Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A.A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

SPARTAN: A META-ALGORITHM FOR REINFORCEMENT LEARNING USING STATE PARTITIONING AND ACTION NETWORK

Kyohong Shin
Taesik Lee

Department of Industrial and Systems Engineering, KAIST
291 Daehak-ro, Yuseong-gu
Daejeon, 34141, REPUBLIC OF KOREA

ABSTRACT

Targeting finite-horizon Markov Decision Process problems, we propose a novel approach with an aim to significantly enhance the scalability of reinforcement learning (RL) algorithms. Our approach, which we call a State Partitioning and Action Network, SPartAN in short, is a meta-algorithm that offers a framework an RL algorithm can be incorporated into. Key ideas in SPartAN are threefold: reducing the size of an original RL problem by partitioning the state space into smaller compartments, using a simulation model to directly obtain values of the terminal states of the upstream compartment, and constructing a quality heuristic policy in the downstream compartment by an action network to use in the simulation. Using temporal difference learning as an example RL algorithm, we show that SPartAN is able to reliably derive a high quality policy solution. Through empirical analysis, we also find that a smaller downstream state subspace in SPartAN yields higher performance.

1 INTRODUCTION

Reinforcement learning (RL) addresses the curse of dimensionality by using value function approximation techniques (Powell 2007). Despite the continuous development of various value function approximation techniques (Boyan and Moore 1995; Mnih et al. 2015), the scalability problem is far from being conquered, and the need for new ideas and techniques is ever increasing. Targeting finite-horizon MDP problems, we propose a novel approach with an aim to significantly enhance the scalability of RL algorithms. Our approach, which we call a State Partitioning and Action Network, SPartAN in short, is a meta-algorithm in that it works as a framework the existing RL algorithms can be incorporated into.

2 STATE PARTITIONING AND ACTION NETWORK (SPARTAN)

SPartAN addresses the scalability problem by taking a different approach to better transfer values of downstream state space to upstream state space. Instead of relying on repeated visits by generating large number of sample paths from top to bottom, we use a simulation to directly obtain approximate value of the interim states, given a well-performing policy in the downstream state subspace.

Key ideas in SPartAN are threefold: reducing the size of an original RL problem by partitioning the state space into smaller compartments, directly obtaining values of the terminal states of the upstream compartment by using a simulation model, and constructing a quality heuristic policy in the downstream compartment by Deep Neural Network (action network) to use in the simulation. We want to emphasize that SPartAN works with any RL algorithm. Any RL algorithm can be used within the framework of SPartAN, and this is why we refer SPartAN to a meta-algorithm. SPartAN addresses the problem associated with a large state space MDP model, which is a common challenge to all RL algorithm.

3 NUMERICAL EXPERIMENT

We use TD learning algorithm as an example of an RL algorithm and the toy model that is developed based on patient transport prioritization and hospital selection problem in a mass-casualty incident. For the experiment, we set $L = \frac{T}{2}$ (T is a length of horizon) as a partitioning step location for the entire state space. We obtain five policy solutions from SPartAN and TD learning, and under each policy we run 100 simulations respectively. As a performance measure, the average number of survivors from the simulations is reported. Figure 1(a) shows that the performance of the policy solutions by SPartAN is higher than TD learning. The results show that SPartAN produces policies with more consistent performance, whereas the variance in the policy performance from TD learning is relatively large, rendering TD learning much less reliable. To further understand how SPartAN works, we test difference of partitioning location by varying $L = \frac{T}{8}, \frac{T}{4}, \frac{T}{2}, \frac{3T}{4}, \frac{7T}{8}$, and T. It seems that large L works to the advantage of SPartAN; that is, when L is large (i.e. the set of states in the downstream is small), the performance of the policy solution is better and the variation in the computed policies is small. As can be seen in Figure 1(b), the best performance setting is $L = \frac{3T}{4}$.

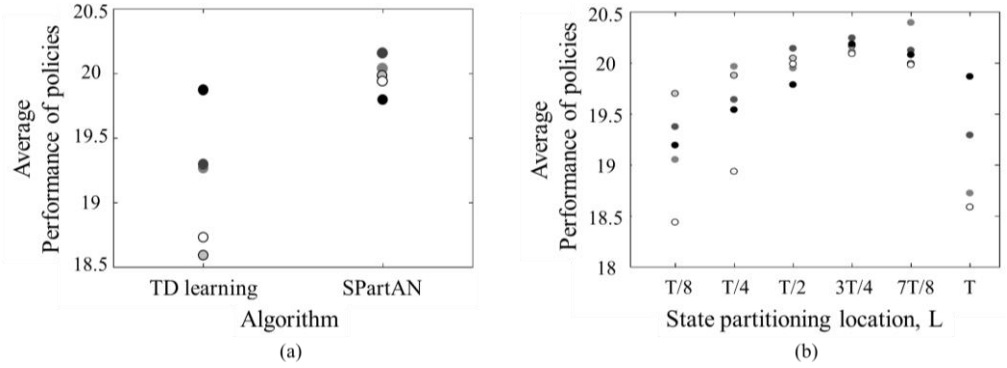

Figure 1. Performance comparison: (a) TD learning and SPartAN, (b) State partitioning location, L.

4 CONCLUSION

We propose a novel approach that can significantly enhance the scalability for solving large-scale finite-horizon MDP models. Using a toy model and TD learning as an example RL algorithm, we showed that SPartAN is able to reliably derive a high quality policy solution. Through empirical analysis, we also found that a small downstream state subspace yields higher performance for SPartAN.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2016R1A2B4014323).

REFERENCES

Boyan, J. A., and A. W. Moore. 1995. "Generalization in Reinforcement Learning: Safely Approximating the Value Function". In *Advances in Neural Information Processing Systems*, 369–376.

Mnih, V., K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, et al. 2015. "Human-Level Control Through Deep Reinforcement Learning". *Nature* 518(7540):529–533.

Powell, W. B. 2007. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Hoboken, New Jersey: John Wiley & Sons.