
# M1 Informatique - UE Projet

# Carnet de bord : les coulisses de la recherche documentaire

**Spécialité :**

**Agents Distribues, Robotique, Recherche Opérationnelle, Interaction, Décision**

**Sujet :**

**Étude de l'apprentissage par renforcement profond sur Pendulum**

**Étudiants :**

**FU Vincent, ANDROIDE**
**LIU Yuhao, ANDROIDE**

# Table des matières

# 1. Introduction

De la reconnaissance faciale à la robotique en passant par les voitures autonomes, toutes ces domaines utilisent des techniques essentielles de l'Intelligence Artificielle : les algorithmes d'apprentissage par renforcement.

L'apprentissage par renforcement consiste à apprendre à un agent (une entité capable d'agir de façon autonome : un robot, etc ...) la meilleure façon d'interagir dans un environnement à partir de ses tentatives d'interaction successives. La qualité des interactions de l'agent avec son environnement est entièrement définie par une valeur de récompense. On cherche alors à maximiser cette récompense en indiquant à l'agent la meilleure façon d'y parvenir qu'on nomme la politique optimale. Le but est alors d'évaluer les performances (la capacité à atteindre une récompense maximale) les algorithmes classiques (dit issus de l'état de l'art de par leurs efficacités) dans un environnement particulier. En effet, selon l'environnement étudié et l'algorithme utilisé, les performances de ces algorithmes peuvent différer.

L'enjeu est donc de comprendre quels sont les différentes caractéristiques que certains algorithmes possèdent amenant à une bonne performance et en quoi ces caractéristiques sont critiques pour les performances de ces derniers par rapport à l'environnement étudié. L'environnement sélectionné dans le cadre du projet se nomme Pendulum où le but est d'apprendre à une machine inamovible tenant une pendule à tige rigide à masse des rotations de tige afin d'équilibrer la tige vers la position verticale vers le haut et d'y rester le plus longtemps possible sous la contrainte de gravité. Des précédents travaux ont montré l'inefficacité de l'algorithme Policy Gradient sur Pendulum alors que les algorithmes issus de l'état de l'art se révèlent performants. Le principal défi est donc d'expliquer les performances de ces derniers sur Pendulum par rapport à l'algorithme Policy Gradient à travers différents tests de performance.

# 2. Mots clés retenus

En raison de la complexité du sujet et du niveau de connaissances à acquérir, nous devions regarder une série de vidéos de notre encadrant (M. Olivier Sigaud) sur les cours d'apprentissage en robotique. De ce fait, l'ensemble des mots clés sélectionnés nous viennent naturellement puisque ces mots clés représentent les bases et les notions principales de l'apprentissage par renforcement.

Voici quelques mots clés importants :

| | | |
|---|---|---|
| Reinforcement Learning | — Markov Decision Process<br>— Policy<br>— Dynamic Programming<br>— Bellman Optimally Operator | — Value Function<br>— Action-Value Function<br>— On policy<br>— Off policy |
| Deep Reinforcement Learning | — Continuous Action Space<br>— Policy Gradient<br>— Gradient descent<br>— State-dependent Baseline<br>— Monte Carlo approach<br>— Bootstrap approach<br>— Temporal Difference | — N-step return<br>— Soft Actor-Critic<br>— Entropy Regularization<br>— Trust Region Policy Optimization<br>— Deep Deterministic Policy Gradient<br>— Stochastic policies<br>— Replay Buffer |

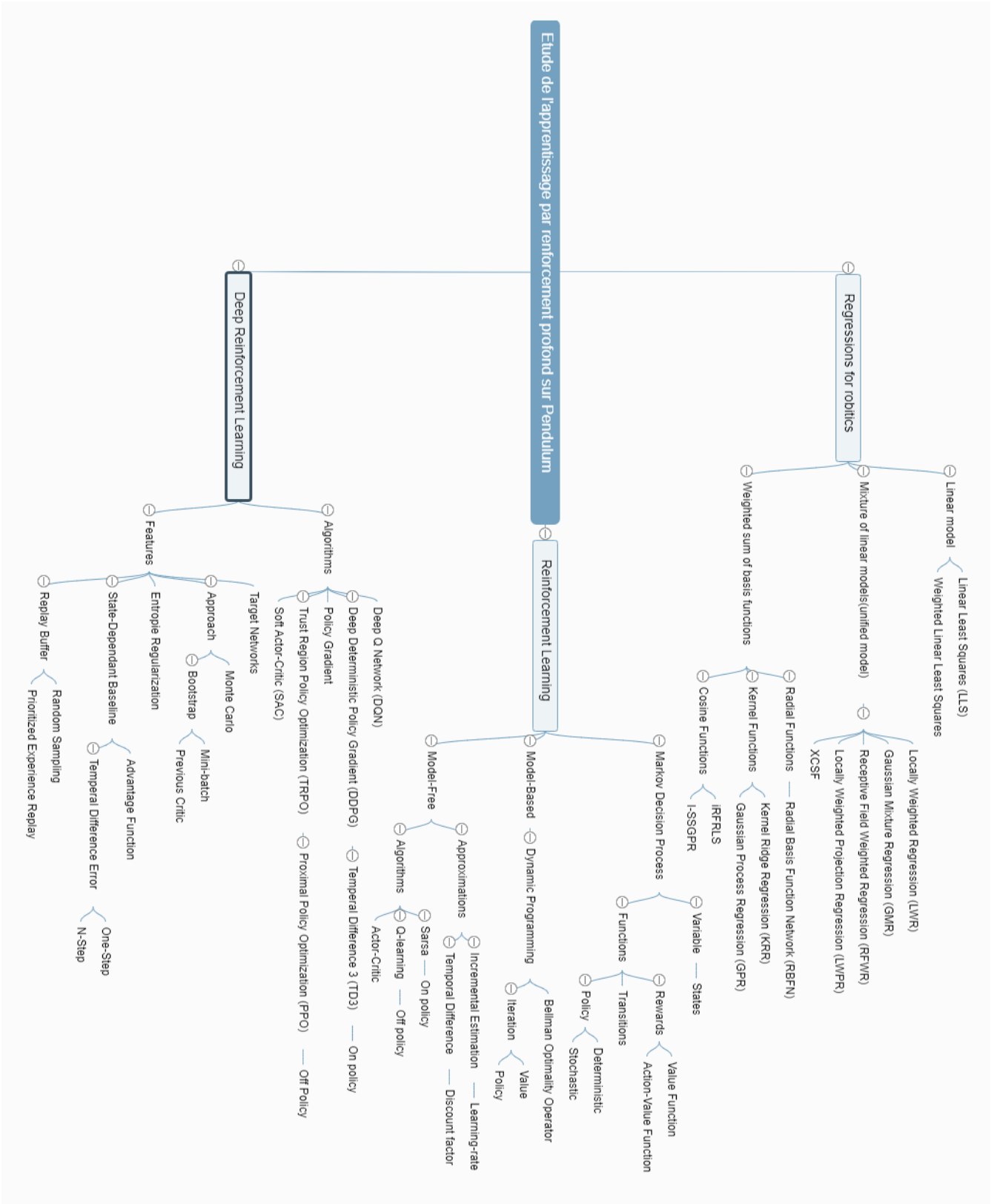


FIGURE *1* – *Carte heuristique*

## 3. Descriptif de la recherche documentaire

Comme cité précédemment, nous devions dès le début du projet assimiler les cours sur l'apprentissage par renforcement. De ce fait, un nombre important de références nous est proposé dont une qui représente la référence sur l'apprentissage par renforcement : le livre de Sutton et Barto [1] nommé Reinforcement Learning : an introduction (deuxième édition) édité en 2018 (la première édition datant de 1998). Ce livre décrit toutes les bases en long et en large de l'apprentissage par renforcement notamment en détaillant toutes les algorithmes de l'état de l'art, leurs caractéristiques et quelques applications comme l'exemple d'AlphaGo. Ainsi, l'ensemble de notre recherche était issu de nos réfléxions sur les algorithmes de l'état de l'art décrit dans ce livre.

Nous avions donc dans un premier temps utilisé les bases de données de ressources en ligne de Sorbonne Université, en particulier la Digital Library de l'Association for Computing Machinery (ACM). Cette base de données regroupe différents types de publications (articles, journaux, revues, compte-rendus de conférences, etc ...) centrés autour de l'informatique. La Digital Library possède également des outils de recherche très efficaces comme la recherche de profils d'auteurs (où on peut observer leurs sujets de recherche, leurs articles, leurs collègues) ou encore la recherche bibliographique par requête (il est possible d'écrire des formules booléennes avec des mots-clés ou noms d'auteurs etc ...). Sachant que le livre de Sutton et Barto est quasiment cité dans les références bibliographiques de tout article sur l'apprentissage par renforcement, nous avons entré leur noms dans leur moteur de recherche puis sélectionné et filtré en fonction de leurs titres et contenus seulement les articles en lien avec notre projet. Les sources en lien avec notre projet sont relativement nombreux (il n'y a pas eu de difficulté à trouver des articles sur le sujet) puisque l'apprentissage par renforcement est un domaine de recherche très actif actuellement. L'avantage de ces articles écrits en anglais est de donner une vision plus approfondie du domaine et requiert donc un minimum de connaissances pour le lecteur. De plus, les articles issus du Digital Library d'ACM sont examinés et vérifiés selon des normes par des pairs compétents assurant la qualité des publications.

Par la suite afin de diversifier notre bibliographie et en observant que certaines références que notre professeur nous a fournies n'étaient pas présentes dans la base de d'ACM, nous avions utilisé une autre base de données : arXiv regroupant des articles du domaine des sciences exactes, effectué une recherche bibliographique par mot-clés et sélectionné les articles pertinents. Cependant, la spécificité de l'archive arXiv est que les publications sont des preprints. C'est-à-dire qu'il n'y a pas d'évaluations par des pairs. Cela permet aux auteurs d'être plus libre sur leurs publications mais à l'inverse le lecteur doit être beaucoup plus critique sur les publications et à lui seul de vérifier la pertinence des articles.

## 4. Bibliographie produite dans le cadre du projet

## Références


[1] Zafarali Ahmed et al. "Understanding the Impact of Entropy on Policy Optimization". en. In : *International Conference on Machine Learning*. ISSN : 2640-3498. PMLR, mai 2019, p. 151-160. URL : http://proceedings.mlr.press/v97/ahmed19a.html (visité le 26/02/2021).

[2] Marcin Andrychowicz et al. "What Matters In On-Policy Reinforcement Learning ? A Large-Scale Empirical Study". In : *arXiv :2006.05990 [cs, stat]* (juin 2020). arXiv : 2006.05990. URL : http://arxiv.org/abs/2006.05990 (visité le 26/02/2021).

[3] Andras Antos, Rémi Munos et Csaba Szepesvari. *Fitted Q-iteration in continuous action-space MDPs*. en. report. 2007, p. 24. URL : https://hal.inria.fr/inria-00185311 (visité le 26/02/2021).

[4] Jonathan Baxter et Peter L. Bartlett. "Infinite-horizon policy-gradient estimation". In : *Journal of Artificial Intelligence Research* 15.1 (nov. 2001), p. 319-350. ISSN : 1076-9757.

[5] Marc Peter Deisenroth. "A Survey on Policy Search for Robotics". en. In : *Foundations and Trends in Robotics* 2.1-2 (2011), p. 1-142. ISSN : 1935-8253, 1935-8261. DOI : 10.1561/2300000021. URL : http://www.nowpublishers.com/articles/foundations-and-trends-in-robotics/ROB-021 (visité le 26/02/2021).

[6] Scott Fujimoto, Herke Hoof et David Meger. "Addressing Function Approximation Error in Actor-Critic Methods". en. In : *International Conference on Machine Learning*. ISSN : 2640-3498. PMLR, juil. 2018, p. 1587-1596. URL : http://proceedings.mlr.press/v80/fujimoto18a.html (visité le 26/02/2021).

[7] Mohammad Ghavamzadeh, Yaakov Engel et Michal Valko. "Bayesian policy gradient and actor-critic algorithms". In : *The Journal of Machine Learning Research* 17.1 (jan. 2016), p. 2319-2371. ISSN : 1532-4435.

[8] Tuomas Haarnoja et al. "Reinforcement Learning with Deep Energy-Based Policies". en. In : *International Conference on Machine Learning*. ISSN : 2640-3498. PMLR, juil. 2017, p. 1352-1361. URL : http://proceedings.mlr.press/v70/haarnoja17a.html (visité le 26/02/2021).

[9] Tuomas Haarnoja et al. "Soft Actor-Critic Algorithms and Applications". In : *arXiv :1812.05905 [cs, stat]* (jan. 2019). arXiv : 1812.05905. URL : http://arxiv.org/abs/1812.05905 (visité le 26/02/2021).

[10] Tuomas Haarnoja et al. "Soft Actor-Critic : Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor". en. In : *International Conference on Machine Learning*. ISSN : 2640-3498. PMLR, juil. 2018, p. 1861-1870. URL : http://proceedings.mlr.press/v80/haarnoja18b.html (visité le 26/02/2021).

[11] Dingcheng Li et al. "Video Recommendation with Multi-gate Mixture of Experts Soft Actor Critic". In : *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '20. New York, NY, USA : Association for Computing Machinery, juil. 2020, p. 1553-1556. ISBN : 978-1-4503-8016-4. DOI : 10.1145/3397271.3401238. URL : https://doi.org/10.1145/3397271.3401238 (visité le 02/04/2021).

[12] Jan Hendrik Metzen et Frank Kirchner. "Model-based direct policy search". In : *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems : volume 1 - Volume 1*. AAMAS '10. Richland, SC : International Foundation for Autonomous Agents et Multiagent Systems, mai 2010, p. 1589-1590. ISBN : 978-0-9826571-1-9. (Visité le 02/04/2021).

[13] Yiming Peng et al. "NEAT for large-scale reinforcement learning through evolutionary feature learning and policy gradient search". In : *Proceedings of the Genetic and Evolutionary Computation Conference*. GECCO '18. New York, NY, USA : Association for Computing Machinery, juil. 2018, p. 490-497. ISBN : 978-1-4503-5618-3. DOI : 10.1145/3205455.3205536. URL : https://doi.org/10.1145/3205455.3205536 (visité le 02/04/2021).

[14] Martin Riedmiller. "Neural Fitted Q Iteration – First Experiences with a Data Efficient Neural Reinforcement Learning Method". en. In : *Machine Learning : ECML 2005*. Sous la dir. de João Gama et al. Lecture Notes in Computer Science. Berlin, Heidelberg : Springer, 2005, p. 317-328. ISBN : 978-3-540-31692-3. DOI : 10.1007/11564096_32.

[15] John Schulman et al. "Gradient Estimation Using Stochastic Computation Graphs". In : *arXiv :1506.05254 [cs]* (jan. 2016). arXiv : 1506.05254. URL : http://arxiv.org/abs/1506.05254 (visité le 26/02/2021).

[16] John Schulman et al. "High-Dimensional Continuous Control Using Generalized Advantage Estimation". In : *arXiv :1506.02438 [cs]* (oct. 2018). arXiv : 1506.02438. URL : http://arxiv.org/abs/1506.02438 (visité le 26/02/2021).

[17] Olivier Sigaud et Freek Stulp. "Policy search in continuous action domains : An overview". en. In : *Neural Networks* 113 (mai 2019), p. 28-40. ISSN : 0893-6080. DOI : 10.1016/j.neunet.2019.01.011. URL : https://www.sciencedirect.com/science/article/pii/S089360801930022X (visité le 26/02/2021).

[18] B. Stapelberg et K. M. Malan. "Global structure of policy search spaces for reinforcement learning". In : *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. GECCO '19. New York, NY, USA : Association for Computing Machinery, juil. 2019, p. 1773-1781. ISBN : 978-1-4503-6748-6. DOI : 10.1145/3319619.3326843. URL : https://doi.org/10.1145/3319619.3326843 (visité le 02/04/2021).

[19] Richard S. Sutton et Andrew G. Barto. *Reinforcement Learning, second edition : An Introduction*. en. Google-Books-ID : uWV0DwAAQBAJ. MIT Press, nov. 2018. ISBN : 978-0-262-35270-3.

[20] Matthew E. Taylor, Shimon Whiteson et Peter Stone. "Transfer via inter-task mappings in policy search reinforcement learning". In : *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*. AAMAS '07. New York, NY, USA : Association for Computing Machinery, mai 2007, p. 1-8. ISBN : 978-81-904262-7-5. DOI : 10.1145/1329125.1329170. URL : https://doi.org/10.1145/1329125.1329170 (visité le 02/04/2021).

[21] Eiji Uchibe. "Efficient sample reuse in policy search by multiple importance sampling". In : *Proceedings of the Genetic and Evolutionary Computation Conference*. GECCO '18. New York, NY, USA : Association for Computing Machinery, juil. 2018, p. 545-552. ISBN : 978-1-4503-5618-3. DOI : 10.1145/3205455.3205564. URL : https://doi.org/10.1145/3205455.3205564 (visité le 02/04/2021).

[22] Herke Van Hoof, Gerhard Neumann et Jan Peters. "Non-parametric policy search with limited information loss". In : *The Journal of Machine Learning Research* 18.1 (jan. 2017), p. 2472-2517. ISSN : 1532-4435.

# 5. Évaluation des sources

Évaluation de la source 9 :

La source a été publiée le 29 janvier 2019 par Tuomas Haarnoja et ses collaborateurs. Au vu de sa date de publication, l'article est d'actualité et donc pertinent dans un domaine comme l'Intelligence Artifielle où les techniques sont en constante évolution. L'auteur principal Tuomas Haarnoja issu de l'Université de Californie Berkeley, est actuellement chercheur pour DeepMind, l'entreprise qui a conçu l'agent à intelligence artificielle AlphaGo entraînant un engouement pour l'intelligence artificielle. Il est notamment avec ses collaborateurs les premiers à définir l'algorithme Soft Actor-Critic, un algorithme performant au point de rivaliser avec les autres algorithmes issus de l'état de l'art. Tuomas Haarnoja est donc compétent dans son domaine de par son contribution.

Dans l'article, les auteurs expliquent notamment les défauts des algorithmes Model-free (sample complexity, brittleness to hyperparameters) et en quoi leur algorithme apporte une solution a cela par l'utilisation d'un outil mathématique : l'entropie.

La ressource est également fiable à travers sa forme en respectant les normes d'un article scientifique correctement rédigé : introduction de l'article par rapport à l'état actuel des connaissances sur le sujet, descriptions des formules et algorithmes utilisés et leurs démonstrations théoriques, comparaisons de différents résultats à travers d'outils efficaces comme les graphes et applications pratiques dans le monde réel dans le cadre de l'Intelligence Artificielle et enfin une bibliographie riche de plus de 30 références.

Cependant, l'article a été publié dans l'archive arXiv et est donc un preprint, il n'a donc pas reçu une relecture par des pairs, il faut donc être plus critique à l'égard de l'article. Néanmoins, l'utilisation de divers environnements classiques de benchmark (présents dans de nombreux articles scientifiques) comme Hooper, Walker2d, HalfCheetah et la cohérence des résultats donne suffisamment de fiabilité à l'article.

Évaluation de la source 17 :

La source a été publiée en mai 2019 par Olivier Sigaud et Freek Stulp, et est donc également pertinent par rapport à date de publication. L'auteur Olivier Sigaud, professeur en Robotique à Sorbonne Université qui est également notre responsable de notre projet, travaille actuellement au sein de l'Institut des Systèmes Intelligents et de la Robotique en tant que chercheur. Il est donc compétent dans le domaine de l'intelligence artificielle.

Concernant l'article, le contenu est extrêmement riche et contrairement à l'article de Haarnoja où on détaillait qu'un seul algorithme en profondeur (le Soft Actor-Critic), on donne ici une vision sur plusieurs algorithmes de recherche de politique (Policy Search). L'article n'entre donc pas dans des détails de raisonnements mathématiques mais explicite plus les concepts des différents algorithmes de Policy Search : les principaux caractéristiques, les liens et différences. L'article est donc très intéressant pour des lecteurs recherchant au-delà des connaissances mathématiques, des connaissances conceptuelles sur le sujet.

Concernant la fiabilité de l'article, l'article possède plus de 3 pages de bibliographie dans lesquels on observe en plus des références sur les algorithmes de Policy Search, d'autres références sur des domaines

de recherche subsidiaires par rapport au sujet de l'article comme les algorithmes évolutionnistes, ou encore l'optimisation bayésienne. Cela montre la diversité des réflexions des auteurs sur le sujet donnant beaucoup plus de fiabilité sur la qualité de l'article.

Enfin, l'article est également transparent sur la position des auteurs. On nous informe que Olivier Sigaud est soutenu par la commission Européenne à travers le projet public Deferred Restructuring of Experience in Autonomous Machines. L'article n'est donc pas biaisé et fiable sur ce point.

Évaluation de la source 19 :

Le livre initial a été écrit en 1998 par Sutton et Barto et a été remis à jour dans cette deuxième édition en 2018 par MIT Press. On remarque donc une évolution de l'état des connaissances sur l'apprentissage par renforcement.

Les auteurs Sutton et Barto, professeurs d'informatique respectivement à l'Université de l'Alberta et à l'Université du Massachusetts Amherst sont considérés comme les pionniers de l'apprentissage par renforcement à travers ce livre qui a été cité plus de 41000 fois. Le livre est donc la référence de base pour tout nouveau lecteur souhaitant apprendre l'apprentissage par renforcement.

Long de plus de 500 pages, le livre décrit tous les algorithmes de l'état de l'art en s'appuyant sur des notions mathématiques et des exemples concrets d'applications. Riche plus d'environ dix pages de bibliographie, ce livre est une ressource fiable non seulement de par sa richesse bibliographique mais aussi par son contenu où on peut notamment retrouver des notions récurrentes de l'apprentissage par renforcement : state, policy, reward, value function, action value function dans un cadre mathématique bien connu : Markov decision process.

Enfin, sachant que le livre a été éditer par MIT Press et donc suivi un circuit de publication (vérification par des pairs avant publication, etc ...) d'une communauté scientifique, la ressource est autant plus fiable.