Reinforcement Learning in Queues

Urtzi Ayesta

Queueing Systems Vol. 100
Special Issue: 100 views on queues

Reinforcement Learning in Queues

U. Ayesta

24 February 2022

Introduction

Control and optimization in queues have been an active area of research for decades, see for instance [10, 11]. Most of the literature up to the present has focused on the *model-based* setting, a term used to describe the situation in which a model is known. In the coming years, we will witness a huge interest from the community in the *model-free* approach, a setting that does not assume knowledge of an exact underlying mathematical model. In this short note I provide a personal view of some of the challenges that lie ahead in the transition from model-based to model-free solutions in a queueing context.

Model-based control in queues

A large body of literature is available within the framework of Markovian decision processes (MDP), also known as stochastic dynamic programming, which provides a rich and powerful modeling tool from an analytical and computational viewpoint. Formally, an MDP is a stochastic control process where the objective is to minimize a long-run cost. At each time step, depending on the state and action taken by the MDP, the decision maker incurs a cost and the process reaches a new random state. The so-called value function captures the total cost of the optimal policy, and classical results show that it is the unique solution of a fixed point equation known as Bellman's Optimality Equation, see [8] for precise details.

When applied to queueing networks, typical objectives have been to minimize the queue length, delays, or a more sophisticated measure like energy. In a wide variety of problems, MDPs have led to the establishment of structural results such as optimality of switching curves and threshold policies, see Figure 1. The key idea to establish such results is to show that the value function enjoys

U. Ayesta
CNRS, IRIT, and IKERBASQUE - Basque Foundation for Science, urtzi.ayesta@irit.fr

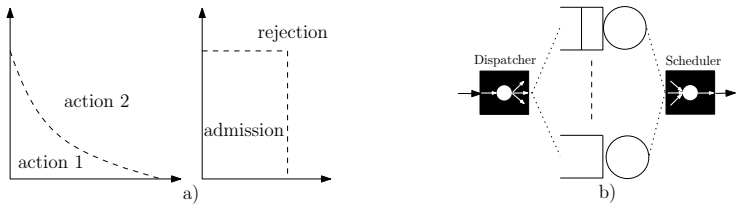

Fig. 1 a) A switching curve *(left)* and a threshold policy *(right)*; b) Load balancing and scheduling in queueing systems, where classes/queues can be cast as *arms* of an MABP/RMABP.

monotonicity properties, see for instance [9,6].

If we focus on resource allocation problems in queueing systems, see Figure (1) b), the class of MDPs known as multi-armed bandit problems (MABP) and restless multi-armed bandit problems (RMABP) have received a lot of attention since they provide a natural framework to study them, see [4]. MABP and RMABP have in common that their solution is given in terms of index policies, i.e., one can define for each bandit an index – that only depends on its own state – and the index policy activates in each time step the bandit with the highest index. Index policies are optimal for MABPs, known in this case as Gittins' index policy, and asymptotically optimal for RMABPs, known as Whittle's index policy, see [5]. Classical results in resource allocation, like optimality of Join the Shortest Queue or $c\mu$-rule, can be interpreted in terms of optimality of Gittins' index in an MABP, and asymptotic optimality of the $c\mu/\theta$-rule in terms of asymptotic optimality of Whittle's index in a RMABP.

Model-free control of queues.

The section above provides a brief illustration of the breadth and depth of the results available in the model-based setting. In the model-free case, a model is substituted by a so-called environment that for every present state and action, returns a sample of the one-step cost, and the next state. By interacting with the environment, reinforcement learning (RL) algorithms based on Q-learning, see [12], are capable of finding the value function that solves the MDP. In the last 5 years, this approach has been extremely successful at solving very complex problems with large state dimensions, obtaining in particular supra-human performance in games [2]. A key feature of this approach is the use of a neural network to approximate the value function. It is important to mention that this success relies on large amounts of computational resources, and a precise tuning and adaptation of the algorithm to the problem under consideration. Many of these successes have been obtained in an episodic setting (like games) in which there is a certainty that the episode will eventually finish, at which moment a reward will be observed. This situation differs dramatically from what we typically encounter in a queueing setting. Without aiming at being exhaustive, I hereby mention a few examples that illustrate the challenges of applying RL in

queues, and that are representative of topics that – I expect – will be studied in detail in the coming years.

As a first example let us consider the case of modulated queues, for which there is a broad literature in the model-based setting. Here, the parameters of the modulated queue change dynamically over time as the state of the modulating process evolves. In the model-based situation, this can be handled by simply adding the state of the modulating queue to the state descriptor of the value function. Interesting problems arise in the hybrid situation in which the state of the modulated queue is known, but not that of the modulating process. As outlined in [1], one could envision the deployment of algorithms that exploit a model-free type of approach to infer the state of the modulating process, and combine it with a model-based formulation regarding the modulated queue. As a second example, we consider queues with blocking in which threshold policies are optimal. In such a queue the probability of visiting blocking states can be extremely small, and the challenge is to improve the exploration so that the RL algorithm can learn the optimal thresholds (see Figure 1 a)). In [7], we explore how Fleming-Viot particle systems combined with RL can help find the optimal thresholds in a model-free setting. As a third example we mention an intermediate situation between the model-based and model-free dichotomy, in which the RL algorithm leverages knowledge on the underlying MDP, for instance existence of switching curves or optimality of index policies, in order to develop more efficient learning algorithms. In [3] the standard Q-learning algorithm is modified in order to propose algorithms that learn (faster) the Whittle indices of an RMABP.

References

1. S. Duran and I.M. Verloop. Asymptotic optimal control of Markov-modulated restless bandits. In *ACM SIGMETRICS 2018*.
2. D. Silver et al. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 2017.
3. F. Robledo et al. QWI: Q-learning with Whittle index. *RLNQ 2021*.
4. M. Larranaga et al. Dynamic control of birth-and-death restless bandits: application to resource-allocation problems. *IEEE/ACM ToN*, 24(6):3812–3825, 2016.
5. J.C. Gittins, K. Glazebrook, and R. Weber. *Multi-armed Bandit Allocation Indices*. Wiley, 2011.
6. G. Koole. *Monotonicity in Markov Reward and Decision Chains: Theory and Applications (Foundations and Trends in Stochastic Systems)*. Now Publishers Inc, 2007.
7. D. Mastropietro, S. Majewski, U. Ayesta, and M. Jonckheere. Boosting reinforcement learning using Fleming-Viot particle systems. *submitted*, 2022.
8. M. L. Puterman. *Markov Decision Processes*. John Wiley & Sons, 2005.
9. K.W. Ross. *Multiservice Loss Models for Broadband Telecommunication Networks*. Springer, 1995.
10. L.I. Sennott. *Stochastic Dynamic Programming*. Wiley, 1999.
11. S. Stidham. *Optimal Design of Queueing Systems*. Chapman and Hall/CRC, 2009.
12. C. Watkins and P. Dayan. Q-Learning. *Machine learning*, 8(3-4):279–292, 1992.