



Optimisation de l'exploitation d'une machine modélisée par un processus de décision markovien partiellement observable (POMDP) aux paramètres inconnus du décideur.

Les POMDP sont des modèles qui prennent en compte les actions, la connaissance (éventuellement erronée) et les motivations d'un décideur. Ainsi, ils modélisent le monde vu par les yeux dudit décideur, ce qui a piqué ma curiosité. L'exploitation de machine (dans une laverie par exemple) en était une application possible.

En supposant qu'une machine soit la même quelle que soit son âge, l'évolution dans le POMDP doit idéalement être statique au niveau d'un ensemble d'états «fonctionnels». Aussi le décideur doit choisir ses actions afin de boucler sur cet ensemble, et faire de toute dérive un cycle afin d'y retourner.

Le candidat atteste avoir travaillé en monôme.

Positionnement thématique (ÉTAPE 1) :

- *INFORMATIQUE (Informatique pratique)*
- *INFORMATIQUE (Informatique Théorique)*
- *MATHEMATIQUES (Autres)*

Mots-clés (ÉTAPE 1) :

Mots-clés (en français)

Processus de Décision Markovien Partiellement Observable

Bandit à plusieurs bras

Optimisation

Politique d'entretien

Apprentissage en ligne

Mots-clés (en anglais)

Partially Observable Markov Decision Process

Multi-armed bandit

Explore/Exploit tradeoff

Maintenance Policy

Online learning

Bibliographie commentée

En 1968, James E. Eckles [1] propose un modèle d'exploitation de machines avec des états infinis et discrets (un nombre fini d'états couplés à une échelle de temps discrète), un nombre fini d'actions possibles et un nombre fini d'observations possibles. Les actions et l'état à un instant donné donnent la distribution de probabilités quant à l'observation faite à cet instant, et à l'état suivant. Le coût engendré dépend de l'état à cet instant, de l'action choisie, et de l'état suivant. Il propose aussi de mesurer la qualité d'une politique d'exploitation à l'espérance du coût qu'elle engendre, pondérée d'un facteur géométrique en fonction de l'horizon où il est engendré, et prouve l'existence d'une politique optimale en ce sens.

Dans un livre de 1998 de la MIT Press, réédité en 2015, R. S. Sutton et A. G. Barto [2] expliquent les problèmes du bandit (machine à sous) à plusieurs bras (leviers), qui traitent de l'équilibre entre exploration de l'environnement et exploitation des connaissances acquises, et ce sans apprentissage préalable ni connaissance de la structure interne du modèle (apprentissage en ligne). Les auteurs expliquent certains algorithmes liés à ces problèmes, comme ε -greedy, UCB et SoftMax, ou encore l'influence du coût moyen estimé ou de l'estimation initiale de la rentabilité d'une action.

Ils présentent aussi les processus des décisions markoviens, et mentionnent leurs homologues partiellement observables.

En 2002, P. Auer, N. Cesa-Bianchi, Y. Freund et R.E. Schapire [3] publient l'algorithme EXP3 pour les problèmes du bandit à plusieurs bras, dans le cas où le gain obtenu en tirant un bras ne suit pas nécessairement une loi de probabilité fixée.

La même année, P. Auer, cette fois-ci avec N. Cesa-Bianchi et P. Fischer [4] proposent des variantes des algorithmes UCB et ε -greedy et étudient la progression du regret lié au fait de ne pas toujours choisir le meilleur levier sur un temps fini.

Finalement en 2022, M. Jafarnia-Jahromi, R. Jain, et A. Nayyar [5] construisent et exposent leur méthode de construction d'un algorithme d'apprentissage en ligne sur POMDP (*PSRL-POMDP*).

Problématique retenue

Comment adapter les algorithmes liés aux bandits à plusieurs bras à l'exploitation sans apprentissage préalable d'une machine plus générale modélisée par un processus de décision markovien partiellement observable ?

Objectifs du TIPE du candidat

- Modéliser une ou plusieurs machines sous la forme de POMDPs
- S'inspirer des algorithmes classiques des bandits à plusieurs bras pour développer des politiques d'exploitation
- Identifier les contraintes supplémentaires apportées par les POMDPs
- Adapter ces politiques à la structure et ces contraintes

Références bibliographiques (ÉTAPE 1)

- [1] JAMES E. ECKLES : Optimum Maintenance with Incomplete Information : <https://doi.org/10.1287/opre.16.5.1058>
- [2] RICHARD S. SUTTON, ANDREW G. BARTO : Reinforcement Learning: An Introduction : <https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf>
- [3] P. AUER, N. CESA-BIANCHI, Y. FREUND, R.E. SCHAPIRE : The nonstochastic multiarmed bandit problem : <https://cseweb.ucsd.edu/~yfreund/papers/bandits.pdf>
- [4] PETER AUER, NICOLÒ CESA-BIANCHI, PAUL FISCHER : Finite-time Analysis of the Multiarmed Bandit Problem : <https://link.springer.com/article/10.1023/A:1013689704352>
- [5] MEHDI JAFARNIA-JAHROMI, RAHUL JAIN, ET ASHUTOSH NAYYAR : Online learning for unknown partially observable MDPs : <https://arxiv.org/pdf/2102.12661v1>