Shortest-Job-First Scheduling in Many-Server Queues with Impatient Customers and Noisy Service-Time Estimates

Jing Dong
Columbia University, 3022 Broadway, New York, NY 10027, jing.dong@gsb.columbia.edu
Rouba Ibrahim
University College London, 1 Canada Square, London E14 5AB, rouba.ibrahim@ucl.ac.uk

Size-based scheduling has been extensively studied, yet almost exclusively in single-server queues with infinitely patient jobs and perfectly known service times. Much less is known about its performance in many-server queues, particularly under noisy service-time information. In this paper, we derive theoretical results that quantify the performance of the non-preemptive Shortest-Job-First (SJF) policy in many-server queues with abandonment and noisy service-time estimates. In particular, we consider the $M/GI/s+GI$ queue and service-time estimates that have either a discrete distribution with finite support, or a continuous distribution. In the discrete case, we prove that the SJF policy asymptotically maximizes the throughput in the system, among all non-preemptive scheduling disciplines that use that noisy service-time information. In the continuous case, we prove that a discretized version of SJF, which classifies customers into a finite number of priority classes based on their noisy service time predictions, asymptotically maximizes the throughput in the system as well, when the number of priority classes increases without bound. By taking this limit, performance under the SJF policy is, asymptotically, indistinguishable from performance under a carefully designed two-class priority rule, where customers with short predicted service times (below a threshold) are served without wait, while customers with long predicted service time (above the threshold) eventually abandon without service.

Key words: SJF, impatience, noisy service times, many-server queues.

1. Introduction

In this paper, we study how to schedule customers based on noisy service-time estimates. In particular, we focus on analyzing system performance under the non-preemptive noisy **s**hortest-**j**ob-**f**irst (SJF) policy, which prioritizes the customer with the smallest service-time *estimate*, in many-server

queues with customer abandonment. Size-based scheduling, e.g., SJF, has been extensively studied, yet almost exclusively in single-server queues and assuming perfectly known service times. In contrast, the performance analysis of SJF with estimation errors remains largely an open problem, even in the relatively simple single-server queue setting (Down 2019, Scully et al. 2021). We are, to the best of our knowledge, the first to derive theoretical results quantifying the performance of SJF with noisy service-time information in a many-server queue setting with impatient customers. Throughout this paper, we use SJF to denote the noisy SJF policy which relies on noisy point estimates of the service times, rather than assuming full knowledge of the service times.

1.1. Size-Based Scheduling in Service Systems

Our focus in this paper is on service systems. In service systems, such as healthcare facilities or contact centers, service requests are typically processed by multiple agents working in parallel. Furthermore, queued customers who are waiting to be served by an available agent do not wait indefinitely for service. In other words, customers have finite patience times, and they abandon the queue if they have to wait for too long. Most importantly, the times needed to process customer service requests are typically not perfectly known before entry to service. For example, in the context of a call center, predictions of future call durations are notoriously imprecise (Ibrahim et al. 2016). Lastly, we assume that customers in service cannot be preempted, as per common practice.

For a realistic queueing-model representation of a service system, we consider here a many-server queueing system with customer abandonment where only noisy information is available, ex-ante, about the service times, e.g., in the form of service-time predictions. It seems natural to exploit service-time information, if it is available, when making customer scheduling decisions. For example, it is well known that, with perfectly known service times, scheduling the shortest job first is typically effective in reducing overall system congestion. However, there are relatively few papers that study system performance when scheduling is based on noisy service-time estimates, even in single-server queues without abandonment; see the literature review in Section 2. Since size-based scheduling holds great promise, there is a need to investigate how much of its superior performance extends to more general queueing models, e.g., models that are practically relevant to the design of service systems. That is the aim of this paper.

1.2. Size-Based Scheduling: Why is it Hard to Analyze?

In broad terms, analyzing size-based scheduling policies is complicated because one must keep track of the service time of each customer in the queue, leading to a curse of dimensionality issue. While asymptotic analysis, e.g., under heavy traffic, generally allows for simpler descriptions of the system, it involves studying suitably scaled measure-valued system-state descriptors, which is technically challenging (Banerjee et al. 2020). In moving from a single server to multiple servers, the main technical challenge in analyzing the Shortest-Remaining-Processing-Time (SRPT) scheduling policy and SJF (which is akin to a non-preemptive version of SRPT) is that many-server queues are not order-conserving so classical arguments in the single-server setting, e.g., the tagged job method, do not readily extend; see Section 4.2 in Grosof et al. (2018). Allowing for customer abandonment complicates the analysis even further. Indeed, scheduling decisions in systems with abandonment is well-known to be difficult, because the optimal scheduling policy can be state-dependent and varies for different patience-time distributions (Puha and Ward 2019).

Dong and Ibrahim (2021) studies the asymptotic performance of SRPT in the $M/GI/s+GI$ queue, in steady state, with perfect service-time information. To overcome the analytical challenges, that paper relies on a coupling between the SRPT queue and an analytically tractable loss queue. With noisy service-time information, as we consider here, such coupling proofs do not readily extend because it is difficult to ensure a strict ordering of the sample paths across coupled systems. For example, assume that we couple the system under noisy SJF with another system under a different scheduling policy. As the scheduling decisions in the SJF system are based on noisy service-time information, mistakes can happen. In particular, it is possible that, at any given decision epoch, the job scheduled to be served, i.e., the one with the smallest predicted service time, does not have the shortest service time among all the jobs waiting in the queue. Thus, when coupling the noisy SJF system with another, it is not possible to ensure that the sample paths of the corresponding stochastic processes are ordered properly. This sets our current setting apart from the one in Dong and Ibrahim (2021), where there was no noise in the service-time predictions. As a result, the setting with noisy service times brings about unique challenges that necessitate the usage of new analytical tools, as we do in this paper.

1.3. This Paper's Contributions

Our theoretical results take steps towards filling some important gaps in the literature. First, the performance analysis of the SRPT or SJF policies with estimation errors remains largely an open problem (Down 2019, Scully et al. 2021). Second, with the exceptions of Grosof et al. (2018), which considers infinitely patient customers, and Dong and Ibrahim (2021), which allows for finite patience times, there are no known theoretical results about the performance of the SRPT or SJF policies in many-server queues, even when service times are perfectly known in advance. We assume that we have a noisy point estimate of the service time of each customer and that a mild assumption on the monotonicity of the mean actual service time conditional on its prediction holds. Noisy service time estimates may have either a discrete or a continuous distribution, depending on how they were developed, e.g., based on either a classification algorithm or a regression model. We focus here on a many-server asymptotic mode of analysis and the overloaded regime. The overloaded regime is appropriate because there is a nontrivial queue to be managed while, under a moderate or light load, queueing times are negligible in large systems with abandonment (Garnett et al. 2002). In the many-server overloaded regime, a non-negligible proportion of customers abandon the queue. Thus, carefully designing the scheduling policy to optimize the throughput is crucial in this setting.

Here are our main theoretical contributions. When the noisy service time has a discrete distribution with finite support, we show that the SJF policy asymptotically maximizes the throughput, among all non-preemptive policies which rely on that noisy service-time information. To wit, this includes blind scheduling policies, such as first-come-first-served and last-come-first-served, which do not exploit the service-time information at all. We quantify the value of that throughput by leveraging results from the extant literature on multi-class priority systems, in particular, Atar et al. (2014). When the noisy service time has a continuous distribution, we demonstrate that a two-class non-preemptive priority policy where customers with small predicted service times (below a carefully designed threshold) are prioritized over customers with large predicted service times

(above the threshold), asymptotically maximizes the throughput. We also prove that a discretized version of SJF, which classifies customers into a finite number of priority classes based on their noisy service-time predictions, asymptotically maximizes the throughput as well, as the number of priority classes increases without bound. Since the SJF policy arises as the limit of the discretized SJF policy, as the number of priority classes increases to infinity, the throughput is asymptotically maximal under SJF too. We also characterize other asymptotic performance measures. In particular, in the limit, under SJF, customers with short predicted service times (below a threshold) are served immediately without waiting while customers with long predicted service times (above the threshold) wait until they abandon the queue.

Finally, we establish a monotonicity property on the asymptotic system throughput under a bivariate stochastic-order relation on the pair of actual and predicted service times. As a corollary, we show that, in the practically relevant case of lognormally distributed service times, the higher the correlation between the actual and predicted service times, the higher the asymptotic throughput under SJF.

The rest of this paper is organized as follows. In Section 2, we review the literature. In Section 3, we describe our model. In Section 4, we state and prove our main results by drawing the connection between the SJF queueing system and the multi-class priority queue. In Section 5, we relate the accuracy of the service-time prediction to the throughput. In Section 6, we describe a data-driven way to derive the threshold in the two-class approximation to SJF with continuous predicted service times. In Section 7, we conduct numerical experiments to further substantiate our understanding of SJF. In Section 8, we conclude. We relegate the proofs of standard results and some numerical results to the appendix.

2. Literature Review

Size-based scheduling policies, such as non-preemptive SJF or preemptive SRPT, have received much attention in the literature due to their attractive properties, e.g., minimizing the mean response time in the system. Most of the current developments focus on single-server queues with infinitely patient customers.

For example, Schrage (1968) and Schrage and Miller (1966) demonstrate optimality properties of SRPT in the $M/G/1$ system. There is a notable stream of works that studies SRPT under heavy traffic (Down et al. (2009), Gromoll et al. (2011), Puha et al. (2015)). Scully et al. (2018) develops a unified framework to analyze several age-based scheduling policies. This framework enables the study of scheduling policies with non-monotone age-based index rules, which specify the order in which jobs are scheduled. Size-based scheduling with noisy service time information is rarely considered in the literature, even in the single-server setting. Notable exceptions are Wierman and Nuyens (2008), Mitzenmacher (2021), Scully et al. (2021), and Scully et al. (2020). All of these papers study single-server queues without customer abandonment, focus on the objective of minimizing mean response time or mean holding cost, and consider a specific form of estimation errors, e.g., bounded error sizes. For example, Wierman and Nuyens (2008) consider additive or multiplicative errors, where the estimate of a job of size s is within $[s-\sigma, s+\sigma]$ or $[s(1-\sigma), s(1-\sigma)]$. Scully et al. (2021) considers more general multiplicative errors, where the estimate of a job of size s is within $[\beta s, \alpha s]$ for $\alpha \geq \beta > 0$. Mitzenmacher (2021) considers a single classifier that predicts whether the job is above a given threshold. We study many-server queues with abandonment, focus on the objective of maximizing system throughput, and consider fairly general estimation errors through a mild assumption on the conditional mean of the actual service time. Our estimation error assumption covers the important case where job-size predictions are based on regression or classification models. We also study the effect of prediction accuracy. We measure the accuracy of the prediction by the positive quadrant dependence order, which is related to the correlation between actual and predicted service times when service times are lognormal.

With multiple servers, we know that SRPT is not necessarily optimal; e.g., see Leonardi and Raz (2007). An important reference is Grosof et al. (2018), which studies the performance of the SRPT policy in a many-server queueing system with Poisson arrivals, general service times, and no abandonment, i.e., the $M/G/k$ system. In this setting, the SRPT policy is shown to achieve an asymptotically optimal mean sojourn time in the conventional heavy-traffic regime. Scully et al.

(2020) extends this result and demonstrates that the Gittins policy is optimal, in heavy traffic, in the $M/G/k$ system. The Gittins policy adapts to any amount of available information about service times (including, e.g., having a noisy point estimate of the service time), and is equivalent to SRPT when the service times are fully known, but the index can be hard to compute in practice when the available service time information is more general. To the best of our knowledge, Dong and Ibrahim (2021) is the first to derive theoretical results about the performance of SRPT in many-server queues with abandonment. The results of that paper are limited in two main dimensions: (1) they are based on the unrealistic assumption that service times are perfectly known, and (2) they allow for preemptions that may be practically infeasible (this simplifies the analysis and enables the coupling proof in that paper). Here, we go beyond those two limiting assumptions. We also use a completely different proof technique based on fluid limits for many-server systems with a finite number of priority classes. In particular, our proof builds on Atar et al. (2014) who derive fluid limits for overloaded multi-class many-server queues with impatient customers when the number of priority classes is finite. However, scheduling decisions in that paper are not based on the noisy individual service-time information.

Overall, given those gaps in the literature, there is a need to investigate the extent to which the superior performance of SJF continues to hold in many-server queues where patience times are finite, and where service times may or may not be perfectly known.

3. Modelling Framework

In this section, we set the stage for our subsequent theoretical development by describing our modeling framework and defining our many-server asymptotic mode of analysis.

3.1. Model Description

We consider the $M/GI/s+GI$ queueing system in steady state, i.e., we assume customers arrive to the system according to a Poisson process, their service times are independent and identically distributed (i.i.d.) continuous random variables with a cumulative distribution function (cdf) G, a probability density function (pdf) g, and mean $1/\mu$, and times to abandon (patience times) are

i.i.d. continuous random variables with a cdf F, a pdf f, and mean $1/\theta$. Each arriving customer is also associated with a predicted service time. We assume the pair of actual and predicted service times are i.i.d. across customers. Let $\zeta_G(x) \coloneqq g(x)/(1-G(x))$ and $\zeta_F(x) \coloneqq f(x)/(1-F(x))$ denote the hazard rate functions for the service time and patience time, respectively. We define $M_G \coloneqq \sup\{x : G(x) < 1\}$ and $M_F \coloneqq \sup\{x : F(x) < 1\}$, and note that we allow M_G and M_F to be infinity. Similar to Assumption 4.2 in Atar et al. (2014), whose fluid-limit results for finite-priority systems we exploit in our analysis, we make the following assumptions on the system primitives:

ASSUMPTION 1. *For the patience-time distribution, we have i)* $\sup_{0\le x<M_F} \zeta_F(x) < \infty$*, i.e.,* ζ_F *is bounded and ii)* $f(x) > 0$ *for* $x \in [0, M_F)$*. For the service-time distribution, we have* $\zeta_G(x)$ *is either bounded or lower semi-continuous on* $(0, M_G)$*.*

We consider the non-preemptive SJF policy. Specifically, a customer who, upon arrival, finds an empty server goes to service immediately. If all servers are busy at the arrival epoch, then the new arriving customer must join the queue. When a server becomes available, the customer in the queue with the smallest predicted service time (point estimate) will be the next to begin service. Service preemptions are not allowed. Customers have finite patience times, generated at the arrival epoch of the customer. We assume that the arrival, service, and abandonment processes are mutually independent. (Note, however, that the actual and predicted service times are correlated.) We define the traffic intensity $\rho \coloneqq \lambda/s\mu$. Because abandonment is allowed in the system, it is not necessary to assume $\rho < 1$ for the system to be stable.

3.2. Many-Server Overloaded Regime

We consider a sequence of $M/GI/s_\lambda + GI$ queues, indexed by the arrival rate λ. We fix the traffic intensity in system λ to $\rho_\lambda = \lambda/(s_\lambda\mu) \equiv \rho > 1$. We hold the service-time and patience-time distributions fixed, i.e., independent of λ, and let λ and s_λ increase without bound.

Let S denote a generic service time and $\hat{S}$ denote the corresponding generic service-time prediction. Define

$$\mu(y) \coloneqq \mathbb{E}[S|\hat{S} = y], \tag{1}$$

to be the conditional mean actual service time conditional on the corresponding $\hat{S} = y$.

We consider two possible forms of predicted service times. First, we assume that $\hat{S}$ is a discrete random variable with finite support, which we denote as $\{a_1, a_2, \ldots, a_K\}$, with $a_1 < a_1 < \cdots < a_K$. This form of predicted service times arises, for example, from a classifier such as a regression tree or a support vector machine. Let h denote the probability mass function (pmf) of $\hat{S}$, and H denote its cdf. Second, we assume that $\hat{S}$ is a continuous random variable. With a mild abuse of notation, let h and H denote the pdf and cdf of $\hat{S}$, respectively. Let $M_H \coloneqq \sup\{x : H(x) < 1\}$. We assume that $h(x) > 0$ for $x \in (0, M_H)$ and $\sup_{0 \le x < M_H} h(x) < \infty$. This form of predicted service time arises, for example, from a Normal or lognormal linear regression.

When $\hat{S}$ is a discrete random variable, we let κ denote the threshold satisfying

$$\kappa = \min\left\{ k \ge 1 : \sum_{i=1}^{k} h(a_i)\mu(a_i) \ge s_\lambda/\lambda \right\}. \tag{2}$$

In other words, we choose κ such that the total workload of customers with predicted service times smaller than or equal to a_κ just exceeds the service capacity of the system. We further define

$$\phi_\kappa = \frac{s_\lambda - \lambda \sum_{i=1}^{\kappa-1} h(a_i)\mu(a_i)}{\lambda h(a_\kappa)\mu(a_\kappa)} = \frac{s_\lambda/\lambda - \sum_{i=1}^{\kappa-1} h(a_i)\mu(a_i)}{h(a_\kappa)\mu(a_\kappa)}, \tag{3}$$

which can be interpreted as the fraction of the workload of customers with predicted service time a_κ which, combined with the workloads of customers with predicted service times strictly less than a_κ, equals the system's service capacity. In particular, note that $\lambda h(a_i)\mu(a_i)$ is the workload of customers with predicted service time a_i, and $\sum_{i=1}^{\kappa-1} \lambda h(a_i)\mu(a_i) + \lambda\phi_\kappa h(a_\kappa)\mu(a_\kappa) = s_\lambda$. By our scaling, κ and ϕ_κ do not depend on λ.

When $\hat{S}$ is a continuous random variable, define the threshold, τ, satisfying

$$\lambda \cdot \mathbb{P}(\hat{S} \le \tau) \cdot \mathbb{E}[S|\hat{S} \le \tau] = \lambda \mathbb{E}[S\mathbf{1}(\hat{S} \le \tau)] = \lambda \int_0^\tau \mu(y)h(y)dy = s_\lambda, \tag{4}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function. That is, we choose τ such that the total workload of customers with predicted service times smaller than or equal to τ matches the system's service capacity. The thresholds κ in (2) and τ in (4) will be useful in Section 4, where we derive theoretical results for the performance of SJF with discrete or continuous predicted service times.

3.3. A Mild Assumption on the Predicted Service Times

Throughout this paper, we make the following assumption on $\mu(y)$ defined in (1) .

ASSUMPTION 2. *The conditional mean actual service time $\mu(y)$ is increasing in y.*

To see when $\mu(y)$ is increasing in y, we recall the concept of positive regression dependence of the actual service time, S, on its predicted value, $\hat{S}$ (Lehmann 1966), which is equivalently known as S is stochastically increasing in $\hat{S}$ (Barlow and Proschan 1975).

DEFINITION 1. S is said to be stochastically increasing in $\hat{S}$ or, equivalently, S is positive regression dependent on $\hat{S}$, if $\mathbb{P}(S > x|\hat{S} = y)$ is increasing in y for all x, where $x, y \geq 0$.

It immediately follows from this definition that $\mu(y)$ is increasing in y under positive regression dependence (this is a sufficient but not necessary condition). Positive regression dependence is a mild assumption on the pair $(S, \hat{S})$. It is satisfied when service-time predictions are obtained from on a regression model, which is quite common in practice.

For example, if $S = \hat{S} + \epsilon$, where the prediction $\hat{S}$ and the noise term ϵ are independent, then S is stochastically increasing in $\hat{S}$. To show this, we note that

$$\mathbb{P}(S > x|\hat{S} = y) = \mathbb{P}(\hat{S} + \epsilon > x|\hat{S} = y) = \mathbb{P}(\epsilon > x - y|\hat{S} = y) = \mathbb{P}(\epsilon > x - y),$$

which is clearly increasing in y for a fixed value of x.

For discrete predicted service times with a finite support, $\hat{S} \in \{a_1, a_2, \ldots, a_K\}$, Assumption 2 implies that $\mu(a_i) < \mu(a_j)$ for $a_i < a_j$. In regression trees, with a large enough sample in each leaf, we have $a_i \approx \mu(a_i)$ with a high probability, so that the order assumed is likely to hold.

4. Many-Server Limits under SJF Scheduling

In this section, we study the steady-state performance of the overloaded $M/GI/s+GI$ queue under the non-preemptive SJF policy with noisy service-time estimates, in the many-server asymptotic limit. Specifically, we consider the asymptotic regime defined in Section 3.2.

Let $X^\lambda(t)$ and $Q^\lambda(t)$ denote the number of customers in the system and in the queue at time t, in the λ-th system, respectively. We also denote $A^\lambda(t)$ as the number of arrivals by time t, $D^\lambda(t)$ as

the number of departures from service by time t, and $R^\lambda(t)$ as the number of departures through abandoment by time t. Note that $D^\lambda(t)$ and $R^\lambda(t)$ depend on the scheduling policy.

Our theoretical analysis is based on relating the SJF policy to a finite-class priority policy. We first review results about the fluid limits of many-server queues with abandonment under a finite-class priority policy, which were established in Atar et al. (2014).

4.1. Fluid Limits for Many-Server Queues with Priorities

In this section, we divide customers into a finite number of priority classes, based on their predicted service times. In particular, let $0 = r_0 < r_1 < r_2 < \cdots < r_{m-1} < r_m = \infty$ denote the class division thresholds. A customer whose predicted service time $\hat{S} \in (r_{i-1}, r_i]$ is classified into the i-th priority class for $1 \le i \le m$. We have m priority classes where class i enjoys a nonpreemptive higher priority over class j if $i < j$. We denote this priority rule by π^p. Due to the thinning property of the Poisson process, class i customers arrive to the system according to a Poisson process with rate $\lambda\mathbb{P}(r_{i-1} < \hat{S} \le r_i) = \lambda(H(r_i) - H(r_{i-1}))$, independently of the other classes. They have i.i.d. service times distributed as $[S|r_{i-1} < \hat{S} < r_i]$ and i.i.d. patience times with cdf F.

Atar et al. (2014) introduce a deterministic fluid model that governs the law of large number behavior of the sequence of stochastic systems under the finite-class priority rule π^p as $\lambda \to \infty$. The fluid model is characterized by a system of equations, (13) – (17), outlined in Appendix A. It is further established that under Assumption 1, for a given fixed initial value, the fluid-model equations have a unique solution (Atar et al. 2014). Let $\bar{\Xi}(t)$ denote the solution to the fluid-model equations with initial value $\bar{\Xi}(0) = \bar{\xi}(0)$. Moreover, Atar et al. (2014) characterize a unique deterministic invariant state of the fluid-model equations, which we denote as $\bar{\Xi}^*$. To understand the long-time behavior of the fluid model, it is also important to understand its invariant distribution. A probability measure μ is said to be an invariant distribution of the fluid-model equations if, given any random element $\bar{\xi}$ whose law is μ, there exists a solution $\bar{\Xi}$ to the fluid-model equations with initial condition $\bar{\xi}$ such that, for any $t \ge 0$, the law of $\bar{\Xi}(t)$ is also μ; see Definition 2.10 in Atar et al. (2023). We make the following assumption to rule out the possibility of random fixed points for the fluid-model equations.

ASSUMPTION 3. *The Dirac delta mass at $\bar{\Xi}^*$ is the unique stationary distribution of the fluid-model equations* (13) – (17).

REMARK 1. Assumption 3 requires that the fluid model has only one invariant distribution, which is concentrated at $\bar{\Xi}^*$. Given that the fluid-model equations are nonlinearly coupled measured-valued equations, it is very hard to verify Assumption 3 as a condition. Atar et al. (2023) shows that Assumption 3 holds for multiclass many-server queues with exponential patience times and class-independent service-time distributions. Extending the result to multiclass queues with class-dependent service times remains an open problem (Atar et al. 2023). However, this assumption is essential to establish that the stationary distribution of the fluid-scaled stochastic system converges to the invariant state of the fluid model as $\lambda \to \infty$ (Theorem 4.4 in Atar et al. (2014)). In particular, note that Atar et al. (2014) only establish the uniqueness of the deterministic fixed point (invariant state). However, they do not rule out the possibility of random fixed points (general invariant distributions).

Define

$$\rho_i = \frac{\lambda(H(r_i) - H(r_{i-1}))\mathbb{E}[S|r_{i-1} < \hat{S} \le r_i]}{s_\lambda} = \frac{\mathbb{E}[S1\{r_{i-1} < \hat{S} \le r_i\}]}{s_\lambda/\lambda}.$$

Based on our scaling, ρ_i does not depend on λ. We also define

$$L = \inf\left\{k : \sum_{i=1}^{k} \rho_i > 1\right\}.$$

Let $D_i^{\lambda,\pi^p}(t)$ denote the number of class i departures from service by time t in the λ-th system under policy π^p.

PROPOSITION 1. *Under Assumptions 1 and 3, for the sequence of systems under policy π^p, we have for $i < L$:*

$$\lim_{\lambda\to\infty} \frac{1}{\lambda} \liminf_{t\to\infty} \frac{1}{t}\mathbb{E}[D_i^{\lambda,\pi^p}(t)] = \lim_{\lambda\to\infty} \frac{1}{\lambda} \limsup_{t\to\infty} \frac{1}{t}\mathbb{E}[D_i^{\lambda,\pi^p}(t)] = H(r_i) - H(r_{i-1});$$

for class L,

$$\begin{aligned}\lim_{\lambda\to\infty} \frac{1}{\lambda} \liminf_{t\to\infty} \frac{1}{t}\mathbb{E}[D_L^{\lambda,\pi^p}(t)] &= \lim_{\lambda\to\infty} \frac{1}{\lambda} \limsup_{t\to\infty} \frac{1}{t}\mathbb{E}[D_L^{\lambda,\pi^p}(t)] \\ &= \frac{s_\lambda/\lambda - \sum_{i=1}^{L-1} \mathbb{E}[S1\{r_{i-1} < \hat{S} \le r_i\}]}{\mathbb{E}[S1\{r_{L-1} < \hat{S} \le r_L\}]}(H(r_L) - H(r_{L-1}));\end{aligned}$$

and for $i > L$,

$$\lim_{\lambda\to\infty}\frac{1}{\lambda}\liminf_{t\to\infty}\frac{1}{t}\mathbb{E}[D_i^{\lambda,\pi^p}(t)] = \lim_{\lambda\to\infty}\frac{1}{\lambda}\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}[D_i^{\lambda,\pi^p}(t)] = 0.$$

Proof. The convergence of the scaled steady-state departure rate to the unique invariant state of the corresponding fluid-model equations follows from Theorem 4.4 in Atar et al. (2014):

$$\lim_{\lambda\to\infty}\frac{1}{\lambda}\liminf_{t\to\infty}\frac{1}{t}\mathbb{E}[D_i^{\lambda,\pi^p}(t)] = \lim_{\lambda\to\infty}\frac{1}{\lambda}\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}[D_i^{\lambda,\pi^p}(t)] = \delta_i,$$

for some $\delta_i \geq 0$. The characterization of the invariant state follows from Theorem 3.3 in Atar et al. (2014): for $i < L$, $\delta_i = H(r_i) - H(r_{i-1})$; for $i > L$, $\delta_i = 0$; and

$$\delta_L = \left(1 - \sum_{i=1}^{L-1}\rho_i\right)\frac{s_\lambda/\lambda}{\mathbb{E}[S|r_{i-1} < \hat{S} \leq r_i]} = \frac{s_\lambda/\lambda - \sum_{i=1}^{L-1}\mathbb{E}[S1\{r_{i-1} < \hat{S} < r_i\}]}{\mathbb{E}[S1\{r_{L-1} < \hat{S} < r_L\}]}(H(r_L) - H(r_{L-1})).$$

■

4.2. Throughput Maximization for Discrete Predicted Service Times

When the predicted service time $\hat{S}$ is discrete with finite support $\{a_1, a_2, \ldots, a_K\}$, where $a_1 < a_1 < \cdots < a_K$, applying SJF scheduling, based on $\hat{S}$, amounts to classifying customers into a finite number of priority classes. The following result follows as a corollary of Proposition 1.

COROLLARY 1. *Under Assumptions 1 and 3, for the sequence of systems under π_{SJF} with discrete predicted service times on a finite support,*

$$\lim_{\lambda\to\infty}\frac{1}{\lambda}\liminf_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi_{SJF}}(t)] = \lim_{\lambda\to\infty}\frac{1}{\lambda}\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi_{SJF}}(t)] = \sum_{i=1}^{\kappa-1}h(a_i) + \phi_\kappa h(a_\kappa),$$

for $\mathbb{P}(\hat{S} = a_i) = h(a_i)$ and ϕ_k given in (3).

Proof. The proof follows directly from Proposition 1 by dividing customers into K classes as follows: If $\hat{S} = a_i$, then the customer is classified into the i-th priority class, $1 \leq i \leq K$. ■

For any scheduling policy π which exploits the noisy discrete service-time information, we define

$$\bar{\mathrm{Th}}^\pi := \limsup_{\lambda\to\infty}\frac{1}{\lambda}\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi}(t)]$$

as the asymptotic throughput in the system. Lemma 1 develops an upper bound for $\bar{\mathrm{Th}}^\pi$.

LEMMA 1. *Under Assumption 2, for any non-preemptive non-anticipative scheduling policy π using the discrete predicted service time $\hat{S}$,*

$$\bar{T}h^{\pi} \leq \sum_{i=1}^{\kappa-1} h(a_i) + \phi_\kappa h(a_\kappa),$$

for κ given in (2) and ϕ_κ given in (3).

Proof. Let $\{t_n\}_{n\geq 1}$ denote a subsequence for which

$$\lim_{t_n\to\infty} \frac{1}{t_n}\mathbb{E}[D^{\lambda,\pi}(t_n)] = \limsup_{t\to\infty} \frac{1}{t}\mathbb{E}[D^{\lambda,\pi}(t)].$$

Note that the above limit exists because $D^{\lambda,\pi}(t) \leq A^{\lambda}(t)$. Let

$$\hat{\gamma}^{\lambda}(a_i) := \lim_{t_n\to\infty} \frac{\mathbb{E}[D_i^{\lambda,\pi}(t_n)]}{\mathbb{E}[A_i^{\lambda}(t_n)]} = \lim_{t_n\to\infty} \frac{\frac{1}{t_n}\mathbb{E}[D_i^{\lambda,\pi}(t_n)]}{\frac{1}{t_n}\mathbb{E}[A_i^{\lambda}(t_n)]}.$$

where $A_i^{\lambda}(t)$ and $D_i^{\lambda,\pi}(t)$ count the number of customers with predicted service time a_i arrived and served by time t, respectively. We can interpret $\hat{\gamma}^{\lambda}(a_i)$ as the long-run average probability of getting service conditional on the customer's predicted service time $\hat{S} = a_i$, along the subsequence $\{t_n\}_{n\geq 1}$. Then,

$$\lim_{t_n\to\infty} \frac{1}{t_n}\mathbb{E}[D^{\lambda,\pi}(t_n)] = \lambda \sum_{i=1}^{K} \hat{\gamma}^{\lambda}(a_i)h(a_i).$$

Note that for any non-anticipative and non-preemptive scheduling policy where the scheduling decision is based on service-time prediction only, we can write $\mathbb{E}[S|\hat{S}=y, \text{Serv}] = \mathbb{E}[S|\hat{S}=y] = \mu(y)$, since conditional on the predicted service time $\hat{S}$, the actual service time is independent of whether the customer is served. We also note that

$$\lambda \sum_{i=1}^{K} \mu(a_i)\hat{\gamma}^{\lambda}(a_i)h(a_i) \leq s^{\lambda}, \tag{5}$$

i.e., the average amount of work served per unit of time is less than the processing capacity.

Based on (5), $\lim_{t_n\to\infty} \frac{1}{t_n}\mathbb{E}[D^{\lambda,\pi}(t_n)]$ is upper bounded by λ multiplied by the objective value of the following optimization problem:

$$\begin{aligned} \max_{\gamma} \ & \sum_{i=1}^{K} \gamma(a_i)h(a_i) \\ \text{s.t.} & \sum_{i=1}^{K} \mu(a_i)\gamma(a_i)h(a_i) \leq \frac{s^{\lambda}}{\lambda} = \sum_{i=1}^{\kappa-1} \mu(a_i)h(a_i) + \phi_\kappa \mu(a_\kappa)h(a_\kappa), \\ & \gamma(a_i) \in [0,1]. \end{aligned} \tag{6}$$

From the first constraint in (6), since $\mu(a)$ is increasing in a, we have

$$\sum_{i=1}^{\kappa-1}\mu(a_i)\gamma(a_i)h(a_i)+\mu(a_\kappa)\sum_{i=\kappa}^{K}\gamma(a_i)h(a_i)\le\sum_{i=1}^{K}\mu(a_i)\gamma(a_i)h(a_i)$$
$$\le\sum_{i=1}^{\kappa-1}\mu(a_i)h(a_i)+\phi_\kappa\mu(a_\kappa)h(a_\kappa).$$

This further implies that

$$\mu(a_\kappa)(1-\phi_\kappa)\gamma(a_\kappa)h(a_\kappa)+\mu(a_\kappa)\sum_{i=\kappa+1}^{K}\gamma(a_i)h(a_i)$$
$$\le\sum_{i=1}^{\kappa-1}\mu(a_i)(1-\gamma(a_i))h(a_i)+\phi_\kappa\mu(a_\kappa)(1-\gamma(a_\kappa))h(a_\kappa)$$
$$\le\mu(a_\kappa)\left(\sum_{i=1}^{\kappa-1}(1-\gamma(a_i))h(a_i)+\phi_\kappa(1-\gamma(a_\kappa))h(a_\kappa)\right).$$

Rearranging the above inequality, we have

$$\sum_{i=1}^{K}\gamma(a_i)h(a_i)\le\sum_{i=1}^{\kappa-1}h(a_i)+\phi_\kappa h(a_\kappa).$$

This implies that

$$\lim_{t_n\to\infty}\frac{1}{t_n}\mathbb{E}[D^{\lambda,\pi}(t_n)]\le\lambda\left(\sum_{i=1}^{\kappa-1}h(a_i)+\phi_\kappa h(a_\kappa)\right).$$

Thus, $\bar{\text{Th}}^{\pi}\le\sum_{i=1}^{\kappa-1}h(a_i)+\phi_\kappa h(a_\kappa)$. ∎

Corollary 1 and Lemma 1 implies that π_{SJF} maximizes the asymptotic throughput in many-server queues with abandonment when the predicted service times are discrete with a finite support.

4.3. Throughput Maximization for Continuous Predicted Service Times

We now consider the case where $\hat{S}$ has a continuous distribution. We begin by considering a special two-class priority rule. We show (Corollary 2 and Lemma 2) that this policy asymptotically maximizes the throughput in the system, among all non-preemptive non-anticipative policies that exploit the noisy service-time information. We then consider a discretized SJF policy, and quantify how the asymptotic throughput changes with the discretization mesh size. We also demonstrate that the throughput under this discretized policy converges to the maximal throughput when the number of priority classes increases without bound. Note that the SJF policy can be viewed as the limit of the discretized SJF policy as the mesh size goes to zero (see Lemma 4). This indicates that SJF maximizes the asymptotic throughput.

4.3.1. Two-class priority policy. We first study the asymptotic throughput in a system operating under the following two-class priority rule, which we denote by π_0. Under π_0, all customers with $\hat{S} \leq \tau$, for τ in (4), are given high non-preemptive priority, and the remaining customers, i.e., the ones with $\hat{S} > \tau$, are given low priority. We next derive the asymptotic throughput under π_0, which follows as a corollary to Proposition 1.

COROLLARY 2. *Under Assumptions 1 and 3, for the sequence of systems under π_0 with continuous predicted service time,*

$$\lim_{\lambda\to\infty}\frac{1}{\lambda}\liminf_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi_0}(t)] = \lim_{\lambda\to\infty}\frac{1}{\lambda}\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi_0}(t)] = \int_0^\tau h(y)dy = \mathbb{P}(\hat{S}\leq\tau).$$

Proof. Based on Proposition 1, we have for class 1,

$$\lim_{\lambda\to\infty}\frac{1}{\lambda}\liminf_{t\to\infty}\frac{1}{t}\mathbb{E}[D_1^{\lambda,\pi_0}(t)] = \lim_{\lambda\to\infty}\frac{1}{\lambda}\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}[D_1^{\lambda,\pi_0}(t)] = \int_0^\tau h(y)dy;$$

and for class 2,

$$\lim_{\lambda\to\infty}\frac{1}{\lambda}\liminf_{t\to\infty}\frac{1}{t}\mathbb{E}[D_2^{\lambda,\pi_0}(t)] = \lim_{\lambda\to\infty}\frac{1}{\lambda}\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}[D_2^{\lambda,\pi_0}(t)] = 0.$$

Adding up the departure rates for the two classes, we have the aggregated departure rate result. ∎

In Lemma 2, we demonstrate that the throughput under any non-anticipative and non-preemptive scheduling policy that exploits the noisy continuous service-time information is upper bounded by the throughput under π_0, i.e., π_0 maximizes the asymptotic throughput.

LEMMA 2. *Under Assumption 2, for any non-preemptive non-anticipative scheduling policy π using the noisy service-time information,*

$$\bar{T}h^\pi \leq \int_0^\tau h(y)dy.$$

Proof. Let $\{t_n\}_{n\geq 1}$ denote the subsequence for which

$$\lim_{t_n\to\infty}\frac{1}{t_n}\mathbb{E}[D^{\lambda,\pi}(t_n)] = \limsup_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi}(t)].$$

Note that the above limit exist, because $D^{\lambda,\pi}(t) \le A^{\lambda}(t)$. Denote $D^{\lambda,\pi}_{\le y}(t)$ as the number of customers with predicted service time less than or equal to y, served by time t. Define

$$\hat{\gamma}^{\lambda}(y) := \frac{\frac{d}{dy}\lim_{t_n\to\infty}\frac{1}{t_n}\mathbb{E}[D^{\lambda,\pi}_{\le y}(t_n)]}{\lambda h(y)}.$$

We can interpret $\hat{\gamma}^{\lambda}(y)$ as the long-run average probability of getting service conditional on the customer's predicted service time $\hat{S}=y$, along the subsequence $\{t_n\}_{n\ge 1}$. Then,

$$\lim_{t_n\to\infty}\frac{1}{t_n}\mathbb{E}[D^{\lambda,\pi}(t_n)] = \lambda\int_0^{\infty}\hat{\gamma}^{\lambda}(y)h(y)dy,$$

Since $\mathbb{E}[S|\hat{S}=y,\text{Serv}] = \mathbb{E}[S|\hat{S}=y] = \mu(y)$,

$$\lambda\int_0^{\infty}\mu(y)\hat{\gamma}^{\lambda}(y)h(y)dy \le s^{\lambda}, \tag{7}$$

i.e., the average amount of work served per unit of time is less than the processing capacity. Based on (7), $\lim_{t_n\to\infty}\frac{1}{t_n}\mathbb{E}[D^{\lambda,\pi}(t_n)]$ is upper bounded by λ multiplied by the objective value of the following optimization problem

$$\begin{aligned}
&\max_{\gamma}\ \int_0^{\infty}\gamma(y)h(y)dy\\
&\text{s.t.}\int_0^{\infty}\mu(y)\gamma(y)h(y)dy \le \frac{s^{\lambda}}{\lambda} = \int_0^{\tau}\mu(y)h(y)dy\\
&\quad\gamma(y)\in[0,1]
\end{aligned} \tag{8}$$

From the first constraint in (8), since $\mu(y)$ is increasing in y, we have

$$\int_0^{\tau}\mu(y)\gamma(y)h(y)dy + \mu(\tau)\int_{\tau}^{\infty}\gamma(y)h(y)dy \le \int_0^{\infty}\mu(y)\gamma(y)h(y)dy \le \int_0^{\tau}\mu(y)h(y)dy.$$

This further implies that

$$\mu(\tau)\int_{\tau}^{\infty}\gamma(y)h(y)dy \le \int_0^{\tau}\mu(y)(1-\gamma(y))h(y)dy \le \mu(\tau)\int_0^{\tau}(1-\gamma(y))h(y)dy.$$

Thus,

$$\int_{\tau}^{\infty}\gamma(y)h(y)dy \le \int_0^{\tau}(1-\gamma(y))h(y)dy.$$

Rearranging the above inequality, we have

$$\int_0^\infty \gamma(y)h(y)dy \le \int_0^\tau h(y)dy.$$

This implies that

$$\lim_{t_n\to\infty} \frac{1}{t_n}\mathbb{E}[D^{\lambda,\pi}(t_n)] \le \lambda \int_0^\tau h(y)dy.$$

Thus, $\bar{\text{Th}}^\pi \le \int_0^\tau h(y)dy$. ∎

4.3.2. Discretized SJF policy. Suppose that $\mathcal{M}$ is a large enough number such that

$$\lambda\mathbb{E}[S\mathbf{1}(\hat{S} \le \mathcal{M})] > s_\lambda; \tag{9}$$

such an $\mathcal{M}$ exists because the system is assumed to be overloaded. We consider a family of discretized SJF policies with mesh size $\Delta \in (0, \mathcal{M})$. In particular, we divide the customers into $N^\Delta = \max\{\lceil \mathcal{M}/\Delta \rceil + 1, \Delta^{-2}\}$ priority classes based on which interval their predicted service time falls into:

$$[0,\Delta], (\Delta, 2\Delta], (2\Delta, 3\Delta], \ldots, ((N^\Delta - 1)\Delta, N^\Delta\Delta], (N^\Delta\Delta, \infty).$$

Let π_{SJF^Δ} denote the priority rule induced by the Δ-segmentation of priority classes defined above. We next derive an expression for the asymptotic throughput under π_{SJF^Δ}, which follows as a corollary of Proposition 1. Let

$$\kappa^\Delta = \min\left\{k \ge 1 : \mathbb{E}[S\mathbf{1}(\hat{S} \le k\Delta)] > s_\lambda/\lambda\right\}.$$

We also define the fraction of served workload from class κ^Δ as

$$\phi_\kappa^\Delta = \frac{s_\lambda/\lambda - \int_0^{\kappa^\Delta\Delta} \mu(y)h(y)dy}{\int_{(\kappa^\Delta-1)\Delta}^{\kappa^\Delta\Delta} \mu(y)h(y)dy},$$

which does not depend on λ based on our scaling (since $s_\lambda/\lambda = \mathbb{E}[S]/\rho$).

COROLLARY 3. *Under Assumptions 1 and 3, for the sequence of $M/GI/s_\lambda + GI$ systems under π_{SJF^Δ} with continuous predicted service time,*

$$\begin{aligned}\lim_{\lambda\to\infty}\frac{1}{\lambda}\liminf_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi_{SJF^\Delta}}(t)] &= \lim_{\lambda\to\infty}\frac{1}{\lambda}\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}[D^{\lambda,\pi_{SJF^\Delta}}(t)] \\ &= \int_0^{(\kappa^\Delta-1)\Delta} h(y)dy + \phi_\kappa^\Delta \int_{(\kappa^\Delta-1)\Delta}^{\kappa^\Delta\Delta} h(y)dy.\end{aligned}$$

Proof. Based on Proposition 1, for class $k \leq \kappa^\Delta - 1$,

$$\lim_{\lambda\to\infty}\frac{1}{\lambda}\liminf_{t\to\infty}\frac{1}{t}\mathbb{E}[D_k^{\lambda,\pi_{\mathrm{SJF}^\Delta}}(t)] = \lim_{\lambda\to\infty}\frac{1}{\lambda}\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}[D_k^{\lambda,\pi_{\mathrm{SJF}^\Delta}}(t)] = \int_{(k-1)\Delta}^{k\Delta} h(y)dy;$$

for class κ^Δ,

$$\lim_{\lambda\to\infty}\frac{1}{\lambda}\liminf_{t\to\infty}\frac{1}{t}\mathbb{E}[D_k^{\lambda,\pi_{\mathrm{SJF}^\Delta}}(t)] = \lim_{\lambda\to\infty}\frac{1}{\lambda}\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}[D_k^{\lambda,\pi_{\mathrm{SJF}^\Delta}}(t)] = \phi_\kappa^\Delta\int_{(\kappa^\Delta-1)\Delta}^{\kappa^\Delta\Delta} h(y)dy;$$

and for class $k > \kappa^\Delta$,

$$\lim_{\lambda\to\infty}\frac{1}{\lambda}\liminf_{t\to\infty}\frac{1}{t}\mathbb{E}[D_k^{\lambda,\pi_{\mathrm{SJF}^\Delta}}(t)] = \lim_{\lambda\to\infty}\frac{1}{\lambda}\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}[D_k^{\lambda,\pi_{\mathrm{SJF}^\Delta}}(t)] = 0.$$

Since

$$D^{\lambda,\pi_{\mathrm{SJF}^\Delta}}(t) = \sum_{k=1}^{N^\Delta} D_k^{\lambda,\pi_{\mathrm{SJF}^\Delta}}(t),$$

we have the aggregated departure rate result. ∎

Lemma 3. *For the asymptotic throughput, we have*

$$\lim_{\Delta\downarrow 0}\bar{T}h^{\pi_{SJF^\Delta}} = \int_0^\tau h(y)dy,$$

i.e., the throughput optimality gap of π_{SJF^Δ} converges to 0 as $\Delta\downarrow 0$.

Proof. An upper bound for the throughput optimality gap is

$$\begin{aligned}&\int_0^\tau h(y)dy - \left(\int_0^{(\kappa^\Delta-1)\Delta} h(y)dy + \phi_\kappa^\Delta\int_{(\kappa^\Delta-1)\Delta}^{\kappa^\Delta\Delta} h(y)dy\right)\\ \leq &\int_{(\kappa^\Delta-1)\Delta}^{\tau} h(y)dy \leq \max_{0\leq y\leq M} h(y)\Delta.\end{aligned}$$

Note that $\max_{0\leq y\leq M} h(y)\Delta \to 0$ as $\Delta\downarrow 0$. ∎

We see in Lemma 3 that as the mesh size Δ approaches 0, i.e., the number of priority classes increases without bound, the optimality gap above converges to 0, i.e., the asymptotic throughput under the sequence of discretized SJF policies converges the maximal throughput which is equal to $\int_0^\tau h(y)dy$.

LEMMA 4. *Under assumption 2, as $\Delta \downarrow 0$, a job $\hat{S}_1$ has a higher priority over another job $\hat{S}_2$ in π_{SJF^Δ} if, and only if, it has a higher priority under π_{SJF}.*

Proof. We map the set of priority classes of π_{SJF^Δ} to the interval $[0,1]$ by identifying each class $i \in \{1,2,\ldots,N^\Delta\}$ with the number

$$H(i\Delta) \in \left\{0, H(\Delta), \ldots, H\left((N^\Delta - 1)\Delta\right), H(N^\Delta \Delta)\right\} \subset [0,1].$$

Let $\bar{C} \coloneqq \sup_{0<x<\infty} h(x) < \infty$. Since $0 \leq H(i\Delta) - H((i-1)\Delta) = \int_{(i-1)\Delta}^{i\Delta} h(x)dx \leq \bar{C}\Delta$ and $\bar{C} < \infty$,

$$\max_{0 \leq i \leq N^\Delta} \{H(i\Delta) - H((i-1)\Delta)\} \to 0 \text{ as } \Delta \to 0.$$

In addition, since $\lim_{\Delta \downarrow 0} H(N^\Delta \Delta) \to 1$, the limit is the continuous interval $[0,1]$. In the limit, priority classes are indexed by $[0,1]$, and for each $x \in [0,1)$, any class within $[0,x]$ has non-preemptive priority over every class within $(x,1]$. In particular, for any two jobs, $\hat{S}_1$ and $\hat{S}_2$ with $0 < \hat{S}_1 < \hat{S}_2 < \infty$, since $H(\hat{S}_1) < H(\hat{S}_2)$, $\hat{S}_1$ has a higher priority over $\hat{S}_2$ under $\pi_{\mathrm{SJF}^\Delta}$ as $\Delta \downarrow 0$. ∎

Combining Lemmas 3 and 4 implies that SJF maximizes the asymptotic throughput. In particular, we prove

$$\lim_{\Delta \downarrow 0} \lim_{\lambda \to \infty} \frac{1}{\lambda} \liminf_{t \to \infty} \frac{1}{t} \mathbb{E}[D^{\lambda, \pi_{SJF^\Delta}}(t)] = \lim_{\Delta \downarrow 0} \lim_{\lambda \to \infty} \frac{1}{\lambda} \limsup_{t \to \infty} \frac{1}{t} \mathbb{E}[D^{\lambda, \pi_{SJF^\Delta}}(t)] = \int_0^\tau h(y)dy.$$

REMARK 2. It would be more interesting to establish an interchange of limits result, i.e.,

$$\lim_{\lambda \to \infty} \frac{1}{\lambda} \liminf_{\Delta \downarrow 0} \liminf_{t \to \infty} \frac{1}{t} \mathbb{E}[D^{\lambda, \pi_{SJF^\Delta}}(t)] = \lim_{\lambda \to \infty} \frac{1}{\lambda} \limsup_{\Delta \downarrow 0} \limsup_{t \to \infty} \frac{1}{t} \mathbb{E}[D^{\lambda, \pi_{SJF^\Delta}}(t)] = \int_0^\tau h(y)dy. \quad (10)$$

Note that since $D^{\lambda, \pi_{SJF^\Delta}}(t) \leq A^\lambda(t)$ and $\frac{1}{t}\mathbb{E}[A^\lambda(t)] = \lambda$, the subsequential limit as $\Delta \downarrow 0$ exists. However, we cannot characterize the limit in closed form, which makes it hard to establish the limits in (10). We leave establishing (10) as a future research direction.

Our analysis indicates that both SJF and the properly designed two-class priority rule maximize the asymptotic throughput. In practice, it can be hard to implement SJF, as it requires keeping track of the predicted service times of everyone in the queue. In contrast, the coarse two-class

priority rule is much simpler as it only requires classifying an incoming customer as having a long or short predicted service time. However, implementing the two-class priority rule, i.e., defining the threshold τ, requires knowing the arrival rate and the joint distribution of the actual and predicted service times. In Section 6, we develop a data-driven method to estimate τ. We also note that when the arrival rate is unknown, SJF has the advantage of being agnostic to the arrival rate. In this case, we may want to implement the discretized SJF policy with a properly chosen mesh size to strike a balance between the implementation ease and the performance of the policy.

4.4. Asymptotic Steady-State Performance

In this section, we explore the steady-state asymptotic performance in the system under SJF. The proof of Theorem 1 is similar to the proof of Theorem 1 in Dong and Ibrahim (2021); we relegate it to Appendix B. Let $\text{Serv}^{\lambda}_{\pi_{\text{SJF}}}$ denote the event that a "tagged" customer arrives at a random system state drawn from the system's steady-state distribution is served, and $W^{\lambda}_{\pi_{\text{SJF}}}$ denote the customer's waiting time. If the steady-state distribution is not unique, then we can look at any one of the steady-state distributions.

THEOREM 1. *Under Assumptions 1 – 3, for the sequence of $M/GI/s_\lambda + GI$ queues, if the predicted service time $\hat{S}$ is discrete with finite support, then, under SJF:*

(a) For $i \le \kappa - 1$, $\lim_{\lambda\to\infty} \mathbb{P}(Serv^{\lambda}_{\pi_{SJF}} | \hat{S} = a_i) = 1$; $\lim_{\lambda\to\infty} \mathbb{P}(Serv^{\lambda}_{\pi_{SJF}} | \hat{S} = a_\kappa) = \phi_\kappa$; and for $i > \kappa$, $\lim_{\lambda\to\infty} \mathbb{P}(Serv^{\lambda}_{\pi_{SJF}} | \hat{S} = a_i) = 0$.

(b) For $i \le \kappa - 1$, $\lim_{\lambda\to\infty} \mathbb{E}[W^{\lambda}_{\pi_{SJF}} | \hat{S} = a_i] = 0$; $\lim_{\lambda\to\infty} \mathbb{E}[W^{\lambda}_{\pi_{SJF}} | \hat{S} = a_\kappa] = \int_0^{\bar{w}} (1 - F(x))dx$, where $\bar{w}$ is the solution $F(\bar{w}) = \frac{\sum_{i=1}^{\kappa} \rho_i - 1}{\rho_\kappa}$; and for $i > \kappa$, $\lim_{\lambda\to\infty} \mathbb{E}[W^{\lambda}_{\pi_{SJF}} | \hat{S} = a_i] = 1/\theta$.

If the predicted service time $\hat{S}$ is continuous then, under discretized SJF with mesh size Δ:

(a) $\lim_{\lambda\to\infty} \mathbb{P}(Serv^{\lambda}_{\pi_{SJF^\Delta}} | \hat{S} \le ((\kappa^\Delta - 1)\Delta) = 1$; $\lim_{\lambda\to\infty} \mathbb{P}(Serv^{\lambda}_{\pi_{SJF^\Delta}} | \hat{S} \in (\kappa^\Delta - 1)\Delta, \kappa^\Delta \Delta]) = \phi^\Delta_\kappa$; and $\lim_{\lambda\to\infty} \mathbb{P}(Serv^{\lambda}_{\pi_{SJF^\Delta}} | \hat{S} > \kappa^\Delta \Delta) = 0$.

(b) $\lim_{\lambda\to\infty} \mathbb{E}[W^{\lambda}_{\pi_{SJF^\Delta}} | \hat{S} \le (\kappa^\Delta - 1)\Delta] = 0$; $\lim_{\lambda\to\infty} \mathbb{E}[W^{\lambda}_{\pi_{SJF^\Delta}} | \hat{S} \in ((\kappa^\Delta - 1)\Delta, \kappa^\Delta \Delta]] = \int_0^{\bar{w}^\Delta} (1 - F(x))dx$, where $\bar{w}^\Delta$ is the solution $F(\bar{w}^\Delta) = \frac{\sum_{i=1}^{\kappa^\Delta} \rho_i - 1}{\rho_{\kappa^\Delta}}$; and $\lim_{\lambda\to\infty} \mathbb{E}[W^{\lambda}_{\pi_{SJF^\Delta}} | \hat{S} > \kappa^\Delta \Delta] = 1/\theta$.

In Theorem 1 (a), we derive key steady-state performance measures under SJF when $\hat{S}$ is discrete with a finite support. In this case, customers with service time predictions that are strictly smaller than a_κ are served immediately, and customers whose predictions are strictly larger than a_κ are never served. Finally, customers whose service-time prediction is equal to a_κ are served with probability ϕ_κ. They experience some wait and only customers with a long enough patience time, i.e., patience time longer than $\bar{w}$, are served in the limit.

In Theorem 1 (b), we derive key steady-state performance measures under the discretized SJF policy when $\hat{S}$ is continuous. In this case, by letting the mesh size Δ decrease to 0, i.e., by increasing the number of priority classes without bound, we approach the performance under the two-priority rule, π_0.

LEMMA 5. *For π_{SJF^Δ}, as $\Delta \to 0$, we have:*

(a) $\lim_{\Delta\downarrow 0}\lim_{\lambda\to\infty}\mathbb{P}(Serv^{\lambda}_{\pi_{SJF^\Delta}}|\hat{S}\le\tau)=1$ *and* $\lim_{\Delta\downarrow 0}\lim_{\lambda\to\infty}\mathbb{P}(Serv^{\lambda}_{\pi_{SJF^\Delta}}|\hat{S}>\tau)=0.$

(b) $\lim_{\Delta\downarrow 0}\lim_{\lambda\to\infty}\mathbb{E}[W^{\lambda}_{\pi_{SJF^\Delta}}|\hat{S}\le\tau]=0$ *and* $\lim_{\Delta\downarrow 0}\lim_{\lambda\to\infty}\mathbb{E}[W^{\lambda}_{\pi_{SJF^\Delta}}|\hat{S}>\tau]=1/\theta.$

Lemma 5 demonstrates a state-space collapse result: In steady state, the SJF queueing system becomes asymptotically indistinguishable from the performance of the two-class priority queue where customers with short predicted service times, below τ, have non-preemptive priority over customers with long predicted service times, above τ. We provide numerical support for this two-class approximation in Section 7.

There remains to further understand, in more depth, how the performance under SJF is affected by the accuracy of the prediction. Next, we establish a monotonicity property on the asymptotic throughput in the system under a bivariate stochastic order on the random pairs of actual and predicted service times.

5. A Comparison of Prediction Models

In this section, our aim is to deepen our understanding of how an improvement in prediction accuracy for service times translates into an improvement in system performance when scheduling according to SJF. In particular, we establish a monotonicity property on the asymptotic throughput

in the system under a bivariate stochastic order on the pair of actual and predicted service times (Theorem 2). As a corollary, we show that, in the practically relevant case of lognormally distributed service times, the higher the correlation between the actual and predicted service times, the higher the asymptotic throughput under SJF.

Note that when $\hat{S}$ is a discrete random variable,

$$\bar{\text{Th}}^{SJF} = \sum_{i=1}^{\kappa-1} h(a_i) + \phi_\kappa h(a_\kappa)$$

and, when $\hat{S}$ is a continuous random variable, we use

$$\bar{\text{Th}}^{SJF} = \int_0^\tau h(y)dy.$$

We write $\bar{\text{Th}}^{SJF}(S, \hat{S})$ to make explicit the dependence of the asymptotic throughput on the distributions of the actual and predicted service times.

5.1. PQD Dependence Order

We let $\mathcal{F}(g_X, g_Y)$ denote the set of all bivariate distributions with the same marginal densities (or probability mass function for discrete random variables) g_X and g_Y. Positive Quadrant Dependence (PQD) is a bivariate stochastic order that is defined as follows; see Chapter 9 in Shaked and Shanthikumar (2007).

Definition 2. (PQD order) Let (X_1, Y_1) and (X_2, Y_2) have joint complementary cumulative distribution functions (ccdf) $\bar{G}_1$ and $\bar{G}_2$ and the same univariate marginals, i.e., both in $\mathcal{F}(g_X, g_Y)$. Then,

$$(X_1, Y_1) \leq_{PQD} (X_2, Y_2) \quad \text{if, and only if,} \quad \bar{G}_1(x, y) \leq \bar{G}_2(x, y) \quad \text{for all } (x, y).$$

Theorem 2. *Let $(S, \hat{S}_1), (S, \hat{S}_2) \in \mathcal{F}(g, h)$. The following holds:*

$$\textit{If} \quad (S, \hat{S}_1) \leq_{PQD} (S, \hat{S}_2) \quad \textit{then} \quad \bar{Th}^{SJF}(S, \hat{S}_1) \leq \bar{Th}^{SJF}(S, \hat{S}_2).$$

5.2. Correlation for Lognormal Service Times

We now consider lognormal service times, which arise a lot in practice. Let $r[X, Y]$ denote the correlation between random variables X and Y. We let Z, $\hat{Z}_1$, and $\hat{Z}_2$ denote normally-distributed

random variables. We assume $\hat{Z}_1$ and $\hat{Z}_2$ have identical marginal distributions, but they can have different correlations with Z. Then, $(Z, \hat{Z}_1)$ and $(Z, \hat{Z}_2)$ each follow a bivariate normal distribution. This guarantees that, for $j = 1, 2$, if $r[Z, \hat{Z}_j] > 0$, then $(S, \hat{S}_j)$ satisfies positive regression dependence as defined in Definition 1. We consider two sets of service-time predictions: $(S, \hat{S}_1) \stackrel{d}{=} (e^Z, e^{\hat{Z}_1})$ and $(S, \hat{S}_2) \stackrel{d}{=} (e^Z, e^{\hat{Z}_2})$.

LEMMA 6. *For $(S, \hat{S}_1)$ and $(S, \hat{S}_2)$ defined as above,*

$$r[Z, \hat{Z}_1] \leq r[Z, \hat{Z}_2] \quad \textit{if, and only if,} \quad (S, \hat{S}_1) \leq_{PQD} (S, \hat{S}_2).$$

We next consider a simple example to quantify the effect of $\rho := r[Z, \hat{Z}]$ on the system's throughput. Consider lognormally distributed $S = \exp(Z)$ and $\hat{S} = \exp(\hat{Z})$, where $Z \sim N(m, \sigma^2)$, $\hat{Z} \sim N(0, 1)$. In this case,

$$(Z | \hat{Z} = \hat{z}) \sim N(m + \sigma\rho\hat{z}, \sigma^2(1 - \rho^2)).$$

This implies that $\mathbb{E}[S | \hat{S} = \exp(z)] = \exp(m + \sigma\rho z + \sigma^2(1 - \rho^2)/2)$. Recall that τ is defined such that

$$\int_0^\tau \frac{1}{\sqrt{2\pi}} \exp(-z^2/2) \exp(m + \sigma\rho z + \sigma^2(1 - \rho^2)/2) dz = s_\lambda/\lambda.$$

This implies that

$$\exp(m + \sigma^2/2) \int_0^\tau \frac{1}{\sqrt{2\pi}} \exp(-(z - \sigma\rho)^2/2) dz = s_\lambda/\lambda.$$

Since $\exp(m + \sigma^2/2) = \mathbb{E}[S]$, we have

$$\int_0^\tau \frac{1}{\sqrt{2\pi}} \exp(-(z - \sigma\rho)^2/2) dz = s_\lambda/(\lambda \mathbb{E}[S]).$$

Let z^* denote the $(s_\lambda/(\lambda\mathbb{E}[S]))$-percentile of $N(0, 1)$. Then, $\tau = z^* + \rho\sigma$ and

$$\bar{\mathrm{Th}}^{SJF}(S, \hat{S}) = P(N(0, 1) \leq z^* + \rho\sigma),$$

which allows us to relate the throughput in the system to the correlation between Z and $\hat{Z}$.

Combining Theorem 2 and Lemma 6 implies that, with lognormal service times, there is an easy way to check which of several sets of service-time predictions leads to a higher asymptotic

throughput under SJF. In particular, provided that these predictions have the same marginal distributions, one would only have to compute correlations with the actual service times: The higher the correlation, the higher the throughput. We also note that the assumption of having the same marginal distribution for alternative service-time predictions may be restrictive. Thus, we consider in Section 7 alternative service-time prediction models where this assumption does not hold, as a robustness check, and we reach consistent conclusions there, i.e., a higher correlation leads to a higher throughput.

6. Data-Driven Method to Estimate τ for Continuous Predicted Service Times

In Section 4.3.2, we established that the discretized SJF policy's asymptotic throughput converges to the maximal throughput as the discretization mesh size goes to zero. In addition, this maximal throughput can also be achieved asymptotically by the two-class priority policy, π_0, which non-preemptively prioritizes customers with $\hat{S} \le \tau$ over those with $\hat{S} > \tau$, for τ defined in (4). However, the derivation of τ in (4) assumes knowledge of the joint distributions of S and $\hat{S}$ and the arrival rate, which may not be readily available. In this section, we explore the derivation of τ based on historical data of actual and predicted service times, and interarrival times.

Suppose that we have historical data recorded as $X_i = (S_i, \hat{S}_i, U_i)_{1 \le i \le n}$, where U_i is the interarrival time between the i-th customer and the $(i-1)$-th customer. Define the function

$$g(t) = \mathbb{E}[S1\{\hat{S} \le t\}] - \frac{s}{\lambda} = \mathbb{E}[S1\{\hat{S} \le t\}] - s\mathbb{E}[U].$$

Note that $g(t)$ is increasing in t, with $g'(t) = \mu(t)h(t)$. Let τ denote the root of $g(t)$, i.e., $g(\tau) = 0$. In practice, we may not be able to evaluate $g(t)$ exactly. Instead, we can approximate $g(t)$ by

$$G(t, X_i) = S_i 1\{\hat{S}_i \le t\} - sU_i.$$

Note that $\mathbb{E}[G(t, X_i)] = g(t)$, i.e., $G(t, X_i)$ is an unbiased estimator of $g(t)$. Then, we have a stochastic root-finding problem.

There are many existing algorithms to solve the stochastic root-finding problem. One algorithm is called stochastic approximation. Under the so-called Polyak-Ruppert averaging, we update τ_k iteratively according to

$$\tau_k = \tau_{k-1} - \gamma_k G(\tau_{k-1}, X_k), \tag{11}$$

where $\gamma_k = ak^{-r}$ for some $a > 0$ and $r \in (1/2, 1)$. We also write $\bar{\tau}_k = \frac{1}{k}\sum_{i=1}^{k} \tau_i$. Then we have the following convergence result for $\bar{\tau}_k$.

PROPOSITION 2. *Suppose $c \le \mu(y)h(y) \le C$ and $\mathbb{E}[S^2|\hat{S} = y]h(y) \le C$ for some $c, C \in (0, \infty)$ for any $y \in (0, M_H)$. Then, the iteration in (11) satisfies*

$$\bar{\tau}_k \to \tau \text{ as } k \to \infty \text{ almost surely}$$

and

$$\sqrt{k}(\bar{\tau}_k - \tau) \Rightarrow N\left(0, \frac{Var(S1\{\hat{S} \le \tau\} - sU)}{(g'(\tau))^2}\right) \text{ as } k \to \infty.$$

Proposition 2 indicates that $\bar{\tau}_k$ is a consistent estimator of τ and it converges at the rate $1/\sqrt{k}$. In Section 7.4.2, we study the performance of the two-class priority queue with threshold $\bar{\tau}_k$ for different values of k using simulation experiments. We find that a relatively small k, i.e., several hundred, already leads to very good performance.

7. Numerical Study

In this section, we describe results from some simulation experiments where we: (i) quantify the impact of increasing the number of classes under the discretized SJF rule (Section 7.2); (ii) study the gap between the throughput under SJF and the throughput under the two-class priority rule, π_0, in finite stochastic systems (Section 7.3); (iii) investigate the importance of selecting the right threshold in the two-class priority rule, including analyzing the performance of the data-driven estimation of the threshold (Section 7.4). We summarize our main results here, and relegate many tables with detailed simulation results to the Appendix G.

7.1. Description of the Experiments

We simulate the $M/GI/s+M$ queueing system. We focus on overloaded systems where the arrival rate exceeds the total service rate. In particular, we consider values of the traffic intensity $\rho = 1.4$ and $\rho = 1.8$. For the number of servers, we consider values ranging from $s = 20$ to $s = 1000$. For each set of simulation results, we report point estimates of performance measures which are based on averaging across 10 independent simulation replications of length 1,000,000 arrivals each. For each point estimate, we calculate the corresponding 95% confidence intervals, but we do not report these in the tables because we found them to be very narrow: The half widths are consistently below 0.05% of the corresponding point estimates.

We focus here on continuous service-time predictions. We do so because the system with discrete service-time predictions can be thought of as a special discretized SJF rule, which we study in Section 7.2. We consider lognormally distributed (actual) service times where we fix, without loss of generality, the mean service time to be equal to 1. Let S_i denote the actual service time of customer i. We let Z_i be a normally-distributed random variable with mean $-\ln(2)/2$ and variance $\ln(2)$. This makes $S_i = e^{Z_i}$ lognormally distributed with mean 1 and variance 1. Let α be a scalar such that $0 \leq \alpha \leq 1$, and we let $\hat{Z}_i(\alpha)$ and $\epsilon_i(\alpha)$ be such that

$$Z_i = \hat{Z}_i(\alpha) + \epsilon_i(\alpha), \tag{12}$$

where $\hat{Z}_i(\alpha)$ is normally distributed with mean $-\ln(2)/2$ and variance $\alpha \ln(2)$, and, independently of $\hat{Z}_i(\alpha)$, $\epsilon_i(\alpha)$ is normally distributed with mean 0 and variance $(1-\alpha)\ln(2)$. We define the service-time prediction as $\hat{S}_i(\alpha) := e^{\hat{Z}_i(\alpha)}$. Note that S_i and $\hat{S}_i$ defined in this manner satisfy Assumption 2. We emphasize that both the marginal distribution of $\hat{S}_i(\alpha)$ and the joint distribution of $(S_i, \hat{S}_i(\alpha))$ depend on α. This is different from the model studied in Section 5.2, where we require the marginal distribution of the predicted service times to be fixed. We consider a different form of service-time prediction deliberately because we would like to test the robustness of our results beyond the model in Section 5.2. We vary α to alter the correlation between Z_i and $\hat{Z}_i(\alpha)$, i.e., between the actual and predicted service times: Smaller values of α correspond to noisier predictions. We consider values of

α ranging from $\alpha = 0.001$ ($r[Z_i, \hat{Z}_i(\alpha)] = 0.032$ and $r[S_i, \hat{S}_i(\alpha)] = 0.028$) to $\alpha = 0.98$ ($r[Z_i, \hat{Z}_i(\alpha)] = 0.99$ and $r[S_i, \hat{S}_i(\alpha)] = 0.99$).

In the appendices (Table 30, Table 31, and Figure 3), we also consider a model for $(S_i, \hat{S}_i)$ which is consistent with our description in Section 5.2. In particular, we fix the marginal distributions of S_i and $\hat{S}_i$ to be lognormally distributed with mean 1 and variance 1. We vary the correlation between Z_i and $\hat{Z}_i$ and consider values ranging from 0.005 to 0.99. The conclusions that we reach based on our simulation study are consistent under both service-time prediction models.

7.2. Performance of the Discretized SJF Policy

In Section 4.3.2, we showed that the discretized SJF policy maximizes the asymptotic throughput as the discretization mesh size goes to 0. In this section, we study the pre-limit performance of the discretized SJF policy as the number of priority classes increases, i.e., the discretization mesh size decreases. Note that the SJF can be thought of as an infinite-class priority rule where each predicted service time constitutes its own priority class.

We consider the discretized SJF policy in Section 4.3.2 and let $\mathcal{M} = 2\tau$ where τ is the threshold in (4). We gradually increase the number of priority classes as follows. For a system with k classes, $k = 2, 3, 5, 9$, we divide $\hat{S}$ into k intervals with decreasing priority: $(0, \mathcal{M}/(k-1)]$, $(\mathcal{M}/(k-1), 2\mathcal{M}/k], \ldots, ((k-2)\mathcal{M}/(k-1), \mathcal{M}]$, and $(\mathcal{M}, \infty)$.

In Table 1, we compare the throughput of the discretized SJF policy with different numbers of priority classes. We also list the throughput of FCFS and SJF as two benchmark policies. First, we observe that as k increases, the throughput under the discretized SJF policy increases and gets closer to the throughput under SJF. Note that SJF achieves the maximum throughput among all the policies tested in Table 1. Second, there is a diminishing return in adding more priority classes. We generally do not see a large improvement in performance in going beyond three classes. Beyond five classes, we see only a small increase in the throughput by adding more priority classes. In addition, there is a larger room for improvement (by adding more priority classes) when the service-time predictions are more accurate.

α	$r[S_i, \hat{S}_i(\alpha)]$	$r[Z_i, \hat{Z}_i(\alpha)]$	FCFS/LCFS	SJF	2 classes	3 classes	5 classes	9 classes
0.001	0.0282	0.0316	0.7143	0.7230	0.7149	0.7220	0.7220	0.7220
0.005	0.0636	0.0707	0.7143	0.7327	0.7150	0.7300	0.7300	0.7300
0.01	0.0877	0.1000	0.7143	0.7397	0.7143	0.7356	0.7356	0.7357
0.02	0.1151	0.1414	0.7143	0.7493	0.7142	0.7433	0.7433	0.7440
0.05	0.1899	0.2236	0.7143	0.7677	0.7140	0.7580	0.7582	0.7621
0.1	0.2688	0.3162	0.7143	0.7870	0.7144	0.7737	0.7755	0.7824
0.2	0.3894	0.4472	0.7143	0.8114	0.7175	0.7936	0.8000	0.8081
0.3	0.4767	0.5477	0.7143	0.8276	0.7231	0.8072	0.8179	0.8250
0.4	0.5691	0.6325	0.7143	0.8403	0.7297	0.8178	0.8319	0.8380
0.5	0.6413	0.7071	0.7143	0.8504	0.7367	0.8264	0.8432	0.8485
0.6	0.7154	0.7746	0.7143	0.8587	0.7432	0.8332	0.8523	0.8570
0.7	0.7878	0.8367	0.7143	0.8659	0.7495	0.8387	0.8596	0.8641
0.8	0.8602	0.8944	0.7143	0.8714	0.7550	0.8429	0.8654	0.8697
0.9	0.9333	0.9487	0.7143	0.8763	0.7603	0.8464	0.8701	0.8747
0.95	0.9641	0.9747	0.7143	0.8787	0.7629	0.8477	0.8719	0.8768
0.98	0.9866	0.9899	0.7143	0.8801	0.7642	0.8487	0.8728	0.8781

Table 1 Long-run throughput in the $M/LGN/100+M$ model with $\rho = 1.4$ under the discretized SJF rule where we let $\mathcal{M} = 2\tau$ and divide $[0, \mathcal{M})$ into equally-sized classes.

We also tried other ways of dividing the predicted service times into more and more granular priority classes. For example in Table 2 in Appendix G, we set $\mathcal{M} = \tau$ and divide $[0, \tau]$ into $(k-1)$, $k = 2, 3, 5, 10$, equally spaced priority classes while keeping the lowest priority class as (τ, ∞). We make similar observations as those discussed above based on Table 1.

7.3. Accuracy of the Two-Priority Approximation

In Section 4.3.2 and 4.4, we showed that the system performance under SJF, which can be viewed as the limit of discretized SJF when the number of priority classes increases without bound, is

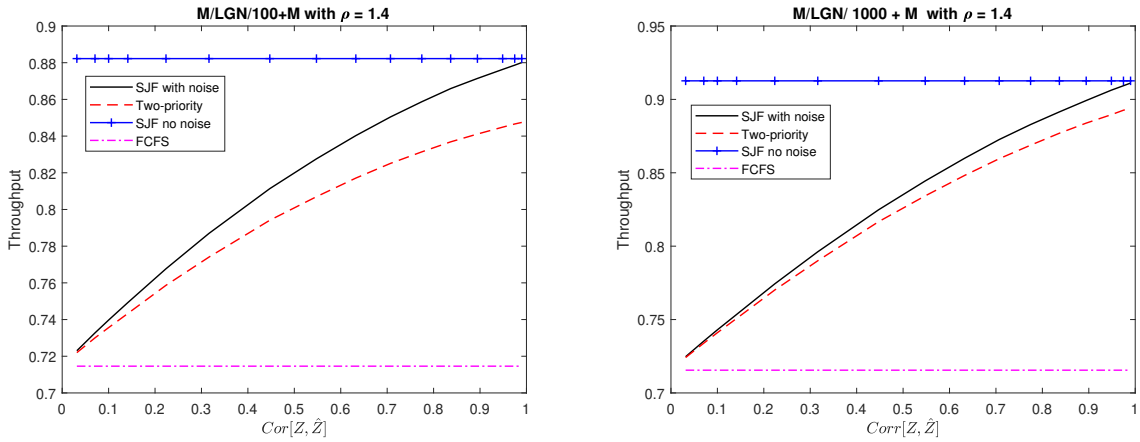

Figure 1 **Long-run throughput, in steady state, in the $M/LGN/n+M$ model.**

asymptotically indistinguishable from the performance under the two-class priority rule π_0. In this section, we investigate the pre-limit performance of π_0 compared to SJF.

We consider the $M/LGN/s+M$ system where we fix $\rho = 1.4$ and vary the number of servers $s = 20, 50, 70, 100, 500, 1000$. The limit in Corollary 2 holds as the number of servers increases without bound. Thus, we expect the gap between π_0 and SJF to be closer in larger systems. Nevertheless, we deliberately consider small values of s too to validate the usefulness of the asymptotic results in relatively small systems. We are also interested in quantifying the effect of the degree of noise in the service-time prediction. As a robustness check, we also consider non-exponential interarrival-time distributions: The Erlang, E_2, distribution and the hyperexponential, H_2, distribution with a squared coefficient of variation $c^2 = 4$ and balanced means, i.e., the two component exponential distributions contribute equally to the mean. We consider these two distributions to vary the degree of variability in the interarrival-time distribution, relative to the exponential distribution which has $c^2 = 1$: The E_2 distribution has $c^2 = 1/2$ and the H_2 distribution has $c^2 = 4$. The numerical results of this section are presented in Tables 3-20 in Appendix G.

7.3.1. Size of the system. Tables 3-8 show that the SJF policy achieves larger throughput, i.e., a smaller abandonment rate, than π_0. And, as expected, the accuracy of the two-class approximations improves as the number of servers increases. For one example, when $\alpha = 0.3$, the relative gaps in the throughput range from 2% for $s = 20$ to 1% for $s = 1000$. For another example, when

$\alpha = 0.98$, those relative gaps range from 5% for $s = 20$ to 2% for $s = 1000$. It is worth noting that the quality of the two-class priority approximations is reasonable for a relatively small number of servers (e.g., $s = 20$). This is important because it implies that the two-class priority rule performs well when the system is not unrealistically large.

7.3.2. Noise in the service-time prediction. We now turn to investigate the impact of noise on the service-time predictions. While it is to be expected that noisier service-time predictions would lead to worse performance in the system under SJF, e.g., smaller throughput, the extent of this degradation in performance is not clear and is worthwhile investigating.

In Figure 1, we plot point estimates of the throughput under SJF and π_0 in the $M/LGN/100+M$ and $M/LGN/1000+M$ models, as a function of the correlation between Z and $\hat{Z}(\alpha)$, for $\rho = 1.4$. We consider values of the correlation ranging from 0.03 ($\alpha = 0.001$) to 0.99 ($\alpha = 0.98$). We also include in the plots, as benchmarks, curves corresponding to the SJF policy assuming perfect knowledge of the service times and the FCFS policy. We note that the throughput for SJF with no noise (top curve in the plot) and FCFS (bottom curve in the plot) are constant as a function of the correlation as they do not depend on the predicted service time. For the SJF and two-class priority rules with noisy service time information, when the service-time prediction is very noisy (low values of α), we expect performance in the system to be close to the FCFS performance. In contrast, when the prediction is very accurate (high values of α), we expect the performance to be close to the performance under SJF with perfect knowledge of service times. This is confirmed by the plots in Figure 1. Overall, the deterioration in throughput, as α decreases, is (loosely) upper bounded by the difference in throughput between FCFS and SJF with no noise, which is around 20% for both $s = 100$ and $s = 1000$.

Interestingly, it is also apparent in Figure 1 that the quality of the two-class priority approximation *degrades* as the correlation increases. For example, Table 6 shows that, for $s = 100$, the relative difference in the throughput ranges from 0.14% for $\alpha = 0.001$ (first row) to 3.7% for $\alpha = 0.98$ (last row). This is practically meaningful because service-time predictions in service systems are usually not very accurate, which is when the two-class priority rule performs very similarly to SJF.

7.3.3. General interarrival-time distributions. In Tables 9 - 20, we investigate the quality of the two-class approximation with a general interarrival-time distribution, namely we consider the E_2 and H_2 distributions. We find that the accuracy of the two-class approximation remains reasonable in these cases and, in general, we obtain results that are directionally consistent with the exponential case. In particular, the quality of the two-class approximation degrades as the correlation between actual and predicted service times increases. Moreover, the quality of the approximation generally improves as the number of servers increases. We also observe that the approximation appears to have slightly superior accuracy with E_2 compared to H_2 interarrival times. For example, for $s = 1000$ and $\alpha = 0.98$, the relative error in the throughput with E_2 is around 18% and it is around 27% with H_2. For another example, for $s = 1000$ and $\alpha = 0.3$, the relative gap in the throughput with E_2 is around 5% and it is around 9% with H_2. Overall, this numerical evidence points to the usefulness of the two-class approximation beyond the Poisson arrival process assumption.

7.4. Selection of the Threshold τ

Due to the good performance of the two-class priority rule π_0 (e.g., achieves maximal throughput asymptotically) and its ease of implementation, in this section, we look further into the implementation of the two-class priority rule.

So far, we have assumed that we can calculate the threshold τ based on (4). However, this calculation requires knowledge of the arrival rate as well as actual and predicted service time distributions, which may not be readily available. As such, we explore the effect of choosing the right threshold in Section 7.4.1 by contrasting system performance with and without the correct threshold value. In Section 7.4.2, we consider the data-driven estimation of the threshold, thus providing numerical support to our results in Section 6.

7.4.1. Choosing the right threshold in the two-class priority rule. In Tables 21 - 23 in Appendix G, we investigate the effect of choosing the right threshold, τ, on performance in the system. In particular, we aim to quantify the performance improvement that results from selecting

a threshold that accounts for the noise in the service-time prediction. To do so, we consider two systems. In the first system, we implement a two-class priority scheduling policy where customers with service-time prediction $\hat{S} \leq \tau_1$ are given non-preemptive priority over customers with $\hat{S} > \tau_1$. We calculate τ_1 by solving (4). In the second system, we consider a two-class priority system with a threshold τ_2 instead. We let τ_2 be the solution of the equation:

$$\lambda \mathbb{P}(S \leq \tau_2)\mathbb{E}[S|S \leq \tau_2] = s;$$

that is, we assume that τ_2 is calculated by assuming out any noise in the service-time prediction.

We consider $s = 100, 1000$ and $\rho = 1.4$ (Tables 21 and 22) and $s = 1000$ and $\rho = 1.8$ (Table 23). We observe that when service-time predictions are extremely noisy (small values of α), there is a negligible advantage from implementing the correct threshold, i.e., based on (4). The reason is that performance in the system is close to performance under FCFS in this case since the service-time information is so noisy that it does not offer a significant advantage over a blind service policy which does not exploit the service-time information at all. On the other hand, when service-time predictions are extremely accurate (large values of α), there is also negligible advantage from implementing the correct threshold. The reason is that there is almost perfect knowledge of the service times in this case, so the two thresholds, τ_1 and τ_2, are very close to each other. In contrast, we see significant improvement from implementing the correct threshold for moderate values of α. For example, for $s = 100$ and $\rho = 1.4$, Table 21 shows that the probability of abandonment reduces by almost 40% when implementing the correct threshold for $\alpha = 0.4$. We make consistent observations for all parameter values tested.

In Figure 2, we investigate the effects of altering the thresholds in the two-class priority rule and plot the corresponding throughput. We consider three different values of $\alpha = 0.005, 0.5, 0.98$ to study the effect of prediction accuracy. In the figure, we use the solid line to indicate the value of τ, calculated based on (4). In general, we observe that there is not much value in further optimizing the threshold value beyond τ. The gap between the throughput at the optimal threshold value and τ is consistently small, at around 1%. In Tables 24 - 26 in Appendix G, we report the corresponding detailed simulation results.

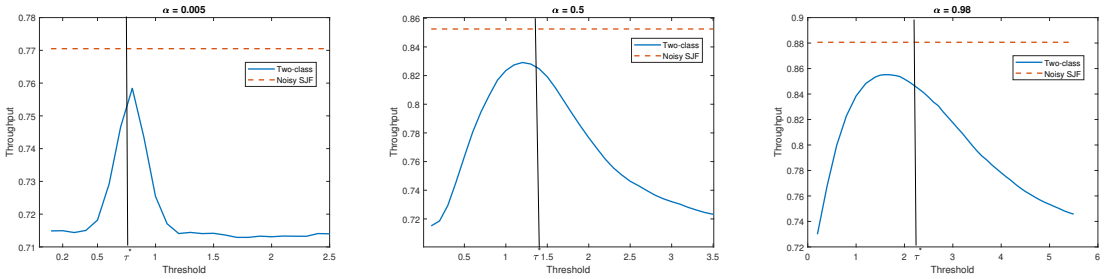

Figure 2 **Throughput, in steady state, in the $M/LGN/100+M$ model for $\rho = 1.4$, different values of the thresholds, and different α values.**

7.4.2. Data-driven estimation of the threshold. In this section, we provide numerical support to our data-driven approach to find τ in Section 6. We set $s = 100$ and $\rho = 1.4$, and consider different values of a in $\gamma_k = ak^{-3/4}$, see (11), along with different sample sizes $k = 500, 1000$, and $10,000$. We assume that the value of the arrival rate λ is known.

We present detailed numerical results in Tables 27-29 in Appendix G. In each table, we fix the value of $a = 0.1, 0.5, 0.7$, and report simulation estimates of different performance measures for each value of α and the sample size. These performance measures are based on a two-class priority rule where the class-division threshold is calculated using the stochastic approximation algorithm with the corresponding parameters. For each value of α, we also report in the tables the performance of the two-class priority rule with threshold τ defined in (4). As expected, the performance improves, i.e., the throughput increases, when the sample size increases. This is because the estimated threshold is more likely to be close to τ as the sample size increases. We also observe that irrespective of the value of a selected, if enough data is collected, the stochastic approximation method will learn the "correct" threshold τ.

The value of tuning the parameter a is more evident when the sample size is smaller. This is especially true for higher values of α, i.e., as the correlation between the actual and predicted service times increases. For example, when $a = 0.1$, for $\alpha = 0.5$ or 0.98, and a sample size of 500, Table 27 shows that there can be as much as around 4 percentage points loss in throughput because of the estimation errors in the threshold. However, this loss in accuracy disappears when a good

value of a is chosen. For example, Table 29 shows that when $a = 0.7$, there is almost no loss in throughput in using the estimated threshold with a sample of size 500.

Thus, our conclusion overall is that the proposed data-driven method to estimate the threshold can perform quite well. In general, we advocate experimenting with a few different values of a and choosing a reasonably large sample size, i.e., a few hundred to a thousand.

8. Conclusions

In this paper, we presented theoretical results quantifying performance in an SJF queueing system with multiple servers, impatient customers, and noisy service-time predictions. We considered an overloaded regime and carried out a many-server asymptotic mode of analysis. We considered both discrete and continuous service-time predictions. We showed that the throughput achieved under SJF is asymptotically maximal among all non-preemptive scheduling policies that exploit the same noisy service-time prediction. For continuous service-time predictions, we further showed that steady-state performance measures converge to their counterparts in a non-preemptive two-class priority system where customers with short predicted service times (below a threshold) have priority over customers with long predicted service times (above a threshold).

We can glean managerial insights based on our theoretical results. Our key theoretical results for continuous predictions show that a service discipline that splits customers into just two properly defined priority classes can yield as good performance as SJF. This is practically important because implementing SJF can be quite challenging in practice, since it involves keeping track of the predicted service requirements and rank ordering every customer in the queue. The accuracy of this two-class approximation is superior in large and congested systems, and performs reasonably well in small systems too, as was substantiated in our numerical study. Thus, a manager may achieve the desired superior performance by implementing a coarse customer classification instead.

Appendix A: Many-Server Queues with a Finite Number of Priority Classes

In this section, we present some relevant developments from Atar et al. (2014) and Atar et al. (2023) which we will utilize in our analysis.

Consider many-server queues with m priority classes. We start by presenting a system descriptor. We then define the fluid-model equations, which govern the law of large number behavior of the scaled stochastic system as the scaling parameter λ goes to infinity (see Theorem 4.3 in Atar et al. (2014)).

The state of the system at time t can be modeled by $2m$ coupled measures: the potential queue measure and the server measure. The m potential queue measures, one for each class, track the time elapsed since the entry into the system for all jobs that have entered the system and whose elapsed time is strictly smaller than their respective patience time. The m server measures, one for each class, track the age of the jobs currently in service. We next provide the mathematical definition of these measures. For the j-th arrival in class i, we denote a_{ij}, r_{ij}, and s_{ij} as the customer's arrival time, patience time, and service time, respectively. Let e_{ij} denote the customer's time of entering service. We also denote A_i as the counting process of class i arrivals and E_i as the counting process of class i customers that enter service. For completeness, we also denote $x_{i,0}$ as the number of class i customers that have arrived before time 0 and we assume these customers indices between $-x_{i,0}+1$ and 0. Then, the age in service measure $\nu_{i,t}(dx)$ is defined as

$$\nu_{i,t}(dx) = \sum_{j=-x_{i,0}+1}^{A_i(t)} 1\{0 \le t - e_{ij} < s_{ij}\}\delta_{t-e_{ij}}(dx),$$

where δ_a is the Dirac delta mass at a. The potential queue measure $\eta_{i,t}(dx)$ is defined as

$$\eta_{i,t}(dx) = \sum_{j=-x_{i,0}+1}^{A_i(t)} 1\{0 \le t - a_{ij} < r_{ij}\}\delta_{t-a_{ij}}(dx).$$

Let $B_i(t)$ denote the number of class i customers in service at time t. Then, $B_i(t) = \int_0^\infty \nu_{i,t}(dx)$. Let $Q_i(t)$ denote the number of class i customers waiting in the queue at time t. Then, the oldest class i job in the queue can be expressed as

$$\inf\left\{y \ge 0 : \int_0^y \eta_{i,t}(dx) = Q_i(t)\right\}.$$

Let $X_i(t) = Q_i(t) + B_i(t)$. We also denote $D_i(t)$, $R_i(t)$ the cumulative number of class i departures on $[0,t]$ from service and from abandonment respectively. Note that

$$X_i(t) = X_i(0) + A_i(t) - D_i(t) - R_i(t) \text{ and } B_i(t) = B_i(0) + E_i(t) - D_i(t).$$

Let $\zeta_{G,i}(x)$ and $\zeta_{F,i}(x)$ denote the hazard rate function of the service time and patience time, respectively, for class i customers. We next introduce the fluid model $(\bar{B}, \bar{X}, \bar{Q}, \bar{D}, \bar{A}, \bar{E}, \bar{R}, \bar{\nu}, \bar{\eta})$ that satisfies the following system of fluid model equations, (13) – (17).

$$\int_0^T \int_0^\infty \zeta_{G,i}(x)\bar{\nu}_{i,t}(dx)dt < \infty \text{ and } \int_0^T \int_0^\infty \zeta_{F,i}(x)\bar{\eta}_{i,t}(dx)dt < \infty \text{ for any } T \ge 0. \tag{13}$$

$\bar{Q}_i$ and $\bar{B}_i$ are nonnegative and

$$\bar{B}_i(t) = \bar{B}_i(0) - \bar{D}_i(t) + \bar{E}_i(t), \quad \bar{X}_i(t) = \bar{X}_i(0) - \bar{D}_i(t) + \bar{A}_i(t) - \bar{R}_i(t), \quad \bar{Q}_i(t) = \bar{X}_i(t) - \bar{B}_i(t). \tag{14}$$

$$\begin{aligned} \bar{B}_i(t) &= \int_0^\infty \bar{\nu}_{i,t}(dx), \quad \bar{D}_i(t) = \int_0^t \int_0^\infty \zeta_{G,i}(x)\bar{\nu}_{i,s}(dx)ds, \\ \bar{R}_i(t) &= \int_0^t \int_0^\infty \zeta_{F,i}(x) 1\left\{\int_0^x \bar{\eta}_{i,s}(du) < Q_i(s)\right\} \bar{\eta}_{i,s}(dx)ds. \end{aligned} \tag{15}$$

The scheduling policy is work-conserving and non-preemptive, which corresponds to that K_i is nonnegative and nondecreasing and

$$1 - \sum_{i=1}^m \bar{B}_i(t) = \left(1 - \sum_{i=1}^m X_i(t)\right)^+ \text{ and } \bar{E}_i(t) = \int_0^t 1\left\{\sum_{l=1}^{i-1} \bar{Q}_l(s) = 0\right\} d\bar{E}_i(s). \tag{16}$$

Lastly, for test functions $\psi(x,t)$ and $\phi(x,t)$ that are continuously differentiable with continuously differentiable first derivatives (x only needs to be defined on the domain of the definition of the service time distribution and patience time distribution respectively), we have

$$\begin{aligned} \int_0^\infty \psi(x,t)\nu_{i,t}(dx) &= \int_0^\infty \psi(x,0)\bar{\nu}_{i,0}(dx) + \int_0^t \int_0^\infty (\partial_x \psi(x,s) + \partial_t \psi(x,s))\bar{\nu}_{i,s}(dx)ds \\ &\quad - \int_0^t \int_0^\infty \zeta_{G,i}(x)\psi(x,s)\bar{\nu}_{i,s}(dx)ds + \int_0^t \psi(0,s)d\bar{E}_i(s) \\ \int_0^\infty \phi(x,t)\bar{\eta}_{i,t}(dx) &= \int_0^\infty \phi(x,0)\bar{\eta}_{i,0}(dx) + \int_0^t \int_0^\infty (\partial_x \phi(x,s) + \partial_t \phi(x,s))\bar{\eta}_{i,s}(dx)ds \\ &\quad - \int_0^t \int_0^\infty \zeta_{F,i}(x)\phi(x,s)\bar{\eta}_{i,s}(dx)ds + \int_0^t \phi(0,s)d\bar{A}_i(s). \end{aligned} \tag{17}$$

Atar et al. (2014) prove that the fluid model equations have a unique solution, which we shall refer to as the fluid model. They further characterize the unique invariant state of the fluid model. Assumption 3 assumes that the Dirac delta mass at the unique invariant state is the unique invariant distribution of the fluid model equations, i.e., the unique probability distribution that is invariant under the flow defined by the fluid model equations. This assumption was not verified in Atar et al. (2014), but verified in Atar et al. (2023) for the special case where the service time distribution is class-independent and the patience times are exponential. How to establish the condition for multiclass queues with class-dependent service time distribution remains an open problem (Atar et al. 2023). Under certain regularity conditions on the service and patience time distribution and Assumption 3, following Atar et al. (2014), we have the stationary distribution of the fluid-scaled, i.e., scaled by $1/\lambda$, dynamics of the stochastic system, the many-server queue with m priority classes, converges to the invariant state of the fluid model (see Proposition 1). This is the main result that we leverage in our development.

Appendix B: Proof of Theorem 1

Proof. We only provide the proof for the case where the predicted service time is discrete with finite support. Since the proof for the discretized SJF with mesh size Δ, $\pi_{\mathrm{SJF}^\Delta}$, follows exactly the same arguments.

We begin by proving part (a). For $i \le \kappa - 1$, from Proposition 1, we have

$$\lim_{\lambda\to\infty}\frac{1}{\lambda}\liminf_{t\to\infty}\frac{1}{t}\mathbb{E}[D_i^{\lambda,\pi_{\mathrm{SJF}}}(t)] = \lim_{\lambda\to\infty}\frac{1}{\lambda}\limsup_{t\to\infty}\frac{1}{t}\mathbb{E}[D_i^{\lambda,\pi_{\mathrm{SJF}}}(t)] = h(a_i).$$

This indicates that $\lim_{\lambda\to\infty}\mathbb{P}(\mathrm{Serv}^\lambda_{\pi_{\mathrm{SJF}}}1\{\hat{S}=a_i\}) = h(a_i)$. Thus,

$$\lim_{\lambda\to\infty}\mathbb{P}(\mathrm{Serv}^\lambda_{\pi_{\mathrm{SJF}}}|\hat{S}=a_i) = \frac{\lim_{\lambda\to\infty}\mathbb{P}(\mathrm{Serv}^\lambda_{\pi_{\mathrm{SJF}}}1\{\hat{S}=a_i\})}{\mathbb{P}(\hat{S}=a_i)} = \frac{h(a_i)}{h(a_i)} = 1.$$

Similarly, for $\hat{S}=a_\kappa$, we have

$$\lim_{\lambda\to\infty}\mathbb{P}(\mathrm{Serv}^\lambda_{\pi_{\mathrm{SJF}}}|\hat{S}=a_\kappa) = \frac{\lim_{\lambda\to\infty}\mathbb{P}(\mathrm{Serv}^\lambda_{\pi_{\mathrm{SJF}}}1\{\hat{S}=a_i\})}{\mathbb{P}(\hat{S}=a_i)} = \frac{\phi_\kappa h(a_\kappa)}{h(a_\kappa)} = \phi_\kappa.$$

For $i > \kappa$, we have

$$\lim_{\lambda\to\infty}\mathbb{P}(\mathrm{Serv}^\lambda_{\pi_{\mathrm{SJF}}}|\hat{S}=a_i) = \frac{\lim_{\lambda\to\infty}\mathbb{P}(\mathrm{Serv}^\lambda_{\pi_{\mathrm{SJF}}}1\{\hat{S}=a_i\})}{\mathbb{P}(\hat{S}=a_i)} = \frac{0}{h(a_i)} = 0.$$

We next prove part (b). Theorem 4.4 in Atar et al. (2014) shows that the stationary distribution of the sequence of stochastic systems (the $M/GI/s_\lambda + GI$ under SJF with predicted service time $\hat{S}$ that is discrete with finite support) converges to the invariant state of the corresponding fluid model as $\lambda\to\infty$. We also note that the stationary queue length for class i (customers with $\hat{S}=a_i$) is stochastically dominated by the stationary number of customers in the system of an $M/GI/\infty$ queue with arrival rate λ_i and service time distribution F, which is the patience time distribution in our model. The latter has a Poisson distribution with rate $\lambda\int_0^\infty(1-F(x))dx = \lambda/\theta$. Thus the fluid-scaled stationary queue length for class i is uniformly integrable. Then from Theorem 3.3 in Atar et al. (2014), which characterizes the invariant state of the fluid model, we have for $i \le \kappa - 1$

$$\lim_{\lambda\to\infty}\frac{1}{\lambda}\limsup_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\int_0^T Q_i^\lambda(t)dt|\hat{S}=a_i\right] = \lim_{\lambda\to\infty}\frac{1}{\lambda}\liminf_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\int_0^T Q_i^\lambda(t)dt|\hat{S}=a_i\right] = 0,$$

$$\lim_{\lambda\to\infty}\frac{1}{\lambda}\limsup_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\int_0^T Q_i^\lambda(t)dt|\hat{S}=a_\kappa\right] = \lim_{\lambda\to\infty}\frac{1}{\lambda}\limsup_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\int_0^T Q_i^\lambda(t)dt|\hat{S}=a_\kappa\right] = h(a_\kappa)\int_0^{\bar{w}}(1-F(x))dx,$$

where $\bar{w}$ is the solution $F(\bar{w}) = \frac{\sum_{i=1}^\kappa \rho_i - 1}{\rho_\kappa}$. and for $i > \kappa$,

$$\lim_{\lambda\to\infty}\frac{1}{\lambda}\limsup_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\int_0^T Q_i^\lambda(t)dt|\hat{S}=a_i\right] = \lim_{\lambda\to\infty}\frac{1}{\lambda}\liminf_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\int_0^T Q_i^\lambda(t)dt|\hat{S}=a_i\right] = h(a_i)/\theta.$$

By Little's law, we have for $i \le \kappa - 1$

$$\lim_{\lambda\to\infty} \mathbb{E}[W_c^\lambda | \hat{S}_c = a_i] = \frac{0}{h(a_i)} = 0,$$

$$\lim_{\lambda\to\infty} \mathbb{E}[W_c^\lambda | \hat{S}_c = a_\kappa] = \frac{h(a_\kappa)\int_0^{\bar{w}}(1-F(x))dx}{h(a_\kappa)} = \int_0^{\bar{w}}(1-F(x))dx,$$

and for $i > \kappa$

$$\lim_{\lambda\to\infty} \mathbb{E}[W_c^\lambda | \hat{S}_c = a_i] = \frac{h(a_i)/\theta}{h(a_i)} = \frac{1}{\theta}.$$

■

Appendix C: Proof of Lemma 5

Proof. We first note that since $(\kappa^\Delta - 1)\Delta \le \tau \le \kappa^\Delta \Delta$ and $\kappa^\Delta\Delta - (\kappa^\Delta - 1)\Delta = \Delta \to 0$ as $\Delta \to 0$,

$$\lim_{\Delta\downarrow 0} \kappa^\Delta \Delta = \tau.$$

For part (a), since

$$\lim_{\lambda\to\infty} \mathbb{P}(\mathrm{Serv}^\lambda_{\pi_{\mathrm{SJF}^\Delta}} 1\{\hat{S} \le \tau\}) \ge \lim_{\lambda\to\infty} \mathbb{P}(\mathrm{Serv}^\lambda_{\pi_{\mathrm{SJF}^\Delta}} 1\{\hat{S} \le ((\kappa^\Delta - 1)\Delta\}) = \mathbb{P}(\hat{S} \le (\kappa^\Delta - 1)\Delta),$$

we have

$$\lim_{\lambda\to\infty} \mathbb{P}(\mathrm{Serv}^\lambda_{\pi_{\mathrm{SJF}^\Delta}} | \hat{S} \le \tau) = \frac{\lim_{\lambda\to\infty} \mathbb{P}(\mathrm{Serv}^\lambda_{\pi_{\mathrm{SJF}^\Delta}} 1\{\hat{S} \le \tau\})}{\mathbb{P}(\hat{S} \le \tau)} \ge \frac{\mathbb{P}(\hat{S} \le (\kappa^\Delta - 1)\Delta)}{\mathbb{P}(\hat{S} \le \tau)}.$$

Next since $(\kappa^\Delta - 1)\Delta \to \tau$ and $\hat{S}$ is a continuous random variable,

$$\lim_{\Delta\downarrow 0} \mathbb{P}(\hat{S} \le ((\kappa^\Delta - 1)\Delta) = \mathbb{P}(\hat{S} \le \tau) > 0,$$

then

$$\lim_{\Delta\downarrow 0} \lim_{\lambda\to\infty} \mathbb{P}(\mathrm{Serv}^\lambda_{\pi_{\mathrm{SJF}^\Delta}} | \hat{S} \le \tau) \ge \lim_{\Delta\downarrow 0} \frac{\mathbb{P}(\hat{S} \le (\kappa^\Delta - 1)\Delta)}{\mathbb{P}(\hat{S} \le \tau)} = 1.$$

Meanwhile, since $\lim_{\Delta\downarrow 0} \lim_{\lambda\to\infty} \mathbb{P}(\mathrm{Serv}^\lambda_{\pi_{\mathrm{SJF}^\Delta}} | \hat{S} \le \tau) \le 1$, we have

$$\lim_{\Delta\downarrow 0} \lim_{\lambda\to\infty} \mathbb{P}(\mathrm{Serv}^\lambda_{\pi_{\mathrm{SJF}^\Delta}} | \hat{S} \le \tau) = 1.$$

Similarly, since

$$\lim_{\lambda\to\infty} \mathbb{P}(\mathrm{Serv}^\lambda_{\pi_{\mathrm{SJF}^\Delta}} 1\{\hat{S} > \tau\}) \le \lim_{\lambda\to\infty} \mathbb{P}(\mathrm{Serv}^\lambda_{\pi_{\mathrm{SJF}^\Delta}} 1\{\hat{S} > (\kappa^\Delta - 1)\Delta\}) = \phi_\kappa^\Delta (H(\kappa^\Delta\Delta) - H((\kappa^\Delta - 1)\Delta)),$$

$$\lim_{\lambda\to\infty} \mathbb{P}(\mathrm{Serv}^\lambda_{\pi_{\mathrm{SJF}^\Delta}} | \hat{S} > \tau\}) = \frac{\lim_{\lambda\to\infty} \mathbb{P}(\mathrm{Serv}^\lambda_{\pi_{\mathrm{SJF}^\Delta}} 1\{\hat{S} > \tau\})}{\mathbb{P}(\hat{S} \le \tau)} \le \frac{\phi_\kappa^\Delta \mathbb{P}(\hat{S} \in ((\kappa^\Delta - 1)\Delta, \kappa^\Delta\Delta])}{\mathbb{P}(\hat{S} \le \tau)}.$$

Next, since

$$\lim_{\Delta\downarrow 0}\phi_\kappa^\Delta \mathbb{P}(\hat{S}\in((\kappa^\Delta-1)\Delta,\kappa^\Delta\Delta])=0,$$

$$\lim_{\Delta\downarrow 0}\lim_{\lambda\to\infty}\mathbb{P}(\mathrm{Serv}^\lambda_{\pi_{\mathrm{SJF}^\Delta}}|\hat{S}>\tau\})\le\lim_{\Delta\downarrow 0}\frac{\phi_\kappa^\Delta\mathbb{P}(\hat{S}\in((\kappa^\Delta-1)\Delta,\kappa^\Delta\Delta])}{\mathbb{P}(\hat{S}\le\tau)}=0.$$

Meanwhile, since $\lim_{\Delta\downarrow 0}\lim_{\lambda\to\infty}\mathbb{P}(\mathrm{Serv}^\lambda_{\pi_{\mathrm{SJF}^\Delta}}|\hat{S}>\tau\})\ge 0$, we have

$$\lim_{\Delta\downarrow 0}\lim_{\lambda\to\infty}\mathbb{P}(\mathrm{Serv}^\lambda_{\pi_{\mathrm{SJF}^\Delta}}|\hat{S}>\tau\})=0.$$

We next prove part (b). Note that

$$\begin{aligned}\lim_{\lambda\to\infty}\mathbb{E}[W^\lambda_{\pi_{\mathrm{SJF}^\Delta}}1\{\hat{S}\le\tau\}]&\le\lim_{\lambda\to\infty}\mathbb{E}[W^\lambda_{\pi_{\mathrm{SJF}^\Delta}}1\{\hat{S}\le\kappa^\Delta\Delta\}\}]\\&=\int_0^{\bar{w}^\Delta}(1-F(x))dx\left(H(\kappa^\Delta\Delta)-H((\kappa^\Delta-1)\Delta)\right).\end{aligned}$$

Then, we have

$$\lim_{\lambda\to\infty}\mathbb{E}[W^\lambda_{\pi_{\mathrm{SJF}^\Delta}}|\hat{S}\le\tau]\le\frac{\int_0^{\bar{w}^\Delta}(1-F(x))dx\left(H(\kappa^\Delta\Delta)-H((\kappa^\Delta-1)\Delta)\right)}{P(\hat{S}\le\tau)}.$$

Since $\int_0^{\bar{w}^\Delta}(1-F(x))dx\le\int_0^\infty(1-F(x))dx=1/\theta$ and $H(\kappa^\Delta\Delta)-H((\kappa^\Delta-1)\Delta)\to 0$ as $\Delta\to 0$,

$$\lim_{\Delta\downarrow 0}\lim_{\lambda\to\infty}\mathbb{E}[W^\lambda_{\pi_{\mathrm{SJF}^\Delta}}|\hat{S}\le\tau]\le\lim_{\Delta\downarrow 0}\frac{\int_0^{\bar{w}^\Delta}(1-F(x))dx\left(H(\kappa^\Delta\Delta)-H((\kappa^\Delta-1)\Delta)\right)}{\mathbb{P}(\hat{S}\le\tau)}=0.$$

Since $\mathbb{E}[W^\lambda_{\pi_{\mathrm{SJF}^\Delta}}|\hat{S}\le\tau]\ge 0$, we have

$$\lim_{\Delta\downarrow 0}\lim_{\lambda\to\infty}\mathbb{E}[W^\lambda_{\pi_{\mathrm{SJF}^\Delta}}|\hat{S}\le\tau]=0.$$

Similarly, since

$$\begin{aligned}\lim_{\lambda\to\infty}\mathbb{E}[W^\lambda_{\pi_{\mathrm{SJF}^\Delta}}1\{\hat{S}>\tau\}]&\ge\lim_{\lambda\to\infty}\mathbb{E}[W^\lambda_{\pi_{\mathrm{SJF}^\Delta}}1\{\hat{S}>\kappa^\Delta\Delta\}\}]\\&=\mathbb{P}(\hat{S}>\kappa^\Delta\Delta)/\theta,\end{aligned}$$

we have

$$\lim_{\Delta\downarrow 0}\lim_{\lambda\to\infty}\mathbb{E}[W^\lambda_{\pi_{\mathrm{SJF}^\Delta}}|\hat{S}>\tau]\ge\lim_{\Delta\downarrow 0}\frac{\mathbb{P}(\hat{S}>\kappa^\Delta\Delta)/\theta}{\mathbb{P}(\hat{S}>\tau)}=1/\theta.$$

Let T denote the patience time of the customer. Then $W^\lambda_{\pi_{\mathrm{SJF}^\Delta}}\le T$ and T is independent of $\hat{S}$. This implies that

$$\mathbb{E}[W^\lambda_{\pi_{\mathrm{SJF}^\Delta}}|\hat{S}>\tau]\le\mathbb{E}[T|\hat{S}>\tau]=\mathbb{E}[T]=1/\theta.$$

Thus,

$$\lim_{\Delta\downarrow 0}\lim_{\lambda\to\infty}\mathbb{E}[W^\lambda_{\pi_{\mathrm{SJF}^\Delta}}|\hat{S}>\tau]=1/\theta.$$

■

Appendix D: Proof of Theorem 2

Proof. Let $(S, \hat{S})$ denote a generic actual and predicted service-time pair. Define $\tilde{\mu}(y) \coloneqq \mathbb{E}[S|\hat{S} = y]$. Let $g(x|y)$ denote the conditional density of S given $\hat{S} = y$, and $G(x|y)$ and $\bar{G}(x|y)$ denote the corresponding cdf and ccdf, respectively. Recall that $h(y)$ is the pdf/pmf of $\hat{S}$ and $H(y)$ is its cdf. We also denote $\bar{H}(y) = 1 - H(y)$. Lastly, we denote $G(x, y) = \mathbb{P}(S \leq x, \hat{S} \leq y)$ as the joint cdf of S and $\hat{S}$ and $\bar{G}(x, y) = \mathbb{P}(S > x, \hat{S} > y)$. Note that

$$\tilde{\mu}(y) = \int_0^\infty x f(x|y) dx = \int_0^\infty \bar{G}(x|y) dx.$$

We first study the case where $\hat{S}$ is a continuous random variable. For a fixed threshold τ, the workload of the higher priority ($\hat{S} \leq \tau$) and lower priority ($\hat{S} > \tau$) classes can be written as follows, where the exchange of integral is due to Tonelli's theorem:

$$\begin{aligned}\lambda \int_0^\tau \tilde{\mu}(y) h(y) dy &= \lambda \int_0^\tau \left(\int_0^\infty \bar{G}(x|y) dx \right) h(y) dy \\ &= \lambda \int_0^\infty \left(\int_0^\tau \bar{G}(x|y) h(y) dy \right) dx = \lambda \int_0^\infty \mathbb{P}(S > x, \hat{S} \leq \tau) dx\end{aligned}$$

and

$$\lambda \int_\tau^\infty \tilde{\mu}(y) h(y) dy = \lambda \int_0^\infty \left(\int_\tau^\infty \bar{G}(x|y) h(y) dy \right) dx = \lambda \int_0^\infty \mathbb{P}(S > x, \hat{S} > \tau) dx = \lambda \int_0^\infty \bar{G}(x, \tau) dx.$$

Next, we consider two different service-time predictions $\hat{S}_1$ and $\hat{S}_2$. We assume $\hat{S}_1$ and $\hat{S}_2$ have the same marginal distribution, h, and $\bar{G}_1(x, y) \leq \bar{G}_2(x, y)$ for all (x, y), i.e., $(S, \hat{S}_1) \leq_{PQD} (S, \hat{S}_2)$. Then, for any fixed τ,

$$\int_0^\infty \bar{G}_1(x, \tau) dx \leq \int_0^\infty \bar{G}_2(x, \tau) dx.$$

We let τ_1 and τ_2 be the thresholds corresponding to the $\hat{S}_1$ and $\hat{S}_2$ service-time predictions, i.e., obtained using (4). Then, we note from (4) that

$$\lambda \mathbb{E}[S \mathbf{1}(\hat{S} \leq \tau)] = \lambda \int_0^\infty \mathbb{P}(S > x, \hat{S} \leq \tau) = \lambda \int_0^\infty \Big(\mathbb{P}(S > x) - \mathbb{P}(S > x, \hat{S} > \tau) \Big) dx = s_\lambda.$$

This leads to:

$$\int_0^\infty \bar{G}_1(x, \tau_1) dx = \int_0^\infty \bar{G}_2(x, \tau_2) dx = \mathbb{E}[S] - \frac{s_\lambda}{\lambda},$$

which must mean that $\tau_1 \leq \tau_2$. Next, since $\hat{S}_1$ and $\hat{S}_2$ have the same marginal distribution h, we must have that $H(\tau_1) \leq H(\tau_2)$, which implies that $\bar{\text{Th}}^{SJF}(S, \hat{S}_1) \leq \bar{\text{Th}}^{SJF}(S, \hat{S}_2)$.

We next consider the case where $\hat{S}$ is a discrete random variable with a support $\{a_1, \ldots, a_K\}$. For any $\kappa > 0$, we have

$$\lambda \sum_{i=\kappa+1}^{K} h(a_i)\tilde{\mu}(a_i) = \lambda \sum_{i=\kappa+1}^{\infty} h(a_i) \int_0^{\infty} \bar{G}(x|a_i)dx = \lambda \int_0^{\infty} \bar{G}(x, a_\kappa)dx.$$

Note that in the discrete case, $G(x, a_\kappa) = \mathbb{P}(S > x, \hat{S} > a_\kappa) = \mathbb{P}(S > x, \hat{S} \geq a_{\kappa+1})$.

Recall that κ is the smallest index such that $\mathbb{E}[S1\{\hat{S} \leq a_\kappa\}] \geq s_\lambda/\lambda$. Since $\mathbb{E}[S] = \mathbb{E}[S1\{\hat{S} \leq a_\kappa\}] + \mathbb{E}[S1\{\hat{S} > a_\kappa\}]$, κ is the smallest index such that

$$\int_0^{\infty} \bar{G}(x, a_\kappa)dx \leq \mathbb{E}[S] - s_\lambda/\lambda.$$

Consider two different discrete service predictors, $\hat{S}_1$ and $\hat{S}_2$, that have the same marginal distribution h and $\bar{G}_1(x, y) \leq \bar{G}_2(x, y)$. Let κ_1 and κ_2 be the thresholds defined in (2) corresponding to the service-time predictions $\hat{S}_1$ and $\hat{S}_2$ respectively. We also write $\tilde{\mu}_j(a_k) = \mathbb{E}[S|\hat{S}_i = a_k]$ for $j = 1, 2$. Since

$$\int_0^{\infty} \bar{G}_1(x, a_\kappa)dx \leq \int_0^{\infty} \bar{G}_2(x, a_\kappa)dx,$$

we have $\kappa_1 \leq \kappa_2$. If $\kappa_1 < \kappa_2$, then we have

$$\sum_{i=1}^{\kappa_1-1} h(a_i) + \phi_{1,\kappa_1} h(a_{\kappa_1}) \leq \sum_{i=1}^{\kappa_2-1} h(a_i) \leq \sum_{i=1}^{\kappa_2-1} h(a_i) + \phi_{2,\kappa_2} h(a_{\kappa_2}),$$

where ϕ_{j,κ_j} is the defined in (3) for the service-time prediction $\hat{S}_j$, $j = 1, 2$. This implies that $\bar{\text{Th}}^{SJF}(S, \hat{S}_1) \leq \bar{\text{Th}}^{SJF}(S, \hat{S}_2)$.

If $\kappa_1 = \kappa_2$, then we first note that since $\int_0^{\infty} \bar{G}_1(x, a_{\kappa_1-1})dx \leq \int_0^{\infty} \bar{G}_2(x, a_{\kappa_1-1})dx$ and $\mathbb{E}[S] = \sum_{i=1}^{\kappa_1-1} h(a_i)\tilde{\mu}_1(a_i) + \int_0^{\infty} \bar{G}_1(x, a_{\kappa_1-1})dx = \sum_{i=1}^{\kappa_1-1} h(a_i)\tilde{\mu}_2(a_i) + \int_0^{\infty} \bar{G}_2(x, a_{\kappa_1-1})dx$,

$$\sum_{i=1}^{\kappa_1-1} h(a_i)\tilde{\mu}_1(a_i) \geq \sum_{i=1}^{\kappa_1-1} h(a_i)\tilde{\mu}_2(a_i).$$

Denote $\delta_1 = \sum_{i=1}^{\kappa_1-1} h(a_i)\tilde{\mu}_1(a_i) - \sum_{i=1}^{\kappa_1-1} h(a_i)\tilde{\mu}_2(a_i) \geq 0$ and $C_1 = \sum_{i=1}^{\kappa_1-1} h(a_i)\tilde{\mu}_2(a_i)$. In addition, since $\int_0^{\infty} \bar{G}_1(x, a_{\kappa_1})dx \leq \int_0^{\infty} \bar{G}_2(x, a_{\kappa_1})dx$,

$$\sum_{i=\kappa_1+1}^{K} h(a_i)\tilde{\mu}_1(a_i) \leq \sum_{i=\kappa_1+1}^{K} h(a_i)\tilde{\mu}_2(a_i).$$

Denote $\delta_2 = \sum_{i=\kappa_1+1}^{K} h(a_i)\tilde{\mu}_2(a_i) - \sum_{i=\kappa_1+1}^{K} h(a_i)\tilde{\mu}_1(a_i) \geq 0$ and $C_2 = \sum_{i=\kappa_1+1}^{K} h(a_i)\tilde{\mu}_2(a_i)$. Next, since

$$h(a_{\kappa_1})\tilde{\mu}_j(\kappa_1) = \mathbb{E}[S] - \sum_{i=1}^{\kappa_1-1} h(a_i)\tilde{\mu}_j(a_i) - \sum_{i=\kappa_1+1}^{K} h(a_i)\tilde{\mu}_j(a_i),$$

we have

$$\begin{aligned}
\phi_{1,\kappa_1} &= \frac{s_\lambda/\lambda - \sum_{i=1}^{\kappa_1-1} h(a_i)\tilde{\mu}_1(a_i)}{h(a_{\kappa_1})\tilde{\mu}_1(a_{\kappa_1})} \\
&= \frac{s_\lambda/\lambda - (C_1+\delta_1)}{\mathbb{E}[S] - (C_1+\delta_1) - (C_2-\delta_2)} \\
&\leq \frac{s_\lambda/\lambda - (C_1+\delta_1)}{\mathbb{E}[S] - (C_1+\delta_1) - C_2} \quad \text{since } \delta_2 \geq 0 \\
&\leq \frac{s_\lambda/\lambda - C_1}{\mathbb{E}[S] - C_1 - C_2} \quad \text{since } \delta_1 \geq 0 \\
&= \phi_{2,\kappa_1}.
\end{aligned}$$

Thus,

$$\sum_{i=1}^{\kappa_1-1} h(a_i) + \phi_{1,\kappa_1} h(a_{\kappa_1}) \leq \sum_{i=1}^{\kappa_1-1} h(a_i) + \phi_{2,\kappa_1} h(a_{\kappa_1}),$$

i.e., $\bar{\text{Th}}^{SJF}(S,\hat{S}_1) \leq \bar{\text{Th}}^{SJF}(S,\hat{S}_2)$. ∎

Appendix E: Proof of Lemma 6

Proof. Assume that $(S,\hat{S}_1) \leq_{PQD} (S,\hat{S}_2)$. Note that the PQD order is preserved under monotonically increasing transformations, so that $(Z,\hat{Z}_1) \leq_{PQD} (Z,\hat{Z}_2)$. By Lemma 1 in (Wu et al. 2019), it follows that $r[Z,\hat{Z}_1] \leq r[Z,\hat{Z}_1]$. For the converse, assume that $r[Z,\hat{Z}_1] \leq r[Z,\hat{Z}_2]$. Then, by Lemma 3 of Wu et al. (2019), it holds that $(Z,\hat{Z}_1) \leq_{PQD} (Z,\hat{Z}_2)$. Thus, $(S,\hat{S}_1) \leq_{PQD} (S,\hat{S}_2)$. ∎

Appendix F: Proof of Proposition 2

Proof. Our proof builds on the strong consistency and Central Limit Theorem for Polyak-Ruppert averaging, i.e., Theorem 2 in Polyak and Juditsky (1992). In particular, we only need to verify that for the stochastic root-finding problem $g(t)=0$, the conditions for Theorem 2 in Polyak and Juditsky (1992) hold.

First, since $g(0) = -\frac{s}{\lambda} < 0$, $g(M_H) = \mathbb{E}[S] - \frac{s}{\lambda} > 0$, and $g'(t) = \mu(t)h(t) \geq c > 0$ for $t \in (0, M_H)$, $g(t)=0$ has a unique solution $\tau \in (0, M_H)$. In addition, since for any $0 < s < t < M_H$,

$$|g(t) - g(s)| = \left|\mathbb{E}[S1\{\hat{S} \leq t\}] - \mathbb{E}[S1\{\hat{S} \leq s\}]\right| = \left|\int_s^t \mu(y)h(y)dy\right| \leq C|t-s|,$$

$g(t)$ is Lipschitz continuous.

Second,

$$\mathbb{E}[(S1\{\hat{S} \leq t\} - U)^2] \leq \mathbb{E}[S^2] + \mathbb{E}[U^2]$$

and for any $0 < s < t < M_H$,

$$\mathbb{E}[(S1\{\hat{S} \le t\} - U)^2] - \mathbb{E}[(S1\{\hat{S} \le s\} - U)^2] = \int_s^t \mathbb{E}[S^2|\hat{S} = y]h(y)dy - 2\mathbb{E}[U]\int_s^t \mu(y)h(y)dy$$
$$\le C(1 + 2\mathbb{E}[U])|t - s|,$$

i.e., $\mathbb{E}[(S1\{\hat{S} \le t\} - U)^2]$ is Lipschitz continuous in t.

Above all, the conditions in Theorem 2 of Polyak and Juditsky (1992) are satisfied. Thus, we have the strong consistency and central limit theorem results. ∎

Appendix G: Supporting Tables and Figures

In this appendix, we present tables and figures with numerical results that provide further support to the numerical study in Section 7. In particular, in Table 2, we present point estimates of the throughput for a multi-class priority policy where the number of classes increases by rendering the high-priority class ($\hat{S} < \tau$) more granular. In Tables 3-20, we provide support to Section 7.3 by comparing the performance between SJF and the two-class priority rule π_0 in the $GI/LGN/s + M$ model with a varying number of servers, s, and fixing $\rho = 1.4$. We let s range from $s = 20$ to $s = 1000$, and consider exponential, Erlang E_2, and hyperexponential H_2 interarrival time distributions. In Tables 21-23, we provide support to Section 7.4.1 by exploring the effect of selecting the "wrong" threshold in the two-class priority rule. In Tables 24 - 26 we present detailed results supporting Figure 2, where we test performance under two-class priority rules with various thresholds. In Tables 27-29, we report supporting numerical results for Section 7.4.2 where we explore our data-driven approach to estimate τ. All the tables mentioned above consider the service-time model of Section 7 in the paper. In Tables 30 and 31 and Figure 3, we present results corresponding to the service-time model of Section 5.2 in the paper.

α	$r[S_i, \hat{S}_i(\alpha)]$	$r[Z_i, \hat{Z}_i(\alpha)]$	FCFS/LCFS	SJF	2 classes	3 classes	5 classes	10 classes
0.001	0.0282	0.0316	0.7143	0.7230	0.7220	0.7220	0.7220	0.7220
0.005	0.0636	0.0707	0.7143	0.7327	0.7300	0.7300	0.7300	0.7305
0.01	0.0877	0.1000	0.7143	0.7397	0.7356	0.7356	0.7357	0.7368
0.02	0.1151	0.1414	0.7143	0.7493	0.7433	0.7433	0.7438	0.7460
0.05	0.1899	0.2236	0.7143	0.7677	0.7587	0.7580	0.7608	0.7633
0.1	0.2688	0.3162	0.7143	0.7870	0.7740	0.7744	0.7796	0.7819
0.2	0.3894	0.4472	0.7143	0.8114	0.7941	0.7975	0.8037	0.8053
0.3	0.4767	0.5477	0.7143	0.8276	0.8070	0.8143	0.8198	0.8213
0.4	0.5691	0.6325	0.7143	0.8403	0.8171	0.8277	0.8325	0.8339
0.5	0.6413	0.7071	0.7143	0.8504	0.8251	0.8386	0.8430	0.8443
0.6	0.7154	0.7746	0.7143	0.8587	0.8314	0.8475	0.8516	0.8527
0.7	0.7878	0.8367	0.7143	0.8659	0.8369	0.8550	0.8590	0.8601
0.8	0.8602	0.8944	0.7143	0.8714	0.8411	0.8610	0.8648	0.8659
0.9	0.9333	0.9487	0.7143	0.8763	0.8449	0.8659	0.8701	0.8712
0.95	0.9641	0.9747	0.7143	0.8787	0.8466	0.8678	0.8725	0.8737
0.98	0.9866	0.9899	0.7143	0.8801	0.8474	0.8688	0.8738	0.8751

Table 2 **Long-run throughput in the $M/LGN/100+M$ model with $\rho = 1.4$ where we divide the high class ($< \tau$) into equally-sized classes.**

			SJF					Two class		
α	r_S	r_Z	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{P}[Serv\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.0282	0.0316	0.1278	7.8799	0.2820	0.8473	0.1519	0.1653	7.9186	0.2834
0.005	0.0636	0.0707	0.1274	7.6873	0.2753	0.8479	0.1515	0.1640	7.7899	0.2792
0.01	0.0877	0.1000	0.1275	7.5972	0.2714	0.8479	0.1520	0.1635	7.6878	0.2757
0.02	0.1151	0.1414	0.1269	7.4058	0.2643	0.8486	0.1517	0.1626	7.5570	0.2707
0.05	0.1899	0.2236	0.1263	7.0310	0.2510	0.8503	0.1497	0.1610	7.3198	0.2621
0.1	0.2688	0.3162	0.1246	6.6656	0.2383	0.8508	0.1492	0.1596	7.0796	0.2534
0.2	0.3894	0.4472	0.1227	6.1890	0.2214	0.8527	0.1472	0.1579	6.7627	0.2413
0.3	0.4767	0.5477	0.1213	5.8858	0.2099	0.8536	0.1467	0.1572	6.5568	0.2341
0.4	0.5691	0.6325	0.1192	5.6276	0.2006	0.8555	0.1449	0.1568	6.4022	0.2288
0.5	0.6413	0.7071	0.1175	5.3845	0.1921	0.8574	0.1431	0.1562	6.2689	0.2236
0.6	0.7154	0.7746	0.1165	5.2308	0.1865	0.8578	0.1423	0.1559	6.1534	0.2195
0.7	0.7878	0.8367	0.1147	5.0559	0.1807	0.8591	0.1409	0.1552	6.0390	0.2154
0.8	0.8602	0.8944	0.1132	4.9112	0.1753	0.8606	0.1394	0.1556	5.9817	0.2131
0.9	0.9333	0.9487	0.1123	4.8034	0.1713	0.8615	0.1387	0.1555	5.9217	0.2110
0.95	0.9641	0.9747	0.1116	4.7389	0.1694	0.8621	0.1381	0.1558	5.8850	0.2102
0.98	0.9866	0.9899	0.1112	4.7026	0.1677	0.8631	0.1373	0.1559	5.8922	0.2102

Table 3 **Accuracy of the two-class approximation in the $M/LGN/20+M$ system with $\rho=1.4$. We let r_S denote $r[S_i,\hat{S}_i(\alpha)]$ and r_Z denote $r[Z_i,\hat{Z}_i(\alpha)]$.**

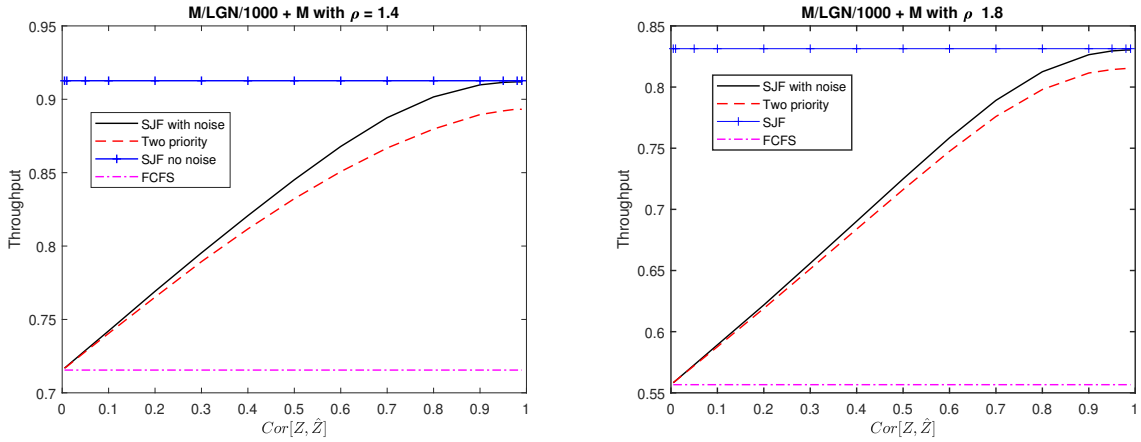

Figure 3 **Long-run throughput, in steady state, in the $M/LGN/1000+M$ model under the service-time model of Section 5.2.**

			SJF					Two class		
α	r_S	r_Z	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{P}[Serv\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.0282	0.0316	0.0936	19.4651	0.2783	0.8976	0.1024	0.1420	19.5063	0.2791
0.005	0.0636	0.0707	0.0940	18.8228	0.2694	0.8980	0.1019	0.1401	18.9993	0.2722
0.01	0.0877	0.1000	0.0943	18.3953	0.2629	0.8984	0.1017	0.1392	18.6868	0.2676
0.02	0.1151	0.1414	0.0944	17.7968	0.2542	0.8986	0.1014	0.1379	18.2297	0.2610
0.05	0.1899	0.2236	0.0945	16.6173	0.2375	0.8992	0.1007	0.1356	17.3616	0.2484
0.1	0.2688	0.3162	0.0947	15.4222	0.2205	0.8995	0.1006	0.1335	16.4520	0.2352
0.2	0.3894	0.4472	0.0943	13.9603	0.1993	0.9000	0.1004	0.1308	15.3454	0.2191
0.3	0.4767	0.5477	0.0936	12.9847	0.1850	0.9002	0.1003	0.1294	14.6476	0.2089
0.4	0.5691	0.6325	0.0925	12.2118	0.1742	0.9002	0.1000	0.1282	14.0640	0.2006
0.5	0.6413	0.7071	0.0915	11.5702	0.1651	0.9002	0.0998	0.1275	13.6136	0.1940
0.6	0.7154	0.7746	0.0905	11.0284	0.1576	0.9005	0.0993	0.1271	13.2339	0.1885
0.7	0.7878	0.8367	0.0894	10.5843	0.1510	0.9012	0.0989	0.1266	12.9298	0.1842
0.8	0.8602	0.8944	0.0885	10.2153	0.1456	0.9014	0.0987	0.1269	12.7138	0.1809
0.9	0.9333	0.9487	0.0876	9.8765	0.1409	0.9017	0.0986	0.1267	12.4845	0.1779
0.95	0.9641	0.9747	0.0870	9.7176	0.1388	0.9020	0.0983	0.1266	12.3774	0.1768
0.98	0.9866	0.9899	0.0864	9.6256	0.1374	0.9025	0.0978	0.1264	12.3194	0.1757

Table 4 **Accuracy of the two-class approximation in the $M/LGN/50+M$ system with $\rho=1.4$. We let r_S denote $r[S_i,\hat{S}_i(\alpha)]$ and r_Z denote $r[Z_i,\hat{Z}_i(\alpha)]$.**

			SJF					Two class		
α	r_S	r_Z	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{P}[Serv\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.0282	0.0316	0.0807	27.1550	0.2773	0.9127	0.0870	0.1323	27.2593	0.2783
0.005	0.0636	0.0707	0.0811	26.2476	0.2681	0.9132	0.0867	0.1303	26.4921	0.2710
0.01	0.0877	0.1000	0.0815	25.6401	0.2614	0.9134	0.0866	0.1292	26.0129	0.2659
0.02	0.1151	0.1414	0.0815	24.7235	0.2522	0.9134	0.0864	0.1280	25.3404	0.2587
0.05	0.1899	0.2236	0.0821	23.0294	0.2348	0.9140	0.0860	0.1255	24.0218	0.2451
0.1	0.2688	0.3162	0.0826	21.2272	0.2164	0.9143	0.0859	0.1233	22.6268	0.2308
0.2	0.3894	0.4472	0.0825	18.9647	0.1934	0.9152	0.0852	0.1201	20.8776	0.2130
0.3	0.4767	0.5477	0.0822	17.4954	0.1783	0.9150	0.0853	0.1181	19.7510	0.2014
0.4	0.5691	0.6325	0.0819	16.3368	0.1666	0.9147	0.0856	0.1170	18.8728	0.1923
0.5	0.6413	0.7071	0.0815	15.3890	0.1570	0.9143	0.0857	0.1163	18.1527	0.1853
0.6	0.7154	0.7746	0.0808	14.6181	0.1490	0.9145	0.0856	0.1158	17.5697	0.1791
0.7	0.7878	0.8367	0.0804	13.9946	0.1425	0.9144	0.0859	0.1152	17.1118	0.1742
0.8	0.8602	0.8944	0.0795	13.4497	0.1371	0.9142	0.0859	0.1154	16.7522	0.1704
0.9	0.9333	0.9487	0.0787	12.9479	0.1320	0.9145	0.0856	0.1155	16.4372	0.1672
0.95	0.9641	0.9747	0.0778	12.6766	0.1297	0.9146	0.0853	0.1156	16.2680	0.1658
0.98	0.9866	0.9899	0.0777	12.5777	0.1284	0.9148	0.0852	0.1155	16.1964	0.1649

Table 5 **Accuracy of the two-class approximation in the $M/LGN/70+M$ system with $\rho=1.4$. We let r_S denote $r[S_i,\hat{S}_i(\alpha)]$ and r_Z denote $r[Z_i,\hat{Z}_i(\alpha)]$.**

			SJF					Two class		
α	r_S	r_Z	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{P}[Serv\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.0282	0.0316	0.0682	38.7296	0.2770	0.9266	0.0731	0.1212	38.8721	0.2780
0.005	0.0636	0.0707	0.0684	37.3598	0.2673	0.9272	0.0728	0.1194	37.6858	0.2700
0.01	0.0877	0.1000	0.0687	36.4471	0.2603	0.9275	0.0726	0.1181	36.9477	0.2644
0.02	0.1151	0.1414	0.0690	35.0869	0.2507	0.9272	0.0726	0.1167	35.9017	0.2567
0.05	0.1899	0.2236	0.0696	32.4679	0.2323	0.9278	0.0719	0.1140	33.7302	0.2413
0.1	0.2688	0.3162	0.0701	29.7250	0.2128	0.9283	0.0716	0.1110	31.5432	0.2260
0.2	0.3894	0.4472	0.0709	26.3390	0.1881	0.9290	0.0713	0.1078	28.7060	0.2059
0.3	0.4767	0.5477	0.0709	23.9961	0.1714	0.9287	0.0715	0.1061	26.9172	0.1930
0.4	0.5691	0.6325	0.0705	22.1714	0.1586	0.9286	0.0717	0.1048	25.5386	0.1829
0.5	0.6413	0.7071	0.0704	20.7567	0.1485	0.9281	0.0719	0.1040	24.4331	0.1749
0.6	0.7154	0.7746	0.0705	19.6973	0.1407	0.9279	0.0722	0.1035	23.5301	0.1686
0.7	0.7878	0.8367	0.0704	18.8018	0.1340	0.9276	0.0727	0.1030	22.7404	0.1631
0.8	0.8602	0.8944	0.0700	18.0100	0.128	0.9271	0.0730	0.1028	22.1475	0.1589
0.9	0.9333	0.9487	0.0698	17.3162	0.1236	0.9268	0.0732	0.1028	21.6270	0.1551
0.95	0.9641	0.9747	0.0695	16.9740	0.1211	0.9268	0.0733	0.1027	21.4076	0.1534
0.98	0.9866	0.9899	0.0692	16.7739	0.1198	0.9270	0.0730	0.1032	21.3123	0.1526

Table 6 Accuracy of the two-class approximation in the $M/LGN/100+M$ system with $\rho=1.4$. We let r_S denote $r[S_i,\hat{S}_i(\alpha)]$ and r_Z denote $r[Z_i,\hat{Z}_i(\alpha)]$.

			SJF					Two class		
α	r_S	r_Z	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{P}[Srv\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|Srv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.0282	0.0316	0.0313	193.0422	0.2760	0.9670	0.0337	0.0760	193.5315	0.2766
0.005	0.0636	0.0707	0.0314	185.5035	0.2654	0.9673	0.0327	0.0743	186.7599	0.2673
0.01	0.0877	0.1000	0.0313	180.6657	0.2580	0.9674	0.0327	0.0733	182.0929	0.2603
0.02	0.1151	0.1414	0.0314	173.4318	0.2474	0.9677	0.0325	0.0721	175.6152	0.2509
0.05	0.1899	0.2236	0.0316	158.5673	0.2268	0.9681	0.0317	0.0693	162.1856	0.2322
0.1	0.2688	0.3162	0.0320	143.2425	0.2051	0.9687	0.0314	0.0676	148.9357	0.2133
0.2	0.3894	0.4472	0.0321	123.8862	0.1770	0.9692	0.0310	0.0639	131.2422	0.1880
0.3	0.4767	0.5477	0.0323	110.3833	0.1576	0.9695	0.0308	0.0618	119.6017	0.1711
0.4	0.5691	0.6325	0.0327	99.7599	0.1425	0.9693	0.0309	0.0611	110.6577	0.1580
0.5	0.6413	0.7071	0.0331	91.2479	0.1302	0.9692	0.0309	0.0600	103.1668	0.1474
0.6	0.7154	0.7746	0.0335	84.4268	0.1204	0.9692	0.0308	0.0588	96.4971	0.1385
0.7	0.7878	0.8367	0.0337	78.7296	0.1120	0.9690	0.0312	0.0583	91.2719	0.1312
0.8	0.8602	0.8944	0.0341	73.5779	0.1048	0.9683	0.0317	0.0584	87.1753	0.1251
0.9	0.9333	0.9487	0.0344	68.8307	0.0983	0.9679	0.0319	0.0584	83.6712	0.1199
0.95	0.9641	0.9747	0.0347	66.9124	0.0954	0.9676	0.0321	0.0586	82.2796	0.1177
0.98	0.9866	0.9899	0.0346	65.6831	0.0937	0.9674	0.0324	0.0589	81.2640	0.1163

Table 7 **Accuracy of the two-class approximation in the $M/LGN/500+M$ system with $\rho=1.4$. We let r_S denote $r[S_i,\hat{S}_i(\alpha)]$ and r_Z denote $r[Z_i,\hat{Z}_i(\alpha)]$.**

			SJF					Two class		
α	r_S	r_Z	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{P}[Srv\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|Srv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.0282	0.0316	0.0221	385.1895	0.2753	0.9771	0.0230	0.0605	385.9085	0.2759
0.005	0.0636	0.0707	0.0222	369.7601	0.2647	0.9770	0.0230	0.0591	371.5782	0.2661
0.01	0.0877	0.1000	0.0222	359.7233	0.2570	0.9772	0.0229	0.0578	362.3138	0.2589
0.02	0.1151	0.1414	0.0222	345.4297	0.2465	0.9773	0.0228	0.0561	348.9148	0.2492
0.05	0.1899	0.2236	0.0228	315.7488	0.2257	0.9776	0.0225	0.0550	321.1919	0.2299
0.1	0.2688	0.3162	0.0228	284.5682	0.2038	0.9781	0.0220	0.0526	293.4705	0.2099
0.2	0.3894	0.4472	0.0226	244.5212	0.1752	0.9783	0.0218	0.0501	256.1455	0.1833
0.3	0.4767	0.5477	0.0231	217.6770	0.1555	0.9784	0.0217	0.0479	231.6201	0.1657
0.4	0.5691	0.6325	0.0233	196.1077	0.1401	0.9785	0.0215	0.0471	212.4476	0.1518
0.5	0.6413	0.7071	0.0236	178.0930	0.1273	0.9784	0.0217	0.0460	196.7703	0.1404
0.6	0.7154	0.7746	0.0239	164.0412	0.1171	0.9785	0.0215	0.0457	183.0561	0.1312
0.7	0.7878	0.8367	0.0241	152.3888	0.1085	0.9783	0.0218	0.0448	171.2986	0.1230
0.8	0.8602	0.8944	0.0244	141.5950	0.1008	0.9778	0.0223	0.0444	161.9419	0.1161
0.9	0.9333	0.9487	0.0245	131.6309	0.0937	0.9778	0.0223	0.0446	154.1588	0.1104
0.95	0.9641	0.9747	0.024	127.2698	0.0907	0.9775	0.0224	0.0441	150.1600	0.1074
0.98	0.9866	0.9899	0.0249	124.7826	0.0889	0.9773	0.0226	0.0446	148.4281	0.1062

Table 8 Accuracy of the two-class approximation in the $M/LGN/1000+M$ system with $\rho=1.4$. We let r_S denote $r[S_i,\hat{S}_i(\alpha)]$ and r_Z denote $r[Z_i,\hat{Z}_i(\alpha)]$.

			SJF					Two class		
α	r_S	r_Z	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{P}[Srv\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|Srv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.028	0.032	0.1203	7.6000	0.2805	0.8574	0.1418	0.1613	7.6472	0.2823
0.005	0.064	0.071	0.1201	7.3771	0.2727	0.8584	0.1411	0.1598	7.4789	0.2764
0.010	0.088	0.100	0.1200	7.2389	0.2674	0.8593	0.1407	0.1589	7.3836	0.2730
0.020	0.115	0.141	0.1195	7.0458	0.2604	0.8595	0.1403	0.1578	7.2380	0.2676
0.050	0.190	0.224	0.1189	6.6677	0.2470	0.8605	0.1394	0.1560	6.9678	0.2579
0.100	0.269	0.316	0.1181	6.2859	0.2334	0.8616	0.1384	0.1538	6.6837	0.2477
0.200	0.389	0.447	0.1158	5.7770	0.2151	0.8640	0.1362	0.1515	6.3259	0.2346
0.300	0.477	0.548	0.1142	5.4469	0.2030	0.8652	0.1353	0.1503	6.1087	0.2268
0.400	0.569	0.632	0.1130	5.1831	0.1936	0.8660	0.1343	0.1498	5.9432	0.2204
0.500	0.641	0.707	0.1116	4.9742	0.1857	0.8669	0.1335	0.1490	5.7870	0.2148
0.600	0.715	0.775	0.1101	4.7940	0.1794	0.8678	0.1325	0.1483	5.6535	0.2101
0.700	0.788	0.837	0.1085	4.6211	0.1733	0.8691	0.1309	0.1483	5.5707	0.2073
0.800	0.860	0.894	0.1075	4.4943	0.1682	0.8700	0.1304	0.1485	5.5082	0.2045
0.900	0.933	0.949	0.1063	4.3651	0.1639	0.8709	0.1293	0.1484	5.4355	0.2019
0.950	0.964	0.975	0.1055	4.3139	0.1620	0.8714	0.1286	0.1479	5.3853	0.2008
0.980	0.987	0.990	0.1051	4.2719	0.1603	0.8723	0.1280	0.1475	5.3586	0.1993

Table 9 **Accuracy of the two-class approximation in the $E_2/LGN/20+M$ system with $\rho=1.4$. We let r_S denote $r[S_i,\hat{S}_i(\alpha)]$ and r_Z denote $r[Z_i,\hat{Z}_i(\alpha)]$.**

			SJF					Two class		
α	r_S	r_Z	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{P}[Srv\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|Srv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.028	0.032	0.0843	19.1663	0.2776	0.9067	0.0932	0.1361	19.2116	0.2784
0.005	0.064	0.071	0.0846	18.5042	0.2685	0.9074	0.0924	0.1342	18.6655	0.2711
0.010	0.088	0.100	0.0846	18.0598	0.2620	0.9078	0.0921	0.1327	18.3237	0.2661
0.020	0.115	0.141	0.0847	17.4125	0.2528	0.9081	0.0915	0.1312	17.8599	0.2593
0.050	0.190	0.224	0.0852	16.2206	0.2356	0.9088	0.0909	0.1286	16.9637	0.2462
0.100	0.269	0.316	0.0854	14.9756	0.2177	0.9094	0.0904	0.1257	15.9828	0.2321
0.200	0.389	0.447	0.0853	13.4269	0.1955	0.9104	0.0898	0.1227	14.7598	0.2145
0.300	0.477	0.548	0.0848	12.3727	0.1805	0.9105	0.0895	0.1205	13.9467	0.2027
0.400	0.569	0.632	0.0844	11.5895	0.1690	0.9106	0.0896	0.1195	13.3314	0.1939
0.500	0.641	0.707	0.0837	10.9035	0.1593	0.9107	0.0894	0.1189	12.8671	0.1868
0.600	0.715	0.775	0.0830	10.3651	0.1519	0.9104	0.0893	0.1181	12.4467	0.1810
0.700	0.788	0.837	0.0823	9.9153	0.1452	0.9107	0.0891	0.1178	12.1143	0.1761
0.800	0.860	0.894	0.0815	9.5312	0.1395	0.9109	0.0892	0.1176	11.8451	0.1722
0.900	0.933	0.949	0.0807	9.1761	0.1346	0.9111	0.0890	0.1172	11.5976	0.1688
0.950	0.964	0.975	0.0802	9.0085	0.1322	0.9115	0.0886	0.1172	11.4867	0.1674
0.980	0.987	0.990	0.0797	8.9177	0.1308	0.9118	0.0883	0.1170	11.4208	0.1664

Table 10 **Accuracy of the two-class approximation in the $E_2/LGN/50+M$ system with $\rho=1.4$. We let r_S denote $r[S_i,\hat{S}_i(\alpha)]$ and r_Z denote $r[Z_i,\hat{Z}_i(\alpha)]$.**

			SJF					Two class		
α	r_S	r_Z	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{P}[Srv\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|Srv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.028	0.032	0.0721	26.8586	0.2771	0.9210	0.0785	0.1258	26.9833	0.2782
0.005	0.064	0.071	0.0725	25.9124	0.2676	0.9214	0.0782	0.1238	26.1820	0.2705
0.010	0.088	0.100	0.0727	25.2677	0.2608	0.9216	0.0781	0.1225	25.6840	0.2650
0.020	0.115	0.141	0.0728	24.3464	0.2513	0.9221	0.0776	0.1207	24.9476	0.2574
0.050	0.190	0.224	0.0731	22.5973	0.2332	0.9228	0.0768	0.1181	23.5485	0.2429
0.100	0.269	0.316	0.0737	20.7317	0.2142	0.9234	0.0765	0.1154	22.0836	0.2278
0.200	0.389	0.447	0.0738	18.3632	0.1902	0.9243	0.0757	0.1115	20.1591	0.2084
0.300	0.477	0.548	0.0739	16.8231	0.1741	0.9245	0.0757	0.1094	18.9330	0.1957
0.400	0.569	0.632	0.0738	15.6130	0.1619	0.9243	0.0759	0.1082	17.9971	0.1861
0.500	0.641	0.707	0.0736	14.6565	0.1519	0.9241	0.0762	0.1076	17.2460	0.1784
0.600	0.715	0.775	0.0732	13.8468	0.1438	0.9239	0.0761	0.1067	16.6088	0.1718
0.700	0.788	0.837	0.0729	13.1600	0.1368	0.9239	0.0762	0.1064	16.1052	0.1665
0.800	0.860	0.894	0.0723	12.6183	0.1312	0.9236	0.0765	0.1063	15.7032	0.1624
0.900	0.933	0.949	0.0717	12.0995	0.1260	0.9237	0.0764	0.1061	15.3265	0.1587
0.950	0.964	0.975	0.0711	11.8269	0.1237	0.9237	0.0762	0.1061	15.1405	0.1571
0.980	0.987	0.990	0.0711	11.7352	0.1224	0.9238	0.0763	0.1063	15.0772	0.1563

Table 11 **Accuracy of the two-class approximation in the $E_2/LGN/70+M$ system with $\rho=1.4$. We let r_S denote $r[S_i,\hat{S}_i(\alpha)]$ and r_Z denote $r[Z_i,\hat{Z}_i(\alpha)]$.**

			SJF					Two class		
α	r_S	r_Z	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{P}[Srv\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|Srv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.028	0.032	0.0609	38.4000	0.2767	0.9336	0.0659	0.1148	38.5521	0.2776
0.005	0.064	0.071	0.0612	36.9966	0.2669	0.9340	0.0657	0.1130	37.3661	0.2694
0.010	0.088	0.100	0.0615	36.0972	0.2600	0.9344	0.0656	0.1116	36.6090	0.2637
0.020	0.115	0.141	0.0614	34.7482	0.2502	0.9347	0.0653	0.1100	35.5108	0.2556
0.050	0.190	0.224	0.0617	32.0798	0.2311	0.9355	0.0643	0.1071	33.3963	0.2403
0.100	0.269	0.316	0.0623	29.3281	0.2115	0.9359	0.0637	0.1040	31.1034	0.2240
0.200	0.389	0.447	0.0627	25.7568	0.1860	0.9370	0.0630	0.1007	28.1755	0.2032
0.300	0.477	0.548	0.0628	23.3824	0.1690	0.9372	0.0627	0.0983	26.2479	0.1892
0.400	0.569	0.632	0.0631	21.5662	0.1560	0.9371	0.0630	0.0968	24.7544	0.1784
0.500	0.641	0.707	0.0631	20.0407	0.1452	0.9368	0.0632	0.0958	23.5435	0.1699
0.600	0.715	0.775	0.0633	18.8622	0.1367	0.9364	0.0635	0.0949	22.5803	0.1630
0.700	0.788	0.837	0.0631	17.8337	0.1293	0.9361	0.0639	0.0948	21.7885	0.1572
0.800	0.860	0.894	0.0627	17.0033	0.1232	0.9359	0.0641	0.0946	21.1502	0.1525
0.900	0.933	0.949	0.0626	16.2702	0.1180	0.9356	0.0644	0.0951	20.6316	0.1489
0.950	0.964	0.975	0.0622	15.9075	0.1157	0.9355	0.0644	0.0948	20.3390	0.1471
0.980	0.987	0.990	0.0622	15.7209	0.1143	0.9354	0.0645	0.0949	20.2036	0.1459

Table 12 **Accuracy of the two-class approximation in the $E_2/LGN/100+M$ system with $\rho=1.4$. We let r_S denote $r[S_i,\hat{S}_i(\alpha)]$ and r_Z denote $r[Z_i,\hat{Z}_i(\alpha)]$.**

			SJF					Two class		
α	r_S	r_Z	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{P}[Srv\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|Srv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.028	0.032	0.0281	192.5986	0.2757	0.9703	0.0297	0.0714	192.9406	0.2764
0.005	0.064	0.071	0.0280	185.0995	0.2653	0.9706	0.0294	0.0693	185.9109	0.2668
0.010	0.088	0.100	0.0280	180.0648	0.2575	0.9712	0.0291	0.0681	181.2633	0.2597
0.020	0.115	0.141	0.0279	172.6901	0.2470	0.9712	0.0287	0.0665	174.6868	0.2501
0.050	0.190	0.224	0.0281	158.3895	0.2266	0.9717	0.0282	0.0640	161.9424	0.2317
0.100	0.269	0.316	0.0284	143.1809	0.2049	0.9721	0.0280	0.0622	148.5127	0.2123
0.200	0.389	0.447	0.0284	123.5693	0.1769	0.9730	0.0271	0.0588	130.7087	0.1871
0.300	0.477	0.548	0.0286	110.0988	0.1574	0.9732	0.0269	0.0562	118.9307	0.1699
0.400	0.569	0.632	0.0288	99.4409	0.1423	0.9731	0.0270	0.0549	109.4350	0.1564
0.500	0.641	0.707	0.0289	90.7548	0.1298	0.9731	0.0270	0.0536	101.5314	0.1453
0.600	0.715	0.775	0.0289	83.3735	0.1194	0.9733	0.0267	0.0530	95.2402	0.1360
0.700	0.788	0.837	0.0293	77.1723	0.1105	0.9730	0.0271	0.0521	89.6548	0.1280
0.800	0.860	0.894	0.0295	71.8486	0.1030	0.9727	0.0272	0.0511	84.9333	0.1214
0.900	0.933	0.949	0.0298	67.1998	0.0965	0.9724	0.0275	0.0511	81.0636	0.1160
0.950	0.964	0.975	0.0298	64.9673	0.0935	0.9723	0.0275	0.0509	79.1162	0.1133
0.980	0.987	0.990	0.0301	63.8889	0.0919	0.9720	0.0278	0.0514	78.3799	0.1122

Table 13 **Accuracy of the two-class approximation in the $E_2/LGN/500+M$ system with $\rho=1.4$. We let r_S denote $r[S_i,\hat{S}_i(\alpha)]$ and r_Z denote $r[Z_i,\hat{Z}_i(\alpha)]$.**

			SJF					Two class		
α	r_S	r_Z	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{P}[Srv\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|Srv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.028	0.032	0.0199	384.4576	0.2751	0.9795	0.0205	0.0563	385.2675	0.2756
0.005	0.064	0.071	0.0200	368.9952	0.2644	0.9795	0.0204	0.0550	370.8819	0.2657
0.010	0.088	0.100	0.0201	358.8449	0.2567	0.9797	0.0204	0.0539	361.2103	0.2585
0.020	0.115	0.141	0.0198	344.1819	0.2461	0.9799	0.0201	0.0520	347.4897	0.2486
0.050	0.190	0.224	0.0199	315.9164	0.2257	0.9802	0.0199	0.0500	320.9830	0.2296
0.100	0.269	0.316	0.0201	285.1776	0.2038	0.9806	0.0194	0.0475	293.0889	0.2094
0.200	0.389	0.447	0.0202	245.0863	0.1753	0.9813	0.0190	0.0450	255.5324	0.1829
0.300	0.477	0.548	0.0202	217.9099	0.1558	0.9812	0.0189	0.0428	231.0379	0.1650
0.400	0.569	0.632	0.0201	195.8986	0.1403	0.9814	0.0187	0.0416	211.0099	0.1508
0.500	0.641	0.707	0.0204	178.3116	0.1275	0.9812	0.0189	0.0407	194.6715	0.1391
0.600	0.715	0.775	0.0205	163.2710	0.1168	0.9813	0.0187	0.0401	180.9760	0.1294
0.700	0.788	0.837	0.0209	150.5013	0.1076	0.9813	0.0189	0.0397	169.5193	0.1211
0.800	0.860	0.894	0.0211	139.8046	0.0998	0.9809	0.0193	0.0388	159.6090	0.1141
0.900	0.933	0.949	0.0212	129.8794	0.0929	0.9808	0.0192	0.0387	151.1977	0.1079
0.950	0.964	0.975	0.0216	125.3307	0.0898	0.9805	0.0195	0.0391	147.3497	0.1054
0.980	0.987	0.990	0.0217	122.8059	0.0880	0.9805	0.0195	0.0396	145.5665	0.1040

Table 14 **Accuracy of the two-class approximation in the $E_2/LGN/1000+M$ system with $\rho=1.4$. We let r_S denote $r[S_i,\hat{S}_i(\alpha)]$ and r_Z denote $r[Z_i,\hat{Z}_i(\alpha)]$.**

			SJF					Two class		
α	r_S	r_Z	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{P}[Srv\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|Srv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.028	0.032	0.1442	9.5566	0.3017	0.8118	0.1876	0.1778	9.5983	0.3037
0.005	0.064	0.071	0.1439	9.3999	0.2963	0.8117	0.1876	0.1774	9.4568	0.2991
0.010	0.088	0.100	0.1441	9.2811	0.2919	0.8127	0.1873	0.1777	9.4043	0.2974
0.020	0.115	0.141	0.1440	9.1114	0.2862	0.8129	0.1871	0.1774	9.2790	0.2925
0.050	0.190	0.224	0.1427	8.7845	0.2744	0.8145	0.1851	0.1774	9.1123	0.2862
0.100	0.269	0.316	0.1405	8.3835	0.2610	0.8176	0.1827	0.1769	8.8642	0.2777
0.200	0.389	0.447	0.1379	7.9298	0.2456	0.8204	0.1801	0.1770	8.6304	0.2694
0.300	0.477	0.548	0.1361	7.6372	0.2352	0.8216	0.1784	0.1780	8.5075	0.2644
0.400	0.569	0.632	0.1342	7.3757	0.2259	0.8243	0.1761	0.1786	8.3757	0.2595
0.500	0.641	0.707	0.1334	7.2127	0.2194	0.8251	0.1752	0.1781	8.2582	0.2557
0.600	0.715	0.775	0.1321	7.0191	0.2130	0.8269	0.1737	0.1777	8.1403	0.2520
0.700	0.788	0.837	0.1302	6.8199	0.2064	0.8294	0.1710	0.1782	8.0890	0.2503
0.800	0.860	0.894	0.1284	6.6641	0.2014	0.8309	0.1691	0.1792	8.0570	0.2486
0.900	0.933	0.949	0.1277	6.5748	0.1979	0.8319	0.1682	0.1793	8.0262	0.2480
0.950	0.964	0.975	0.1264	6.4769	0.1957	0.8327	0.1668	0.1794	8.0013	0.2469
0.980	0.987	0.990	0.1268	6.4743	0.1946	0.8332	0.1670	0.1789	7.9582	0.2460

Table 15 **Accuracy of the two-class approximation in the $H_2/LGN/20+M$ system with $\rho=1.4$. We let r_S denote $r[S_i,\hat{S}_i(\alpha)]$ and r_Z denote $r[Z_i,\hat{Z}_i(\alpha)]$.**

			SJF					Two class		
α	r_S	r_Z	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{P}[Srv\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|Srv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.000	0.000	0.1202	20.9803	0.2813	0.8645	0.1355	0.1574	21.0334	0.2829
0.005	0.000	0.000	0.1201	20.4762	0.2746	0.8643	0.1354	0.1562	20.6647	0.2782
0.010	0.000	0.000	0.1198	20.0842	0.2690	0.8649	0.1350	0.1557	20.4151	0.2744
0.020	0.000	0.000	0.1196	19.5496	0.2612	0.8653	0.1348	0.1552	20.0450	0.2686
0.050	0.000	0.000	0.1185	18.5245	0.2467	0.8661	0.1337	0.1540	19.3796	0.2589
0.100	0.000	0.000	0.1175	17.5000	0.2322	0.8670	0.1331	0.1530	18.6745	0.2491
0.200	0.000	0.000	0.1152	16.1860	0.2138	0.8681	0.1319	0.1520	17.7926	0.2366
0.300	0.000	0.000	0.1136	15.3766	0.2020	0.8684	0.1313	0.1520	17.3132	0.2289
0.400	0.000	0.000	0.1121	14.6763	0.1922	0.8694	0.1306	0.1525	16.9645	0.2237
0.500	0.000	0.000	0.1104	14.0933	0.1837	0.8705	0.1295	0.1513	16.5340	0.2175
0.600	0.000	0.000	0.1089	13.5920	0.1770	0.8711	0.1286	0.1511	16.2363	0.2138
0.700	0.000	0.000	0.1074	13.1906	0.1713	0.8721	0.1277	0.1515	16.0247	0.2108
0.800	0.000	0.000	0.1062	12.8519	0.1662	0.8727	0.1272	0.1520	15.8930	0.2087
0.900	0.000	0.000	0.1051	12.5684	0.1621	0.8734	0.1266	0.1518	15.7354	0.2067
0.950	0.000	0.000	0.1047	12.4545	0.1604	0.8735	0.1263	0.1518	15.6609	0.2056
0.980	0.000	0.000	0.1040	12.3443	0.1590	0.8741	0.1257	0.1519	15.6352	0.2053

Table 16 **Accuracy of the two-class approximation in the $H_2/LGN/50+M$ system with $\rho=1.4$. We let r_S denote $r[S_i,\hat{S}_i(\alpha)]$ and r_Z denote $r[Z_i,\hat{Z}_i(\alpha)]$.**

			SJF					Two class		
α	r_S	r_Z	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{P}[Srv\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|Srv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.000	0.000	0.1093	28.5403	0.2783	0.8815	0.1181	0.1493	28.6561	0.2797
0.005	0.000	0.000	0.1090	27.6998	0.2700	0.8816	0.1180	0.1481	28.0307	0.2739
0.010	0.000	0.000	0.1091	27.1867	0.2644	0.8821	0.1180	0.1473	27.5927	0.2692
0.020	0.000	0.000	0.1093	26.4241	0.2564	0.8819	0.1180	0.1470	27.1056	0.2636
0.050	0.000	0.000	0.1092	24.9610	0.2412	0.8821	0.1178	0.1457	26.0772	0.2527
0.100	0.000	0.000	0.1083	23.4060	0.2255	0.8828	0.1173	0.1443	25.0008	0.2417
0.200	0.000	0.000	0.1067	21.4613	0.2058	0.8834	0.1170	0.1433	23.7146	0.2283
0.300	0.000	0.000	0.1049	20.1737	0.1928	0.8836	0.1164	0.1428	22.9102	0.2200
0.400	0.000	0.000	0.1033	19.1303	0.1824	0.8841	0.1157	0.1421	22.2151	0.2131
0.500	0.000	0.000	0.1019	18.3156	0.1740	0.8847	0.1152	0.1418	21.6616	0.2072
0.600	0.000	0.000	0.1006	17.6233	0.1670	0.8852	0.1147	0.1412	21.1741	0.2025
0.700	0.000	0.000	0.0996	17.0540	0.1611	0.8856	0.1142	0.1415	20.8658	0.1989
0.800	0.000	0.000	0.0985	16.6125	0.1564	0.8858	0.1141	0.1420	20.6432	0.1965
0.900	0.000	0.000	0.0972	16.1842	0.1520	0.8864	0.1136	0.1425	20.5054	0.1948
0.950	0.000	0.000	0.0969	16.0049	0.1502	0.8865	0.1135	0.1421	20.3421	0.1936
0.980	0.000	0.000	0.0965	15.8451	0.1486	0.8869	0.1129	0.1421	20.2513	0.1927

Table 17 **Accuracy of the two-class approximation in the $H_2/LGN/70+M$ system with $\rho=1.4$. We let r_S denote $r[S_i,\hat{S}_i(\alpha)]$ and r_Z denote $r[Z_i,\hat{Z}_i(\alpha)]$.**

			SJF					Two class		
α	r_S	r_Z	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{P}[Srv\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|Srv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.000	0.000	0.0972	39.9466	0.2763	0.8980	0.1016	0.1405	40.0690	0.2774
0.005	0.000	0.000	0.0972	38.6817	0.2676	0.8981	0.1016	0.1394	39.1220	0.2708
0.010	0.000	0.000	0.0974	37.8927	0.2614	0.8985	0.1017	0.1386	38.5434	0.2664
0.020	0.000	0.000	0.0977	36.7268	0.2527	0.8985	0.1018	0.1378	37.6521	0.2599
0.050	0.000	0.000	0.0981	34.5031	0.2367	0.8984	0.1015	0.1362	36.0083	0.2477
0.100	0.000	0.000	0.0977	32.1095	0.2198	0.8986	0.1014	0.1345	34.2889	0.2353
0.200	0.000	0.000	0.0968	29.1708	0.1989	0.8986	0.1016	0.1326	32.1778	0.2202
0.300	0.000	0.000	0.0958	27.2400	0.1850	0.8983	0.1017	0.1324	30.9556	0.2111
0.400	0.000	0.000	0.0944	25.6410	0.1741	0.8983	0.1014	0.1317	29.8297	0.2031
0.500	0.000	0.000	0.0934	24.4467	0.1653	0.8984	0.1013	0.1312	28.9752	0.1969
0.600	0.000	0.000	0.0922	23.3802	0.1579	0.8988	0.1008	0.1311	28.3206	0.1923
0.700	0.000	0.000	0.0912	22.5526	0.1517	0.8990	0.1009	0.1312	27.7582	0.1881
0.800	0.000	0.000	0.0899	21.8434	0.1465	0.8993	0.1005	0.1313	27.3787	0.1852
0.900	0.000	0.000	0.0889	21.2800	0.1423	0.8994	0.1006	0.1317	27.1024	0.1831
0.950	0.000	0.000	0.0886	21.0090	0.1404	0.8993	0.1006	0.1318	26.9356	0.1822
0.980	0.000	0.000	0.0883	20.7917	0.1388	0.8996	0.1004	0.1315	26.7635	0.1810

Table 18 **Accuracy of the two-class approximation in the $H_2/LGN/100+M$ system with $\rho=1.4$. We let r_S denote $r[S_i,\hat{S}_i(\alpha)]$ and r_Z denote $r[Z_i,\hat{Z}_i(\alpha)]$.**

			SJF					Two class		
α	r_S	r_Z	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{P}[Srv\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|Srv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.000	0.000	0.0464	192.5039	0.2739	0.9529	0.0470	0.0962	192.7617	0.2746
0.005	0.000	0.000	0.0467	185.1853	0.2637	0.9532	0.0467	0.0947	186.4578	0.2658
0.010	0.000	0.000	0.0469	180.2191	0.2563	0.9533	0.0467	0.0937	182.3211	0.2594
0.020	0.000	0.000	0.0473	173.0731	0.2460	0.9533	0.0468	0.0924	176.3980	0.2507
0.050	0.000	0.000	0.0481	159.4629	0.2264	0.9533	0.0466	0.0905	164.9280	0.2341
0.100	0.000	0.000	0.0494	145.3271	0.2057	0.9532	0.0471	0.0889	153.0399	0.2168
0.200	0.000	0.000	0.0499	126.8800	0.1795	0.9533	0.0470	0.0861	137.6744	0.1950
0.300	0.000	0.000	0.0503	114.9555	0.1621	0.9527	0.0474	0.0842	128.4582	0.1813
0.400	0.000	0.000	0.0508	105.2577	0.1484	0.9522	0.0478	0.0839	120.7713	0.1702
0.500	0.000	0.000	0.0514	97.6453	0.1374	0.9516	0.0485	0.0829	114.3697	0.1611
0.600	0.000	0.000	0.0519	91.0014	0.1281	0.9510	0.0489	0.0824	109.0296	0.1535
0.700	0.000	0.000	0.0522	85.8681	0.1205	0.9504	0.0498	0.0821	104.9968	0.1475
0.800	0.000	0.000	0.0523	81.5812	0.1146	0.9493	0.0506	0.0827	102.0708	0.1433
0.900	0.000	0.000	0.0522	77.8739	0.1093	0.9488	0.0510	0.0828	99.2851	0.1393
0.950	0.000	0.000	0.0521	75.8341	0.1066	0.9488	0.0509	0.0824	97.5674	0.1372
0.980	0.000	0.000	0.0517	74.7900	0.1051	0.9488	0.0508	0.0825	96.7648	0.1358

Table 19 **Accuracy of the two-class approximation in the $H_2/LGN/500+M$ system with $\rho=1.4$. We let r_S denote $r[S_i,\hat{S}_i(\alpha)]$ and r_Z denote $r[Z_i,\hat{Z}_i(\alpha)]$.**

			SJF					Two class		
α	r_S	r_Z	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{P}[Srv\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|\hat{S}\leq\tau]$	$\mathbb{E}[W\|Srv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.000	0.000	0.0326	382.2934	0.2731	0.9673	0.0327	0.0782	383.8840	0.2741
0.005	0.000	0.000	0.0331	367.4388	0.2628	0.9671	0.0328	0.0770	370.6655	0.2649
0.010	0.000	0.000	0.0331	357.0457	0.2550	0.9675	0.0325	0.0765	361.5121	0.2579
0.020	0.000	0.000	0.0334	342.6493	0.2445	0.9673	0.0326	0.0753	348.6045	0.2486
0.050	0.000	0.000	0.0344	315.2138	0.2247	0.9671	0.0328	0.0731	323.6669	0.2308
0.100	0.000	0.000	0.0347	285.6612	0.2035	0.9675	0.0327	0.0705	297.6611	0.2120
0.200	0.000	0.000	0.0353	247.3018	0.1762	0.9675	0.0324	0.0675	264.3535	0.1883
0.300	0.000	0.000	0.0357	222.8869	0.1583	0.9669	0.0331	0.0655	243.6374	0.1729
0.400	0.000	0.000	0.0359	201.9231	0.1433	0.9669	0.0331	0.0648	226.2017	0.1605
0.500	0.000	0.000	0.0366	185.6467	0.1315	0.9663	0.0338	0.0644	212.5150	0.1507
0.600	0.000	0.000	0.0373	171.4987	0.1214	0.9659	0.0343	0.0641	200.4547	0.1421
0.700	0.000	0.000	0.0376	160.0602	0.1130	0.9654	0.0347	0.0641	191.4851	0.1355
0.800	0.000	0.000	0.0379	150.6434	0.1063	0.9647	0.0354	0.0636	183.6398	0.1299
0.900	0.000	0.000	0.0384	142.3597	0.1004	0.9643	0.0358	0.0641	177.6162	0.1255
0.950	0.000	0.000	0.0386	138.1450	0.0976	0.9640	0.0361	0.0644	174.1809	0.1232
0.980	0.000	0.000	0.0386	135.8387	0.0960	0.9640	0.0360	0.0647	172.4279	0.1218

Table 20 **Accuracy of the two-class approximation in the $H_2/LGN/1000+M$ system with $\rho=1.4$. We let r_S denote $r[S_i,\hat{S}_i(\alpha)]$ and r_Z denote $r[Z_i,\hat{Z}_i(\alpha)]$.**

			Two class right threshold			Two class wrong threshold		
α	$r[S_i, \hat{S}_i(\alpha)]$	$r[Z_i, \hat{Z}_i(\alpha)]$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.0282	0.0316	0.1212	38.8721	0.2780	0.2361	39.7960	0.2851
0.005	0.0636	0.0707	0.1194	37.6858	0.2700	0.2361	39.7820	0.2850
0.01	0.0877	0.1000	0.1181	36.9477	0.2644	0.2366	39.8898	0.2857
0.02	0.1151	0.1414	0.1167	35.9017	0.2567	0.2367	39.9327	0.2858
0.05	0.1899	0.2236	0.1140	33.7302	0.2413	0.2370	40.0037	0.2860
0.1	0.2688	0.3162	0.1110	31.5432	0.2260	0.2372	40.0213	0.2860
0.2	0.3894	0.4472	0.1078	28.7060	0.2059	0.2362	39.9446	0.2848
0.3	0.4767	0.5477	0.1061	26.9172	0.1930	0.2281	38.9449	0.2772
0.4	0.5691	0.6325	0.1048	25.5386	0.1829	0.2120	36.8546	0.2622
0.5	0.6413	0.7071	0.1040	24.4331	0.1749	0.1896	33.9810	0.2420
0.6	0.7154	0.7746	0.1035	23.5301	0.1686	0.1639	30.6543	0.2182
0.7	0.7878	0.8367	0.1030	22.7404	0.1631	0.1401	27.4672	0.1956
0.8	0.8602	0.8944	0.1028	22.1475	0.1589	0.1225	24.9204	0.1773
0.9	0.9333	0.9487	0.1028	21.6270	0.1551	0.1111	22.9721	0.1637
0.95	0.9641	0.9747	0.1027	21.4076	0.1534	0.1076	22.2185	0.1585
0.98	0.9866	0.9899	0.1032	21.3123	0.1526	0.1057	21.8478	0.1557

Table 21 **Effect of choosing the wrong threshold in the $M/LGN/100+M$ system with $\rho = 1.4$.**

			Two class right threshold			Two class wrong threshold		
α	$r[S_i, \hat{S}_i(\alpha)]$	$r[Z_i, \hat{Z}_i(\alpha)]$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.0282	0.0316	0.0605	385.9085	0.2759	0.2386	397.2644	0.2845
0.005	0.0636	0.0707	0.0591	371.5782	0.2661	0.2384	396.8136	0.2844
0.01	0.0877	0.1000	0.0578	362.3138	0.2589	0.2389	397.9682	0.2849
0.02	0.1151	0.1414	0.0561	348.9148	0.2492	0.2391	398.3045	0.2850
0.05	0.1899	0.2236	0.0550	321.1919	0.2299	0.2392	398.5709	0.2850
0.1	0.2688	0.3162	0.0526	293.4705	0.2099	0.2395	398.9648	0.2850
0.2	0.3894	0.4472	0.0501	256.1455	0.1833	0.2384	398.2404	0.2841
0.3	0.4767	0.5477	0.0479	231.6201	0.1657	0.2307	388.8506	0.2766
0.4	0.5691	0.6325	0.0471	212.4476	0.1518	0.2142	367.6738	0.2615
0.5	0.6413	0.7071	0.0460	196.7703	0.1404	0.1910	338.1445	0.2405
0.6	0.7154	0.7746	0.0457	183.0561	0.1312	0.1617	301.7280	0.2149
0.7	0.7878	0.8367	0.0448	171.2986	0.1230	0.1283	261.3477	0.1862
0.8	0.8602	0.8944	0.0444	161.9419	0.1161	0.0911	217.5845	0.1550
0.9	0.9333	0.9487	0.0446	154.1588	0.1104	0.0540	172.5530	0.1230
0.95	0.9641	0.9747	0.0441	150.1600	0.1074	0.0465	157.9092	0.1127
0.98	0.9866	0.9899	0.0446	148.4281	0.1062	0.0453	152.7485	0.1090

Table 22 **Effect of choosing the wrong threshold in the $M/LGN/1000+M$ system with $\rho = 1.4$.**

			Two class right threshold			Two class wrong threshold		
α	$r[S_i, \hat{S}_i(\alpha)]$	$r[Z_i, \hat{Z}_i(\alpha)]$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.001	0.0282	0.0316	0.0554	779.1574	0.4330	0.3244	795.0558	0.4432
0.005	0.0636	0.0707	0.0550	755.7675	0.4205	0.3244	794.9826	0.4432
0.01	0.0877	0.1000	0.0537	741.2371	0.4118	0.3250	796.9677	0.4438
0.02	0.1151	0.1414	0.0524	719.1483	0.3995	0.3254	798.0134	0.4440
0.05	0.1899	0.2236	0.0516	674.0468	0.3745	0.3255	798.4892	0.4439
0.1	0.2688	0.3162	0.0493	624.7043	0.3477	0.3247	797.0073	0.4429
0.2	0.3894	0.4472	0.0480	559.9949	0.3112	0.3123	777.9405	0.4319
0.3	0.4767	0.5477	0.0467	512.1316	0.2846	0.2887	740.5255	0.4102
0.4	0.5691	0.6325	0.0456	474.2025	0.2632	0.2588	692.0790	0.3830
0.5	0.6413	0.7071	0.0448	442.1810	0.2455	0.2245	636.0262	0.3519
0.6	0.7154	0.7746	0.0438	414.8870	0.2303	0.1862	574.2843	0.3182
0.7	0.7878	0.8367	0.0432	392.1642	0.2173	0.1452	510.4045	0.2830
0.8	0.8602	0.8944	0.0433	370.1393	0.2055	0.1007	443.6124	0.2461
0.9	0.9333	0.9487	0.0428	350.5303	0.1944	0.0535	373.5181	0.2076
0.95	0.9641	0.9747	0.0426	341.1217	0.1893	0.0398	345.1128	0.1923
0.98	0.9866	0.9899	0.0424	336.0697	0.1864	0.0404	336.4412	0.1872

Table 23 **Effect of choosing the wrong threshold in the $M/LGN/1000+M$ system with $\rho = 1.8$.**

Threshold	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.1	0.2359	39.7825	0.2851
0.2	0.2358	39.7678	0.2850
0.3	0.2364	39.8739	0.2856
0.4	0.2358	39.7654	0.2850
0.5	0.2323	39.3386	0.2818
0.6	0.2161	37.8413	0.2708
0.7	0.1786	35.4548	0.2533
0.8	0.1193	33.8284	0.2416
0.9	0.1477	35.9088	0.2566
1.0	0.2076	38.5402	0.2746
1.1	0.2296	39.7715	0.2829
1.2	0.2367	40.2363	0.2859
1.3	0.2361	39.9034	0.2856
1.4	0.2369	39.9970	0.2859
1.5	0.2373	40.0262	0.2859
1.6	0.2379	40.1740	0.2864
1.7	0.2388	40.3732	0.2871
1.8	0.2389	40.3934	0.2871
1.9	0.2386	40.3303	0.2867
2.0	0.2388	40.3484	0.2869
2.1	0.2386	40.3266	0.2867
2.2	0.2387	40.3336	0.2867
2.3	0.2384	40.2790	0.2868
2.4	0.2375	40.0982	0.2859
2.5	0.2373	40.0923	0.2860

Table 24 Performance of the two-class priority rule with different thresholds in the $M/LGN/100+M$ system with $\rho = 1.4$ and $\alpha = 0.05$.

Threshold	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.5	0.1904	33.0804	0.2365
0.6	0.1727	30.6904	0.2195
0.7	0.1571	28.8038	0.2054
0.8	0.1426	27.1573	0.1938
0.9	0.1296	25.6631	0.1831
1	0.1191	24.6835	0.1765
1.1	0.1112	24.1149	0.1725
1.2	0.1058	23.8837	0.1709
1.3	0.1030	24.0103	0.1719
1.4	0.1042	24.4662	0.1752
1.5	0.1090	25.2790	0.1808
1.6	0.1179	26.4570	0.1887
1.7	0.1291	27.7108	0.1977
1.8	0.1417	29.0181	0.2066
1.9	0.1539	30.2144	0.2152
2	0.1652	31.3631	0.2233
2.1	0.1754	32.4260	0.2310
2.2	0.1848	33.4725	0.2383
2.3	0.1922	34.3007	0.2445
2.4	0.1981	34.9255	0.2494
2.5	0.2034	35.5592	0.2537
2.6	0.2070	35.8318	0.2568
2.7	0.2109	36.3029	0.2602
2.8	0.2144	36.7563	0.2634
2.9	0.2172	37.1143	0.2659
3	0.2197	37.4571	0.2680

Table 25 Performance of the two-class priority rule with different thresholds in the $M/LGN/100+M$ system with $\rho=1.4$ and $\alpha=0.5$.

Threshold	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.2	0.2232	37.6952	0.2701
0.6	0.1600	27.9883	0.2001
0.8	0.1372	24.7636	0.1771
1	0.1209	22.6296	0.1615
1.4	0.1026	20.5076	0.1466
1.7	0.0980	20.2193	0.1449
1.8	0.0976	20.2818	0.1454
2	0.0994	20.7442	0.1483
2.1	0.1012	21.1173	0.1506
2.2	0.1035	21.4915	0.1533
2.4	0.1093	22.3028	0.1590
2.5	0.1131	22.8139	0.1625
2.8	0.1260	24.4249	0.1740
3	0.1357	25.6474	0.1825
3.1	0.1407	26.2374	0.1868
3.4	0.1557	28.1426	0.2003
3.6	0.1646	29.2736	0.2086
3.7	0.1675	29.5110	0.2117
4.1	0.1821	31.4576	0.2249
4.2	0.1853	31.8707	0.2277
4.5	0.1944	33.2430	0.2365
4.7	0.1977	33.6131	0.2410
4.8	0.1996	33.8863	0.2431
5.2	0.2071	34.9718	0.2499
5.4	0.2089	35.2672	0.2531
5.5	0.2104	35.4780	0.2543

Table 26 Performance of the two-class priority rule with different thresholds in the $M/LGN/100+M$ system with $\rho = 1.4$, $\alpha = 0.98$.

			Two class			SJF		
α	$r[Z_i, \hat{Z}_i]$	Sample	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.05	0.2236	500	0.1983	36.5354	0.2614	0.0697	32.4614	0.2323
0.05	0.2236	1,000	0.1718	35.0211	0.2505	0.0697	32.4614	0.2323
0.05	0.2236	10,000	0.1175	33.7846	0.2414	0.0697	32.4614	0.2323
0.05	0.2236	–	0.1140	33.7302	0.2413	0.0697	32.4614	0.2323
0.5	0.7071	500	0.1623	29.3776	0.2101	0.0705	20.7530	0.1485
0.5	0.7071	1,000	0.1468	27.5163	0.1967	0.0705	20.7530	0.1485
0.5	0.7071	10,000	0.1116	24.1868	0.1728	0.0705	20.7530	0.1485
0.5	0.7071	–	0.1040	24.4331	0.1749	0.0705	20.7530	0.1485
0.98	0.9899	500	0.1498	26.5691	0.1894	0.0690	16.7138	0.1194
0.98	0.9899	1,000	0.1363	24.7321	0.1762	0.0690	16.7138	0.1194
0.98	0.9899	10,000	0.1066	21.0335	0.1503	0.0690	16.7138	0.1194
0.98	0.9899	–	0.1032	21.3123	0.1526	0.0690	16.7138	0.1194

Table 27 Performance of the two-class priority rule with difference threshold estimated using SGD with $a = 0.1$ and for different values of α and sample size. The last row in each table block uses the threshold calculated based on (4).

			Two class			SJF		
α	$r[Z_i, \hat{Z}_i]$	Sample	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.05	0.2236	500	0.1113	34.0356	0.2437	0.0697	32.0745	0.2295
0.05	0.2236	1,000	0.1121	33.7786	0.2418	0.0697	32.0745	0.2295
0.05	0.2236	10,000	0.1115	32.8210	0.2348	0.0697	32.0745	0.2295
0.05	0.2236	–	0.1140	33.7302	0.2413	0.0697	32.4614	0.2323
0.5	0.7071	500	0.1045	23.8488	0.1708	0.0705	20.6156	0.1475
0.5	0.7071	1,000	0.1034	23.9176	0.1715	0.0705	20.6156	0.1475
0.5	0.7071	10,000	0.1029	23.9198	0.1711	0.0705	20.6156	0.1475
0.5	0.7071	–	0.1040	24.4331	0.1749	0.0705	20.7530	0.1485
0.98	0.9899	500	0.0977	20.3024	0.1451	0.0690	16.7167	0.1194
0.98	0.9899	1,000	0.0979	20.4706	0.1464	0.0690	16.7167	0.1194
0.98	0.9899	10,000	0.1027	21.3673	0.1523	0.0690	16.7167	0.1194
0.98	0.9899	–	0.1032	21.3123	0.1526	0.0690	16.7138	0.1194

Table 28 Performance of the two-class priority rule with difference threshold estimated using SGD with $a = 0.5$ and for different values of α and sample size. The last row in each table block uses the threshold calculated based on (4).

			Two class			SJF		
α	$r[Z_i, \hat{Z}_i]$	Sample	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.05	0.2236	500	0.1109	33.9341	0.2429	0.0697	32.4614	0.2323
0.05	0.2236	1000	0.1153	33.7294	0.2415	0.0697	32.4614	0.2323
0.05	0.2236	10000	0.1118	33.8589	0.2421	0.0697	32.4614	0.2323
0.05	0.2236	–	0.1140	33.7302	0.2413	0.0697	32.4614	0.2323
0.5	0.7071	500	0.1034	23.8908	0.1713	0.0705	20.7530	0.1485
0.5	0.7071	1000	0.1030	24.0014	0.1719	0.0705	20.7530	0.1485
0.5	0.7071	10000	0.1048	24.5669	0.1760	0.0705	20.7530	0.1485
0.5	0.7071	–	0.1040	24.4331	0.1749	0.0705	20.7530	0.1485
0.98	0.9899	500	0.0990	20.6545	0.1478	0.0690	16.7138	0.1194
0.98	0.9899	1000	0.0980	20.5035	0.1465	0.0690	16.7138	0.1194
0.98	0.9899	10000	0.1026	21.3496	0.1521	0.0690	16.7138	0.1194
0.98	0.9899	–	0.1032	21.3123	0.1526	0.0690	16.7138	0.1194

Table 29 Performance of the two-class priority rule with difference threshold estimated using SGD with $a = 0.7$ and for different values of α and sample size. The last row in each table block uses the threshold calculated based on (4).

		SJF			Two class		
$r[S,_i \hat{S}_i]$	$r[Z,_i \hat{Z}_i]$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.0025	0.005	0.0220	395.3176	0.2832	0.0612	395.5416	0.2831
0.0071	0.01	0.0220	393.6909	0.2819	0.0609	394.4173	0.2821
0.0366	0.05	0.0222	379.7860	0.2711	0.0592	381.0778	0.2719
0.0711	0.1	0.0222	361.5732	0.2578	0.0576	363.7636	0.2595
0.1501	0.2	0.0223	324.0653	0.2308	0.0531	329.4904	0.2348
0.2312	0.3	0.0226	287.2849	0.2046	0.0489	295.5902	0.2106
0.3196	0.4	0.0230	251.8599	0.1793	0.0455	264.6596	0.1883
0.4161	0.5	0.0232	217.6817	0.1548	0.0439	235.7478	0.1677
0.5137	0.6	0.0234	185.7213	0.1321	0.0452	209.6101	0.1493
0.6233	0.7	0.0237	157.6626	0.1125	0.0467	186.6144	0.1332
0.7402	0.8	0.0243	138.1429	0.0984	0.0465	168.6685	0.1203
0.8662	0.9	0.0247	126.5476	0.0902	0.0445	155.0831	0.1104
0.9318	0.95	0.0251	123.9812	0.0885	0.0444	151.4413	0.1079
0.9724	0.98	0.0251	123.2799	0.0881	0.0442	149.5975	0.1066
0.9862	0.99	0.0254	123.2587	0.0882	0.0449	149.7034	0.1069

Table 30 **Accuracy of the approximation in the $M/LGN/1000+M$ system with $\rho=1.4$ under service-time model of Section 5.2.**

		SJF			Two class		
$r[S_{,i}\,\hat{S}_i]$	$r[Z_i, \hat{Z}_i]$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$	$\mathbb{E}[W\|Serv]$	$\mathbb{E}[Q]$	$\mathbb{P}(Ab)$
0.0025	0.005	0.0171	792.5415	0.4417	0.0551	792.9762	0.4417
0.0071	0.01	0.0170	790.2275	0.4402	0.0548	790.5204	0.4401
0.0366	0.05	0.0169	768.3630	0.4272	0.0539	770.1773	0.4278
0.0711	0.1	0.0173	739.7211	0.4108	0.0531	742.8040	0.4123
0.1501	0.2	0.0177	681.5804	0.3782	0.0505	686.7148	0.3810
0.2312	0.3	0.0181	620.6161	0.3441	0.0479	629.1391	0.3487
0.3196	0.4	0.0185	558.1607	0.3097	0.0452	570.1010	0.3162
0.4161	0.5	0.0189	496.6495	0.2751	0.0427	512.6401	0.2839
0.5137	0.6	0.0192	435.2636	0.2415	0.0419	455.2422	0.2525
0.6233	0.7	0.0195	380.2486	0.2109	0.0409	404.3327	0.2241
0.7402	0.8	0.0200	337.9368	0.1875	0.0407	364.2818	0.2020
0.8662	0.9	0.0203	312.3646	0.1735	0.0405	339.4761	0.1885
0.9318	0.95	0.0209	306.9997	0.1705	0.0413	334.3451	0.1857
0.9724	0.98	0.0210	305.1819	0.1697	0.0418	332.6789	0.1848
0.9862	0.99	0.0210	304.5051	0.1697	0.0426	332.5934	0.1852

Table 31 **Accuracy of the approximation in the $M/LGN/1000+M$ system with $\rho=1.8$ under service-time model of Section 5.2 .**

References

Atar, R., W. Kang, H. Kaspi, and K. Ramanan (2023). Long-time limit of nonlinearly coupled measure-valued equations that model many-server queues with reneging. *SIAM Journal on Mathematical Analysis 55*(6), 7189–7239.

Atar, R., H. Kaspi, and N. Shimkin (2014). Fluid limits for many-server systems with reneging under a priority policy. *Mathematics of Operations Research 39*(3), 672–696.

Banerjee, S., A. Budhiraja, and A. L. Puha (2020). Heavy traffic scaling limits for shortest remaining processing time queues with heavy tailed processing time distributions.

Barlow, R. E. and F. Proschan (1975). Statistical theory of reliability and life testing: probability models. Technical report, Florida State Univ Tallahassee.

Dong, J. and R. Ibrahim (2021). SRPT scheduling discipline in many-server queues with impatient customers. *Management Science 67*(12), 7708–7718.

Down, D. G. (2019). Open problem—size-based scheduling with estimation errors. *Stochastic Systems 9*(3), 295–296.

Down, D. G., H. C. Gromoll, and A. L. Puha (2009). Fluid limits for shortest remaining processing time queues. *Mathematics of Operations Research 34*(4), 880–911.

Garnett, O., A. Mandelbaum, and M. Reiman (2002). Designing a call center with impatient customers. *Manufacturing & Service Operations Management 4*(3), 208–227.

Gromoll, H. C., Ł. Kruk, and A. L. Puha (2011). Diffusion limits for shortest remaining processing time queues. *Stochastic Systems 1*(1), 1–16.

Grosof, I., Z. Scully, and M. Harchol-Balter (2018). SRPT for multiserver systems. *Performance Evaluation 127*, 154–175.

Ibrahim, R., H. Ye, P. L'Ecuyer, and H. Shen (2016). Modeling and forecasting call center arrivals: A literature survey and a case study. *International Journal of Forecasting 32*(3), 865–874.

Lehmann, E. L. (1966). Some concepts of dependence. *The Annals of Mathematical Statistics 37*(5), 1137–1153.

Leonardi, S. and D. Raz (2007). Approximating total flow time on parallel machines. *Journal of Computer and System Sciences 73*(6), 875–891.

Mitzenmacher, M. (2021). Queues with small advice. In *SIAM Conference on Applied and Computational Discrete Algorithms (ACDA21)*, pp. 1–12. SIAM.

Polyak, B. T. and A. B. Juditsky (1992). Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization 30*(4), 838–855.

Puha, A. L. et al. (2015). Diffusion limits for shortest remaining processing time queues under nonstandard spatial scaling. *The Annals of Applied Probability 25*(6), 3381–3404.

Puha, A. L. and A. R. Ward (2019). Scheduling an overloaded multiclass many-server queue with impatient customers. In *Operations Research & Management Science in the Age of Analytics*, pp. 189–217. INFORMS.

Schrage, L. (1968). Letter to the editor-a proof of the optimality of the shortest remaining processing time discipline. *Operations Research 16*(3), 687–690.

Schrage, L. E. and L. W. Miller (1966). The queue M/G/1 with the shortest remaining processing time discipline. *Operations Research 14*(4), 670–684.

Scully, Z., I. Grosof, and M. Harchol-Balter (2020). The gittins policy is nearly optimal in the M/G/k under extremely general conditions. *Proceedings of the ACM on Measurement and Analysis of Computing Systems 4*(3), 1–29.

Scully, Z., I. Grosof, and M. Mitzenmacher (2021). Uniform bounds for scheduling with job size estimates. *arXiv preprint arXiv:2110.00633*.

Scully, Z., M. Harchol-Balter, and A. Scheller-Wolf (2018). Soap: One clean analysis of all age-based scheduling policies. *Proceedings of the ACM on Measurement and Analysis of Computing Systems 2*(1), 1–30.

Shaked, M. and J. G. Shanthikumar (2007). *Stochastic orders*. Springer Science & Business Media.

Wierman, A. and M. Nuyens (2008). Scheduling despite inexact job-size information. In *Proceedings of the 2008 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, pp. 25–36.

Wu, C., A. Bassamboo, and O. Perry (2019). Service system with dependent service and patience times. *Management Science 65*(3), 1151–1172.