Research Statement

I am currently working on creating approximations and bounds for complex queueing models. These models have proven valuable to study communication and computer systems. Many organizations, such as call centers, banks, airlines routinely utilize queueing models to manage customer-facing environments efficiently. Additionally, queueing models also have important applications for emerging applications such as cloud computing and ride-sharing.

Motivation

Unfortunately, it is complicated to analyze queueing models. First, the arrival and service processes which mainly determine a queueing model are uncertain and so must be represented as stochastic processes which are difficult to analyze. Second, we only can obtain imperfect information from data collected from service systems. To address these challenges in the engineering perspectives, it is important to propose some tractable approximation techniques such as easy algorithms or simple formulas to approximate unknown queueing characteristics. Over the years, many useful queueing approximations have been developed, which depend on limited information, such as the first two moments of the interarrival-time and service-time distributions. Therefore, a natural and important research question is how accurate these approximations are. Given approximations, people traditionally evaluate approximations by simulating one system and reporting accuracy. But the approach is hard to work when only partial information is given. I am working on new ways in my thesis research that goes beyond just simulating one or two concrete cases.

The original interest in the thesis research was primarily motivated by parametric-decomposition approximations for Non-Markovian open queueing networks, as in [Whitt(1983)] where each queue is approximated by a single-server queue or multi-server queue. The tool Queueing Network Analyzer (QNA) treats both single-server and multi-server to approximate the steady-state performance of a whole queueing network. For the single-server case, each queue is approximated by a $GI/GI/1$ queue with unlimited waiting room and the first-come-first-served discipline, where the inter-arrival time and service time are each approximately characterized by their mean and squared coefficient of variation. The procedure in QNA is, first, to approximate each arrival process by a renewal process partially characterized by the first two moments of the renewal interval and, second, for each node to apply approximation formula for the congestion measures in a single-server queue partially characterized by the first two moments of the inter-arrival time and service time distribution. Therefore, evaluating the approximation formulas via effective mechanism can serve to avoid remarkable errors. Some existing results from [Whitt(1984a)], [Klincewicz and Whitt(1984)] and [Whitt(1984b)] focusing on bounds for $GI/M/1$ and $K_2/GI/1$ to provide useful insight into examining approximation for non-Markov networks of queues and other complex queueing systems. These initial results for some special models motivate us to treat more complicated queueing models. Even though devising high-quality approximations and tight bounds no doubt will continue to be largely an art, there are a need and a real possibility for more supporting theory.

Doctoral Research

**Dissertation Research**: We propose an effective mechanism (using tight upper bound and tight lower bound) to examine approximations of queueing models. In my dissertation work, I am investigating new ways to obtain useful upper and lower bounds on performance measures given limited information. These new ways would be introduced in my thesis research “Extremal Queueing Theory”, which considers the class of extremal queueing problems given partial information of arrival and service process as well as the class of queueing approximations and bounds problems under regularity
assumptions. Most of them are long-standing open problems in the history of queueing theory. The dissertation research will play an essential role in not only addressing mathematical challenges in theoretical perspectives but also yielding important applications in the views of engineering, i.e., designing high quality queueing approximations for actual service systems.

**Details of Thesis Research:** I start my thesis study via firstly addressing the classic queueing model $GI/GI/1$ with unlimited waiting room and FCFS discipline. In [Chen and Whitt(2018)] "Extremal $GI/GI/1$ Queues Given Two Moments" (submitted to *Operations Research*), we study tight upper and lower bounds for the mean (transient and steady-state) waiting time in the $GI/GI/1$ queue given the first two moments of the interarrival-time and service-time distributions. For distributions with nonempty compact support, we show that these bounds (with one distribution given and overall) are attained at extremal distributions with support on at most three points. The proof exploits theory for the moment problem and penalty functions in addition to standard stochastic theory for the model. We derive an alternative fixed-point characterization for the steady-state mean that is promising for deriving the additional structure of the extremal distributions. We then apply relatively tractable numerical algorithms to identify the optimal distributions within the three-point distributions. For the overall upper bound with unbounded support sets, we propose a simple approximation formula and provide a numerical comparison of the approximations and bounds, showing that the new approximate bound is very accurate. However, the upper bound is attained asymptotically by two-point distributions as the upper mass point of the service-time distribution increases and the probability decreases, which lead to difficulty in some classical simulation and numerical approaches such as Monte-Carlo simulation and Numerical Inversion Transform [Abate and Whitt(1995)]. To address the impact of the rare event associated with the large service time, [Chen and Whitt(2019a)] "Algorithms for the Upper Bound Mean Waiting Time in the $GI/GI/1$ queue" (submitted to *Queueing Systems*) proposes effective numerical and simulation algorithms. A key step is to exploit the Marshall (1968) representation of the mean waiting time in terms of the idle-time distribution, which is insensitive to the rare event of the large service time. The algorithms are aided by reductions of these special queues to $D/GI/1$ and $GI/D/1$ models. One numerical algorithm exploits a negative binomial recursive formula, while another exploits a discrete-time Markov chain recursion. After completing these papers for $GI/GI/1$ queues, we look into $GI/GI/K$ models. Motivated by established results in Kingman-Lundberg bound in [Kingman(1964)], large deviation asymptotic theory in [Asmussen(2003)] and Tchebycheff Systems (see [Karlin and Studden(1966)]), we exploit a new performance analysis approach for $GI/GI/K$ queues given partial information of interarrival time and service time process. We have found the decay rates of $GI/GI/K$ can be optimized by extremal distributions which are solutions of Tchebycheff systems. Even though queueing decay rates are not always consistent with the steady-state mean waiting time (larger decay rates does not imply smaller waiting time), these extremal models can be applied to estimate the mean waiting time to generate set-valued approximations for queueing models with practical levels of accuracy. Overall, [Chen and Whitt(2019e)] "Set-valued performance of queues given partial information" (submitted to *Management Science*) contributes to a better understanding of simple queueing approximations for $GI/GI/K$ based on decay rate extremal queues. More generally, we present a case for general set-valued performance approximations, given partial information about the model. We show that with appropriate partial information it is possible to give a better idea of the range of likely values.

**Future Research**

After finishing papers on bounds and set-valued approximation for $GI/GI/1$, I plan to establish related results for other models such as $GI/GI/K$. The tight upper bound and lower bound for $GI/GI/K$ have been unknown for many years. For that purpose, the fixed point theorems would play crucial roles. The key observation is tight bounds of steady-state of queueing models are attained by a sort of fixed point solutions of interarrival-time and service-time distributions in optimization. According to some specific fixed-point representations, intractable extremal queueing problems can be formulated into tractable optimization problems. I plan to utilize this to establish the extremal theory for $GI/GI/K$ models and improve existing approximations for that. It is also the first time to apply the unusual technique to
address extremal queueing problems. I have seen great prospectives of fixed point theory for simple models GI/GI/1 queues in our recent working paper (see [Chen and Whitt(2019b)]). It is even possible to design proper optimization algorithms to determine extremal models in practice.

Moreover, I plan to consider models with dependence on arrival processes, different service disciplines, and time-varying arrival rates. In particular, it is important to understand dependence. People usually treat queues under independent assumptions, i.e., inter-arrival time is independent and service time is independent. Incorporating dependence into models will make models more close to real situations. Thus what happens when there is no independent condition is interesting. On the basis, it is important to capture the impact of dependence from the arrival process and service process for approximations and bounds.

There are opportunities for exploiting on-line learning techniques to develop real-time performance descriptions. In real service models, the data stream of customers arrival and service mechanism usually capture the main features of whole queueing systems such that current data can be used to predict future arrival and service levels (see [Whitt and Zhang(2019)]). According to this, it is interesting to predict the performance measures such that waiting time in the future via learning current distributions. Based on learning and prediction, it might be tractable to create data-driven approximations based on online information to provide good insights for decision-makers. It is also interesting to use machine learning technology to do decision marking such as scheduling customers, stuffing for service levels and so forth.

I expect that these new innovative methodologies will appear to address extremal queueing problems to design high-quality approximations and bounds during my doctoral study. Nobody prefers to waiting in lines, even though it is inevitable in real lives. With the help of understanding queueing models, many things can be managed in short and efficient ways. I would contribute to creating a simpler understanding of the frontier of queueing theory.

References


