
Bistable Recurrent Cells and Belief Filtering for Q-learning in Partially Observable Markov Decision Processes

Auteur : Lambrechts, Gaspard

Promoteur(s) : Ernst, Damien

Faculté : Faculté des Sciences appliquées

Diplôme : Master : ingénieur civil en science des données, à finalité spécialisée

Année académique : 2020-2021

URI/URL : <http://hdl.handle.net/2268.2/11474>

Avertissement à l'attention des usagers :

Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.

Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.



SUMMARY

Bistable Recurrent Cells and Belief Filtering for Q-learning in Partially Observable Markov Decision Processes

Master's thesis carried out to obtain the degree of Master of Science in Data Science and
Engineering.

GASPARD LAMBRECHTS

supervised by

DAMIEN ERNST

Academic year 2020 - 2021

Summary

Title: Bistable Recurrent Cells and Belief Filtering for Q-learning in Partially Observable Markov Decision Processes

Student: Gaspard Lambrechts

Section: Master in Data Science and Engineering

Supervisor: Damien Ernst

In this master's thesis, reinforcement learning (RL) methods are used to learn (near-)optimal policies to act in several Markov decision processes (MDPs) and partially observable Markov decision processes (POMDPs). More precisely, Q-learning and recurrent Q-learning techniques are used. Some of the considered POMDPs require a high-memorisation ability in order to achieve optimal decision making. In POMDPs, RL techniques usually rely on approximating functions that take as input sequences of observations with variable length. Recurrent neural networks (RNNs) are thus a clever choice of such approximators. This work is based on the recently introduced bistable recurrent cells, which have been empirically shown to provide a significantly better long term memory than standard cells, such as the long short-term memory (LSTM) and the gated recurrent unit (GRU). These cells are named the bistable recurrent cell (BRC) and the recurrently neuromodulated BRC (nBRC). First, by importing these cells for the first time in the RL setting, it is empirically shown that they also provide a significant advantage in memory-demanding POMDPs, in comparison to LSTM and GRU. Second, the ability of the RNN to represent a belief distribution over the states of the POMDP is studied. It is achieved by evaluating the mutual information between the hidden states of the RNN and the belief filtered on the successive observations. This analysis is thus strongly anchored in the theory of information and the theory of optimal control for POMDPs. Third, as a complement to this research project, a new target update is proposed for Q-learning algorithms with target networks, for both reactive and recurrent policies. This new update speeds up learning, especially in environments with sparse rewards.

Résumé

Titre: Cellules récurrentes bistables et filtrage de la distribution sur les états pour le *Q-learning* dans les processus de décisions markoviens partiellement observables

Student: Gaspard Lambrechts

Section: Master ingénieur civil en science des données

Promoteur: Damien Ernst

Dans ce mémoire, des méthodes d'apprentissage par renforcement (RL) sont utilisées pour apprendre des politiques décisionnelles (quasi) optimales dans plusieurs processus de décisions markoviens (MDPs) et processus de décisions markoviens partiellement observables (POMDPs). Plus précisément, les techniques de *Q-learning* et de *Q-learning* récurrent sont utilisées. Certains des POMDPs considérés nécessitent une capacité de mémorisation importante afin d'atteindre une politique décisionnelle optimale. Dans les POMDPs, les techniques du RL sont habituellement basées sur l'approximation de fonctions qui prennent en entrée des séquences d'observations de tailles variables. Les réseaux neuronaux récurrents (RNNs) constituent donc un choix naturel pour de tels approximateurs. Ce travail repose notamment sur les cellules récurrentes bistables qui ont été inventées récemment et qui offrent une bien meilleure mémoire à long terme que les réseaux classiques comme le *long short-term memory* (LSTM) et le *gated recurrent unit* (GRU). Ces cellules bistables sont appelées *bistable recurrent cell* (BRC) et *recurrently neuromodulated BRC* (nBRC). Tout d'abord, en important ces cellules dans le RL pour la première fois, il est démontré qu'elles offrent un avantage significatif dans des environnements nécessitant une mémoire importante, en comparaison du LSTM ou du GRU. Ensuite, la capacité des RNNs à représenter une distribution sur les états du POMDP est étudiée. Cette étude est réalisée en évaluant l'information mutuelle entre les états récurrents du RNN et la distribution sur les états filtrée depuis les observations successives. Cette analyse est donc ancrée à la fois dans la théorie de l'information et la théorie du contrôle optimal pour les POMDPs. Finalement, en complément de ce projet de recherche, une nouvelle mise à jour de la *target* est proposée pour les algorithmes de *Q-learning* avec *target network*, aussi bien pour les politiques stationnaires que récurrentes. Cette nouvelle mise à jour de la *target* accélère l'apprentissage, surtout dans les environnements avec récompenses rares.