David FILLIAT
Professeur
Unité Informatique et Ingénierie des Systèmes
ENSTA Paris – INRIA FLOWERS
Tel : +33 1 81 87 20 34
david.filliat@ensta-paris.fr

Palaiseau, le 22/11/2022

ENSTA Paris / U2IS

Objet : Rapport sur le manuscrit de thèse de Thomas CHAFFRE

Le manuscrit proposé par Thomas CHAFFRE, intitulé « Reinforcement learning and transfer of adaptive control parameters for improved robustness to unobservable current disturbance », comporte 6 chapitres, y compris l'introduction et la conclusion, sur 111 pages, suivis d'une bibliographie fournie, qui couvre bien l'état de l'art de l'apprentissage par renforcement appliqué à la robotique réelle, et d'une annexe présentant les articles publiés dans le cadre de ces travaux. Le manuscrit est rédigé en anglais, très agréable à lire, avec une présentation de très bonne qualité. Les explications sont claires et précises, et les illustrations nombreuses et de bonne qualité complètent très bien le texte et facilitent la lecture.

L'introduction présente le contexte général des véhicules autonomes et de leur contrôle et se focalise sur les applications aux véhicules sous-marins. Elle détaille en particulier les défis tels que les dynamiques peu ou mal connues, les non-linéarités ou les conditions environnementales changeantes et les grandes approches de contrôle pouvant s'adapter à ces challenges, que ce soit les approches de contrôle adaptatif ou les méthodes par apprentissage, ainsi que les approches en boucles ouvertes ou fermées. Il illustre ensuite très rapidement les limites des approches 'classiques' sur deux exemples simples. Ce contexte très clair permet alors de définir des problématiques et des objectifs de recherche sur le développement d'approches par apprentissage pour le contrôle adaptatif. Elle se termine par une présentation de la structure du manuscrit et de la liste des publications. Cette introduction très pédagogique positionne très bien le travail, en particulier par un positionnement très équilibré entre les approches contrôle et les approches par apprentissage.

Le chapitre deux présente ensuite les éléments de fond utilisés pour conduire les travaux. Il commence par les différentes approches de contrôle adaptatif, avec ou sans modèle et par apprentissage. Cette partie est assez courte, mais permet de bien présenter les principales approches auxquelles il est intéressant de comparer les méthodes d'apprentissage par renforcement. Il présente ensuite de manière particulièrement complète et pédagogique l'apprentissage par renforcement, partant des bases et allant progressivement jusqu'aux méthodes acteur-critique utilisées dans les travaux (algorithme SAC), en insistant sur les points les plus pertinents pour ses travaux (le compromis exploration/exploitation, l'utilisation de « replay buffer », …). Il discute également des difficultés majeures dans l'utilisation de l'apprentissage par renforcement qui seront abordées dans la thèse, en particulier la définition de récompenses, l'exploitation intelligente du rejeu de transitions mémorisées, et les problèmes de glissement de domaines qui conduisent à la surestimation des fonctions de valeurs dans des zones mal explorées. Le chapitre présente enfin le simulateur et les modèles des deux robots sous-marins utilisés dans cette thèse. Ce chapitre est clair, très pédagogique et couvre très bien les bases utiles pour apprécier les travaux présentés dans le manuscrit.

Le chapitre trois détaille ensuite le nouveau système de contrôle adaptatif proposé comme contribution principale de cette thèse. Il débute par un état de l'art des travaux appliquant des méthodes d'apprentissage au contrôle de robot sous-marins, qui sont finalement assez peu nombreux. Il présente ensuite trois études préliminaires, qui permettent de mettre cette contribution en contexte en en présentant la genèse. La première étude, conduite en collaboration avec un autre doctorant, porte sur l'application directe d'une approche d'apprentissage par renforcement, avec la puissance des moteurs comme espace d'action et sa comparaison avec un contrôleur PID classique. La tâche visée est celle de ralliement de

point dans un environnement perturbé par des courant inconnus. Cette approche (comme toute celles décrites dans la suite de la thèse) est décrite de manière très claire et précise, avec tous les détails nécessaires, et les résultats expérimentaux sont présentés et commentés de manière très pertinente. Ces résultats montrent clairement que cette approche 'naïve' offre des performances très faibles par rapport à un contrôle classique, même très simple, et qu'il y a donc nécessité de développer des approches plus pertinentes. La deuxième étude présentée, appliquée cette fois à un drone aérien, introduit plusieurs idées clés exploitées dans la thèse. En premier lieu, elle propose une paramétrisation des pôles du contrôleur PID par l'apprentissage par renforcement, ce qui permet d'exploiter plus finement les avantages de ce contrôleur que lorsque l'apprentissage contrôle directement ses gains. Elle propose également de prendre plus d'historique des états en compte, en prenant les états précédents et les différences entre états précédents en entrée, afin de permettre une meilleure observation des perturbations qui ne sont pas explicitement mesurées. Les résultats expérimentaux montrent clairement les bénéfices de cette approche par rapport à l'approche sans modèle et à l'approche de contrôle PID avec des gains fixes en termes de qualité du contrôle et de taux de succès des tâches. La troisième étude porte ensuite sur une étude empirique de la stabilité des contrôleurs appris, partant du constat qu'il n'y a pas de garanties théoriques simples à cause de l'utilisation des réseaux de neurones au sein du contrôleur et des variations du processus contrôlé. Une structure de contrôle à base de PID conçus à partir de la stabilité de Lyapunov permet de montrer que les contrôleurs appris, même s'il de fournissent pas de garanties théoriques, conduisent à des états stables de manière très proches du contrôleur adaptatif garanti. Les résultats montrent cependant que la stabilité des paramètres du contrôleur appris est très différente du contrôleur de référence, et il aurait été intéressant de discuter plus en détails les raisons et les conséquences de cette différence. Suite à ces trois études, le chapitre formule la contribution principale, celle d'une nouvelle méthode de contrôle adaptatif. Il reprend ainsi l'idée de paramétriser les pôles d'un contrôleur PID, avec une approche directe un peu différente de l'étude précédente qui est bien justifiée. Cette approche présente notamment un fort intérêt en termes d'interprétabilité car la position des pôles permet de garantir une certaine stabilité et de borner les dynamiques de réponse du système. Il propose ensuite une approche d'apprentissage exploitant l'algorithme Soft Actor Critic, en détaillant plusieurs choix d'implémentation et une nouvelle méthode d'échantillonnage de la mémoire de rejeu s'inspirant de caractéristiques connues du rejeu d'expérience observé chez l'animal. Cette dernière proposition est intéressante, car elle introduit des motivations assez différentes de celles des travaux présentés jusque-là qui étaient surtout issus de théorie du contrôle et de l'apprentissage par renforcement. Ces motivations sont parfois contradictoires, par exemple l'idée d'exploiter un rejeu corrélé temporellement est observée en biologie, alors que c'est justement en partie pour lutter contre ces corrélations temporelles que les mémoires de rejeu ont été introduites. Ce point est discuté dans le manuscrit, mais il aurait été intéressant d'affiner les éléments qui permettraient de trouver le meilleur compromis entre indépendance statistique et corrélation temporelle.

Le chapitre quatre présente ensuite les expérimentations permettant de valider l'approche en simulation. La tâche retenue est de stabiliser la direction et la vitesse d'un AUV dans un environnement avec des courants inconnus. Une première comparaison du contrôleur appris avec le PID de référence, en restant dans le domaine d'entrainement, montre un gain de performance très important, en particulier dans les cas les plus difficiles de perturbations et d'objectifs changeant au cours du temps, même si le contrôleur appris montre aussi ses limites dans ce scénario. Une deuxième évaluation s'intéresse ensuite à la généralisation des contrôleurs pour des perturbations allant au-delà du domaine d'entrainement. Quelques éléments de la méthode sont affinés dans ce cadre (historique des entrées, limites des paramètres, métrique d'évaluation), et il est dommage de ne pas avoir évalué les apports de ces modifications par rapport à la première évaluation. Les évaluations sur des scénarios avec des niveaux de généralisation croissants montrent une certaine capacité de généralisation des approches d'apprentissage. Cependant, les performances baissent assez fortement dans les scénarios les plus difficiles, alors que les performances du contrôleur de référence restent relativement plus élevées (scénarios 4 et 5). Ce n'est que pour les perturbations non stationnaires (scénarios 6) que l'approche par apprentissage reprend le dessus. Il aurait ainsi été intéressant d'analyser plus finement les raisons des différences entre ces deux cas et les éléments qui sont les plus limitatifs pour les performances des deux approches. Cette évaluation est aussi l'occasion d'évaluer la nouvelle approche de rejeu de données inspirée de la biologie. Cette approche montre des gains assez clairs de stabilité et de vitesse de l'apprentissage. Cependant, il aurait été intéressant d'analyser l'importance des différents paramètres de l'approche proposée (taille relative des

deux mémoires, importance du rejeu corrélé temporellement, …) dans ces gains de performances, et de l'évaluer sur des benchmarks standard d'apprentissage par renforcement.

Le chapitre cinq présente ensuite des évaluations réalisées sur plateforme réelle. La description des moyens expérimentaux est très détaillée et montre très bien l'ampleur des travaux réalisés pour cette validation, avec un AUV en piscine incluant un générateur de perturbations et des moyens de mesure permettant de fournir les trajectoires réelles pour l'évaluation. L'approche choisie est de réaliser l'entrainement en simulation, puis d'évaluer les contrôleurs appris sur plateforme réelle. Pour cela, la méthode d'apprentissage utilisée apporte de nouveau quelques modifications par rapport aux méthodes employées précédemment, en particulier l'algorithme SAC intègre un réglage automatique de paramètres, et une méthode de « domain randomization » originale est intégrée en tirant aléatoirement un contexte de simulation parmi trois de difficulté différentes. L'évaluation des contrôleurs obtenus est très complète et montre bien le gain important obtenu par rapport au contrôleur de référence, à la fois dans les situations avec et sans perturbations. En particulier, le contrôleur adaptatif appris est beaucoup moins sensible aux perturbations que le contrôleur fixe.

Pour conclure le manuscrit, le dernier chapitre propose un résumé des différentes contributions. Cette présentation montre bien le recul pris sur les travaux car elle présente les travaux de manière assez différente du corps du manuscrit, en reprenant les grandes problématiques abordées (mécanismes d'ajustement, transfert simulation-réalité, efficacité en termes de nombre d'échantillons…). Il propose également un ensemble de perspectives de prolongation des travaux réalisés qui montrent une bonne compréhension des enjeux et des pistes pertinentes.

En conclusion de ce rapport, ce manuscrit présente des contributions très pertinentes sur l'utilisation de l'apprentissage par renforcement pour le contrôle adaptatif de robots sous-marins. Il montre également une très bonne connaissance de l'état de l'art et l'ensemble des travaux est très bien positionné par rapport à cet existant, à la fois du point de vue du contrôle et de l'apprentissage par renforcement. Un point fort de ce manuscrit est clairement la qualité de la rédaction et du travail de validation expérimental, à la fois sur des simulations variées et sur un robot réel. Les méthodes présentées sont pertinentes, mais on peut cependant regretter l'absence de comparaison de toutes les approches proposées sur une même tâche car chaque évaluation apporte des modifications à l'algorithme utilisé pour l'évaluation précédente. Chaque contribution est ainsi très bien positionnée individuellement par rapport aux contrôleurs « classiques », mais elles ne sont pas directement comparées entre elles. Il serait de même intéressant de conduire des études d'ablation (notamment pour la méthode de rejeu bio-inspirée) afin de montrer clairement les éléments les plus importants des approches. Enfin, ces travaux ont mené à cinq publications dans conférences et journaux internationaux, ce qui atteste également de leur grande qualité.

De mon point de vue, le travail proposé remplis donc parfaitement les objectifs d'un travail de thèse et pour toutes ces raisons, je donne un avis entièrement favorable à la soutenance de la thèse de Thomas Chaffre.

David Filliat

Professeur ENSTA Paris