2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)
Allerton Park and Retreat Center
Monticello, IL, USA, October 2-5, 2018

On Learning the $c\mu$ Rule in Single and Parallel Server Networks

Subhashini Krishnasamy*, Ari Arapostathis†, Ramesh Johari‡ and Sanjay Shakkottai†
*Tata Institute of Fundamental Research
Email: subhashini.kb@utexas.edu
†The University of Texas at Austin
Email: ari@ece.utexas.edu, shakkott@mail.utexas.edu
‡Stanford University
Email: rjohari@stanford.edu

***Abstract*—We consider learning-based variants of the $c\mu$ rule for scheduling in single and parallel server settings of multi-class queueing systems.**

In the single server setting, the $c\mu$ rule is known to minimize the expected holding-cost (weighted queue-lengths summed over classes and a fixed time horizon). We focus on the problem where the service rates μ are unknown with the *holding-cost regret* (regret against the $c\mu$ rule with known μ) as our objective. We show that the greedy algorithm that uses empirically learned service rates results in a *constant* holding-cost regret (the regret is independent of the time horizon). This *free exploration* can be explained in the single server setting by the fact that any work-conserving policy obtains the same number of samples in a busy cycle.

In the parallel server setting, we show that the $c\mu$ rule may result in unstable queues, even for arrival rates within the capacity region. We then present sufficient conditions for geometric ergodicity under the $c\mu$ rule. Using these results, we propose an almost greedy algorithm that explores only when the number of samples falls below a threshold. We show that this algorithm delivers constant holding-cost regret because a free exploration condition is eventually satisfied.

***Index Terms*—queueing systems, learning, $c\mu$ rule, stability**

I. Introduction.

We consider a canonical scheduling problem in a discrete-time, multi-class, multi-server parallel server queueing system. In particular, we consider a system with U distinct queues, and K distinct servers. Each queue corresponds to a different class of arrivals; arrivals queue i are Bernoulli(λ_i), i.i.d across time. Service rates μ_{ij} are heterogeneous across every pair of queue i and server j (i.e., a "link"). At each time step, a central scheduler may match at most one queue to each server. Services are also Bernoulli; thus jobs may fail to be served when matched, and in this case the policy is allowed to choose a different server for the same job in subsequent time step(s). Jobs in queue i incur a holding cost c_i per time step spent waiting for service. Letting $Q_i(t)$ denote the queue length of queue i at time t, the performance measure of interest up to time T is the cumulative expected holding cost incurred up to time T:

$$\sum_{t=1}^{T} \sum_{i \in [U]} c_i \mathbb{E}\left[Q_i(t)\right].$$

(All our analysis extends to the case where the objective of interest is a time-discounted cost, i.e., where the t'th term is scaled by β^t, where the discount factor satisfies $0 < \beta < 1$.)

Our emphasis in this paper is on solving this problem when the link service rates are *a priori unknown*; the scheduler only learns the link service rates by matching queues to servers, and observing the outcomes. We use as our benchmark the *$c\mu$ rule* for scheduling, when link service rates are known. The $c\mu$ rule operates as follows: at each time step, each link from a nonempty queue i to server j is given a weight $c_i\mu_{ij}$; all other links are given weight zero. The server then chooses a maximum weight matching on the resulting graph as the schedule for that time step. It is well known that when there is only a single server, this rule delivers the optimal expected holding cost among all feasible scheduling policies. Further, there has been extensive analysis of the performance and optimality properties of this rule even in multiple server settings.

When service rates are unknown, we measure the performance of any policy using (expected) *regret* at T: this is the expected difference between the cumulative cost of the policy, and the cumulative cost of the $c\mu$ rule. Our goal is to characterize policies that minimize regret. In typical learning problems such as the stochastic multiarmed bandit (MAB) problem, optimal policies must resolve an *exploration-exploitation* tradeoff. In particular, in order to minimize regret, the policy must invest effort to learn about unknown actions, some of which may later prove to be suboptimal—and thus incur regret in the process. In such settings, any optimal policy incurs regret that increases without bound as $T \to \infty$; for example, for the standard MAB problem, it is well known that optimal regret scales as $O(\ln T)$ [1], [2], [3].

In this paper, we show a striking result: in a wide range of settings, the *empirical $c\mu$* rule—i.e., the $c\mu$ rule applied using the current estimates of the mean service rates—is regret optimal, and further, the resulting optimal regret is bounded by a *constant* independent of T. Thus, in such settings there is no tradeoff between exploration and exploitation. The scheduler can simply execute the optimal schedule given

978-1-5386-6596-1/18/$31.00 ©2018 IEEE

its current best estimate of the services rates of the links. In other words, the empirical $c\mu$ rule benefits from *free exploration*.

We make three main contributions: (1) regret analysis of the empirical $c\mu$ rule in the single server setting; (2) stability analysis of the $c\mu$ rule in the multi-server setting; and (3) subsequent regret analysis of the empirical $c\mu$ rule in the multi-server setting.

For a full version of this paper with the formulation, related work and all the technical proofs, please see [4].

ACKNOWLEDGMENTS.

This work was partially supported by NSF grants CNS-1343383 and DMS-1715210, Army Research Office grant W911NF-17-1-0359 and W911NF-17-1-0019, Office of Naval Research grant N00014-16-1-2956, and the US DoT supported D-STOP Tier 1 University Transportation Center.

REFERENCES

[1] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. in Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[3] S. Agrawal and N. Goyal, "Analysis of Thompson sampling for the multi-armed bandit problem," in *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, 2012, pp. 39.1–39.26.

[4] S. Krishnasamy, A. Arapostathis, R. Johari, and S. Shakkottai, "On learning the cμ rule: Single and multiserver settings," *CoRR*, vol. abs/1802.06723, 2018. [Online]. Available: http://arxiv.org/abs/1802.06723