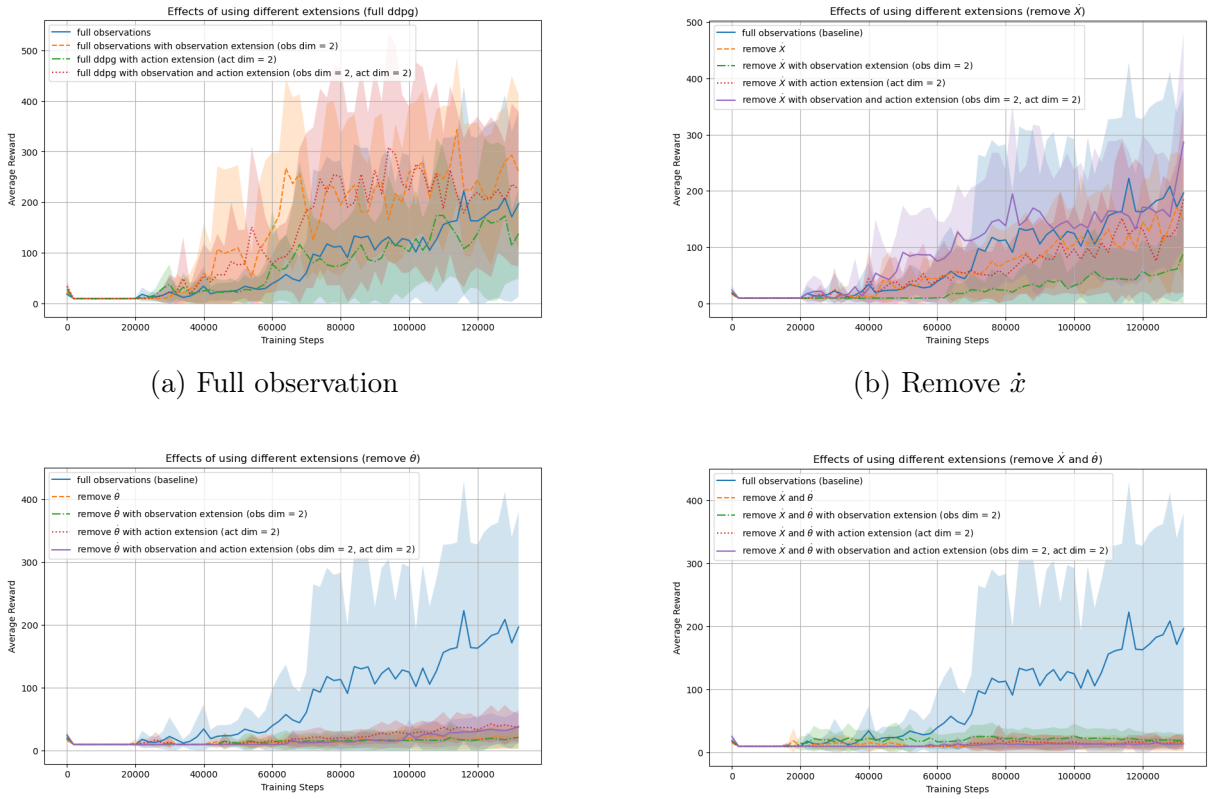


FIGURE 4 – L'impact de l'extension combinée (dim =2)

Ce graphique montre les résultats expérimentaux obtenus en combinant les deux wrappers. L'amélioration des performances d'apprentissage correspond essentiellement à la combinaison des gains apportés par chaque wrapper, sans qu'aucun phénomène inattendu ne soit observé.

2.5 Comparaison locale des effets de l'extension

Dans cette section, nous comparons séparément les effets de l'utilisation de différents wrappers dans chaque situation.



(a) Full observation

(b) Remove $\dot{x}$

(c) Remove $\dot{\theta}$

(d) Remove both

FIGURE 5 – Comparaison locale des effets de l'extension

Pour l'observation complète, nous avons constaté que l'extension des observations améliore les performances d'apprentissage. Cependant, l'extension des actions a légèrement réduit la performance.

Pour le cas de la suppression de $\dot{x}$, nous avons observé un résultat similaire. Cependant, de manière inattendue, l'utilisation des deux wrappers combinés dans cette situation a permis d'obtenir des performances d'apprentissage supérieures à celles de l'observation complète. Cela suggère que, même en présence de features inconnues dans l'environnement, l'extension du nombre d'états fournis au modèle ainsi que l'augmentation du nombre d'actions introduites dans le processus du critic peuvent permettre d'atteindre des performances d'apprentissage comparables à celles d'une observation complète.

Pour le cas de la suppression de $\dot{\theta}$, bien que les wrappers améliorent les performances d'apprentissage, le résultat final reste néanmoins inférieur à celui obtenu avec l'observa-

tion complète. Ainsi, en cas de manque d'informations cruciales, même si l'utilisation de wrappers peut améliorer les performances d'apprentissage, le résultat global ne pourra jamais égaler celui obtenu avec une observation complète.

Pour le cas de la suppression des deux caractéristique, étant donné que le résultat final est nettement inférieur à celui de l'observation complète, les courbes associée sont difficilement discernables. Cependant, sur la base des conclusions tirées des expériences précédentes, nous pouvons appliquer les mêmes à ce cas.

En résumé, bien qu'a priori l'extension de l'observation et de l'action semble avoir un impact positif sur les performances d'apprentissage, les résultats expérimentaux montrent que le wrapper d'action a parfois un effet négatif. Cela indique qu'il est nécessaire d'utiliser ces wrappers de manière sélective et réfléchie, en fonction des besoins spécifiques du modèle et de l'environnement.

3 Étude supplémentaire

Pour les études supplémentaires, nous avons étendu l'horizon temporel de la mémoire des états et des séquences d'actions au-delà de 2, afin d'étudier l'impact sur les performances d'apprentissage.

3.1 L'impact de l'extension étendue des observations

Dans cette section, nous étudions l'impact de l'extension multi-étapes des observations sur les performances de l'algorithme. Nous avons analysé les effets d'une extension de 2 et 3 étapes des observations dans différents environnements, qu'ils soient complètement observables ou partiellement observables.Cette analyse vise à comprendre si l'ajout d'informations temporelles supplémentaires permet d'améliorer la performance de l'algorithme en favorisant une meilleure prise de décision et en réduisant l'incertitude liée à la dynamique de l'environnement.

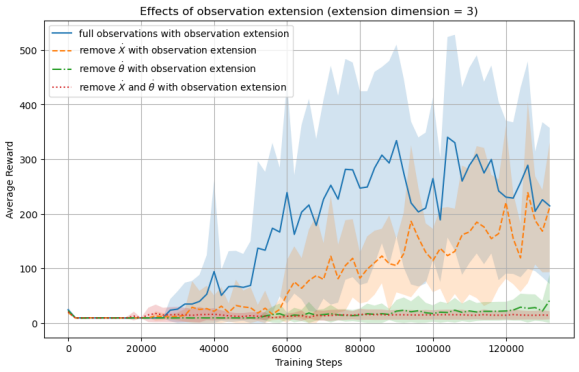


FIGURE 6 – dim=3

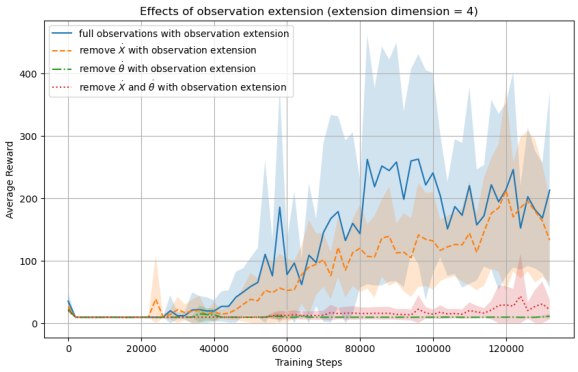
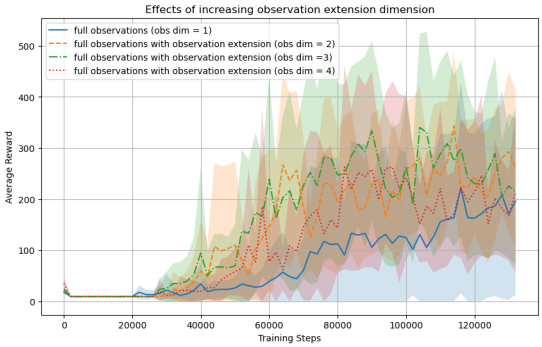


FIGURE 7 – dim=4

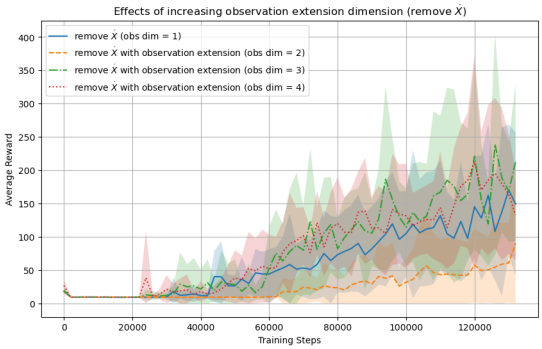
FIGURE 8 – L'impact de l'extension étendue d'observation

Dans l'ensemble, les résultats de l'extension multi-pas n'ont pas significativement amélioré les performances, ni réduit les écarts entre les différents environnements partiellement observables. Cela pourrait être dû au fait que, bien que les extensions temporelles ajoutent des informations passées, elles n'apportent pas de nouvelles données suffisamment perti-

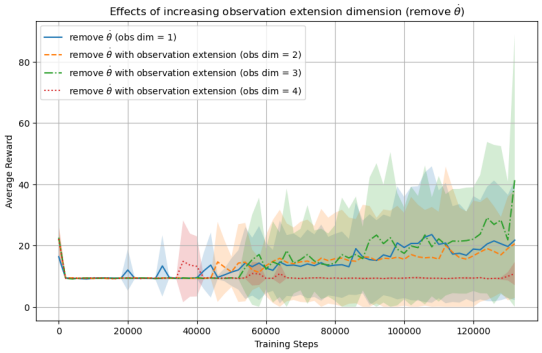
nentes pour compenser les caractéristiques manquantes. De plus, l'extension sur plusieurs étapes pourrait introduire de la redondance ou du bruit, compliquant ainsi le processus d'apprentissage et rendant plus difficile l'exploitation efficace de ces informations supplémentaires.



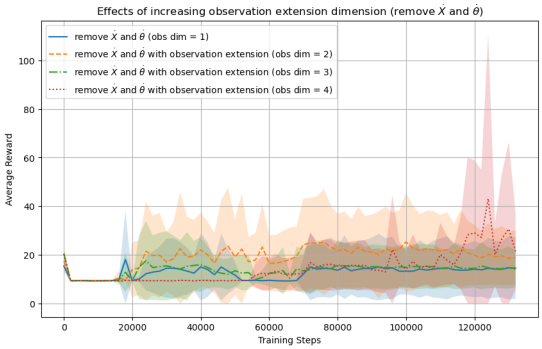
(a) Full observation



(b) Remove $\dot{x}$



(c) Remove $\dot{\theta}$



(d) Remove both

FIGURE 9 – L'impact local de l'extension d'observation

Localement, nous pouvons voir que dans certains environnements (comme par exemple lorsque $\dot{x}$ est supprimé ou lorsque l'environnement est complètement observé), la récompense moyenne du modèle s'améliore nettement. Cependant, lorsque la dimension de l'extension devient trop élevée, bien que la performance soit globalement bonne, on observe une plus grande volatilité. Cela indique que l'ajout de dimensions d'observation supplémentaires apporte certes des informations, mais peut également introduire des redondances ou du bruit, rendant ainsi le modèle moins stable.

Dans certains environnements (par exemple, lorsque $\dot{x}$ et $\dot{\theta}$ sont supprimés ou lorsque seulement $\dot{\theta}$ est supprimé), les performances du modèle n'ont pas montré d'amélioration significative. Même avec la dimension d'observation maximale (dim=4), les résultats restent relativement faibles, suggérant que la perte d'information dans ces cas ne peut pas être facilement compensée par l'ajout d'historique d'observations, voire que ces informations supplémentaires peuvent parfois accroître l'incertitude durant l'entraînement.

3.2 L'impact de l'extension étendue des actions

De la même manière, nous avons analysé, dans cette partie, les effets de l'extension d'actions sur 2 et 3 étapes, afin de déterminer si l'ajout d'historique d'actions aide l'agent à mieux apprendre et à prendre des décisions plus efficaces.

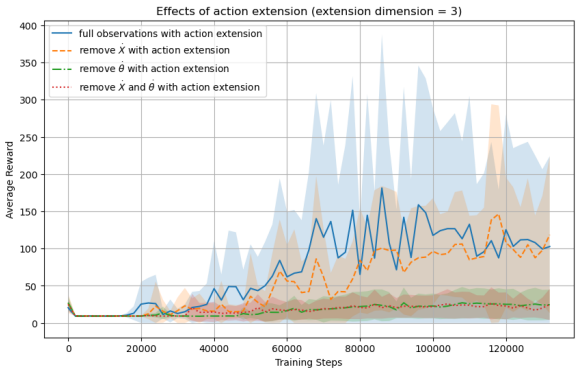


FIGURE 10 – dim=3

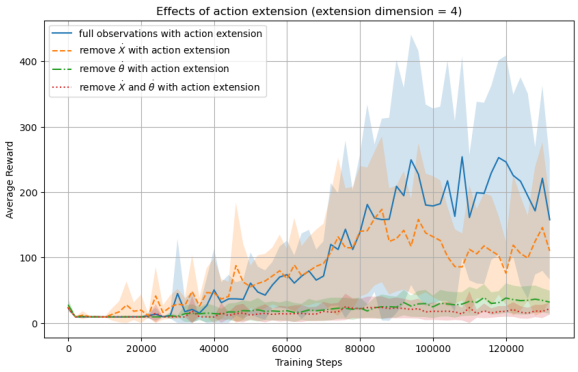


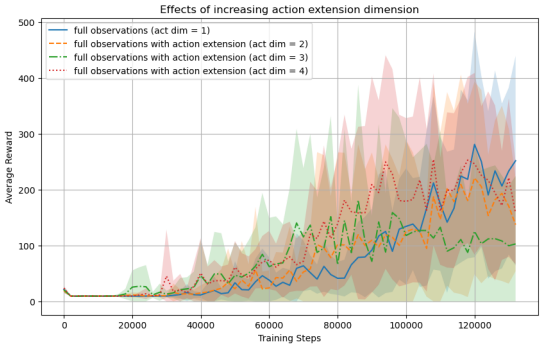
FIGURE 11 – dim=4

FIGURE 12 – L'impact de l'extension étendue d'action

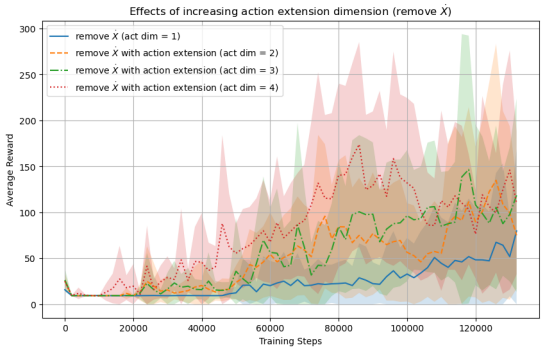
D'après ces deux graphiques, les courbes pour les dimensions d'extension d'action de 3 et 4 suivent des tendances globalement similaires, avec des différences peu marquées. Que ce soit avec une dimension de 3 ou de 4, l'extension d'action a apporté une amélioration des performances, notamment dans les environnements d'observation complète et ceux où $\dot{x}$ a été supprimé. Dans certaines zones, les récompenses moyennes atteignent des niveaux élevés pour ces deux dimensions.

Cependant, bien que la tendance globale soit similaire, la dimension 4 semble présenter une plus grande volatilité, notamment dans les environnements où $\dot{\theta}$ ou à la fois $\dot{x}$ et $\dot{\theta}$ sont supprimés. Cette volatilité pourrait s'expliquer par une complexité ou une instabilité accrue liée à l'introduction de dimensions d'action supplémentaires, rendant difficile l'exploitation efficace de ces informations additionnelles par le modèle.

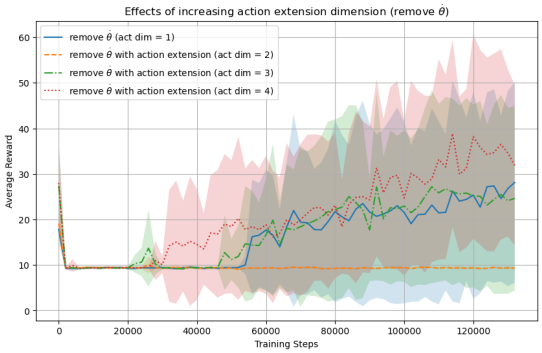
En résumé, l'extension d'action contribue probablement à l'amélioration des performances, mais une augmentation trop importante de la dimension peut entraîner davantage d'incertitudes ou de fluctuations, particulièrement dans les environnements partiellement observables.



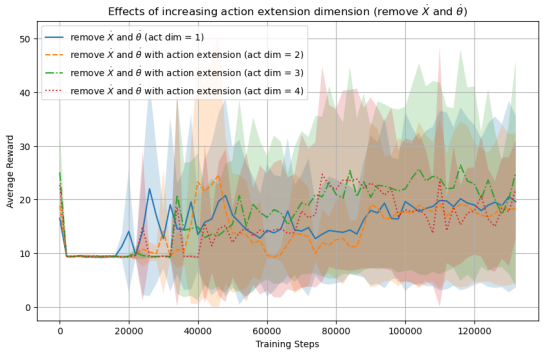
(a) Full observation



(b) Remove $\dot{x}$



(c) Remove $\dot{\theta}$



(d) Remove both

FIGURE 13 – L'impact local de l'extension d'action

Pour l'observation complète, quelle que soit l'augmentation du nombre d'actions fournies, le résultat final reste inférieur à celui obtenu en n'introduisant qu'une seule action dans le critic.

Pour le cas de la suppression de $\dot{x}$, nous pouvons observer sur le graphique que l'utilisation de l'action extension wrapper permet d'améliorer légèrement les performances d'apprentissage par rapport à l'absence de ce wrapper. Cependant, il est difficile de conclure à une corrélation claire entre le nombre d'actions étendues et l'amélioration des performances.

Pour le cas de la suppression de $\dot{\theta}$, lorsque nous étendons le nombre d'actions à un seul supplémentaire, les performances d'apprentissage diminuent considérablement. Cependant, lorsque nous augmentons le nombre d'actions à 3 ou 4, l'effet négatif semble se stabiliser, et les performances sont comparables à celles obtenues avec une seule action.

Pour le cas de la suppression de deux features, les résultats ne montrent pas de différence significative, quelle que soit l'extension du nombre d'actions.

En résumé, bien que l'extension des actions présente théoriquement un potentiel pour améliorer les performances d'apprentissage, son effet s'avère instable dans les expériences actuelles. Il est donc nécessaire d'utiliser l'action extension de manière sélective. De plus, un fine-tuning du nombre d'actions est essentiel pour définir un paramètre optimal qui maximise les performances sans nuire à l'apprentissage. Par ailleurs, il est possible qu'en dehors des valeurs testées dans nos expériences, un autre nombre d'actions pourrait potentiellement améliorer les performances d'apprentissage.

4 Difficultés rencontrées

Temps de l'apprentissage et consommation de mémoire : En raison du besoin de répéter de nombreuses expériences, le temps d'apprentissage est multiplié. De plus, notre code ne prend pas en compte la consommation de mémoire, car nous avons sauvegardé chaque modèle après l'entraînement au lieu de simplement conserver les résultats d'apprentissage. Cela entraîne une consommation excessive de mémoire et provoque des plantages fréquents du notebook lorsque la mémoire est saturée. Dans les projets futurs, il sera essentiel de prendre en compte ces deux problèmes.

5 Conclusion

Nous avons vérifié à travers nos expériences que, dans des situations spécifiques, l'observation extension et l'action extension peuvent améliorer les performances d'apprentissage. Cependant, cet effet est instable, en grande partie en raison de la forte composante aléatoire des processus d'exploration des modèles de RL, ce qui entraîne une variance élevée dans les résultats obtenus. Ainsi, il est nécessaire d'effectuer un fine-tuning afin de déterminer un ensemble de paramètres appropriés, adaptés à l'environnement spécifique. Cela permettra de mieux stabiliser les résultats et de maximiser l'efficacité de l'observation extension et de l'action extension dans le contexte d'apprentissage.