Data-driven Queueing Systems (QS) Capacity Mining

Yawo Kobara [1] Opher Baron [2] Dmitry Krass [2]

[1]Odette School of Management, University of Windsor [2]Rotman School of Management, University of Toronto

Introduction

Data-driven capacity estimation is one of the most significant issues in practical queuing theory.

- Motivation and Background:
  - Queueing systems (QSs), especially healthcare data such as patient flow data is often incomplete in the record.
  - Healthcare system processes are either unobservable or difficult to record.
  - Determining the right number of servers in a system is impossible with queueing models due to the lack of all the model parameters.
- Problem Description: Suppose we have multi-server queueing system that outputs a dataset $(A_k, D_k), k = 1, \ldots, K$, where A_k and D_k are the arrival and departure times of the k^{th} customer.

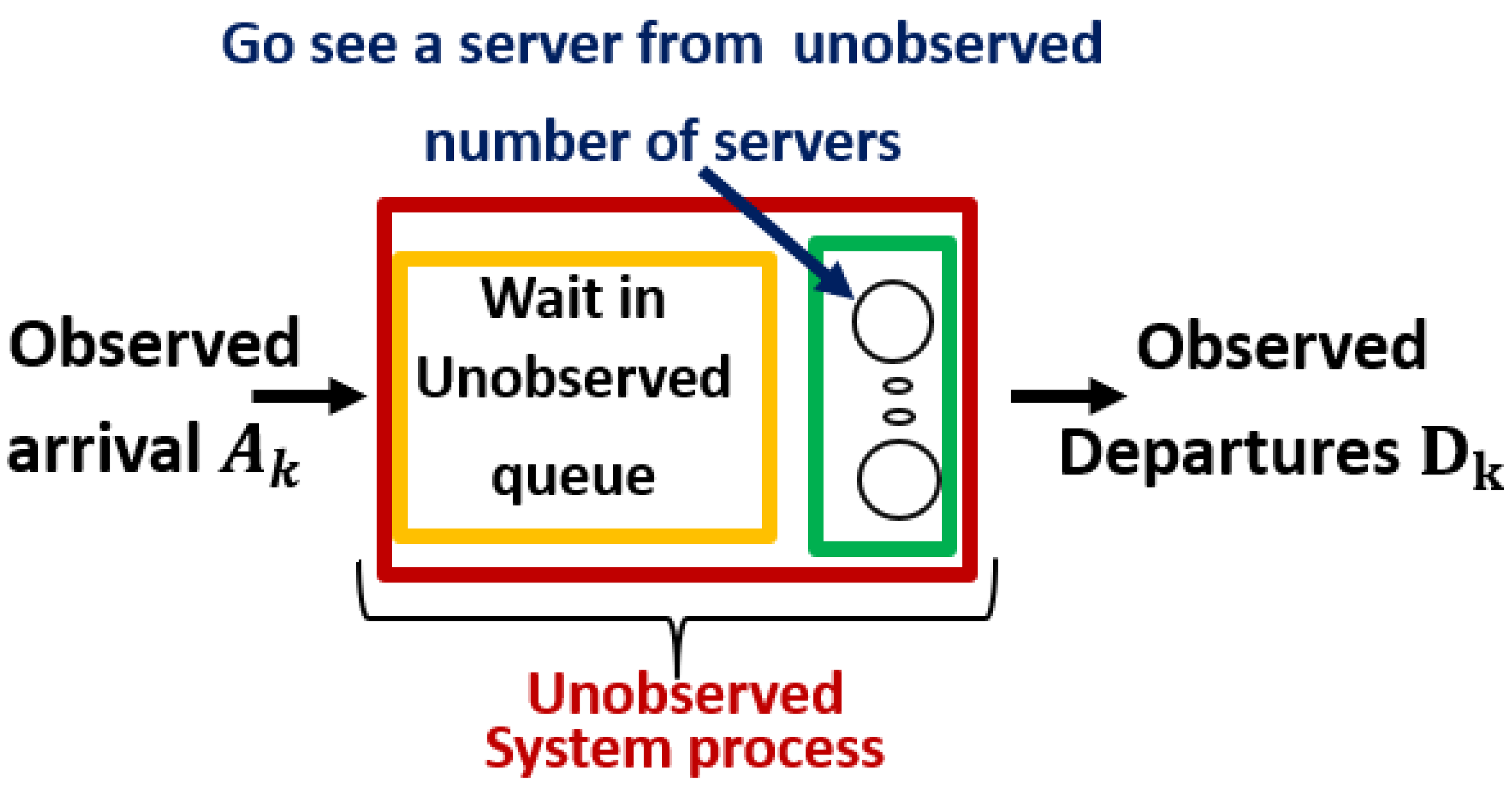


Figure 1. Problem setting visualization

- Question: Determine the effective capacity of a QS using only the arrival and departure timestamps of customers.what is the number of servers (service time distribution) of the queueing system?

Methodology

@ **Krivulin based estimator** The effective number of servers in the system is estimated as the smallest number that keeps all Krivulin estimated service times positive. Mathematically, it is given as ??

$$\hat{c}_1 = \min\{c : S_k = D_k - E_k(c) > 0, \text{ for } k = 1, \ldots, K, c = 1, 2, \ldots\} \quad (1)$$

Where $E_k(c) = \max(A_k, D_{(j-c)}$ is the Krivulin (1994) estimate of the entering service time of k^{th} customer, supposing there are c servers.

@ **Overtake based estimator** Recursively count people who arrived before a customer but departed after her and select the maximum.

$$\hat{c}_2 = \max\{C^+\} + 1 \quad (2)$$

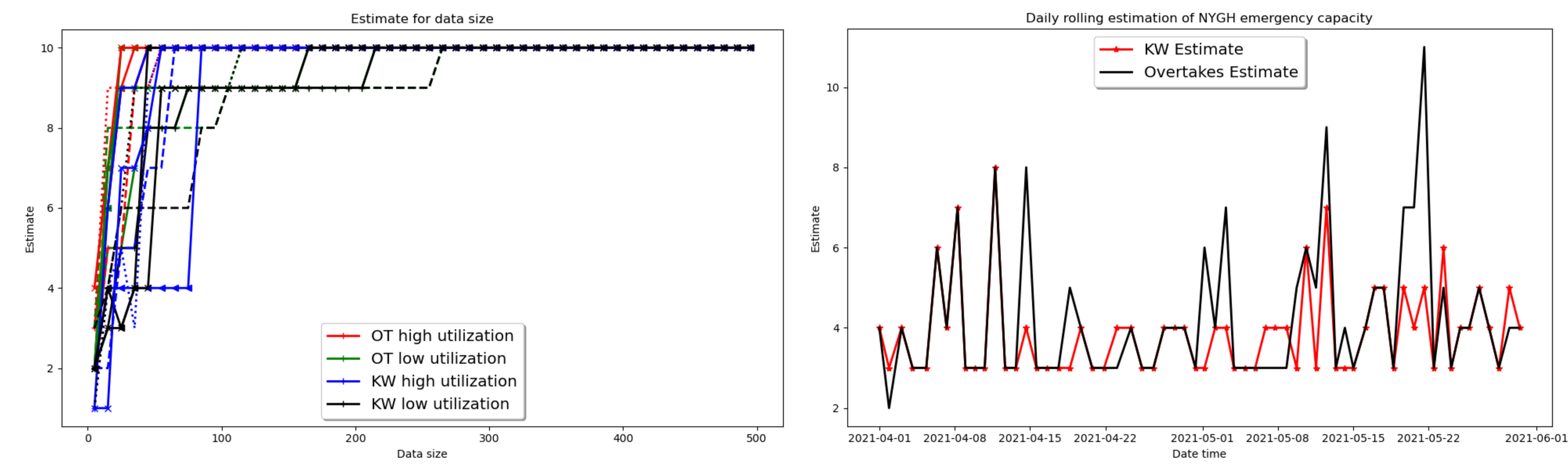
where C^+ is the number of customers an arriving customer overtakes. $C^+ \in \{0, 1, \ldots, c-1\}$.

@ Krivulin based Zero variance deterministic service algorithm
- Variance of the estimated service times of all customers are evaluated for all possible numbers of servers, c.
- Service time is deterministic, for the true number of servers.
- Estimate service time for all patients varying the number of servers.
- The number of servers, c that return a 0 (minimal) variance is the true number of servers.

@ Linear Model Based Algorithm
- First compute the sojourn time, S_k, of each customer k as $s_k = D_k - A_k$.
- Then compute the number of patients in the system when customer k arrived as n_k = count all customers j, with arrival time $A_j < A_k$ and their departure time $D_j > A_k$.
- Assume there are c^* servers in the system.
- Select only observations with $n_k \leq c^* + 1$.
- Run a simple linear regression between s_k and n_k, $(s_k = \alpha + \beta n_k)$.
- If β is insignificant, (that is pvalue >0.05, $\beta = 0$) then $c^* < c$,
- Now select observations with number in the system $n_k \leq c^* = c^* + 1$ and repeat 1 - 3. Continue to increase until you obtain a significant β. Then the true number of servers is the first $c^* - 1$ for the first c^* value that gives a significant β.

Results for the first two methods

- **Theorem 1** if we observe that people depart from the system in the same order they entered, then, we can't estimate the capacity.
- **Theorem 2** if we observe that at least one person overtook someone, then, we can estimate the capacity.



(a) Capacity estimation in the simulated queueing system.

(b) Estimation of the NYGH capacity using daily rolling data size

Figure 2. Sample Results

Table 1. χ^2 values of the χ^2 tests comparing the observed number of overtakes in the NYGH data to the simulated number of overtakes using the capacity estimated.

Data Size	KW Est	OT Est	1^{st} Simul	2^{nd} Simul	3^{rd} Simul	4^{th} Simul	C-value (0.05)
25	2	2	10	13	10	10	38
75	2	2	21	62	48	45	96
125	4	4	180	234	171	243	152
250	4	4	427	393	459	382	288
550	4	4	1028	1188	909	1135	606
1000	4	4	1874	2118	2046	1948	1075
2000	7	7	6456	8130	6687	8292	2105
5000	8	8	16824	23826	17213	23175	18014

Conclusions

- Determined sufficient conditions to mine the capacity of a QS under incomplete information.
  - Randomness
  - Continuity of service time distribution

.
- Two algorithms to estimate the effective number of servers were developed and they worked well.
- Only a little bit of data is needed with a small time for the algorithm to perform. The small data size required for convergence shows that the algorithms are cost-effective and fast.
- Other application includes
  - Competitive advantage; you can only observe your competitor from the outside and not from the inside and estimate their capacity to make your decisions in a business game.
  - Compliance auditing: you can audit the efficiency, effectiveness and compliance of your servers to see if they work as expected.

References

[1] Krivulin, N. K. (1994). A recursive equations-based representation for the g/g/m queue. Applied Mathematics Letters, 7(3):73–77.

[2] iefer, J. and Wolfowitz, J. (1956). On the characteristics of the general queue- ing process, with applications to random walk. The Annals of Mathematical Statistics, pages 147–161.

[3] Kiefer and Wolfowitz, 1955] Kiefer, J. d. and Wolfowitz, J. (1955). On the theory of queues with many servers. Transactions of the American Mathematical Society, 78(1):1–18