WORST-CASE APPROXIMATIONS FOR ROBUST ANALYSIS IN MULTISERVER QUEUES AND QUEUING NETWORKS

Hyung-Khee Eun[1], Sara Shashaani[1], and Russell R. Barton[2]

[1]Fitts Dept. of Industrial and Systems Eng., North Carolina State University, Raleigh, NC, USA
[2] Smeal College of Business, Pennsylvania State University, University Park, PA, USA

ABSTRACT

This study explores strategies for robust optimization of queueing performance in the presence of input model uncertainty. Ambiguity sets for Distributionally Robust Optimization (DRO) based on Wasserstein distance is preferred for general DRO settings where the computation of performance given the distribution form is straightforward. For complex queueing systems, distributions with large Wasserstein distance (from the nominal distributions) do not necessarily provide extreme objective values. Thus, the calculation of performance extremes must be done via an inner level of maximization, making DRO a compute-intensive activity. We explore approximations for queue waiting time in a number of settings and show how they can provide low-cost guidance on extreme objective values, allowing for more rapid DRO. Approximations are provided for single- and multi-server queues and queueing networks, each illustrated with an example. We also show in settings with small number of solution alternatives that these approximations lead to robust solutions.

1 INTRODUCTION

The Wasserstein distance, sometimes referred to as transport distance, Mallow's distance, or earthmover distance, is a metric for the difference between two probability distributions (Panaretos and Zemel 2019). There has been recent interest in the use of the Wasserstein distance for robust analysis and distributionally robust optimization (DRO) of stochastic models (Kuhn et al. 2019; Blanchet et al. 2022). In many optimization settings, computing a worst-case subproblem by maximizing the performance as a function of the input distribution subject to a Wasserstein distance constraint is computationally feasible. This is not generally the case for robust simulation-optimization as was shown by Eun et al. (2024). That work demonstrated the superiority of the Kingman approximation, which can be viewed as a moment-based distance function, for identifying worst-case performance for G/G/1 and the capacitated G/G/1/k examples.

This work expands that examination to consider G/G/m queues and networks of G/G/m queues using Kingman, Wasserstein and other moment-based approximations. Several moment-based approximations for the G/G/m queue have been proposed, and a number are summarized by Whitt (1993). Note that approximations based on the first two moments can fail if the coefficient of variation for service time is very large (Gupta et al. 2010). Even though two service time distributions share the identical first two moments, if their higher moments are different, there is a "gap": a range between the largest and smallest mean waiting times. Gupta et al. (2010) show substantial over-estimation for squared coefficients of variation of service time (c_s^2) values of 19 and 99.

Our goal is to assess the reliability of these approximation methods in capturing extreme performance behaviors, thereby improving robust analytical strategies for sophisticated queueing environments.

1.1 Input Uncertainty

The motivation for this study is to develop efficient methods to support robust simulation-optimization in the presence of input model uncertainty. Simulations of real-world systems are typically driven by input

979-8-3315-8726-0/25/$31.00 ©2025 IEEE

distribution models that are fitted to finite data. The finite data result in discrepancy between fitted and true distributions, leading to errors in predicted system performance. The focus of input model uncertainty research is to characterize this error and its impact on simulation-predicted system performance.

Discrete-event simulation models focus on tracking entities accessing a sequence of resources, and queues are common in these models. For queuing models, the input uncertainty often arises from error in the estimated distributions of interarrival and service times. Let ν be the input model (distribution), and suppose a functional of the stochastic performance of the stochastic system is represented by $f(\nu)$, e.g., f can be the expectation of the stochastic simulation output, a quantile of the system performance, etc. When ν is a member of a parametric family, estimating ν is reduced to estimating its parameters θ, i.e., ν_θ. Otherwise, its estimate is the empirical distribution (Barton et al. 2022). We denote ν_0 as the *true* distribution and $\hat{\nu}$ as the fitted input distribution model. A natural point estimate of $f(\nu_0)$ is

$$\bar{f}(\hat{\nu}) = \frac{1}{r}\sum_{j=1}^{r} F_j(\hat{\nu}),$$

where $F_j(\hat{\nu}) := F(\xi_j \sim \hat{\nu})$ is the output from the j-th identically distributed simulation replication, with random variates ξ_j generated from the input model $\hat{\nu}$. Assuming that the simulation replications $j = 1, 2, \cdots, r$ are independent, the decomposition shows

$$F_j(\hat{\nu}) - f(\nu_0) = [F_j(\hat{\nu}) - f(\hat{\nu})] + [f(\hat{\nu}) - f(\nu_0)],$$

where the first term is the simulation error and the second term is the input uncertainty error. The law of total variance gives

$$\mathrm{Var}(\bar{f}(\hat{\nu})) = \frac{\mathbb{E}[\mathrm{Var}(F_j(\hat{\nu})|\hat{\nu})]}{r} + \mathrm{Var}(f(\hat{\nu})),$$

where the variance and expectation on the right hand side are with respect to the probability distribution of the random $\hat{\nu}$ (uncertainty in the fitted input model) and the inner variance in $\mathbb{E}[\mathrm{Var}(F_j(\hat{\nu})|\hat{\nu})]$ is with respect to the simulation outputs' probability distribution. Similar to the decomposition above, we see that the estimator's variance contains the variance due to the simulation error and variance due to uncertainty in the fitted input model. The mean square error

$$\mathbb{E}[(f(\hat{\nu}) - f(\nu_0))^2] = (\mathbb{E}[f(\hat{\nu})] - f(\nu_0))^2 + \mathrm{Var}(f(\hat{\nu})) = \mathrm{Bias}(f(\nu))^2 + \mathrm{Var}(f(\hat{\nu}))$$

shows that the error of the performance measure under the input model $\hat{\nu}$ includes the bias induced by input uncertainty. Bias is harder to quantify but there are recent studies that tackle that in parametric (Morgan et al. 2019) and nonparametric (Vahdat and Shashaani 2023) settings.

Thus input uncertainty can affect the reliability of decision-making processes through not only the variance but also the bias of the estimated performance measures obtained from simulation (Barton et al. 2022; Lam 2016). Therefore, accurate characterization of input uncertainty in the optimization models is essential for the validity of the simulation results.

1.2 Distributionally Robust Optimization for Queues

For discrete-event simulation-optimization, uncertainty often arises from input model specification. When such uncertainty is substantial (i.e., sensitive system performance to the input distribution), a solution that appears optimal under the assumed input model may be suboptimal (He and Song 2024). DRO seeks to find optimal decisions (represented by a decision vector, x) that are insensitive to the value of uncertain parameters in the optimization model formulation. DRO hedges against risk of selecting an inferior system due to small or corrupted input samples or other errors in the model parameters. The main application of DRO has been in data-driven settings to avoid overfitting and failing to generalize on out-of-sample data (Bertsimas and Van Parys 2022) or to take an adversarial learning approach (Blanchet et al. 2022).

Consider the stochastic system performance $F(x,\xi)$ to be a function of $x \in \mathcal{X} \subseteq \mathbb{R}^d$ and a random input $\xi \in \Xi$ that follows probability distribution ν. Suppose that the performance function of interest is its expectation:

$$f(x,\nu) := \mathbb{E}_\nu[F(x,\xi)].$$

When addressing uncertainty in the input data, it is natural to consider a set of distributions with some perturbation from the distribution fitted empirically to the data at hand, which we henceforth refer to as the *nominal* distribution, and consider the performance of the system (decision) being evaluated under all possible input distributions. The system under study is then better (say less costly) with alternative x_1 than x_2 if $f(x_1,\nu) < f(x_2,\nu)$ for all possible input distributions ν, which is to say under the worst possible input distribution. This is why DRO can be viewed as a way to hedge the risks associated with input uncertainty (Rahimian and Mehrotra 2022).

In DRO, an ambiguity set $\mathcal{P}$ is constructed that includes distributions within a certain discrepancy from the nominal distribution. The implicit assumption is that the true input distribution ν_0 resides within the ambiguity set. Suppose the goal is to minimize a performance measure. Decisions are then made to ensure robustness against the variations within this set that minimizes the maximum value of the objective function across all the distributions in the ambiguity set (Blanchet et al. 2022):

$$\inf_{x\in\mathcal{X}} \sup_{\nu\in\mathcal{P}} f(x,\nu)$$

where $f(x,\nu)$ may be an expectation, quantile, etc. of $F(x,\xi)$.

Since DRO finds a decision that minimizes the worst-case objective value among all probability measures in $\mathcal{P}$, the key to DRO is how to construct $\mathcal{P}$. In queuing systems, one of the recent advancements by Van Eekelen et al. (2022) introduces the use of Mean Absolute Deviation (MAD) as a measure of dispersion instead of variance, significantly simplifying extremal queue analysis. For a vector of independent random variables $X = (X_1,\ldots,X_n) \sim \nu \in \mathcal{P}$ and some convex function $h(\cdot)$, the optimization problem under MAD constraints is formulated as

$$\max_{\nu\in\mathcal{P}(\mu,d,a,b)} \mathbb{E}_\nu[h(X)],$$

where the ambiguity set $\mathcal{P}(\mu,d,a,b)$ with $\mu = (\mu_1,\cdots,\mu_n), d = (d_1,\cdots,d_n), a = (a_1,\cdots,a_n)$, and $b = (b_1,\cdots,b_n)$ consists of distributions defined by known mean, MAD, and bounded support such that

$$\mathcal{P}(\mu,d,a,b) = \{\nu : \mathrm{supp}(X_i) \subseteq [a_i,b_i], \mathbb{E}_\nu[X_i] = \mu_i, \mathbb{E}_\nu|X_i - \mu_i| = d_i \; \forall i = 1,\cdots,n, X_i \perp X_j, \forall i \neq j\}.$$

It is shown that the extremal distribution for each random variable X_i is a three-point distribution specified by parameters a_i, b_i, μ_i, and d_i, greatly simplifying the identification of worst-case scenarios.

Specifically, in the G/G/1 case with i.i.d. interarrival times $\{U_i\}$ and i.i.d service times $\{V_i\}$, the extremal queue problem is given as

$$\max_{\nu\in\mathcal{P}(\mu_v,d_v,a_v,b_v)\times\mathcal{P}(\mu_u,d_u,a_u,b_u)} \mathbb{E}_\nu[W],$$

where $\mathcal{P}(\mu_v,d_v,a_v,b_v)\times\mathcal{P}(\mu_u,d_u,a_u,b_u)$ is the set containing all product measures of feasible marginal distributions for V and U. The random variables V and U follow the extremal three-point distributions $\mathcal{P}(\mu_v,d_v,a_v,b_v)$ and $\mathcal{P}(\mu_u,d_u,a_u,b_u)$, respectively. Then the tight upper bounds follow from V and U.

1.3 Discrepancy measures for setting $\mathcal{P}$

The discrepancy-based approach to setting $\mathcal{P}$ includes all possible distributions that have statistical distance (discrepancy) from the nominal distribution $\hat{\nu}$ that is less than some threshold δ. The ambiguity set $\mathcal{P}_\delta$ is defined as

$$\mathcal{P}_\delta = \{\nu_1 : d(\hat{\nu},\nu_1) \leq \delta\},$$

where $d(\hat{\nu}, \nu_1)$ represents the distance or discrepancy between two distributions $\hat{\nu}$ and ν_1.

The Wasserstein distance is a well-established metric in DRO (Kuhn et al. 2019; Blanchet et al. 2022). It is an effective metric for probability measures with a finite p-th moment. Unlike ϕ-divergence, it can quantify the distance between two distributions whose supports do not overlap (Peyré and Cuturi 2019). While Wasserstein distance is commonly used in DRO, its usefulness in queueing settings is reduced by the poor relationship between Wasserstein distance and difference in system performance.

1.4 Bootstrap determination of $\mathcal{P}$

Bertsimas and Van Parys (2022) proposed an alternative approach for construction of ambiguity sets. The authors use bootstrap samples to characterize $\mathcal{P}$ for robust optimization, to estimate the fraction of bootstrap-resample cost estimates (for a chosen decision) exceeding some threshold. In their setting, evaluation of the cost function can be done inexpensively. In robust simulation-optimization, however, function evaluations are costly and the existence of a proxy (or metamodel) can be helpful so that not all bootstrap input distributions need be simulated.

1.5 Moment-based approximations for queues

Moment-based approximations for queue performance have the ability to screen for worst-case candidates from any set of distributions within a set $\mathcal{P}$ (Eun et al. 2024). The discrepancy metric between two distributions is the difference in predicted system performance. Because they are based on the first few moments, these approximations are pseudo-metrics only: two different distributions having the same first few moments will give identical estimated system performance. In the next sections we discuss several moment-based discrepancy pseudo-metrics, and examine their performance in more general settings than in (Eun et al. 2024): multiserver queues and multiserver queueing networks.

The next section presents moment-based approximations. Their performance and comparison with Wasserstein discrepancy are examined in the sections that follow.

2 WORST-CASE APPROXIMATIONS FOR SINGLE- AND MULTI-SERVER QUEUES

2.1 Kingman (G/G/1)

Eun et al. (2024) demonstrated that the Kingman distance, a pseudo moment-based metric derived from Kingman's approximation, is effective in identifying worst-case input models within sets of distributions reflecting input uncertainty. Kingman (1961) published the following approximation of waiting time $W = W(\xi)$ for a G/G/1 queue with a parametric input model ν_θ with parameters $\theta = (\rho, c_a^2, c_s^2, \mu)$:

$$\mathbb{E}_{\nu_\theta}[W]_{G/G/1} \approx \left(\frac{\rho}{1-\rho}\right)\left(\frac{c_a^2 + c_s^2}{2}\right)\left(\frac{1}{\mu}\right),$$

where ν is the joint probability distribution of the interarrival times A and service times S, i.e., $\xi = (A, S)$, which is the product of two marginal distributions given their independence, parameterized by the utilization ρ, the average service rate μ, and the coefficients of variation for interarrival time and service time, c_a and c_s, respectively. The approximation for G/G/1 is precise at high utilization, and becomes exact in the M/M/1 case. For M/G/1 queue, it reduces to the exact Pollaczek-Khinchine formula.

Beyond these cases, one can consider using this approximation for G/G/m queues despite their potential deviation from the true mean waiting time. We will explore this later and show that even in G/G/m cases, Kingman provides a desirable monotonic property. Therefore, even if not accurate for prediction of the true mean waiting time, it can still be useful for comparison and selection of the worst case.

The existing literature also entails other metrics, bounds or approximations that can be suitable for more complex queuing systems. We will review Goldberg bounds and Whitt approximations for G/G/m application in this section, and later in Section 3 for queuing networks.

2.2 Goldberg (G/G/m)

Li and Goldberg (2025) develop the general tail bounds for the steady-state queue length in G/G/m with $\frac{1}{1-\rho}$ scaling such that $\mathbb{P}\left(L \geq \frac{x}{1-\rho}\right)$ can be bounded only using finite moments of interarrival and service times. In a special case of tail bounds using the second moment of interarrival times and a slightly larger moment of service times, i.e., $2+\epsilon$ for some $\epsilon > 0$, the expected queue length is obtained as

$$\mathbb{E}_{\nu_\theta}[L]_{G/G/m} \leq \left(\frac{1}{1-\rho}\right)\left(2.1\times 10^{21}\epsilon^{-4}\mathbb{E}[(S\mu)^2]\left(\mathbb{E}[(S\mu)^2]^{1+\epsilon} + \mathbb{E}[(S\mu)^{2+\epsilon}] + 49\mathbb{E}[(A\lambda)^2]\right)\right),$$

where $\lambda = \mathbb{E}[A]^{-1}$ and $\mu = \mathbb{E}[S]^{-1}$ are the rate of interarrival and service, respectively. Using Little's law $\mathbb{E}_{\nu_\theta}[W]_{G/G/m} = \lambda \mathbb{E}_{\nu_\theta}[L]_{G/G/m}$, these queue-length bounds yield bounds on the expected waiting time.

2.3 Whitt (G/G/m)

Whitt (1993) proposed that in G/G/m queues, the expected waiting time can be approximated by

$$\mathbb{E}_{\nu_\theta}[W]_{G/G/m} \approx \frac{c_a^2 + c_s^2}{2}\mathbb{E}_{\nu_\theta}[W]_{M/M/m}. \tag{1}$$

The exact waiting time for $M/M/m$ system $\mathbb{E}_{\nu_\theta}[W]_{M/M/m}$ is computed as (Banks et al. 2004)

$$\mathbb{E}_{\nu_\theta}[W]_{M/M/m} = \left(\frac{1}{1-\rho}\right)\left(\frac{(m\rho)^m P_0}{m!(1-\rho)}\right)\left(\frac{1}{\mu}\right),$$

where $\rho = \frac{\lambda}{m\mu}$ is the utilization and P_0 is the probability of the system is empty and computed by

$$P_0 = \left\{\left[\sum_{n=0}^{m-1}\frac{(m\rho)^n}{n!}\right] + \left[(m\rho)^m\left(\frac{1}{m!}\right)\left(\frac{1}{1-\rho}\right)\right]\right\}^{-1}.$$

This approximation performs well when $c_a^2 \geq 1$, $c_s^2 \geq 1$. However, it may overestimate the waiting time when they are relatively small. To address this, Whitt (1993) also proposes a more general approximation for the G/G/m model that account for a broader range of variability conditions as follows:

$$\mathbb{E}_{\nu_\theta}[W]_{G/G/m} \approx \phi(\rho, c_a^2, c_s^2, m)\left(\frac{c_a^2 + c_s^2}{2}\right)\mathbb{E}_{\nu_\theta}[W]_{M/M/m}, \tag{2}$$

where $\phi(\rho, c_a^2, c_s^2, m)$ is an interpolating approximation function we do not list here due to space limit.

2.4 Simulated v. Approximated Waiting Times in G/G/m Queues

We run an experiment to validate whether each approximation method can estimate the mean waiting time. Since the purpose of using these approximation methods is optimization, even if the approximated mean waiting time value is not accurate, as long as it preserves a monotonic mapping, it can be used to identify the worst-case distribution. Figures 1 and 2 illustrate this property for each method in two case of G/G/1 and G/G/3. The details of the experiments, including the input parameters are listed in each figure. We use $c_s^2 = 1$ is all cases, gamma distributed interarrivals and lognormal service times. The analysis used a warm-up period of 500K and run-length of 1M.

These results particularly highlight that

- The Whitt and heavy-traffic Whitt provide fairly accurate approximations of the mean waiting time. Here we note, as long as the simulation in run to steady-state, the monotonic mapping appears near exact with only small deviations under larger c_a^2 in G/G/m.

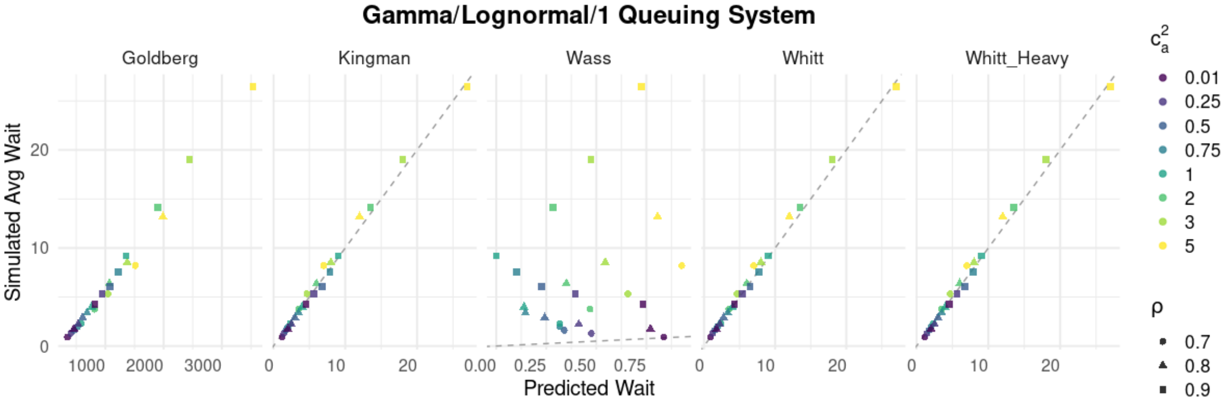

Figure 1: Comparison of simulated and analytical waiting times for a G/G/1 queue.

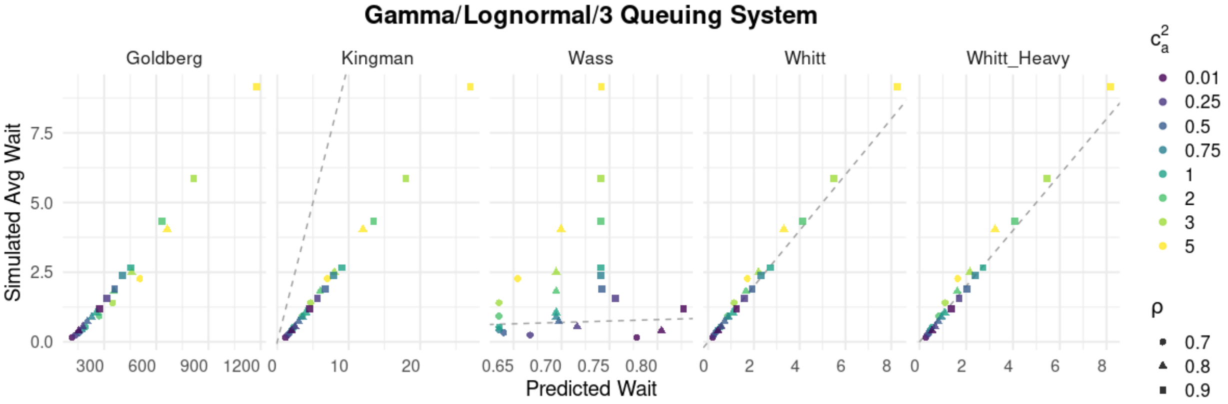

Figure 2: Comparison of simulated and analytical waiting times for a G/G/3 queue.

- Kingman loses accuracy in the G/G/m, yet maintains the monotonic property fairly well.
- Goldberg exhibits a relatively monotonic mapping albeit with significantly different scaling (still after the x-axis of Goldberg approximations is scaled down 1000 to 1) in both cases.
- Wasserstein provides neither a good approximation nor a monotonic mapping in either case.

The conclusion from these experiments is that the worst-case approximation methods can be well-poised for use in DRO by deterministically identifying the worst-case performance. One can use these methods to identify, from the set of distributions in the uncertainty set, which moment values would correspond to such worst-case performance.

3 WORST-CASE APPROXIMATIONS FOR MULTI-SERVER QUEUEING NETWORKS

We extend our analysis to the more complex setting of multi-class, multi-server open queueing networks, with a particular focus on job-shop systems. To model these systems, we employ the Queueing Network Analyzer (QNA) proposed by Whitt (1983), which models each node independently using two parameters: the mean rate and the squared coefficient of variation. QNA approximates internal flows by applying transformations—merging, splitting, and departure—to propagate these parameters throughout the network (Figure 3). While it assumes approximate independence between nodes, it incorporates flow variability to capture internal dependencies. In our experiment, each node is modeled as a G/G/m queue.

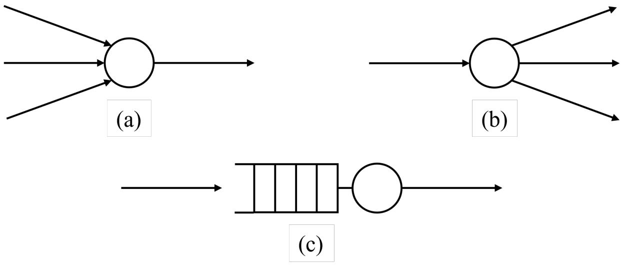

Figure 3: The three flow transformations used in QNA to propagate rate and variability parameters: (a) merging (superposition) combines parameters from multiple input streams; (b) splitting (thinning) modifies parameters as a stream is divided; and (c) departure (flow through a queue) approximates the output stream's parameters based on arrivals, service, and queueing effects. This figure is adapted from (Whitt 1993).

Suppose our queuing network consists of n nodes and k classes, $\lambda_k,\ k = 1, \cdots, r$ is the external arrival rate of class k, and n_{kj} is the j-th node visited by job class k. In QNA, if τ_{kl} is the mean service time of class k at the l-th node of its route and λ_k is the external arrivate rate of class k, the mean service time at node j is

$$\tau_j = \frac{\sum_{k=1}^{r} \sum_{l=1}^{n_k} \lambda_k \tau_{kl} \mathbf{1}\{n_{kl} = j\}}{\sum_{k=1}^{r} \sum_{l=1}^{n_k} \lambda_k \mathbf{1}\{n_{kl} = j\}}.$$

Then if m_j is the number of servers at node j, the node traffic intensity is computed as $\rho_j = \lambda_{\cdot j} \tau_j / m_j$. Here $\lambda_{\cdot j}$ is the total arrival rate to node j and computed by $\lambda_{\cdot j} = \lambda_{0j} + \sum_{i=1}^{n} \lambda_i q_{ij}$ where λ_{ij} and q_{ij} denote the flow rate between nodes and the proportion of the jobs completing service at node i that go next to node j, respectively. Furthermore, $\lambda_{0j} = \sum_{k=1}^{r} \lambda_k \mathbf{1}\{n_{k1} = j\}$ is the external arrival rate to node j.

The coefficient of variation for service time at node j is

$$c_{sj}^2 = \left(\frac{\sum_{k=1}^{r} \sum_{l=1}^{n_k} \lambda_k \tau_{kl}^2 (c_{skl}^2 + 1) \mathbf{1}\{n_{kl} = j\}}{\sum_{k=1}^{r} \sum_{l=1}^{n_k} \lambda_k \mathbf{1}\{n_{kl} = j\}} \right) \frac{1}{\tau_j^2} - 1,$$

where c_{skl}^2 is the variability parameter of the service-time distribution of class k at the l-th node of its route.

The coefficient of variation for arrivals at node j is

$$c_{aj}^2 = 1 - w_j + w_j \left[p_{0j} c_{0j}^2 + \sum_{i=1}^{n} p_{ij} q_{ij} \left[1 + (1 - \rho_i^2)(c_{ai}^2 - 1) + \rho_i^2 m_i^{-0.5} (\max\{c_{si}^2, 0.2\} - 1) + 1 - q_{ij} \right] \right],$$

where c_{0j} is the variability parameter of the external arrival to node j, w_j are appropriate weights (details skipped due to space limit), and p_{ij} are proportion of entering jobs into node j that come from node i.

From here, one can approximate the mean waiting time in the queue of each node j, denoted $\mathbb{E}_{\nu_{\theta_j}}[W_j]_{G/G/m}$, with $\theta_j = (\rho_j, c_{aj}^2, c_{sj}^2, m_j, \mu_j)$ following the heavy-traffic Whitt approximation in (1) or the general case of (2). The last step is to compute the waiting time of the queuing network following the open Jackson networks computation, i.e., $\mathbb{E}[W] = \lambda^{-1} \sum_{j=1}^{n} \lambda_{\cdot j} \mathbb{E}_{\nu_{\theta_j}}[W_j]_{G/G/m}$, with λ denoting the total external arrival rate to the network.

3.1 Job-Shop Example

We use a multi-class job-shop example from Law (2015) with adjustments detailed below. There are $n = 5$ stations and $r = 3$ routes (for three classes of jobs), with $(n_1, n_2, n_3) = (4, 3, 5)$ number of nodes on each route with each route specified as

$$n_{kj} = \begin{bmatrix} 3 & 1 & 2 & 5 & \\ 4 & 1 & 3 & & \\ 2 & 5 & 1 & 4 & 3 \end{bmatrix}.$$

The external arrival rates for each class is $(\lambda_1, \lambda_2, \lambda_3) = (1.2, 2, 0.8)$. The mean service time of class k at the j-th node of its route is

$$\tau_{kj} = \begin{bmatrix} 0.50 & 0.60 & 0.85 & 0.50 & \\ 0.80 & 0.50 & 0.75 & & \\ 0.70 & 0.25 & 0.50 & 0.70 & 1.00 \end{bmatrix}.$$

In this system, the number of servers as each station is $(m_1, m_2, m_3, m_4, m_5) = (3, 2, 4, 3, 1)$. These values lead to station utilizations $(\rho_1, \rho_2, \rho_3, \rho_4, \rho_5) = (0.707, 0.79, 0.725, 0.72, 0.8)$.

3.2 Simulated v. Approximated Waiting Times in a Job-Shop Example

Similar to Section 2.4, here we compare how each approximation method can estimate the mean waiting time or at least provide a monotonic mapping for use during optimization. In Figures 4 and 5, we maintain the set-up of the job-shop and use and uncertainty set for external interarrival times with mean values $\lambda_{\cdot}^{-1} \in \{0.25, 0.26, \dots, 0.4\}$. In Figure 4 the mean waiting times per station are shown. The most accurate method is the heavy-traffic Whitt approximation. The general Whitt approximation underestimates the mean waiting time in station 5. Kingman and Goldberg preserve monotonicity for each station but lose accuracy. And the Wasserstein does not exhibit promising approximation. Looking at the overall network's mean waiting time in Figure 5, we see similar patterns.

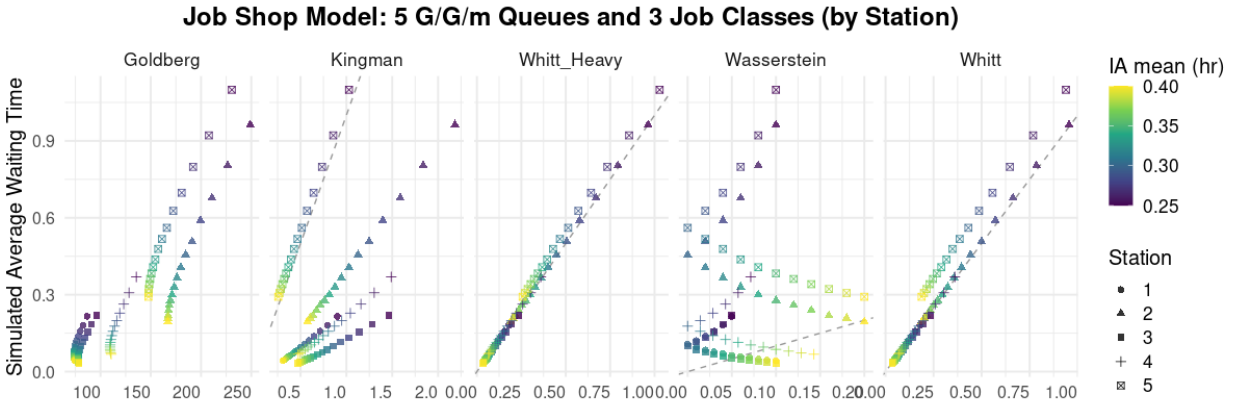

Figure 4: Simulated v. approximated mean waiting times at each station in the job-shop example.

4 COMPARING SYSTEMS UNDER INPUT UNCERTAINTY USING WORST-CASE APPROXIMATION

In this section, we experiment with the worse-case approximations and the Wasserstein distance metric to identify a robust system when comparing a small number of systems. The results here suggest that in larger scale comparisons that occur over the course of DRO, for example, these worse-case approximations will be effective in reducing computation without jeopardizing solution accuracy.

In each case of G/G/m and a job-shop example, our set up will be as follows. We assume that we have a data set of 1000 arrival times coming from a nominal distribution ν_0 unknown to us. DRO can be used here to find an optimal solution that is robust to this input uncertainty. Suppose our goal is to find

$$\min_{x \in \mathcal{X}} \max_{\nu \in \mathcal{P}} f(x, \nu) := \mathbb{E}_\nu[W(x)] + cx,$$

where ν is the interarrival distribution, x is the number of servers (m in the G/G/m case and a vector for each station in the job-shop case). We denote $W(x)$ as the waiting time random variable for a system with configuration x and $\mathbb{E}_\nu[W(x)]$ denotes its expected value when arrivals follows ν. Finally, we assume there

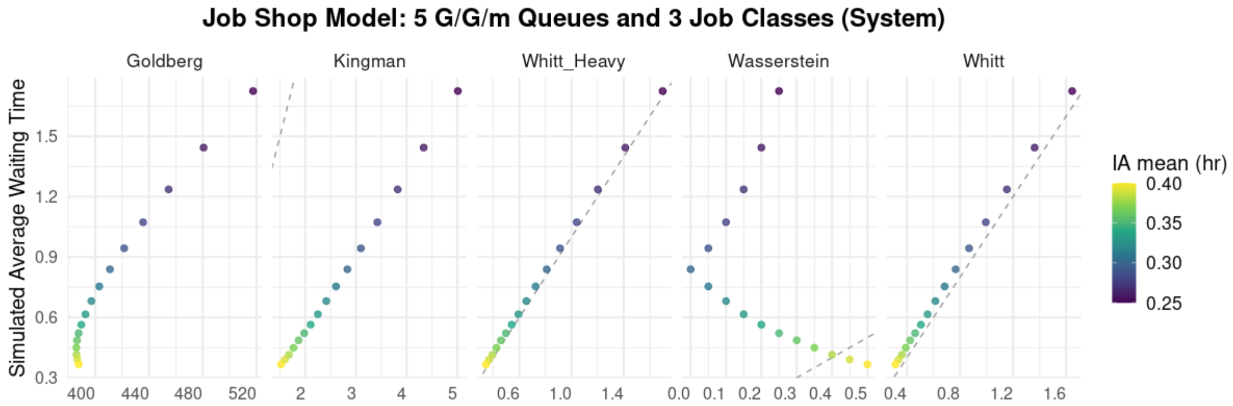

Figure 5: Simulated v. approximated mean system waiting times in the job-shop example.

is a cost of c associated with each server. We therefore seek a system that reduces this objective function when we have uncertainty about ν_0. The complete procedure is listed in Algorithm 1.

```
Algorithm 1: Finding a robust system in a finite set with worst-case approximation
Input: Dataset $\mathcal{D}$, set of systems to compare $\mathcal{X}$
Construct $b$ bootstrapped datasets $\mathcal{D}_i,\ i=1,2,\cdots,b$ and fit to each bootstrap an input model
  denoted $\nu_i,\ i=1,2,\cdots,b$. Moreover, fit the nominal distribution $\hat{\nu}$ to the original dataset $\mathcal{D}$.
for System $j \in \{1,2,\cdots,|\mathcal{X}|\}$ do
    Compute $d_{\mathrm{K}}(x_j,\nu_i), d_{\mathrm{G}}(x_j,\nu_i), d_{\mathrm{H}}(x,\nu_i), d_{\mathrm{Wh}}(x_j,\nu_i)$, and $d_{\mathrm{W}}(x_j,\nu_i)$ denoting Kingman
      approximation, Goldberg bound, heavy traffic and generalized Whitt approximations, and the
      Wasserstein distance for system $x_j$ under input model $\nu_j,\ i=1,2,\cdots,b$.
    Identify the worst-case distribution for each method as $\hat{\nu}_j^{*\mathrm{method}} = \arg\max_i d_{\mathrm{method}}(x_j,\nu_i)$ for
      method $\in$ {K, G, H, Wh, W} —assuming all $\mathcal{D}_i$ lead to stable systems.
    Simulate system $x_j$ to estimate the mean waiting times under each method's worst-case
      distribution $\mathbb{E}_{\hat{\nu}_j^{*\mathrm{K}}}[W(x_j)], \mathbb{E}_{\hat{\nu}_j^{*\mathrm{G}}}[W(x_j)], \mathbb{E}_{\hat{\nu}_j^{*\mathrm{H}}}[W(x_j)], \mathbb{E}_{\hat{\nu}_j^{*\mathrm{Wh}}}[W(x_j)]$, and $\mathbb{E}_{\hat{\nu}_j^{*\mathrm{W}}}[W(x_j)]$.
Output: $j^{*\mathrm{method}} = \arg\min_j \mathbb{E}_{\hat{\nu}_j^{*\mathrm{method}}}[W(x_j)],$ method $\in$ {K, G, H, Wh, W, N} where N stands
  for the nominal distribution $\hat{\nu}_j^{*\mathrm{N}} = \hat{\nu},\ \forall j$.
```

To generate an uncertainty set $\mathcal{P}$, we use bootstrapping of the dataset. We fit parameters for each bootstrap using maximum likelihood estimation assuming that we know the interarrival distribution is Gamma, call the fitted distributions $\hat{\nu}_i,\ i=1,2,\cdots,b$ for b bootstraps. Here it is worthy of note that for heavy-traffic systems or when data size is too small relative to the variability of the interarrival times, it is possible that the bootstrapped dataset leads to an unstable system with $\rho \geq 1$. In this paper, we remove those bootstraps. But to apply such a method in uncapacitated queues, one has to have a way to deal with these cases. For example, with objective the fraction of bootstrap evaluations exceeding some threshold as in Bertsimas and Van Parys (2022), unstable systems would be included in the fraction. In capacitated queues, the method needs no modification, as demonstrated in (Eun et al. 2024).

Once we remove the unstable bootstrapped data for each system, we use the corresponding moments of each bootstrap to deterministically compute its Kingman, Goldberg, heavy-traffic Whitt, generalized Whitt, and the Wasserstein approximations. Based on these approximations, we identify for each method the worst-case distribution. That is, we identify the bootstrap that leads to the largest approximation and use its fitted distribution—denote that $\hat{\nu}^{*\mathrm{K}}, \hat{\nu}^{*\mathrm{G}}, \hat{\nu}^{*\mathrm{H}}, \hat{\nu}^{*\mathrm{Wh}}$, and $\hat{\nu}^{*\mathrm{W}}$ for each method—to a run a steady-state simulation and estimate the objective function under that input distribution. We also compare the performance of the nominal distribution $\hat{\nu}$ for arrivals that fits a parameter to the empirical data. The latter case will demonstrate the impact of input uncertainty on misguided selection.

For the steady-state analysis, we use a warm-up value of 500K and run-length of 1M. This process is repeated for each $x \in \mathcal{X}$, which in this case consists four distinct values. We then compare the estimated worst-case objectives with the one under the true parameters to assess whether each method can correctly find the true best system.

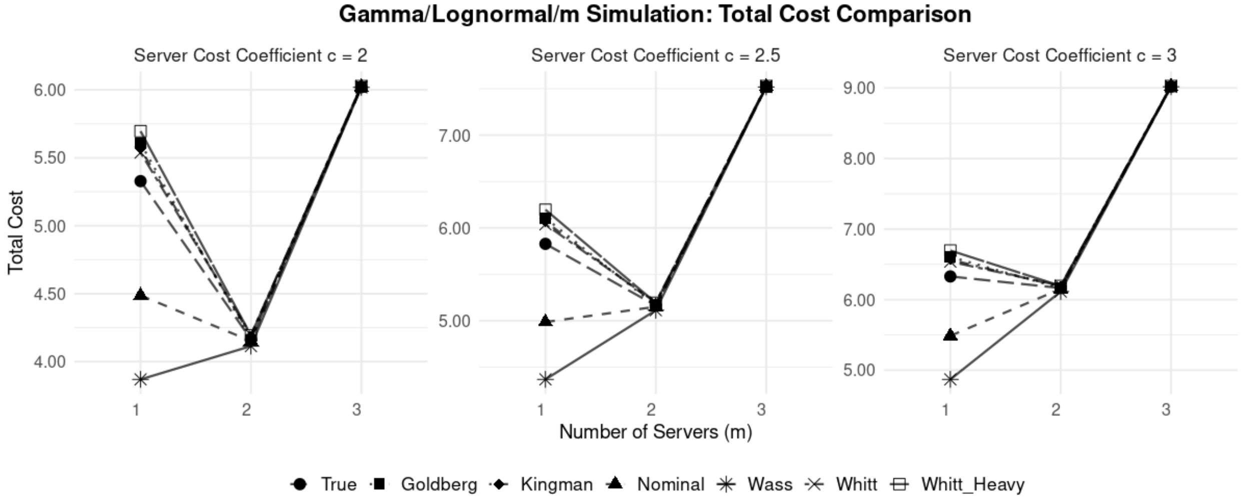

Figure 6: Total cost includes the mean waiting time and fixed cost of $\$c$ per server. The approximation methods identify the true best system, yet the nominal and worst Wasserstein distributions are prone to failing in doing so.

Figure 6 presents results for a G/G/m system in which interarrival times follow a gamma distribution and service times follow a lognormal distribution. The distributions are parameterized to yield a mean interarrival time of $\lambda^{-1} = 1.3$ with $c_a^2 = 1$, and a mean service time of $\mu^{-1} = 1$ with $c_s^2 = 1$. The server costs are set to $c \in \{2, 2.5, 3\}$ (in dollars). The results are based on $b = 100$ bootstrap datasets, each of size 1000. If the true input distribution was known, system 2 with $m = 2$ would be the best system with the lowest objective value. And important observation here is the risk of using the nominal distribution (and ignoring the input uncertainty). It is evident that at least in two of the three cost scenarios (when the server cost is \$2.5 or \$3), the nominal distribution can lead to sub-optimal selection of system 1. Note the significant underestimation of cost for system 1 using the nominal distribution is all c scenarios. This shows the seriousness of input uncertainty and its ability to misguide the comparison and search (in an optimization setting).

To hedge against the risk of input uncertainty, we can look for the worst-case performance using the approximation methods in a bootstrap-based uncertainty set (as detailed in Algorithm 1). Remarkably, all approximation methods except the Wasserstein correctly select the true best system. For the inferior system 1, all worst-case approximation methods except Wasserstein overestimate the total cost by less than 1 unit. In contrast, Wasserstein approximation hugely underestimates the total cost of system 1 and incorrectly identifies it as the best system.

Figure 7 presents a similar experiment for the job-shop example described in Section 3.1 with server unit cost of $c \in \{0.2, 0.3, 0.4\}$ (in dollars) and evaluating four systems with varying number of servers per station (as shown on the x-axis). Under the true distribution, the fourth system yields the minimum objective value. We observe that the nominal distribution notably underestimates the performance of each system in all cost scenarios. Importantly this effect of input uncertainty puts the correct selection of the best system at peril when the server cost is \$0.3 or \$0.4, highlighting a critical risk it poses in decision-

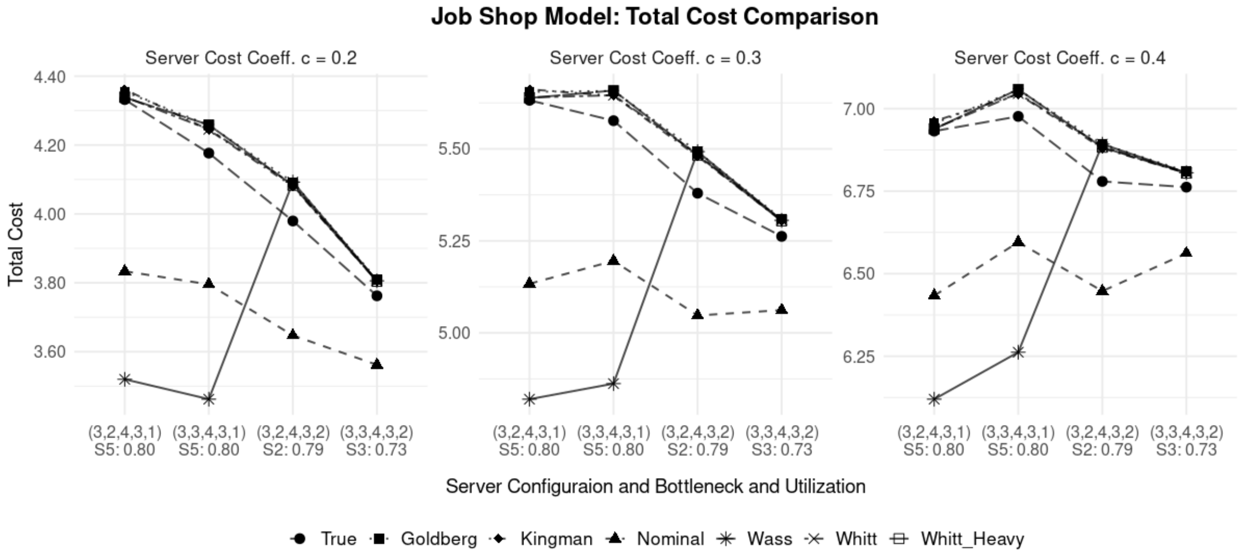

Figure 7: Total cost in the job-shop network for different number of servers per each of the 5 stations. Under each system, the utilization of the bottleneck station is printed. In all cost scenarios, system 4 appears optimal under the true interarrival distribution. Yet, the failure of nominal distribution (in the latter two scenarios) and the worst Wasserstein approximation in identifying this system is evident.

making and optimization. As in the previous case, the worst Wasserstein approximation fails in identifying the best system, whereas the other methods succeed. We also note that across the three scenarios, the worst Wasserstein approximation does lead to the same worst-case input distribution as other methods for systems 3 and 4 (in which the bottleneck station's utilization is less then 80%—suggesting that Wasserstein approximation may work well in steadier and light-traffic systems). This observation is consistent with the G/G/m experiment in Figure 6 as well. Not surprisingly, in all non-Wasserstein approximations, the corresponding worst-case distributions are identical. We expect this due to the monotonic property of these methods.

5 CONCLUSION

In this paper, we propose use of moment-based approximation methods, including the Kingman approximation, Goldberg bound, heavy-traffic Whitt approximation and the generalized Whitt approximation to identify the worst-case distribution of a G/G/m queueing system as well as a multi-server multi-class open queuing network. Our experiments reveal a desirable property, i.e., monotonicity, in these approximation methods. In other words, these approximation methods can be maximized deterministically to suggest input distribution parameters that result in worst-case mean waiting time in these systems. A direct impact of the proposed method is in handling of input uncertainty via DRO, where we seek to compare queuing systems and optimization of queuing systems to make robust decisions. We show the widely-used Wasserstein metric does not provide such properties and cannot be used as effectively for solving the inner maximization of the DRO. Future work will be dedicated to establishing convergent optimization algorithms that use these approximations to solve the underlying robust optimization with minimal simulation run.

ACKNOWLEDGMENTS

The authors acknowledge support by the National Science Foundation under grant CMMI-2226347 and Office of Naval Research under grant N000142412398.

REFERENCES

Banks, J., J. S. Carson II, B. L. Nelson, and D. M. Nicol. 2004. *Discrete-Event System Simulation*. 4th ed. Upper Saddle River, New Jersey: Prentice Hall.

Barton, R. R., H. Lam, and E. Song. 2022. "Input Uncertainty in Stochastic Simulation". In *The Palgrave Handbook of Operations Research*, edited by S. Salhi and J. Boylan, 573–620. Cham: Palgrave Macmillan.

Bertsimas, D., and B. Van Parys. 2022. "Bootstrap Robust Prescriptive Analytics". *Mathematical Programming* 195(1):39–78.

Blanchet, J., K. Murthy, and F. Zhang. 2022. "Optimal Transport-based Distributionally Robust Optimization: Structural Properties and Iterative Schemes". *Mathematics of Operations Research* 47(2):1500–1529.

Eun, H. K., S. Shashaani, and R. R. Barton. 2024. "Comparative Analysis of Distance Metrics for Distributionally Robust Optimization in Queuing Systems: Wasserstein vs. Kingman". In *2024 Winter Simulation Conference (WSC)*, 3368–3379 https://doi.org/10.1109/WSC63780.2024.10838888.

Gupta, V., M. Harchol-Balter, J. G. Dai, and B. Zwart. 2010. "On the Inapproximability of M/G/K: Why Two Moments of Job Size Distribution are Not Enough". *Queueing Systems* 64(1):5–48.

He, L., and E. Song. 2024. "Introductory Tutorial: Simulation Optimization Under Input Uncertainty". In *2024 Winter Simulation Conference (WSC)*, 1338–1352 https://doi.org/10.1109/WSC63780.2024.10838862.

Kingman, J. F. C. 1961. "The Single Server Queue in Heavy Traffic". *Mathematical Proceedings of the Cambridge Philosophical Society* 57(4):902–904.

Kuhn, D., P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. 2019. "Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning". In *Operations Research & Management Science in the Age of Analytics*, edited by S. Netessine, D. Shier, and H. J. Greenberg, 130–166. INFORMS TutORials in Operations Research.

Lam, H. 2016. "Advanced Tutorial: Input Uncertainty and Robust Analysis in Stochastic Simulation". In *2016 Winter Simulation Conference (WSC)*, 178–192 https://doi.org/10.1109/WSC.2016.7822088.

Law, A. M. 2015. *Simulation Modeling and Analysis*. 5th ed. New York: McGraw-Hill Education.

Li, Y., and D. A. Goldberg. 2025. "Simple and Explicit Bounds for Multiserver Queues with 1/1-ρ Scaling". *Mathematics of Operations Research* 50(2):813–837.

Morgan, L. E., B. L. Nelson, A. C. Titman, and D. J. Worthington. 2019. "Detecting Bias due to Input Modelling in Computer Simulation". *European Journal of Operational Research* 279(3):869–881.

Panaretos, V. M., and Y. Zemel. 2019. "Statistical Aspects of Wasserstein Distances". *Annual Review of Statistics and Its Application* 6(1):405–431.

Peyré, G., and M. Cuturi. 2019. "Computational Optimal Transport: With Applications to Data Science". *Foundations and Trends® in Machine Learning* 11(5-6):355–607.

Rahimian, H., and S. Mehrotra. 2022. "Frameworks and Results in Distributionally Robust Optimization". *Open Journal of Mathematical Optimization* 3:1–85.

Vahdat, K., and S. Shashaani. 2023. "Robust Output Analysis with Monte-Carlo Methodology". *arXiv preprint arXiv:2207.13612*.

Van Eekelen, W., D. Den Hertog, and J. S. Van Leeuwaarden. 2022. "MAD Dispersion Measure Makes Extremal Queue Analysis Simple". *INFORMS Journal on Computing* 34(3):1681–1692.

Whitt, W. 1983. "The Queueing Network Analyzer". *The Bell System Technical Journal* 62(9):2779–2815.

Whitt, W. 1993. "Approximations for the GI/G/m Queue". *Production and Operations Management* 2(2):114–161.

AUTHOR BIOGRAPHIES

HYUNG KHEE EUN is a third year Ph.D. student in the Edward P. Fitts Department of Industrial and System Engineering with research interests in simulation and stochastic optimization. His research includes roust optimization and bias correction. His email address is heun@ncsu.edu.

SARA SHASHAANI is an Associate Professor and Bowman Faculty Scholar in the Edward P. Fitts Department of Industrial and System Engineering at North Carolina State University. Her research interests are simulation optimization and probabilistic data-driven models. She is a co-creator of SimOpt library. Her email address is sshasha2@ncsu.edu and her homepage is https://shashaani.wordpress.ncsu.edu/.

RUSSELL R. BARTON is Distinguished Professor of Supply Chain and Information Systems in the Smeal College of Business and Professor of Industrial Engineering at the Pennsylvania State University. His research interests include applications of statistical and simulation methods to system design and to product design, manufacturing and delivery. His email address is rbarton@psu.edu and his homepage is https://sites.psu.edu/russellbarton/.