Analysing Congestion Problems in Multi-agent Reinforcement Learning*

(Extended Abstract)

Roxana Rădulescu
Vrije Universiteit Brussel
Belgium
rradules@vub.ac.be

Peter Vrancx
Vrije Universiteit Brussel
Belgium
pvrancx@vub.ac.be

Ann Nowé
Vrije Universiteit Brussel
Belgium
anowe@vub.ac.be

ABSTRACT

We extend the study of congestion problems to a more realistic scenario, the Road Network Domain (RND), where the resources are no longer independent, but rather part of a network, thus choosing one path will also impact the load of another one having common road segments. We demonstrate the application of state-of-the-art multi-agent reinforcement learning methods for this new congestion model and analyse their performance. RND allows us to highlight an important limitation of resource abstraction and show that the difference rewards approach manages to better capture and inform the agents about the dynamics of the environment.

Keywords

Multi-agent reinforcement learning; Congestion problems; Resource abstraction

1. INTRODUCTION

Current benchmark congestion problems present in the literature often make unrealistic assumptions regarding the independence between the available resources. In complex network management domains, such as smart grids and traffic networks, resources are connected and interdependent, such that using one resource impacts the load of others as well. For this purpose we introduce the Road Network Domain (RND), a problem that models the resources as a system of interconnected roads. We proceed to demonstrate the application of state-of-the-art multi-agent reinforcement learning (MARL) methods on this problem and analyse their capacity of capturing the newly introduced dynamics in the environment.

Reinforcement Learning (RL) [3] is a machine learning approach which allows an agent to learn how to solve a task by interacting with the environment. The solution consists in finding a policy, i.e., a mapping between states and actions that maximizes the received reward signal. When transitioning to the multi-agent case, we consider the scenario of self-interested independent Q-learners [5] interacting in the same environment. In MARL a central concern is to provide a reward signal that will offer a beneficial collective behaviour at the system level. Two straightforward approaches are: the *local reward* (L) which reflects information about the parts of the system the agent is involved in, or the *global reward* (G) which reflects the global system utility and should stimulate agents to perform actions beneficial for the system.

A congestion problem from a multi-agent learning perspective is defined by a set of n available resources $\Psi = \{\psi_1, ..., \psi_n\}$. Each resource ψ is defined by three properties: $\psi = \langle w_\psi, c_\psi, x_{\psi,t} \rangle$, where $w_\psi \geq 0$ represents the weighting of the resource, $c_\psi > 0$ is the capacity of ψ and finally $x_{\psi,t} \geq 0$ is the consumption of ψ at time t. A resource ψ is congested when $x_{\psi,t} > c_\psi$. One benchmark congestion problem present in the literature is the *beach problem domain* (BPD) [4], where all the available resources are considered beach sections with the same weight equal to 1 and the same capacity c: $L(\psi, t) = x_{\psi,t} e^{\frac{-x_{\psi,t}}{c}}$. The global utility is defined as the sum over all the local utility functions at time t: $G(t) = \sum_{\psi \in \Psi} L(\psi, t)$. If the number of agents exceeds the total capacity of the system, the configuration achieving the *highest global utility* for this benchmark problem is that one that overcrowds one of the resources and leaves the rest at optimum capacity.

Difference rewards (D) [6] is a MARL reward signal that informs the agents about their individual contribution to the system. Under a global system utility G, the difference rewards for agent i is defined as: $D_i(z) = G(z) - G(z_{-i})$, where z denotes a general term for state, or state-action pair, and $G(z_{-i})$ is the global utility of a virtual system lacking the effect of agent i.

The fourth MARL approach considered here is *resource abstraction* (RA) [2], i.e., grouping the set of resources into disjoint subsets, and modifying the local reward function after reaching the congestion point of a resource, such that agents using it will get a higher penalty for overcrowding. An abstract group is defined by aggregating the properties of the composing resources: consumption $X_{b,t} = \sum_{\psi \in b} x_{\psi,t}$, capacity $C_b = \sum_{\psi \in b} c_\psi$ and weight $W_b = \frac{1}{|b|} \sum_{\psi \in b} w_\psi$.

2. ROAD NETWORK DOMAIN

We propose the *Road Network Domain* (RND), a problem in which the resources are not independent, as using one path introduces additional load for others as well. Each road segment is modelled as a resource, corresponding to the description presented in Section 1. The RND can be used with the utility function of BPD. The local reward of a path

*A full version of this paper is available at https://arxiv.org/abs/1702.08736

Appears in: *Proc. of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017), S. Das, E. Durfee, K. Larson, M. Winikoff (eds.), May 8–12, 2017, São Paulo, Brazil.*

P is then simply the sum over all the local rewards of the composing road segments ψ (e.g., Figure 1, roads AB and BD for the path ABD): $L_{path}(P,t) = \sum_{\psi \in P} L(\psi,t)$. We compute the global system utility by summing over all the local rewards of the roads segments present in the network.

As the impact of agent i on the system is limited to the composing road segments of his chosen path P, we can define the *difference rewards* for the RND as follows, where f is a local reward function:

$$D_i(t) = L_{path}(P,t) - L_{path}(P_{-i},t) \quad (1)$$

For the *resource abstraction* method, we consider here two approaches for defining the abstract groups: over *road segments* or over *paths of the network*. As a road segment is a resource, the properties of an abstract group over a set of segments coincide with the ones defined in section 1. The abstract reward for each road segment ψ and its corresponding group b is defined as:

$$A(b,\psi,t) = \begin{cases} L(\psi,t), & x_{\psi,t} \leq c_\psi \\ -X_{b,t} e^{\frac{-x_{b,t}}{C_b}}, & x_{\psi,t} > c_\psi \end{cases} \quad (2)$$

The abstract reward for choosing a path P at time t then becomes the sum over the abstract reward of each composing road segment.

The extension for an abstract group over a set of paths is straightforward, if we define a path P as a resource with the properties: consumption $x_{P,t}$ as the number of agents that choose path P, capacity $c_P = min_{\psi \in P}(c_\psi)$ and weight $w_P = \frac{1}{|P|} \sum_{\psi \in P} w_\psi$. We can now define the abstract reward for a selected path P at time t:

$$A(b,P,t) = \begin{cases} L_{path}(P,t), & \forall \psi \in P \ : \ x_{\psi,t} \leq c_\psi \\ -X_{b,t} e^{\frac{-x_{b,t}}{C_b}}, & \exists \psi \in P \ : \ x_{\psi,t} > c_\psi \end{cases} \quad (3)$$

where b is the corresponding abstract group of P.

We perform two experiments on the RND instance depicted in Figure 1: with RA defined over paths and RA over road segments. Each agent uses the Q-learning algorithm with an exploration parameter $\epsilon = 0.05$ and an exploration decay rate of 0.9999. We match the resource abstraction RA parameters to the one used in [1]: learning rate $\alpha = 0.1$, decay rate for α is 0.9999 and discount factor $\gamma = 1.0$. The parameters used for L, G and D are: $\alpha = 0.1$, with no decay, and $\gamma = 0.9$.

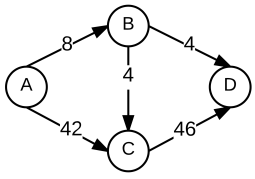

Figure 1: Example of an optimum distribution of 50 agents over the network under the BPD local utility ($c = 5$,$w = 1$).

3. RESULTS AND DISCUSSION

Figures 2 and 3 show that none of the RA settings manage to converge to an optimum configuration for the selected RND scenario. To better understand these results, we can turn to Figure 1. Notice that even though the capacity of the road segments is 5, the optimum configuration does not include any segments having reached this value. We conclude that we cannot express the solution as 'overcrowd these segments and keep the rest at optimum capacity', thus being

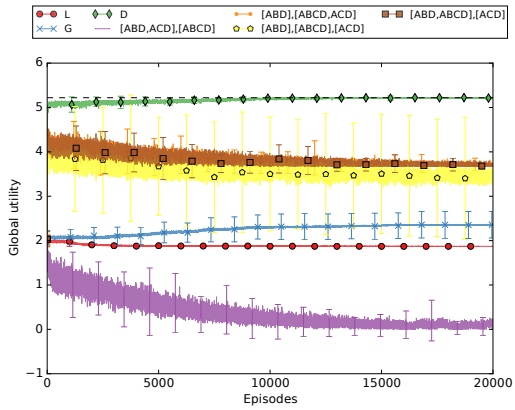

Figure 2: RND with BPD local utility, 50 agents, RA over paths.

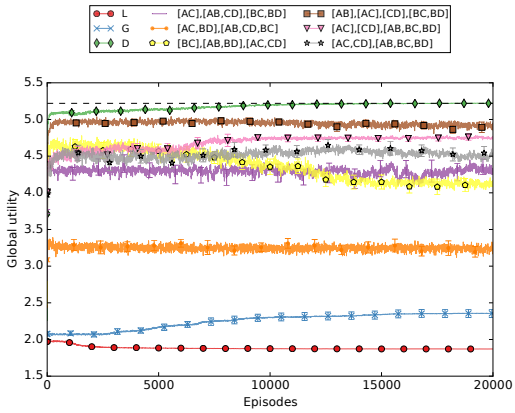

Figure 3: RND with BPD local reward, 50 agents, RA over road segments.

unable to properly express the desired solution using the RA approach. Additionally, it seems that having disjoint abstract groups is not a sufficient condition for being able to reach an optimum solution using RA and that the necessity of having independent resources goes beyond having segments not belonging to the same abstract group. On the other hand, D manages to achieve the optimal performance in this scenario, demonstrating its capacity to allow agents to adapt to more difficult environment dynamics.

The Road Network Domain presents a novel challenge for resource selection congestion problems, introducing the realistic aspect of interconnected resources as we often find in real-word application such as: electricity grids or traffic networks. We note that the network topology used here is a small one, yet sufficient to illustrate the additional challenge, and that more research is necessary in order to evaluate scenarios that closely model real-world situations.

Acknowledgments

This work is supported by Flanders Innovation & Entrepreneurship (VLAIO), SBO project 140047: Stable MultI-agent LEarnIng for neTworks (SMILE-IT).

REFERENCES

[1] K. Malialis, S. Devlin, and D. Kudenko. Intrusion response using difference rewards for scalability and online learning. In *Workshop on Adaptive and Learning Agents at AAMAS (ALA-14)*, 2014.

[2] K. Malialis, S. Devlin, and D. Kudenko. Resource abstraction for reinforcement learning in multiagent congestion problems. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 503–511. International Foundation for Autonomous Agents and Multiagent Systems, 2016.

[3] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press Cambridge, 1998.

[4] K. Tumer and S. Proper. Coordinating actions in congestion games: impact of top–down and bottom–up utilities. *Autonomous agents and multi-agent systems*, 27(3):419–443, 2013.

[5] C. J. C. H. Watkins. *Learning from delayed rewards*. PhD thesis, University of Cambridge England, 1989.

[6] D. H. Wolpert and K. Tumer. Optimal payoff functions for members of collectives. *Advances in Complex Systems*, 4(2/3):265–279, 2001.