The slow-server problem with multiple slow servers

Koole, Ger

published in
Queueing Systems
2022

DOI (link to publisher)
10.1007/s11134-022-09761-y

document version
Publisher's PDF, also known as Version of record

citation for published version (APA)
Koole, G. (2022). The slow-server problem with multiple slow servers. *Queueing Systems*, *100*(3-4), 469-471. https://doi.org/10.1007/s11134-022-09761-y

Queueing Systems (2022) 100:469–471
https://doi.org/10.1007/s11134-022-09761-y

The slow-server problem with multiple slow servers

Ger Koole[1]

Received: 3 January 2022 / Accepted: 28 February 2022 / Published online: 5 May 2022

1 Introduction

Stochastic dynamic programming is a method to find numerically optimal policies in Markov decision chains and processes. It can also be used to prove the structure of optimal policies. Sometimes the optimal policy can be completely characterized, which is, for example, the case when routing to parallel queues with equal service rates. In other cases, the characterization is only partial, as for admission control to a queue: under certain conditions, the optimal policy is of threshold type, but we do not know the value of the threshold.

To prove these results, certain properties of the dynamic programming *value function*, such as convexity, are shown to hold inductively. The choice of properties to propagate is a crucial step in the method: the set should contain the right properties to obtain the desired results, and the set should be closed under the dynamic programming operator. For example, in the case of admission control to a single queue, the value function typically is non-decreasing and convex in the queue length. Convexity is required to show that the policy is of threshold type, but convexity alone cannot be propagated, the value function needs to be non-decreasing as well. These inequalities can be 1-dimensional, such as convexity, and multi-dimensional, such as *submodularity*. For quite a number of models, the optimal policies can be characterized this way, see [5] for a systematic overview.

Of course it also occurs that different models have the same inequalities. This means that these models can be combined into new models. In fact, the crucial question is not whether a set of inequalities is propagated by a model, but by all "building blocks" of the model, such as an arrival event, the decision to move a customer, a departure, etc.

One particular set of 2-dimensional inequalities, consisting of *multimodularity* and increasingness in both dimensions, allows us to solve the model introduced in Lin and Kumar [6]. Customers arrive to a queue with 2 heterogeneous servers. It is obvious that we always use the faster one, but when it is optimal to use the slower one depends on the number of customers in the queue. Using a complicated argument, it is shown in

✉ Ger Koole
ger.koole@vu.nl

[1] Department of Mathematics, Vrije Universiteit, Amsterdam, The Netherlands

[6] that the optimal policy is of threshold type: above a certain level of customers, we should use the slower server. This can also be shown by propagating the inequalities mentioned above [4]. Note that, although the optimal policies look similar, this problem is of a different type than the 1-dimensional admission control problem: because we have to know whether the second server is occupied, the state space is 2-dimensional.

An obvious question is what happens if there is more than 1 slow server. It is to be expected that some form of threshold policy would still be optimal, and numerical experiments confirm that. However, it proves to be impossible to do so using the method just described: some of the inequalities do not propagate for multi-server queues and for more than 2 dimensions. Despite many efforts, this problem remains open for nearly 40 years.

In the next section, we introduce the problem more formally and we finish with a discussion of implications.

2 Problem statement

Consider arrivals to a queue according to some general process modeled as a Markovian arrival stream [1]. There are multiple servers with exponentially distributed handling times, not all having the same rate. For simplicity, we assume that there is a fast server with rate μ_1 and 2 slow servers with rate μ_2, $\mu_2 < \mu_1$. We are interested in finding the policy that minimizes the long-run average or discounted costs. As direct costs we take the number of customers in the system (if we exclude the customers in service from the costs the problem becomes trivial). We will denote the states with (x, y), where x refers to the number of customers in the queue including the one at the fast server, and y the number of customers in service at the slow servers.

Conjecture If sending a customer to a slow server is optimal in (x, y), then it is also optimal in $(x + 1, y)$.

We could make other conjectures, for example, about monotonicity in y, but this is the most fundamental one. As stated in the introduction, it is still open.

3 Discussion

Mathematics has a strong publication bias: hardly ever it is published that someone failed to prove something. Perhaps we see this as a personal failure, a sign that we have not tried hard enough, instead of an important insight about the problem at hand which is valuable to share. In the case of the current problem however, there is evidence of other people having tried to prove this result. In fact, Rykov, in [9], claims to have proven it. However, it was shown in de Véricourt and Zhou [2] that this claim is incorrect and that the proof is incomplete. Indeed, a claim of this type is easily made, but constructing a proof consists of tediously verifying all inequalities for all operators for all states. A crucial case close to the boundary of the state space is easily forgotten, and that was indeed the case. It is interesting to note that moving from single-server to multi-server queues the number of useful inequalities is often reduced (see [5]).

Also [7] claim to have proven the optimality of a threshold policy, but their proof is lengthy and not very structured. The authors of [2] had an elaborate email exchange with the authors of [7] about some parts of the proof, but this did not take away their doubts concerning its completeness.

Unfortunately, the monotonicity literature is known for its highly complicated proofs, some of which later proved to be invalid. An example is [11], on routing to queues with general service times. Its main result was shown to be incorrect by a counterexample in [10]. Another example deals with server assignments in tandems of parallel queues where a claim in [8] was shown to be incorrect by a counterexample in [3].

In conclusion, "it still remains to find an intuitive argument as to why a threshold policy is optimal" [2], probably using a method different from propagating value functions, or to construct a counterexample to the claim. This problem is too fundamental to be left open.

References

1. Asmussen, S., Koole, G.M.: Marked point processes as limits of Markovian arrival streams. J. Appl. Probab. **30**, 365–372 (1993)
2. de Véricourt, F., Zhou, Y.-P.: On the incomplete results for the heterogeneous server problem. Queueing Syst. **52**, 189–191 (2006)
3. Hordijk, A., Koole, G.M.: The μc-rule is not optimal in the second node of the tandem queue: a counterexample. Adv. Appl. Probab. **24**, 234–237 (1992)
4. Koole, G.M.: A simple proof of the optimality of a threshold policy in a two-server queueing system. Syst. Control Lett. **26**, 301–303 (1995)
5. Koole, G.M.: Structural results for the control of queueing systems using event-based dynamic programming. Queueing Syst. **30**, 323–339 (1998)
6. Lin, W., Kumar, P.R.: Optimal control of a queueing system with two heterogeneous servers. IEEE Trans. Autom. Control **29**, 696–703 (1984)
7. Luh, H.P., Viniotis, I.: Threshold control policies for heterogeneous server systems. Math. Methods Oper. Res. **55**, 121–142 (2002)
8. Nain, P., Tsoucas, P., Walrand, J.: Interchange arguments in stochastic scheduling. J. Appl. Probab. **27**, 815–826 (1989)
9. Rykov, V.V.: Monotone control of queueing systems with heterogeneous servers. Queueing Syst. **37**, 391–403 (2001)
10. Sparaggis, P.D.,Towsley, D.: Optimal routing in systems with ILR service time distributions. CMPSCI Technical Report 93–13, University of Massachusetts at Amherst (1993)
11. Weber, R.R.: On the optimal assignment of customers to parallel queues. J. Appl. Probab. **15**, 406–413 (1978)