Learning in Queues

Bruno Gaujal

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

Learning in Queues

Bruno Gaujal

February 18, 2022

1 Introduction

From the early days, queueing systems have been associated with decision problems. Natural questions emerge such as: Should the arriving packet be accepted in the buffer or be rejected? Which queue should the next customer be sent to? Which customer should be served next among all pending ones? The early works of Naor [8] or Sobel [12] in the late 1960s as well as many others investigate these natural problems, and, as mentioned by Puterman in his reference book [11], queueing control represents one of the main areas of application of *Markov decision process* (MDP) methodology.

Let us consider a general Markov decision process $\mathscr{M} = (\mathscr{S}, \mathscr{A}, r, P)$ where the state (action) space $\mathscr{S}$ ($\mathscr{A}$) is of size S (A), $r : \mathscr{S} \times \mathscr{A} \to [0,1]$ is an unknown random reward function and P is the transition kernel: for all $s, s' \in \mathscr{S}, a \in \mathscr{A}, P(s'|s,a)$ is the unknown probability to go from s to s' under action a.

Reinforcement Learning (RL) in MDPs consists in designing a learning algorithm that takes online actions while exploring the states of the MDP. Its choice of action a_t in state s_t improves over time by exploiting the samples of the rewards $r_t(s_t, a_t)$ and of the transition kernel $P(s_{t+1}|s_t, a_t)$ that the learner collects. Here, we focus on one of the most popular measures of learning efficiency, namely the *regret* of the first T steps of the algorithm: $Reg(T) = \sum_{t=1}^{T}(r^*(s_t) - r_t(s_t, a_t))$ where r^* is the value of the unknown optimal policy in state s_t. The advent of Q-learning (see the reference book [13]) has triggered a fast growing interest in reinforcement learning for MDPs. Learning algorithms in *multi-armed bandits*, such as *Upper Confidence Bound* (UCB) [7] have also inspired another type of learning algorithms for MDPs (called model based). One can cite the variants of UCRL, from I to V and counting, started in [5] or UCB-VI [3]. The Bayesian counterpart of UCB for bandits is *Thompson sampling*. Several extensions to MDPs have also been proposed such as posterior sampling [9] or TSDE [10]. The

Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000 Grenoble, France E-mail: bruno.gaujal@inria.fr

best regret bound of all these algorithms is

$$Reg(T) \leq c\sqrt{DSAT}, \tag{1}$$

where c is a constant and D is the diameter of the MDP. Conversely, there exist MDPs for which any learning algorithm has a regret at least $c'\sqrt{DSAT}$ (where $c' < c$). These worst case bounds may give a sense of optimality that can be misleading, as seen later.

2 Reinforcement Learning in Queueing Systems

In spite of this explosive growth of research on online learning, few papers are dedicated to reinforcement learning in queues. One can argue that learning in queues does not reduce to RL/MDP/regret and other approaches have emerged as surveyed in [2]. In this note we focus on regret minimization in queueing control and argue that improvements can and must be made, both on the analysis and algorithmic aspects. So what is so specific to MDPs modeling control problems on queues?

- *No discount*. Discounts are rampant in the reinforcement learning literature, especially in Q-learning algorithms [13]. However, discounts are not natural in queues because the unit of the cost is often milliseconds and not dollars[1].

- *Very large state space and diameter.* Queueing systems suffer from the curse of dimensionality: The size of the state space S is exponential in the number of queues. In this case, the classical bound (1) on the regret can be unacceptable. Furthermore, the diameter D can be exponential in the state space size for many queue control problems. Here again, the bound (1) becomes irrelevant, even when the state space size is small.

- *Structured transition matrices.* In most queueing systems, the transition matrices are sparse and structured. The structure of the kernel has already been successfully exploited in Markovian bandits in [4] for example, and this line of work should extend to the exploitation of the structure of the transition kernel in queues.

Parametric learning. This last item opens the first promising track for queue-specific learning: exploit the parametric nature of the kernel. Indeed, most queueing systems are defined by a small number of parameters (*e.g.* the arrival and service rates in an admission control problem in an M/M/1/K queue). A d-linear additive model, inspired by parametric bandits, was introduced in [6] and assumes that $P(\cdot|s,a) = \langle \phi(s,a), \theta(\cdot)\rangle$, where $\phi(s,a)$ is a known feature mapping and θ is an unknown measure on $\mathbb{R}^d$, and has been popular ever since. However it implies that the transition kernel is of rank d and does not apply in most queueing systems where the kernel has almost full rank. The *linear mixture model* introduced in [14] assumes instead that $P(s'|s,a) = \langle \phi(s'|s,a), \theta\rangle$, $\theta \in \mathbb{R}^d$. This is more adapted to queues. For example, the admission control problem in a M/M/1/K queue is a linear mixture of dimension $d = 2$. [14] shows that, with discount γ, the regret bound of their optimist algorithm becomes $Reg(T) \leq d\sqrt{T}/(1-\gamma)^2$, replacing S and A by d, the number of parameters (the diameter remains under the form

[1] although this goes against the precept that time is money.

$(1-\gamma)^{-1}$). However, discounting is not natural in queueing control. The design of an efficient learning algorithm for undiscounted parametric MDPs is still open.

Instance specific regret bounds. Another promising track is to move away from the minimax approach used in the construction of regret upper and lower bounds given above. We argue that *instance specific* analysis will provide much tighter bounds and more insight of the actual performance of a learning algorithm in queueing control.

The main idea underlying this belief is that, under the optimal policy, some states will be rarely visited (geometric stationary distributions are typical, as in a $M/M/1/K$ queue under optimal admission control). Therefore, learning the optimal decisions in those rare states is quite useless to get a good performance. Currently, this is not taken into account (all states are treated uniformly in the regret analysis). For learning in admission control and speed scaling problems as studied in [1], we can show that exploiting the stationary measure in the analysis of classical learning algorithms yields $Reg(T) \leq K\sqrt{T}$ where K only depends on the load of the system under the optimal policy. The dependence in the size of the state space, the action space and, more importantly, in the diameter of the MDP disappears. We believe that this type of results can be generalized to many other cases (optimal routing and allocation problems) where the stationary distribution under all policies is extremely uneven between states.

A lot is still to be done to understand the behavior of existing learning algorithms in queues as well as to design new learning techniques adapted to queueing control.

References

1. J. Anselmi, B. Gaujal, and L.-S. Rebuffi. Optimal speed profile of a DVFS processor under soft deadlines. *Performance Evaluation, Elsevier*, 2021. to appear.
2. A. Asanjarani, Y. Nazarathy, and P. Taylor. A survey of parameter and state estimation in queues. *Queueing Systems*, 97:39–80, 2021.
3. M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.
4. N. Gast, B. Gaujal, and K. Khun. Reinforcement learning for Markovian bandits: Is posterior sampling more scalable than optimism? Technical Report hal-03262006, HAL-Inria, June 2021.
5. T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4):1563–1600, 2010.
6. C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. Provably efficient reinforcement learning with linear function approximation. Technical Report arXiv:1907.05388, arXiv preprint, 2019.
7. T. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.
8. P. Naor. On the regulation of queue size by levying tolls. *Econometrica*, 37:15–24, 1969.
9. I. Osband, D. Russo, and B. Van Roy. More efficient reinforcement learning via Posterior Sampling. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3003–3011. Curran Associates, Inc., 2013.
10. Y. Ouyang, M. Gagrani, A. Nayyar, and R. Jain. Learning unknown Markov decision processes: A Thompson sampling approach. *arXiv preprint arXiv:1709.04570*, 2017.
11. M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1 edition, Apr. 1994.
12. M. J. Sobel. Optimal average cost policy for a queue with start-up and shut-down costs. *Operations Research*, 17:145–162, 1969.
13. R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.
14. D. Zhou, J. He, and Q. Gu. Provably efficient reinforcement learning for discounted MDPs with feature mapping. *CoRR*, abs/2006.13165, 2020.