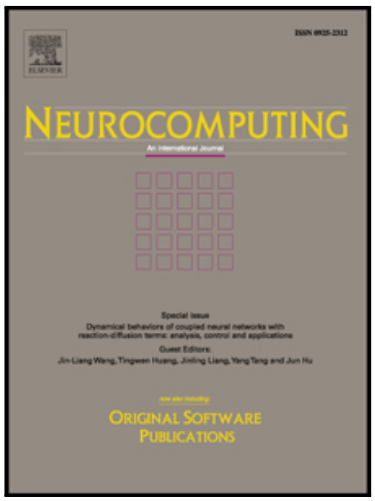

Data-Driven Adaptive Dynamic Programming for Continuous-Time Fully Cooperative Games With Partially Constrained Inputs

Qichao Zhang, Dongbin Zhao, Yuanheng Zhu

PII: S0925-2312(17)30243-6
DOI: 10.1016/j.neucom.2017.01.076
Reference: NEUCOM 18039

To appear in: *Neurocomputing*

Received date: 15 July 2016
Revised date: 20 January 2017
Accepted date: 23 January 2017

Please cite this article as: Qichao Zhang, Dongbin Zhao, Yuanheng Zhu, Data-Driven Adaptive Dynamic Programming for Continuous-Time Fully Cooperative Games With Partially Constrained Inputs, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2017.01.076

Data-Driven Adaptive Dynamic Programming for Continuous-Time Fully Cooperative Games With Partially Constrained Inputs

Qichao Zhang[a,b], Dongbin Zhao[a,b,*], Yuanheng Zhu[a,b]

[a] *The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China*
[b] *The University of Chinese Academy of Sciences, Beijing, 100049, China,*

Abstract

In this paper, the fully cooperative game with partially constrained inputs in the continuous-time Markov decision process environment is investigated using a novel data-driven adaptive dynamic programming method. First, the model-based policy iteration algorithm with one iteration loop is proposed, where the knowledge of system dynamics is required. Then, it is proved that the iteration sequences of value functions and control policies can converge to the optimal ones. In order to relax the exact knowledge of the system dynamics, a model-free iterative equation is derived based on the model-based algorithm and the integral reinforcement learning. Furthermore, a data-driven adaptive dynamic programming is developed to solve the model-free equation using generated system data. From the theoretical analysis, we prove that this model-free iterative equation is equivalent to the model-based iterative equations, which means that the data-driven algorithm can approach the optimal value function and control policies. For the implementation purpose, three neural networks are constructed to approximate the solution of the model-free iteration equation using the off-policy learning scheme after the available system data is collected in the online measurement phase. Finally, two examples are provided to demonstrate the effectiveness of the proposed scheme.

Keywords:
Adaptive dynamic programming, Optimal control, Neural network, Fully cooperative games, Data-driven, Constrained input.

1. Introduction

Recently, a newly developed technique, multi-agent reinforcement learning (MARL) which integrates the developments of reinforcement learning (RL) and game theory, has been widely applied to various fields including robotic control, traffic light control, battery management, distributed sensor network, etc [1, 2, 3]. In MARL, an agent is usually a computational entity which can perceive its environment, make decisions, and act upon its environment through actuators. Generally, the agent is not isolated but connected to its neighbour agents for multi-agent systems (MAS), and the mutual links between each agent can be expressed through a communication diagram. Their behaviors are adopted to optimize some performance indexes based on their own information and the shared one from their neighbors to affect the environment together. However, due to the complexity and variability of the environment, it is difficult to design the agents' behaviors relying on the prior knowledge. That is, it is necessary to learn appropriate behaviors for each agent.

For a single-agent environment, RL provides a method to learn to behave in an unknown or known environment. Through interactions with the environment, the agent adapts its behaviors continually based on a received reward signal, to finally achieve an optimal or near-optimal policy that maximizes the long-term accumulated reward. The accumulated reward is known as the value function [4]. In most cases, RL considers the Markov decision process (MDP). And some well-understood RL algorithms with good convergence such as Q-learning, Sarsa and adaptive dynamic programming (ADP) are proposed and widely used to tackle the single-agent RL task without full information of system

*Corresponding author
Email addresses: zhangqichao2013@163.com (Qichao Zhang), dongbin.zhao@ia.ac.cn (Dongbin Zhao), zyh7716155@163.com (Yuanheng Zhu)