Proceedings of the 2019 Winter Simulation Conference
N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, eds.

USING SIMULATION TO STUDY THE LAST TO ENTER SERVICE DELAY ANNOUNCEMENT IN MULTISERVER QUEUES WITH ABANDONMENT

Aditya Shah
Anders Wikum
Jamol Pender

Operations Research and Information Engineering
Cornell University
228 Rhodes Hall
Ithaca, NY 14850, USA

ABSTRACT

The Last to Enter Service (LES) delay announcement is one of the most commonly used delay announcements in queueing theory because it is quite simple to implement. Recent research has shown that using a convex combination of LES and the conditional mean delay are optimal under the mean squared error and the optimal value depends on the correlation between LES and the virtual waiting time. To this end, we show using simulation that it is important to be careful when using finite queue sizes, especially in a heavy traffic setting. Using simulation we demonstrate that the correlation between LES and the virtual waiting time can differ from heavy traffic results and can therefore have a large impact on the optimal announcement. Finally, we use simulation to assess the value of giving future information in computing correlations with virtual waiting times and show that future information is helpful in some settings.

1 INTRODUCTION

Estimating the delays that customers will experience in service systems is a hard and important problem of study. Often many service systems use delay announcements as a way of improving the customer experience and the overall efficiency of the system. In the context of healthcare systems, many hospitals are choosing to display the average waiting times of recent patients via smartphone apps, the internet, or via a display message signs (DMS). This information may help reduce peak congestion by encouraging incoming patients to visit a less busy emergency facility. In telephone call centers, the waiting queue is generally invisible to the caller, unlike physical queues in department stores and supermarkets. Forecasting delay announcements can provide very useful information to customers.

Most of the current research that analyzes the impact of providing queue length or waiting time information to customers tends to focus on the impact of delay announcements. Delay announcements are useful tools that enable managers of service systems to provide customers with valuable waiting time information. For the most part, the literature only explores how customers respond to the delay announcements or the actual number that is displayed. Previous work by Whitt (1999), Armony and Maglaras (2004), Guo and Zipkin (2007), Hassin (2007), Armony et al. (2009), Jouini et al. (2009), Jouini et al. (2011), Allon and Bassamboo (2011), Allon et al. (2011), Massey and Pender (2018), Bassamboo and Ibrahim (2018),and references therein focus on this aspect of the announcements. However, there is a stream of literature that develops new predictors for providing delay announcements to customers, see for example Ibrahim and Whitt (2008), Ibrahim and Whitt (2009), Ibrahim and Whitt (2011), Ibrahim et al. (2016), Nirenberg et al. (2018), Dong et al. (2018), and Novitzky et al. (2019).

Delay announcements can be helpful when queues are invisible to customers, as in many telecommunications centers. However, the analysis and comparison of alternative delay estimators is very complicated and difficult. In this context, we would like a prediction that can estimate the actual delay that each customer experiences. To this end, we quantify the performance of a delay estimator by using the mean squared error (MSE). Since the MSE is not known in closed form for even simple queueing models, we estimate the expected MSE via simulation by computing the average squared error (ASE), averaging over a large number of customers in steady state. However, we do not believe that the MSE is the best performance measure to use. Thus, one contribution of this work is to assess the performance of the WA-LES(α) defined in the work of Bassamboo and Ibrahim (2018) under more general performance measures that capture the asymmetry of customer attitudes about waiting.

1.1 Contributions of Our Work

By using stochastic simulation for the M/M/C+M queue, we provide new insights to the following questions:

- What is the impact of finite queue sizes on the correlation of LES and actual waiting times?
- What is the affect of using different measures of error between LES and waiting times?
- What is the value of knowing future information about LES?

1.2 Organization of the Paper

The remainder of the paper is organized as follows. Section 2 presents the queueing model studied and how we evaluate the performance of delay announcements. In Section 3, we present our simulation results and explain how various parameters of the queueing model affect the performance of the delay announcements. Finally, a conclusion is given in Section 4.

2 THE MULTISERVER QUEUE WITH LES DELAY ANNOUNCEMENTS

2.1 Queueing Model

In this paper we consider the $M/M/C+M$ queue or the Erlang-A queueing model, which is a Markovian queueing model that allows for customer abandonment. We assume that customers arrive to the queue at arrival rate λ, they have the potential receive service from C identical servers at rate μ, or they can abandon the queue if they are waiting at rate θ. With this information, we can represent the queue length by the following expression

$$Q(t) = Q(0)+\Pi^a\left(\int_0^t \lambda ds\right)-\Pi^d\left(\int_0^t \mu\cdot(Q(s)\wedge C)ds\right)-\Pi^b\left(\int_0^t \theta\cdot(Q(s)-C)^+ds\right)$$

where each $\Pi(\cdot)$ is a unit rate Poisson process. In this model, we also have that $\Pi^a\left(\int_0^t \lambda ds\right)$ counts the number of customers that arrive to the queue in the time interval $(0,t]$. Similarly, $\Pi_i^d\left(\int_0^t \mu\cdot(Q(s)\wedge C)ds\right)$ counts the number of customers that depart the queue having completed their service from a server in the time interval $(0,t]$. Finally, $\Pi^b\left(\int_0^t \theta\cdot(Q(s)-C)^+ds\right)$ counts the number of customers that are impatient and depart the queue via abandonment in the time interval $(0,t]$. There are many results about the Erlang-A model and one can gain more insight by reading the work of Garnett et al. (2002), Zeltyn and Mandelbaum (2005), Talreja et al. (2009), and Daw and Pender (2019).

2.2 Performance Measures

In this work, we want to explore how much LES we should use for different performance measures. We are inspired by the work of Bassamboo and Ibrahim (2018), which uses a convex combination of LES and a static announcement given by the conditional mean delay i.e. $E[W_\infty|W_\infty>0]$ where W_∞ is the steady state waiting time of a customer who does not abandon the queue. Thus, the announcement that a customer would receive known as the weighted average LES or WA-LES(α) is given by

$$\mathrm{WA-LES}(\alpha) = \alpha LES_k + (1-\alpha)E[W_\infty | W_\infty > 0].$$

One standard way to quantify the accuracy of a delay announcement is to use the mean-squared error (MSE), which is defined as the expected value of the square of the difference between delay announcement for the k^{th} customer (P_k) and the corresponding actual waiting time of the k^{th} customer (W_k) i.e.

$$MSE \;=\; \mathbb{E}[(W_k - P_k)^2] = \mathbb{E}\left[(W_k - \alpha(LES_k) - (1-\alpha)E[W_\infty|W_\infty > 0])^2\right].$$

Often it is intractable to explicitly determine the MSE of a delay predictor such as LES. Thus, we exploit stochastic simulation to quantify the accuracy of our delay announcements. In our simulation experiments, we quantify the accuracy of a delay announcement by computing the average squared error (ASE), defined by:

$$ASE \;=\; \frac{1}{N}\sum_{k=1}^{N}(W_k - P_k)^2 = \frac{1}{N}\sum_{k=1}^{N}(W_k - \alpha(LES_k) - (1-\alpha)E[W_\infty|W_\infty > 0])^2.$$

However, we are also inspired by the work of Jouini et al. (2015), where the authors explore the use of a newsvendor or quantile type performance measure to evaluate the performance of the delay announcement. We believe that what we call a Quantile Error (QE) performance measure is more relevant to evaluating the performance of delay announcements than the standard ASE. We define the QE performance measure as follows:

$$\begin{aligned} QE \;&=\; \frac{1}{N}\sum_{k=1}^{N}\left(\gamma_1 \cdot (W_k - P_k)^+ + \gamma_2 \cdot (P_k - W_k)^+\right) \\ &=\; \frac{1}{N}\sum_{k=1}^{N}\gamma_1 \cdot (W_k - \alpha(LES_k) - (1-\alpha)E[W_\infty|W_\infty > 0])^+ \\ &\quad + \frac{1}{N}\sum_{k=1}^{N}\gamma_2 \cdot (\alpha(LES_k) + (1-\alpha)E[W_\infty|W_\infty > 0] - W_k)^+. \end{aligned}$$

The QE performance measure is like an absolute value measure, however, it has the ability to capture the asymmetry between customers receiving a prediction that is larger or smaller than their actual wait. Customers tend to perceive waits larger than their predicted wait to be worse than when their wait time is smaller than the predicted wait. We are able to model this case when we set $\gamma_1 \geq \gamma_2$. Thus, for part of this paper, we will evaluate the proportion of LES we should use when making a delay announcement under the QE performance measure. We intend to understand how the optimal amount of LES differs from the MSE performance measure.

3 SIMULATON RESULTS

In this section, we explain the simulation setup for the simulations we conducted both in the future information context and the LES correlation context. In each of the situations we simulate an $M/M/C+M$ queue, which is known as the Erlang-A queue. Although one can use standard techniques like uniformization for the simulation of the queueing process, we implemented a full discrete event simulation that was also able to capture the virtual waiting times of all of the customers as they pass through the queue, see for example Massey and Pender (2013), Pender and Massey (2017). In addition to the virtual waiting times, we also stored information about the person who last entered service thereby creating a vector of LES wait times.

For the future information part of the paper, we shifted this LES vector k slots to the right to get a future information vector of LES wait times. For the correlation part of the paper, we computed our weighted average LES estimator with the LES vector and the conditional mean waiting time, which is known from our simulation experiments. Note that the weighted average estimator depends on the parameter α and we used this parameter to see the effect of weighting the two pieces of information in a variety of cost function settings. In the next two subsections, we outline in detail the results of the simulation experiments.

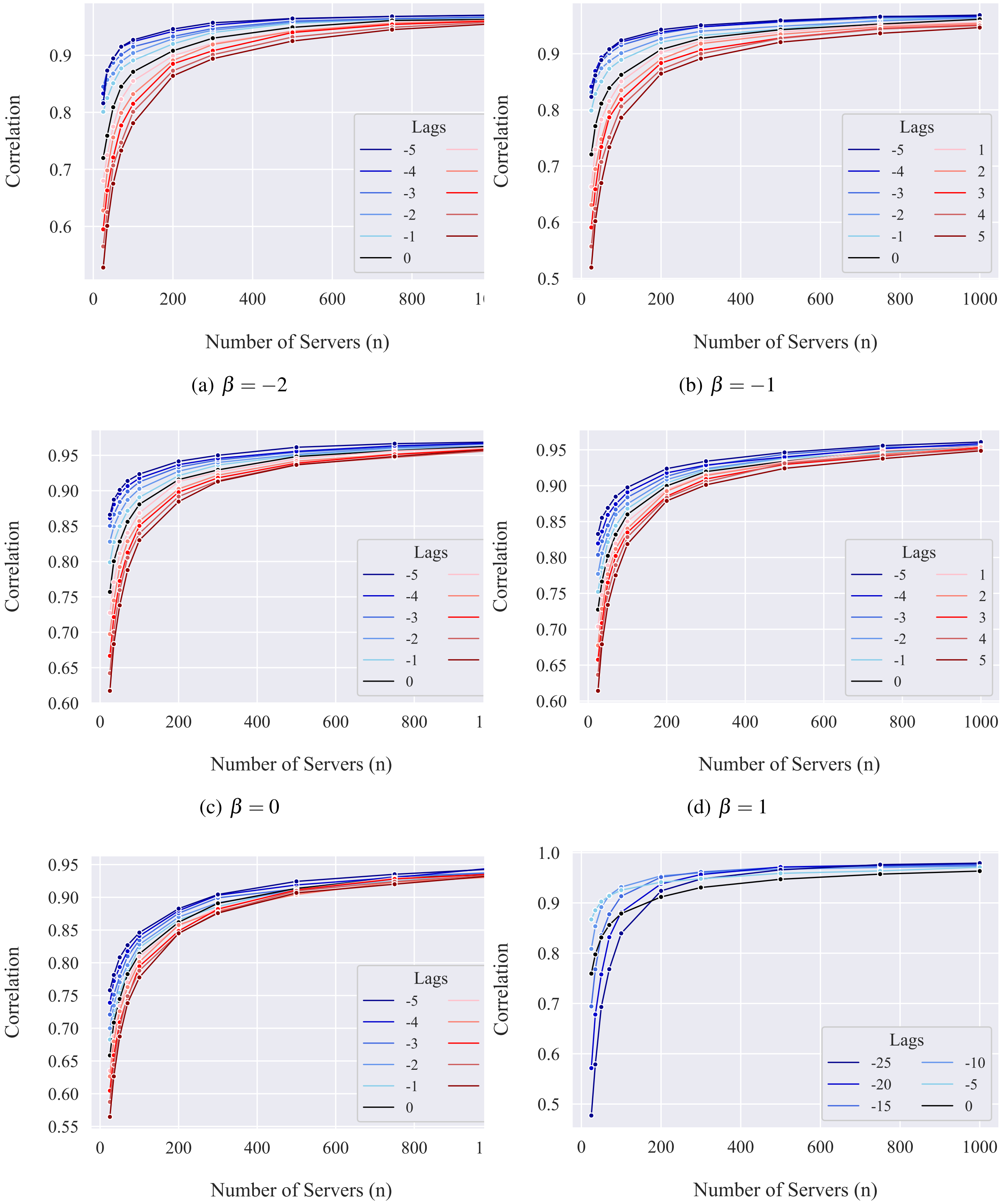
3.1 Future Information

Figure 1 consists of a series of correlation plots between the WA-LES announcement and virtual wait times. Notably, it shows that the WA-LES announcement with minor extrapolation (negative lags) consistently outperforms the LES announcement in the few servers case. It is clear, however, that the effect of future information becomes less significant as the number of servers in the system increases. Moreover, while small amounts of future information are beneficial to WA-LES announcement, large lags have the opposite effect, see the bottom right of Figure 1. In the few servers case, we see that the correlation sees a significant reduction from the base LES predictor under large lags. This effect too becomes less significant as the number of servers in the queuing system increases. Thus, we observe that we might be able to give better predictions if we wait to tell customers about their wait in line.

3.2 LES Correlations with Virtual Waiting Times

In this section, we simulate the multiserver queue with abandonment (Erlang-A) and derive the optimal amount of LES that a manager would use in a delay announcement to customers. In Figures 2 - **??**, we simulated the error between the actual waiting time for customers versus the predicted waiting time using the delay announcement $\alpha LES + (1-\alpha)E[W|W>0]$.

In Figure 2, we simulate the Erlang-A queue with C=25 servers. We plot the **mean squared error** as a function of the parameter α, which determines how much of LES one should use in the delay announcement to customers. For each of the individual plots, we vary the arrival rate $\lambda = 25+5\beta$ where $\beta \in \{-2,0,2\}$ along with the abandonment rate $\theta \in \{.5,1,2\}$ for a total of nine plots. We observe that as we increase the abandonment rate, the amount of LES that we should use in our delay announcement is increasing. This phenomenon is not captured in the work by Bassamboo and Ibrahim (2018), however, to be fair it is not explictly studied there. We also observe that $\beta = 0$ uses the largest amount of LES across all of the scalings. Moreover, we observe that as the absolute value of the β is increased, the smaller amount of LES we should use in our delay announcement. This is consistent with other findings where the correlation is most strong when $\lambda = \mu C$.



(a) $\beta = -2$

(b) $\beta = -1$

(c) $\beta = 0$

(d) $\beta = 1$

(e) $\beta = 2$

(f) Larger Lags

Figure 1: Simulation experiments featuring $\lambda = n + \beta\sqrt{n}$, $\mu = 1$, $C = n$.

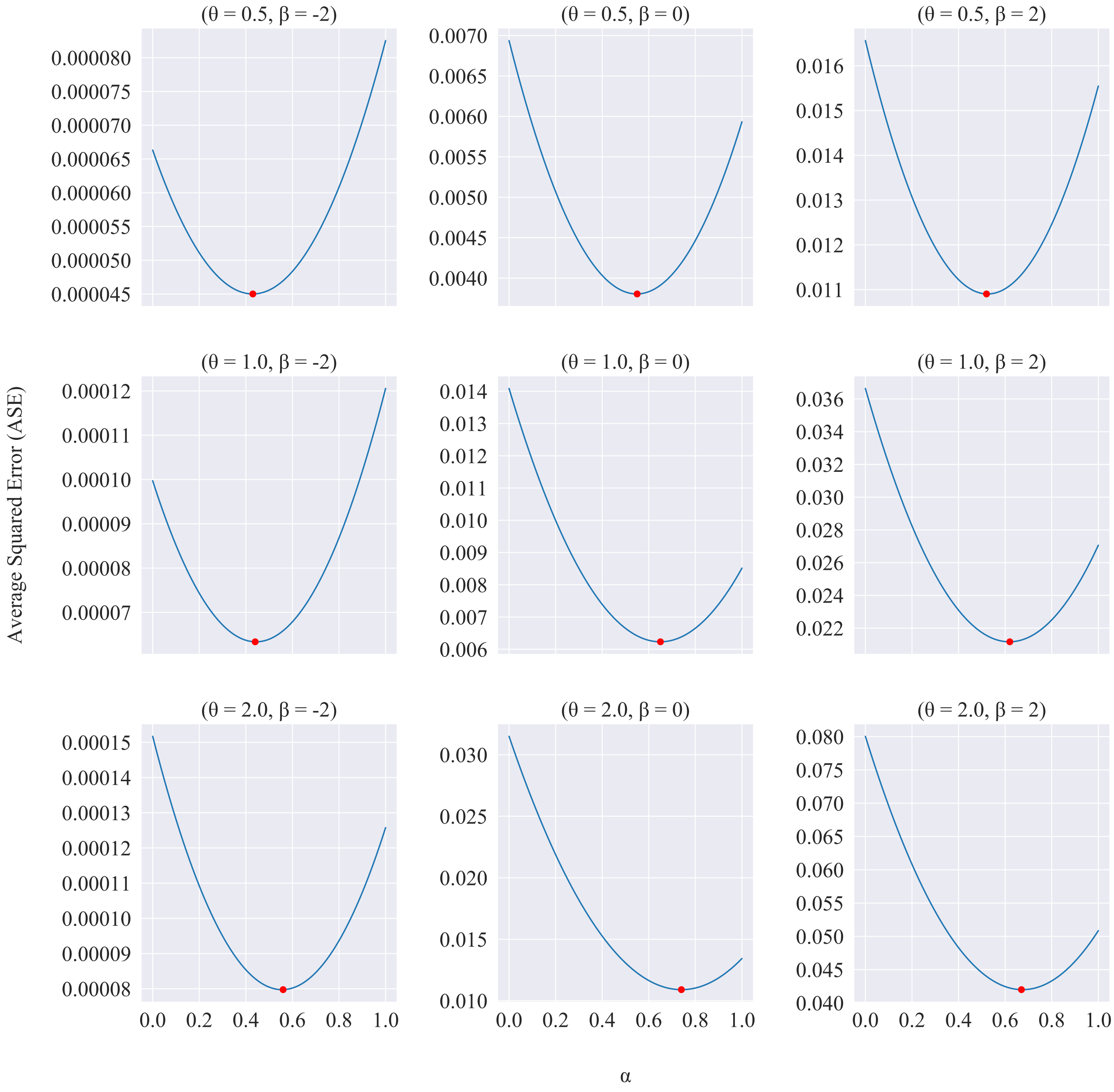


Figure 2: MSE vs. α. $\lambda = n + \beta\sqrt{n}$, $\mu = 1$, $C = n = 25$.

In Figure 3, we simulate the Erlang-A queue with C=25 servers. We plot the **quantile error** with $\gamma_1 = .5$ and $\gamma_2 = 1$ as a function of the parameter α, which determines how much of LES one should use in the delay announcement to customers. For each of the individual plots, we vary the arrival rate $\lambda = 25 + 5\beta$ where $\beta \in \{-2, 0, 2\}$ along with the abandonment rate $\theta \in \{.5, 1, 2\}$ for a total of nine plots. We observe that as we increase the abandonment rate, the amount of LES that we should use in our delay announcement is increasing. We also observe that $\beta = 0$ uses the largest amount of LES across all of the scalings. Moreover, we observe that as the absolute value of the β is increased, the smaller amount of LES we should use in our delay announcement.

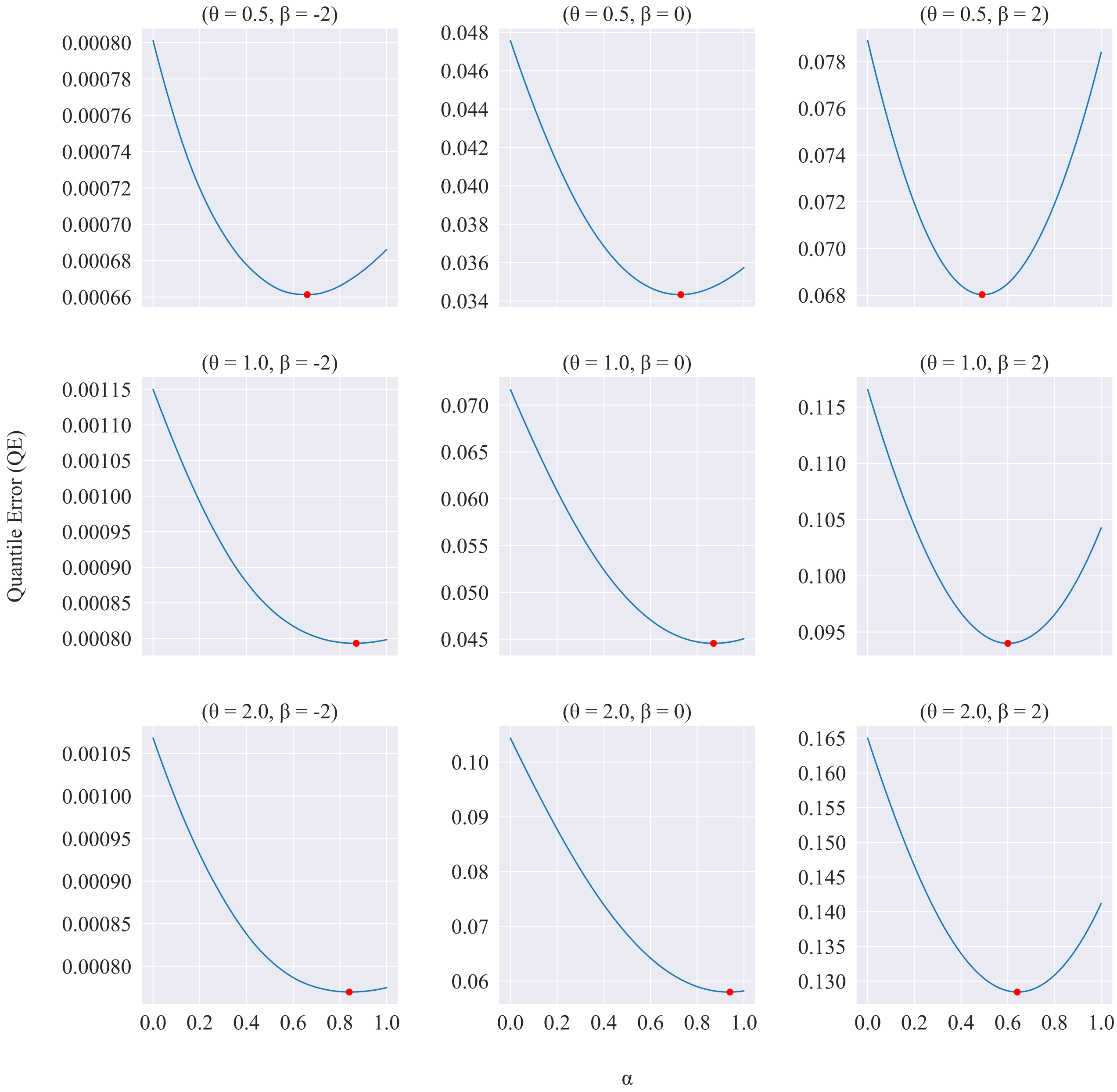


Figure 3: QE vs. α. $\lambda = n + \beta\sqrt{n}$, $\mu = 1$, $C = n = 25$, $\gamma_1 = .5$, $\gamma_2 = 1$.

In Figure 4, we simulate the Erlang-A queue with C=25 servers. We plot the **quantile error** with $\gamma_1 = 1$ and $\gamma_2 = 1$ as a function of the parameter α, which determines how much of LES one should use in the delay announcement to customers. For each of the individual plots, we vary the arrival rate $\lambda = 25 + 5\beta$ where $\beta \in \{-2, 0, 2\}$ along with the abandonment rate $\theta \in \{.5, 1, 2\}$ for a total of nine plots. We observe that as we increase the abandonment rate, the amount of LES that we should use in our delay announcement is increasing. We also observe that $\beta = 0$ uses the largest amount of LES across all of the scalings. Moreover, we observe that as the absolute value of the β is increased, the smaller amount of LES we should use in our delay announcement.

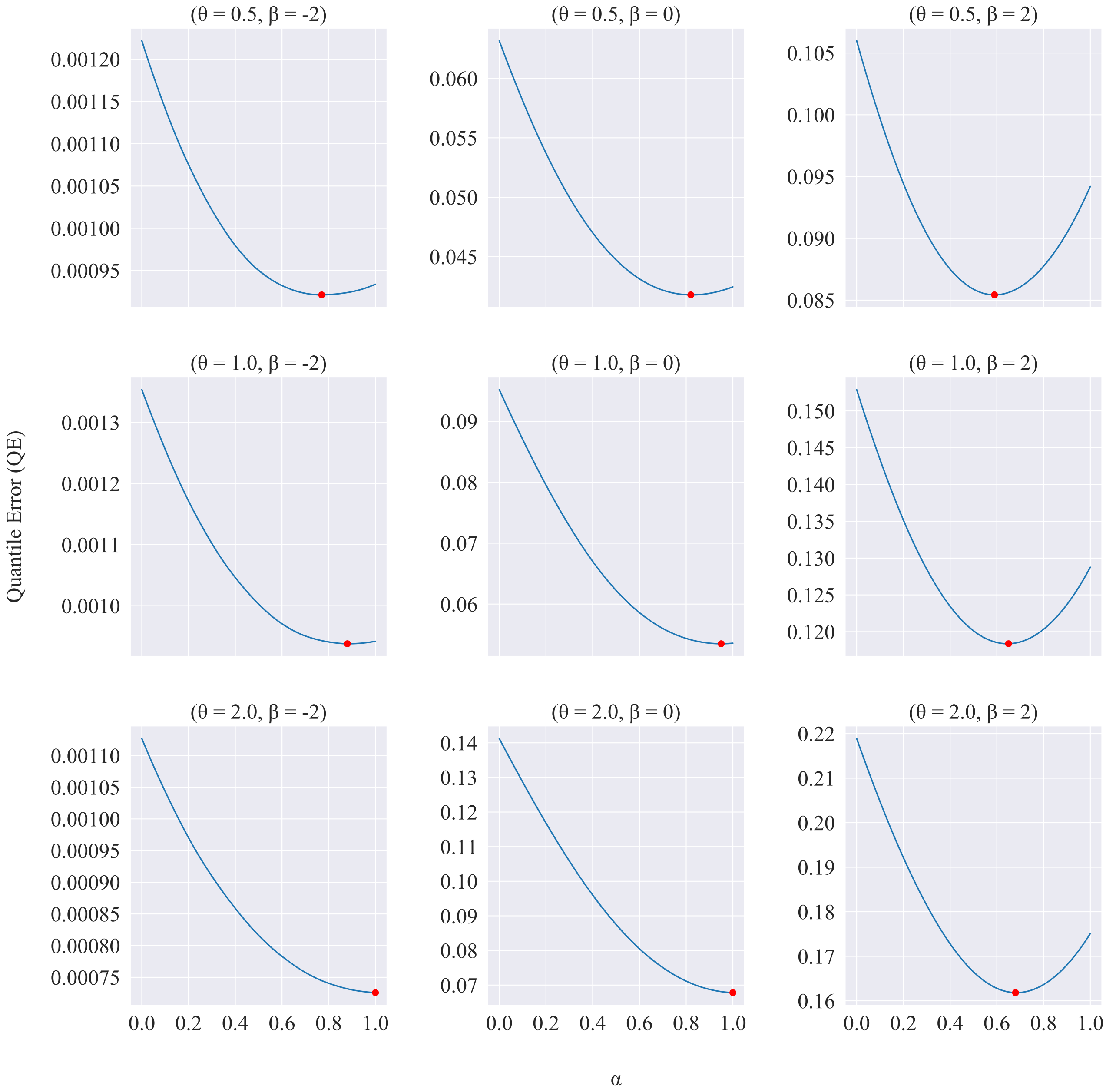


Figure 4: QE vs. α. $\lambda = n + \beta\sqrt{n}$, $\mu = 1$, $C = n = 25$, $\gamma_1 = 1$, $\gamma_2 = 1$.

In Figure 5, we simulate the Erlang-A queue with C=25 servers. We plot the **quantile error** with $\gamma_1 = 2$ and $\gamma_2 = 1$ as a function of the parameter α, which determines how much of LES one should use in the delay announcement to customers. For each of the individual plots, we vary the arrival rate $\lambda = 25 + 5\beta$ where $\beta \in \{-2, 0, 2\}$ along with the abandonment rate $\theta \in \{.5, 1, 2\}$ for a total of nine plots. We observe that as we increase the abandonment rate, the amount of LES that we should use in our delay announcement is increasing. We also observe that $\beta = 0$ uses the largest amount of LES across all of the scalings. Moreover, we observe that as the absolute value of the β is increased, the smaller amount of LES we should use in our delay announcement.

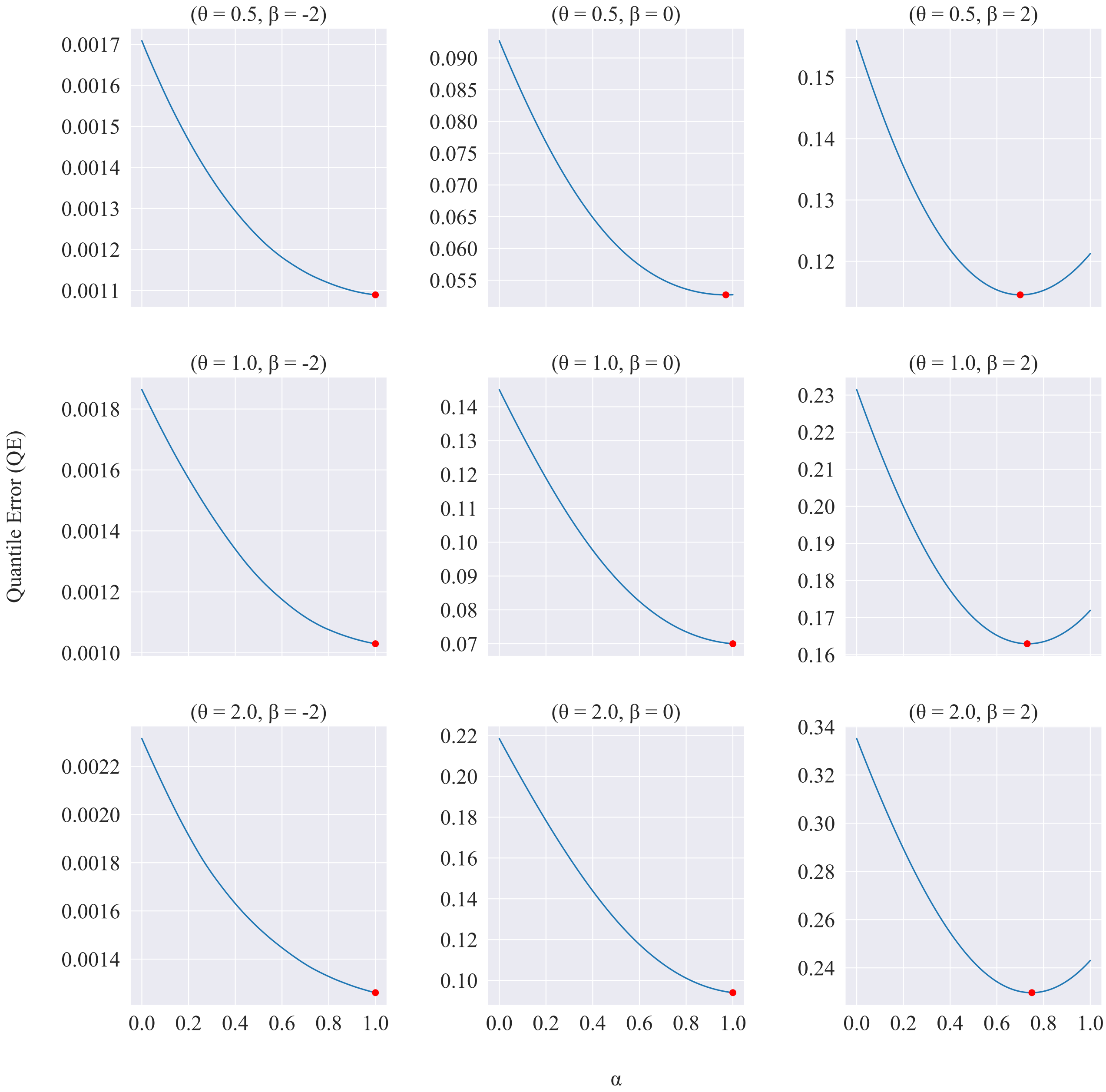


Figure 5: QE vs. α. $\lambda = n + \beta\sqrt{n}$, $\mu = 1$, $C = n = 25$, $\gamma_1 = 2$, $\gamma_2 = 1$.

3.2.1 Observations Across Different Figures

We find that the optimal MSE allocates a smaller proportion of LES to the announcement than any of the other QE. More specifically, we find that as we increase the value of γ_1 the larger proportion of LES is allocated to the delay announcement. Thus, we find that the MSE tends to underutilize the LES in reporting to customers. We also find that as the number of servers is decreased, the errors are more sensitive to changes in the parameters. We also observe that the more impatient customers, the larger proportion of LES is used in the announcement.

3.3 Non-exponential Distributions

Although not explicitly discussed in this work for space considerations, we have performed additional simulations for non-exponential arrival, service, and abandonment distributions. Qualitatively, we see the same performance as the Erlang-A in terms of the mean arrival, service, and abandonment rates.

4 CONCLUSION AND FUTURE RESEARCH

In this paper, we have investigated the impact of alternative error functions on using a combination of LES and the mean waiting time as a delay announcement. Previous work has shown that the optimal proportion of LES under the MSE depends on the correlation between the waiting time and the LES. However, we observe two important realities when using alternative error functions. Under a quantile type of error, we observe that the optimal LES proportion is higher than the MSE. Second, we observe that the proportion increases as we weight the importance of the underestimating the actual waiting time.

We find several directions for future research. For one, we hope to understand the impact of non-stationary arrival rates, non-constant numbers of servers, and self-exciting arrivals like in Daw and Pender (2018b) and Daw and Pender (2018a), which are all very relevant for real-world queues. Finally, in the spirit of Pender et al. (2017) and Pender et al. (2018), we would like to understand the impact of delayed LES in a rigorous manner. This would provide insight for the delay announcement literature when the announcements are delayed themselves.

ACKNOWLEDGMENTS

Aditya Shah and Anders Wikum acknowledge the gracious support of the School Operations Research and Information Engineering at Cornell University. Anders Wikum also acknowledges the support of the Engineering Learning Initiatives program at Cornell University, which provided gracious funding for this research.

REFERENCES

Allon, G., and A. Bassamboo. 2011. “The Impact of Delaying the Delay Announcements”. *Operations Research* 59(5):1198–1210.

Allon, G., A. Bassamboo, and I. Gurvich. 2011. “We Will Be Right With You”: Managing Customer Expectations With Vague Promises and Cheap Talk”. *Operations Research* 59(6):1382–1394.

Armony, M., and C. Maglaras. 2004. “On Customer Contact Centers With a Call-back Option: Customer Decisions, Routing Rules, and System Design”. *Operations Research* 52(2):271–292.

Armony, M., N. Shimkin, and W. Whitt. 2009. “The Impact of Delay Announcements in Many-server Queues With Abandonment”. *Operations Research* 57(1):66–81.

Bassamboo, A., and R. Ibrahim. 2018. “A General Framework to Compare Announcement Accuracy: Static vs LES-based Announcement”. *Working Paper*.

Daw, A., and J. Pender. 2018a. “Exact Simulation of the Queue-Hawkes Process”. In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 4234–4235. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Daw, A., and J. Pender. 2018b. “Queues Driven by Hawkes Processes”. *Stochastic Systems* 8(3):192–229.

Daw, A., and J. Pender. 2019. “New Perspectives on the Erlang-A Queue”. *Advances in Applied Probability* 51(1):268–299.

Dong, J., E. Yom-Tov, and G. B. Yom-Tov. 2018. “The Impact of Delay Announcements on Hospital Network Coordination and Waiting Times”. *Management Science* 65(5):1969–1994.

Garnett, O., A. Mandelbaum, and M. Reiman. 2002. “Designing a Call Center with Impatient Customers”. *Manufacturing & Service Operations Management* 4(3):208–227.

Guo, P., and P. Zipkin. 2007. “Analysis and Comparison of Queues With Different Levels of Delay Information”. *Management Science* 53(6):962–970.

Hassin, R. 2007. “Information and Uncertainty in a Queuing System”. *Probability in the Engineering and Informational Sciences* 21(03):361–380.

Ibrahim, R., M. Armony, and A. Bassamboo. 2016. “Does the Past Predict the Future? The Case of Delay Announcements in Service Systems”. *Management Science* 63(6):1762–1780.

Ibrahim, R., and W. Whitt. 2008. “Real-time Delay Estimation in Call Centers”. In *Proceedings of the 40th Conference on Winter Simulation*, edited by S. J. Mason, R. R. Hill, L. Moench, and O. Rose, 2876–2883. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Ibrahim, R., and W. Whitt. 2009. “Real-time Delay Estimation in Overloaded Multiserver Queues with Abandonments”. *Management Science* 55(10):1729–1742.

Ibrahim, R., and W. Whitt. 2011. “Real-Time Delay Estimation Based on Delay History in Many-Server Service Systems with Time-Varying Arrivals”. *Production and Operations Management* 20(5):654–667.

Jouini, O., O. Z. Akşin, F. Karaesmen, M. S. Aguir, and Y. Dallery. 2015. “Call Center Delay Announcement Using a Newsvendor-Like Performance Criterion”. *Production and Operations Management* 24(4):587–604.

Jouini, O., Z. Aksin, and Y. Dallery. 2011. “Call Centers With Delay Information: Models and Insights”. *Manufacturing &amp; Service Operations Management* 13(4):534–548.

Jouini, O., Y. Dallery, and Z. Akşin. 2009. “Queueing Models for Full-flexible Multi-class Call Centers With Real-time Anticipated Delays”. *International Journal of Production Economics* 120(2):389–399.

Massey, W. A., and J. Pender. 2013, February. “Gaussian Skewness Approximation for Dynamic Rate Multi-server Queues With Abandonment”. *Queueing Systems* 75(2-4):243–277.

Massey, W. A., and J. Pender. 2018. “Dynamic Rate Erlang-A Queues”. *Queueing Systems* 89(1-2):127–164.

Nirenberg, S., A. Daw, and J. Pender. 2018. “The Impact of Queue Length Rounding and Delayed App Information on Disney World Queues”. In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 3849–3860. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Novitzky, S., J. Pender, R. H. Rand, and E. Wesson. 2019. “Nonlinear Dynamics in Queueing Theory: Determining the Size of Oscillations in Queues with Delay”. *SIAM Journal on Applied Dynamical Systems* 18(1):279–311.

Pender, J., and W. A. Massey. 2017. “Approximating and Stabilizing Dynamic Rate Jackson Networks with Abandonment”. *Probability in the Engineering and Informational Sciences* 31(1):1–42.

Pender, J., R. H. Rand, and E. Wesson. 2017. “Queues with Choice via Delay Differential Equations”. *International Journal of Bifurcation and Chaos* 27(04):1730016.

Pender, J., R. H. Rand, and E. Wesson. 2018. “An Analysis of Queues with Delayed Information and Time-varying Arrival Rates”. *Nonlinear Dynamics* 91(4):2411–2427.

Talreja, R., W. Whitt et al. 2009. “Heavy-traffic Limits for Waiting Times in Many-server Queues with Abandonment”. *The Annals of Applied Probability* 19(6):2137–2175.

Whitt, W. 1999. “Improving Service by Informing Customers About Anticipated Delays”. *Management science* 45(2):192–207.

Zeltyn, S., and A. Mandelbaum. 2005. “Call Centers with Impatient Customers: Many-server Asymptotics of the M/M/n+ G Queue”. *Queueing Systems* 51(3-4):361–402.

AUTHOR BIOGRAPHIES

ADITYA SHAH is a junior undergraduate student in Operations Research and Information Engineering at Cornell University in Ithaca, NY. He is very passionate about applied mathematics and algorithmic thinking. He also has done extensive research in data science and machine learning during his undergraduate career. His e-mail address is as2564@cornell.edu.

ANDERS WIKUM is a sophomore undergraduate student in Operations Research and Information Engineering at Cornell University in Ithaca, NY. He is interested in the underlying mathematics in machine learning and broader data science. He has worked on multiple data science, simulation, and algorithmic trading projects. His e-mail address is aew236@cornell.edu.

JAMOL PENDER is an Assistant Professor of Operations Research and Information Engineering at Cornell University in Ithaca, NY. His research interests are in applied probability and he studies how information affects the dynamics of queueing systems and applies his work in healthcare, transportation, and real-world settings. His e-mail address is jjp274@cornell.edu.