Queueing Network Controls via Deep Reinforcement Learning

J. G. Dai[1,2] and Mark Gluzman[3]

Introduction

For more than 30 years, one of the most difficult problems in applied probability and operations research is to find a scalable algorithm for approximately solving the optimal control of stochastic processing networks, particularly when they are heavily loaded. In our work we focus on multiclass queueing networks with Poisson arrival and exponential service time distribution. The control problems of such networks can be modeled within the framework of Markov decision processes (MDPs) via uniformization. Unfortunately, the corresponding MDPs suffer from the curse of dimensionality: there are a large number of job classes, and the buffer capacity for each class is assumed to be infinite.

Modern RL algorithms can successfully overcome the curse of dimensionality and achieve state-of-the-art results in both the model-free and model-based MDP problems. In our paper [1], we demonstrate that a class of deep reinforcement learning algorithms known as proximal policy optimization (PPO), generalized from [5] to our setting, can generate control policies that consistently beat the performance of control policies known in the queueing literature.

Reinforcement learning approach for queueing network control

Originally, PPO algorithm has been developed for MDPs with finite state space and infinite horizon discounted total cost objectives. A conventional setup for queueing network control problems is an MDP with infinite state space, unbounded costs, and a long-run average cost objective. We extend the theoretical framework of PPO algorithm for such MDP problems using Lyapunov drift functions framework [4]. We show that starting from a stable policy it is possible to improve long-run average performance with sufficiently small changes to the initial policy.

The success of PPO algorithm implementation largely depends on the accuracy of Monte Carlo estimates of the relative value function in each policy iteration. We discuss a new way to estimate a relative value function if transition probabilities are known. We adopt the approximating martingale-process (AMP) method [2] which, to the best of our knowledge, has not been used in simulation-based approximate policy improvement setting. We also use a biased estimator of the relative value function through discounting the future costs [3]. In our numerical experiments, the discounting combined with the AMP method has been demonstated to reduce the variance of the relative value function estimation at the cost of a tolerable bias. The use of these two techniques combined has been a decisive factor in the success of our PPO algorithm implementation.

Experimental results for multiclass queueing networks

In our PPO implementation we use two separate feedforward neural networks, one for parametrization of the control policy and the other for value function approximation, a setup common to actor-critic algorithms. We propose to choose architectures of neural networks automatically as the size of a queueing network varies. We demonstrate the effectiveness of these choices as well as other hyperparameter choices.

We have conducted extensive computational experiments for multiclass queueing networks. The PPO algorithm applied for the criss-cross network control optimization obtained policies with long-run average performance within 1% from the optimal. For extended six-class queueing networks (from 6 to 21 classes), PPO policies outperform the best heuristics on average by 10%.

References

[1] J. G. Dai and Mark Gluzman. Queueing network controls via deep reinforcement learning, 2020. arXiv:2008.01644.
[2] Shane G. Henderson and Peter W. Glynn. Approximating martingales for variance reduction in Markov process simulation. *Mathematics of Operations Research*, 27(2):253–271, 2002.
[3] Peter Marbach and John N. Tsitsiklis. Simulation-based optimization of Markov reward processes. *IEEE Transactions on Automatic Control*, 46(2):191–209, 2001.
[4] Sean Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, Cambridge, 2nd edition, 2009.
[5] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. 2017. arXiv:1707.06347.

[1]School of Data Science, Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen
[2]School of Operations Research and Information Engineering, Cornell University, Ithaca, NY
[3]Center for Applied Mathematics, Cornell University, Ithaca, NY