Ling Zhang Paper Reading Record

Ling Zhang

August 14, 2018

1 2018

1.1 Queueing

1. Bassamboo and Randhawa [1]
   - Comments on 05/23/2018 :
     - Accuracy of fluid model of $M/M/n+G$ depends on the state of the system, overloaded, critical-loading, underloaded
     - On the optimization problem:(1) very important to "linearize" the original optimization problem by a sanity check on the range of ρ; (2) the change of variable by applying $G^{-1}((1-1/\rho)^+) = w$ is absolutely brilliant; (3) the analysis thereafter is standard but it bring out the importance of monotonicity of the hazard in determining the optimality of system state, i.e., critical-loading, underloaded, or overloaded.
2. Zeltyn and Mandelbaum [2]
   -
   - Comment: seems difficult to understand.
3. Ward and Glynn [3]
   -
4. Halfin and Whitt [4]
   -
5. Whitt [5]
   -
6. Koçağa and Ward [6]
   -
7. Ward and Kumar [7]
   -

1.2 Fluid Model

1. Who the fuck first proposed the fluid approximation?! Maybe I can find this in Dr. Liu's thesis.

1.3 Time-varying Queueing Models

1.

1.4 Optimal Control

1. Cudina [8]
   -
2. Aravindakshan and Naik [9]
   - **Delayed/delay differential equations**, the most important modeling feature in the paper.
   - Related citations
     (a) Kharatishvili [10]
     (b) Avrani [11]
3. Avrani [11]
   - Used HJB method
   - In the title it also mentioned **Stochastic Corrections**.

1.5 Call-Back

1. Armony and Maglaras [12]
   - Comments on 05/29/2018:
     – Use HW regime to obtain performance approximation, propose a threshold routing policy based on steady state information, and prove that the rationalized regime is indeed the natural equilibrium point.
     – Use $M/M/N$ model, a two-class model, one with "best-effort" service and the other one with guaranteed quality-of-service.
2. Armony and Maglaras [13]
   - Comment on 05/29/2018
     –

1.6 Others

1. Gans et al. [14]
   -
2. Avramidis et al. [15]
   -

2 Dissertation Literature Review

We will include the review of the following stream of literature, ordered by priority.

- Optimal control of queueing systems
- Queueing systems
  - Time-varying Queueing
  - Asymptotic Approximations, fluid and diffusion
  - Queueing Networks

2.1 Optimal Control of Queueing System

There are three major streams of control problem constructed for a queueing system. Markov decision pocesses, which solves the exact solution to the original stochastic optimization problem. Brwonian control problem. Last but not the least, which is also the focus of this dissertation, the fluid control problem.

- MDP: George and Harrison [16] (see also Wijngaard and Stidham Jr [17],Wijngaard and Stidham [18] and Jo [19]), Stidham [20], Stidham and Weber [21] (a survey of Markov decision model for the control of networks), Weber and Stidham [22], Sennott [23] (a book considers a wide vari- ety of dynamic control problems associated with queueing models, and it develops a general computational method for such problems), Ata and Shneorson [24]
- Fluid Control
- Brownian Control: Ghamami and Ward [25]

1.
2. Harrison and Wein [26]
3. Bassamboo et al. [27]
4. Ward and Kumar [7]
5. Harrison [28]
6. Harrison and López [29]
7. Bäuerle [30]

2.2 Service Systems Analyzed by Queueing Theory

We look into three most relevant industries, sharing-economy, health-care and manufacturing. We also add a few more papers on block-chain analysis with queueing theory. See survey Stidham and Weber [21].

- Manufacturing Systems: Buzacott and Yao [31]

2.3 Queueing Networks

- Closed Queueing Networks: Gordon and Newell [32], Yao and Schechner [33], Buzacott and Yao [31], Kelly [34] (for background of Queueueing networks)
- Tandem Queueing: [22] (The global optimal control model of series and cycles of networks was studied),
- Other Queuing Networks: Jackson [35]

2.4 A Review Adapted from Dr. Liu's Dissertation

2.4.1 Time-varying Queueing Models

2.4.2 Network Queueing Models

- A First page of summary of variables

2.5 Time-varying System

2.5.1 Time-varying arrival rates

There ought to be some more on specific systems

- 26, provides background for the fact that how time-varying arrival rates, which commonly occurs in application but makes performance analysis difficult.
- 17
- 50,53
- 14, 15, service systems typically have arrival rates that vary significantly over time, and the results dramatically reveal the consequence, e.g., showing how the peak congestion lags behind the peak arrival rate, as discussed for the $M_t/GI/\infty$ stochastic model.

2.5.2 Time-varying staffing (decisions?)

This is about the decisions. Staffing is of course one of the most obvious decision in call-centers. What are the other decisions? production rate, pricing, etc.

2.5.3 Time-varying performance functions

2.6 Non-exponential informations

- 79, queueing model in patient flow management
- 20, how impatient customer will significantly alter system performance. This translates to how production expiration/ demand abandonment will effect system decisions. It proposed MSHT ED and QD regimes.
- 20, 81, customer abandonment is now recognized as an important feature in service systems. Review that production expiration and demand abandonment is an important feature. Splitting is an important feature we considered in the model.

- 7, service system often do have non-exponential service and patience distributions. (*need to specify in each cases?* **Due to the complexity in each practical problem, we need to balance problem praticality and model generality**)
- 46,76,77,81 shows that the patience distribution beyond its mean has a significant impact. However, 76, 77 also showns that the steady-state performance in the stationary $G/GI/s+GI$ model is relatively insensitive to the service-time cdf beyond its mean.
- In Dr. Liu's 2012 paper, they have shwon the importance of information beyond its mean can have a significant impact as well as for the transient performance.
- 31, an example of application of the result in Dr. Liu's 2012 paper(chapter 2), create new effective real-time delay predictors for arriving customers in a service system with time-varying arrivals.

2.6.1 Classical Erlang Models

Since we have a lot erlang assumptions in two of my models. It is important to review that and **clarify that due to the complexity of our model and the complexity of the associated optimization problem we used Erlang model.**

2.6.2 Non-exponential

Specifically in the car-sharing project, I need to justify the importance of including *non-exponential* and *distributional* information.

- Cathy Xia Paper, the differences is that we considered a new problem and included distributional information

---

2.7 Fluid Model

To discuss the fluid approximation, one needs to discuss both the single-server and many-server model. The former one is referred to as the *conventional* heavy traffic regime. The many-server model, from a math point of view, has *infinitely* many servers in the pre-limit as system *scale* increases to infinity.

- In his third chapter, he mentioned the following. *Fluid model is tractable, we are providing the basis for creating a performance-analysis tool for large-scale service systems, like the Queueing Network Analyzer(QNA) in 73 and 8. Algorithms based on performance formulas are appealing to supplement and complement computer simulation, because the models can be created and solved much more quickly. Thus they can be applied quickly in what if studies. They also can be efficiently embedded in optimization algorithms to systematical determine design and control parameters to meet performance objectives.* Therefore, my work follows closely to the development of time-varying fluid model and furthermore to incorporate it in a optimization problem/ optimal control problem.

2.7.1 Application of Deterministic Fluid Models

- 57, insight of fluid model as a legitimate model itself
- 54, 29, shown a long history of applying deterministic fluid models
- *_In this thesis, we primarily do not concern with establishing limits for sequences of scaled queueing processes. We directly concerned with the fluid model itself. _*

- We do not concern establishing limits. We construct stochastic optimization problem and use fluid model as an approximation. More importantly, we show the stochastic optimization problem equipped with fluid optimal control is a lower bound and asymptotically optimal

2.7.2 Heavy-traffic single-server

- 74 Heavy-traffic fluid and diffusion approximation become helpful. Due to the fact that we used single-server fluid model. We may need to dig a little deeper. Specifically, $\cdot/G*t/*infty$, $*cdot/M*t/1$, double-ended queue model, $M_t/M/1$.
- 74 has been mentioned at the very start as an extensive account
- 38, introduced the $GI/GI/1$ model.
- Who developed the $M*t/M/1$ mode?_ Dr. Liu in his 2012 paper. **Need to check that!**
- 5,32 extended the conventional heavy-traffic to queues with multiple servers. Furthermore, in 32 the convergence of the entire queue-length process is established.

2.7.3 Heavy-traffic many-server

Do I need to discuss the differences between my policy and the QED regmine? Although these two different policies are proposed under different framework.

- 28, Halfin and Whitt regime
- 20, extened Halfin and Whitt regime to Erlang A model
- 77
- 20,46,55,56, Fluid model arises in MSHT

2.7.4 Heavy-traffic many-server with time-varying rate

- 46, 47, 48, a theoretical basis for the MSHT limits for models with time-varying arrival rates and staffing
- Dr. Liu's 2012 paper and the one he has on $G*t/GI/s*t+GI$

2.8 Queueing Network (as is mentioned by Dr. Cao)

Claim that single queue is apparently not capturing all cases. There are problems can be modeled with different structures. Need to review at least, *double-ended queue*, *queueing network*, and *tandem queue*.

- 10, Jackson Network (That is all that he reviewed. Maybe we could review a bit more, the *V*, *reverse V* and the *N* model and pay close attention if any control problem is done.)
-

2.8.1 Transient and Asymptotic Performance

Isn't transient in our definition the same as time-varying. And I think we should separate the review on transient and asymptotic performance.

- 46-48, 56, 58, analyzed transient dynamics, fluid and diffusion approximation.
- Specifically, 58, $M*t/M*t/k*t+M*t$ queue is analyzed
- 56, MSHT limits or infinite-server queue
- 3, for a system with constant parameters, we care about the steady-state models ad they provide evaluation of the average system costs and revenue.

2.8.2 Introduction to Chapter 2

- 56, MSHT limit for the infinite server model
- 45, MSHT limit for the model with general service distribution
- 36, 37, 56, 62, new limits in Chapter 5 are consistent with recent

---

2.9 A network Generalization

2.9.1 Dr. Liu's network review

- Need to check with Dr. Liu what does his chapter three turns into. The difference between his model and mine are the following: *closed network* and *routing*. The first-come first-server can be regard as a proportional routing, however we have shown that it is under some assumptions not always the best policy to implement.
- 50,

2.9.2 Mine

I should include a slightly more comprehensive review, including the *V model*, *reverse V model*, and *N model*, *tandem model*.

References

[1] Achal Bassamboo and Ramandeep S Randhawa. On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Operations research*, 58(5):1398–1413, 2010.

[2] Sergey Zeltyn and Avishai Mandelbaum. Call centers with impatient customers: many-server asymptotics of the $M/M/n+G$ queue. *Queueing Systems*, 51(3-4):361–402, 2005.

[3] Amy R Ward and Peter W Glynn. A diffusion approximation for a $GI/GI/1$ queue with balking or reneging. *Queueing Systems*, 50(4):371–400, 2005.

[4] Shlomo Halfin and Ward Whitt. Heavy-Traffic Limits for Queues with Many Exponential Servers. *Operations Research*, 29(3), 1981. doi: 10.1287/opre.29.3.567.

[5] Ward Whitt. Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters*, 24:205–212, 1999.

[6] Yaşar Levent Koçağa and Amy R Ward. Admission control for a multi-server queue with abandonment. *Queueing Systems*, 65(3):275–323, 2010.

[7] Amy R Ward and Sunil Kumar. Asymptotically optimal admission control of a queue with impatient customers. *Mathematics of Operations Research*, 33(1):167–202, 2008.

[8] Milica Cudina. Asymptotically optimal control for some time-varying stochastic networks, 2006.

[9] Ashwin Aravindakshan and Prasad A Naik. Understanding the memory effects in pulsing advertising. *Operations Research*, 63(1):35–47, 2015.

[10] GL Kharatishvili. A maximum principle in extremal problems with delays. *Mathematical Theory of Control*, pages 26–34, 1967.

[11] Florin Avrani. Optimal control of fluid limits of queuing networks and stochasticity corrections. In *Mathematics of Stochastic Manufacturing Systems: AMS-SIAM Summer Seminar in Applied Mathematics, June 17-22, 1996, Williamsburg, Virginia*, volume 33, page 1. American Mathematical Soc., 1997.

[12] Mor Armony and Constantinos Maglaras. On Customer Contact Centers with a Call-Back Option: Customer Decisions, Routing Rules, and System Design. *Operations Research*, 2004. ISSN 0030-364X. doi: 10.1287/opre.1030.0088.

[13] Mor Armony and Constantinos Maglaras. Contact Centers with a Call-Back Option and Real-Time Delay Information. *Operations Research*, 52(4):527–545, 2004.

[14] Noah Gans, Haipeng Shen, Yong-Pin Zhou, Nikolay Korolev, Alan McCord, and Herbert Ristock. Parametric Forecasting and Stochastic Programming Models for Call-Center Workforce Scheduling. *Manufacturing & Service Operations Management*, 17(4):571–588, 2015. ISSN 1523-4614. doi: 10.1287/msom.2015.0546.

[15] Athanassios N Avramidis, Alexandre Deslauriers, Pierre L ' Ecuyer, and Pierre L 'ecuyer. Modeling Daily Arrivals to a Telephone Call Center. *Management Science*, 50(7):896–908, 2004. doi: 10.1287/mnsc.1040.0236.

[16] Jennifer M George and J Michael Harrison. Dynamic control of a queue with adjustable service rate. *Operations research*, 49(5):720–731, 2001.

[17] Jacob Wijngaard and Shaler Stidham Jr. Forward recursion for markov decision processes with skip-free-to-the-right transitions, part i: theory and algorithm. *Mathematics of Operations Research*, 11(2):295–308, 1986.

[18] J Wijngaard and S Stidham. Forward recursion for markov decision processes with skip-free-to-the-right transitions part ii: non-standard applications. *Statistica Neerlandica*, 54(2):160–174, 2000.

[19] Kyung Y Jo. A lagrangian algorithm for computing the optimal service rates in jackson queuing networks. *Computers & operations research*, 16(5):431–440, 1989.

[20] Shaler Stidham. Optimal control of admission to a queueing system. *IEEE Transactions on Automatic Control*, 30(8):705–713, 1985.

[21] Shaler Stidham and Richard Weber. A survey of markov decision models for control of networks of queues. *Queueing systems*, 13(1-3):291–314, 1993.

[22] Richard R Weber and Shaler Stidham. Optimal control of service rates in networks of queues. *Advances in applied probability*, 19(1):202–218, 1987.

[23] Linn I Sennott. *Stochastic dynamic programming and the control of queueing systems*, volume 504. John Wiley & Sons, 2009.

[24] Barış Ata and Shiri Shneorson. Dynamic control of an m/m/1 service system with adjustable arrival and service rates. *Management Science*, 52(11):1778–1791, 2006.

[25] Samim Ghamami and Amy R Ward. Dynamic scheduling of a two-server parallel server system with complete resource pooling and reneging in heavy traffic: Asymptotic optimality of a two-threshold policy. *Mathematics of Operations Research*, 38(4):761–824, 2013.

[26] J Michael Harrison and Lawrence M Wein. Scheduling networks of queues: Heavy traffic analysis of a two-station closed network. *Operations research*, 38(6):1052–1064, 1990.

[27] Achal Bassamboo, J Michael Harrison, and Assaf Zeevi. Design and control of a large call center: Asymptotic analysis of an lp-based method. *Operations Research*, 54(3):419–435, 2006.

[28] J Michael Harrison. Brownian models of open processing networks: Canonical representation of workload. *Annals of Applied Probability*, pages 75–103, 2000.

[29] J Michael Harrison and Marcel J López. Heavy traffic resource pooling in parallel-server systems. *Queueing systems*, 33(4):339–368, 1999.

[30] Nicole Bäuerle. Optimal control of queueing networks: An approach via fluid models. *Advances in Applied Probability*, 34(2):313–328, 2002.

[31] John A Buzacott and David D. Yao. On queueing network models of flexible manufacturing systems. *Queueing Systems*, 1(1):5–27, 1986.

[32] William J. Gordon and Gordon F. Newell. Closed queueing systems with exponential servers. *Operation Research*, 15(2):145–155, 1967.

[33] David D Yao and Zvi Schechner. Decentralized control of service rates in a closed jackson network. *IEEE Transactions on Automatic Control*, 34(2):236–240, 1989.

[34] Frank P Kelly. *Reversibility and stochastic networks*. Cambridge University Press, 2011.

[35] James R Jackson. Networks of waiting lines. *Operations research*, 5(4):518–521, 1957.